volume VIII♦ **issue** 4
2017

Logos & Episteme

an international journal
of epistemology

Romanian Academy
Iasi Branch
♦♦♦
“Gheorghe Zane” Institute
for Economic and Social
Research

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

www.logos-and-episteme.proiectsbc.ro

ISSN 2069-053

Logos & Episteme

TABLE OF CONTENTS

RESEARCH ARTICLES

Peter BAUMANN, If You Believe, You Believe. A Constitutive Account of Knowledge of One's Own Beliefs.. 389
David COSS, Contextualism and Context Internalism...................................... 417
Patrick GRIM and Nicholas RESCHER, Limitations and the World Beyond.. 425
Richard PETTIGREW, Epistemic Utility and the Normativity of Logic.......... 455
Maura PRIEST, Why Anti-Luck Virtue Epistemology has No Luck with Closure.. 493
Michael J. SHAFFER, An Argument for the Safety Condition on Knowledge.. 517

Notes on the Contributors…………………………………………………... 521
Logos and Episteme. Aims and Scope………………………………………...... 523
Notes to Contributors……………………………………………………...... 525

RESEARCH ARTICLES

IF YOU BELIEVE YOU BELIEVE, YOU BELIEVE. A CONSTITUTIVE ACCOUNT OF KNOWLEDGE OF ONE'S OWN BELIEFS

Peter BAUMANN

ABSTRACT: Can I be wrong about my own beliefs? More precisely: Can I falsely believe that I believe that *p*? I argue that the answer is negative. This runs against what many philosophers and psychologists have traditionally thought and still think. I use a rather new kind of argument, – one that is based on considerations about Moore's paradox. It shows that if one believes that one believes that *p* then one believes that *p* – even though one can believe that *p* without believing that one believes that *p*.

KEYWORDS: self-knowledge, Moore's paradox, second-order beliefs

Can I be wrong about my own beliefs? More precisely: Can I falsely believe that I believe that *p*? Can I have a false second-order belief that I believe that *p* (where the belief that *p* is a first-order belief)? The question is whether a sentence of the following form can be true:

(1) S believes that he believes that *p*, but he does not believe that *p*.[1]

If all instantiations of the scheme (1) are false, then the following holds:

(2) If S believes that he believes that *p*, then he does believe that *p*.

In other words, all our second-order beliefs are true: $BBp \rightarrow Bp$.[2] This is the claim I will argue for.

However, *prima facie* it seems that it is possible to have a false second-order belief with the following content:

[1] For the sake of simplicity, I am not adding temporal indices except where clarity demands it. I assume here that S is attributing a belief to herself as a present one, not a past or future one.

[2] "B*p*" stands for "S believes that *p*." The scope of "B" is the narrowest possible one: B(B*p*) and B(*p*). I will omit parentheses in the following. The claim that $BBp \rightarrow Bp$ is (like some other claims here) one of necessity but I won't mention this below, just for the sake of simplicity.

(3) I believe that *p*.[3]

Why should the fact that someone believes something of the form of (3) entail anything about the truth of that belief? This idea runs against what many philosophers and psychologists have traditionally thought and still think.[4]

I will use a rather new kind of argument for the main thesis here, – one that involves considerations about Moore's paradox and amounts to a constitutive view of self-knowledge of one's beliefs.[5] The main argument will be developed in

[3] Here, (3) is meant as a report of a belief state, not as its expression (cf. Ludwig Wittgenstein, *Philosophical Investigations* (2.ed.) (Oxford: Blackwell, 1958), pp.190-192).

[4] See, e.g., among the psychologists: Richard Nisbett and Timothy D. Wilson, "Telling More than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* 84 (1977): 231-259; Richard Nisbett and Lee Ross, *Human Inference: Strategies and Shortcomings of Social Judgment*, (Englewood Cliffs: Prentice Hall, 1980), 195ff.; Timothy D. Wilson, "Strangers to Ourselves: The Origins and Accuracy of Beliefs about One's Own Metnal States," in *Attribution. Basic Issues and Applications*, eds. John H. Harvey and Gifford Weary (Orlando: Academic Press, 1985), 9-36; Alison Gopnik, "How We Know Our Minds: The Illusion of First-Person Knowledge of Intentionality," *Behavioral and Brain Sciences* 16 (1993): 9ff.; Daryl J. Bem, *Beliefs, Attitudes, and Human Affairs* (Belmont, CA: Brooks/ Cole, 1970); for the "anti-Cartesian" attitude against another transparency thesis see amongst philosophers Timothy Williamson, *Knowledge and its Limits* (Oxford: Oxford University Press, 2000), ch.4.

[5] In a very general way, I am inspired by a paper by Sydney Shoemaker (see his "Moore's Paradox and Self Knowledge," *Philosophical Studies* 77 (1995): 211-228, his "Moore's Paradox and Self-Knowledge," in Sydney Shoemaker, *The First Person Perspective and Other Essays* (Cambridge: Cambridge University Press, 1996), and also his "On Knowing One's Own Mind," *Philosophical Perspectives* 2 (1988): 183-209; see also David M. Rosenthal, "Self-Knowledge and Moore's Paradox," *Philosophical Studies* 77 (1995): 195-209, and Rogers Albritton, "Comments on Moore's Paradox and Self-Knowledge," *Philosophical Studies* 77 (1995): 229-239). He mainly argues that if S believes that *p*, then S believes or even knows he believes that *p*. I, however, argue for the converse claim (see also Byeong D. Lee, "Moore's Paradox and Self-Ascribed Belief," *Erkenntnis* 55 (2001): 359-370). Furthermore, Shoemaker's thesis is restricted to the case of rational people. Shoemaker's "Moore's Paradox" (1995): 225-226 also makes the converse claim, but much more tentatively, and with restriction to rational people (see with even more reservations, his "Moore's Paradox" (1996), 92, 93); the argument presented here does not rely on ideas about rationality. For a similar approach see Tyler Burge, "Our Entitlement to Self-Knowledge," *Proceedings of the Aristotelian Society* 96 (1996): 91-116. The thesis that $BBp \rightarrow Bp$ is much stronger than Burge's earlier claim that "Cartesian" thoughts of the form "I am thinking the thought that water is wet" are always true (see his "Individualism and Self-Knowledge," *The Journal of Philosophy* 85 (1988): 649-663). Jaakko Hintikka, *Knowledge and Belief. An Introduction to the Logic of the Two Notions* (Ithaca, NY: Cornell University Press, 1962), 123ff. goes into a similar direction as I do here. He, however, does not rely upon Moore's paradox (even though he has a lot to say about it; see Hintikka, *Knowledge and Belief*, 64ff.). For more recent constitutive views which differ considerably from the proposal here see, e.g., Richard Moran,

sections 2-4. Section 5 discusses objections. But first I need to say more about the notion of belief and related notions in order to clarify the main thesis and set the stage.

1. Beliefs

I take beliefs to be dispositional mental states that can be both manifest and latent, – dispositions for occurrent thought and more indirectly also for behavior based on such occurrent thoughts.[6] Beliefs do not always express themselves in occurrent thoughts. In my dreamless sleep I still believe that 2+2=4 even though I am sleeping and not thinking at all about numbers. One of the characteristics that distinguish beliefs from other mental states is a specific relation to truth: Their contents are held true by the subject. Desires and other mental states are different in that respect. Beliefs are "cognitive" in this sense; one could also say that a belief is a cognitive attitude towards some content.

In the case of a self-attributing second-order belief the notion of "I" lies within the scope of the second-order belief; it is not sufficient for such beliefs to attribute a belief to someone who happens to be me if I don't think of that person as myself. We are dealing with *de dicto*-beliefs about oneself here,[7] not with *de re*-beliefs. Similarly, the notion of a belief, too, lies within the scope of the second-order belief. If somebody ascribes a belief to herself, then she must be clear about the type of attitude she ascribes to herself. One cannot, for conceptual reasons, believe that (3) is true of oneself and not believe it is a *belief* (that *p*) that one has here. Believing the latter presupposes that one possesses the concept of a belief and that one knows certain basic things about beliefs. One need not have a psychological theory of belief but one needs to know, say, that there is a difference between beliefs and other kinds of attitudes (like desires, for instance). If one does not know these basic things then one does not possess the concept of a belief and thus cannot have second-order (*de dicto*) beliefs.[8] All this will become important below.

Authority and Estrangement. An Essay on Self-Knowledge (Princeton: Princeton University Press, 2001), Fordi Fernández, "Self-Knowledge, Rationality and Moore's Paradox," *Philosophy and Phenomenological Research* 71 (2005): 533-556, and Mathiey Doucet, "Can We Be Self-Deceived about What We Believe? Self-Knowledge, Self-Deception, and Rational Agency," *European Journal of Philosophy* 20 (2012): E1-25.

[6] If not indicated otherwise, I will use "thought" for "occurrent thought."

[7] One could add: with *de se*-beliefs (see David Lewis, "Attitudes de Dicto and de Se," *Philosophical Review* 88 (1979): 513-543).

[8] This holds even given an externalist account of mental or semantic content.

To avoid misunderstandings: By "second-order beliefs" I do not mean beliefs that one does not have a certain first-order belief. It is certainly possible for people to have repressed beliefs that they think they don't have. Jack might just laugh at the thought that his parents abandoned him when he was 4 years old but psychotherapy might uncover that he has a repressed belief that this was indeed the case. This example is of the following form:

> S believes that he does not believe that *p*, but he does believe that *p*.

This is certainly possible but I am not dealing with this case here (see also section 5.1 below).[9]

2. The Argument: First Part

Suppose that

> (4) S believes at t-1 that he believes that *p*.[10]

What does this entail?
Dispositional beliefs are often latent. However, there is a condition on dispositional beliefs which seems very plausible:

> (5) If S believes at t-1 that *p*, then S manifests that belief as an occurrent belief at t* (which is either before or at t-1).[11]

A few remarks on (5) are necessary before we can make the next step in the argument. - The idea behind (5) is that one cannot believe, say, that dogs bark without ever manifesting that belief up to then, that is, without ever occurrently thinking and holding true (up to then) that dogs bark. It doesn't matter whether the thought is a conscious or an unconscious, or even a "Freudian" one (where a thought is "unconscious" in case the person is not aware of having the thought, and "Freudian" if the person cannot (easily) become aware of it). I don't see any reason to deny that one can have unconscious occurrent thoughts; otherwise each thought

[9] See Shoemaker, "Moore's Paradox" (1995), 226; cf. against this Byeong D. Lee, "Shoemaker on Second-Order Belief and Self-Deception," *Dialogue* 41 (2002): 279-289.

[10] In (4) as well as in the antecedent of (5) "believes" is used in the full dispositional sense, covering both latent and manifest belief.

[11] A related reverse principle might seem more uncontroversial: If S has an occurrent belief at t that *p* then S has a dispositional belief at t that *p*. The dispositional belief might be as short-lived as an occurrent belief which comes and goes. But the disposition is still there as long as the occurrent thought is there. Something made the subject think that *p*, and if the same conditions were to hold again the subject would think again that *p* – even if, as a matter of fact, these conditions never come up again.

would be accompanied by the thought that one is having it.

To be sure: One can have a disposition to form a certain belief and for this disposition one does not need any antecedent manifest thought with the content of that belief. Jack might have never thought about the question whether zebras in the wild wear raincoats.[12] He might not have a belief that they don't (nor any alternative view on the matter). However, we can still assume that he has a disposition to form the belief that zebras don't wear raincoats in the wild; for instance, when asked whether they do he might well form such a belief (see Robert Audi's useful distinction between dispositional beliefs and dispositions to believe;[13] Audi, however, holds that one can have a dispositional belief without ever having had the corresponding occurrent belief). Another example: Someone can have a dispositional belief that he is 6 feet tall. Even given knowledge that 6 feet is much less than 12 miles, it does not follow that the subject has a dispositional belief that he is less than 12 miles tall. However, we might very well need to ascribe a disposition to form the relevant belief to the subject. If we gave up on the distinction between dispositions to believe and dispositional beliefs we would get an "inflation of beliefs" and would have to attribute implausibly many beliefs to subjects. For instance, there are many propositions which a person does not accept in the present but will come to accept in the future; this is a basic fact of life. If, as seems very plausible, the future acceptance of some proposition results from the triggering of a relevant prior disposition, and if that disposition is not a disposition to believe but rather a dispositional belief then the person would already count as believing a proposition way before they accept it. This seems very odd and strongly suggests that we need to distinguish between dispositions to believe and dispositional beliefs.[14] Here is another way to mark the difference. Dispositional beliefs are first-order dispositions of thought and subsequent behavior while dispositions to believe are second-order dispositions to develop and

[12] See Daniel Dennett, "A Cure for the Common Code?" in his *Brainstorms. Philosophical Essays on Mind and Psychology* (Cambridge, MA: MIT Press, 1978), 90-108, especially 104; I am using Dennett's example contrary to his own purposes.

[13] See Robert Audi, "Dispositional Beliefs and Dispositions to Believe," *Noûs* 28 (1994): 419-434, and especially 420-421.

[14] Here is one more example. Does a newborn baby believe that there is no greatest prime? If there is no difference between dispositional beliefs and dispositions to believe then it won't be easy to deny this kind of belief to newborns. Sure, the baby doesn't have the notion of a number yet but given the right circumstances (including normal development) it will acquire it plus the belief that there is no greatest prime. So, we need to draw a line somewhere between dispositional beliefs and dispositions to believe – whether we use these expressions or others. The claim (5) above formulates a very plausible criterion to that effect.

form such first-order dispositional beliefs. This distinction is very important und useful in the case of cognitive attitudes. The first-order disposition, the dispositional belief, is a cognitive attitude towards some content while the second-order disposition to form such a belief does not involve any cognitive attitude towards that content (though, perhaps, cognitive attitudes towards other contents).[15]

But isn't it possible that a belief manifests itself in action but not in thought? Can't there be "unthought beliefs" driving our behavior? I don't think so. First, as explained above, I take beliefs to be cognitive attitudes and states. A merely behavioral disposition to act or behave as if *p* without there being or having been any kind of occurrent thought that *p* in the subject's mind is not a cognitive state. It is therefore not clear at all why one should call such a state a "belief" (one can certainly redefine terms as one likes but in this case this would be misleading). Second, as pointed out above, an occurrent belief need not be conscious: One can have it without being aware of having it. Hence, that there has been no conscious occurrent belief at t* does not mean that there hasn't been any occurrent belief at t*. One should not mistake the presence of an unconscious (occurrent) belief for the lack of (occurrent) belief altogether.[16] Third, as pointed out above, one needs to take the distinction between dispositions to believe and dispositional beliefs very seriously; a mere disposition to form some belief that *p* does not constitute a belief that *p*. Finally, even if one still has doubts about (5) one should keep in mind that I am only dealing with the application of (5) to the case of second-order beliefs here (see below). Even if one could make sense of "unthought belief" in general, it would still be very hard to imagine how it should be possible that a second-order belief can express itself in action but not in thought. Can Jack believe that he has the belief that he is good looking but never manifest that second-order belief in thought but only in action? In what kind of action? I doubt there are any "real life"-explanations of behavior in terms of unthought higher-order beliefs. To assume that there can be unthought second-order beliefs (in contrast to unthought first-order beliefs) thus seems very hard to justify.

Now, (4) and (5) entail

[15] Interestingly, there does not seem to be that much of a place for this distinction in the case of non-cognitive dispositions. The behavior of sugar cubes in water can be very usefully described and, perhaps, even explained in terms of the first-order disposition of water-solubility of sugar; there does not seem to be much need for notions of second-order dispositions here; it seems talk about second-order dispositions can be easily replaced here by talk about first-order dispositions.

[16] That all occurrent beliefs are conscious is a very strong claim anyway. One might wonder whether there is a threat of an infinite regress here (is the subject conscious that they are conscious that…?). Apart from that: What counts in favor of such strong "Cartesianism"?

(6) S occurrently believes at t* that he believes that *p*.

A necessary condition for belief becomes important now: a condition sometimes called a condition of "minimal rationality."[17] Here is an example.[18] Someone who thinks that McKinley has been assassinated cannot ignore (that is, not spend any thought on or be unaware of) the question whether McKinley is dead. This does not mean that the subject needs to have the notion of a question or must ask himself the question explicitly whether McKinley is dead ("Hey, is McKinley dead?"). Rather, the subject must somehow be aware of the issue whether McKinley is dead. In the case of this example, the subject must also grasp and agree with the idea that McKinley is dead. Otherwise, he does not even count as someone who believes that McKinley has been assassinated. For analogous reasons, I want to argue in a moment, someone who thinks and believes that he believes that *p*, cannot ignore and must be aware of the question whether *p*. He must have some thought (though not necessarily a positive view one way or another) about that question, too. Otherwise, he wouldn't even count as someone who believes that he believes that *p*. He just wouldn't be grasping the concept of a belief. But the latter is, as we have seen above, necessary for having second-order beliefs.

Thinking that I believe that *p* is a way of thinking about *p;* here one thinks about *p* as something towards which one can have certain attitudes like belief or disbelief. It is thus *ipso facto* a way of being aware of the question whether *p*, given that such attitudes can only be thought of as ways of settling questions.[19] To put it differently: Thinking that I believe that *p* is about the particular propositional attitude of belief. One can think of belief only as something that settles a question. Thus one is then *ipso facto* aware of the question. In other words: One doesn't ignore the question. Therefore, one cannot think that one believes that *p* without thinking in some form that there is a question as to whether *p*; without the latter one's own thought would not be intelligible to oneself (per impossibile). Sure, one does not have to have a conscious belief of a form like "There is the question as to

[17] See, e.g., Christopher Cherniak, *Minimal Rationality* (Cambridge, MA: MIT Press, 1986). To call it that can be a bit misleading. A subject who fails to meet that kind of condition fails to have or fails to be able to have certain concepts and beliefs. However this does not mean that such a subject is irrational (not even minimally) but simply that it does not have what one needs in order to master a given concept and entertain certain beliefs involving those concepts.

[18] See Stephen Stich, *From Folk Psychology to Cognitive Science: The Case against Belief* (Cambridge, MA: MIT Press, 1983), 56.

[19] Settling a question can also be very easy and even trivial, like, e.g., in the case of 1 plus 1 equaling 2.

whether *p* and I have answered it by endorsing *p* so that I believe that *p*" when one thinks that one believes that *p*. Similarly, one does not have to have a conscious belief of a form like "McKinley is dead" when one thinks that McKinley has been assassinated. But as someone who thinks that McKinley has been assassinated cannot and does not ignore (is aware of) the question whether he is dead, so someone who thinks that he believes that *p* cannot and does not ignore (is aware of) the question as to whether *p*. It is important to stress that awareness comes in degrees; a subject need not be maximally aware of a question in order to be aware of it. Something does not need to be in the central focus of one's awareness; it could be closer to the periphery of awareness but the subject would then still count as being aware of it.

Hence, given (6) and given that having beliefs requires this kind of awareness we have to accept

> (7) S does not at t* ignore the question whether *p*.

Since S has a thought (even if it is only the thought or awareness that there is a question here; thoughts need not take the form of explicit deliberation) about the question whether *p* insofar as he thinks that he believes that *p*, we can put (6) and (7) together and say that

> (8) S occurrently believes at t* that he believes that *p*; this involves thoughts about the question whether *p*.

If one does not ignore or if one is aware of a question, then one has thoughts about it.

But does S, in addition, have to believe that *p*? What kinds of attitudes can S have towards *p* when he occurrently believes that he believes that *p* and is aware of the question whether *p*? More precisely: What kinds of attitudes can S have towards *p* when he is aware of and has thoughts about the question whether *p*? There are exactly three options, three stances the subject can take or have (with no other alternative option):

> i. the positive stance: to believe occurrently that *p*,
>
> ii. the negative stance: to believe occurrently that not *p*, or
>
> iii. the indifference stance: to leave it open (occurrently) whether *p*.

Leaving it open is a residual category here. It involves everything between simple lack of a positive view one way or the other about (but still involving awareness of the question) whether *p* on the one hand and explicit suspension of belief about whether *p* on the other hand.[20]

Hence, S – who occurrently believes at t* that he believes that *p* – can only be in one of the following three situations (at that time):

> I. occurrently believe at t* that he believes that *p* and occurrently believe at t* that *p*,
>
> II. occurrently believe at t* that he believes that *p* and occurrently believe at t* that not *p*, or
>
> III. occurrently believe at t* that he believes that *p* and (occurrently) leave it open at t* whether *p* (while being aware of the question whether *p*).

It appears then that either S has pairs of thoughts here - a second-order and a first-order thought – or S has a pair of a second-order thought and an indifference stance on the content of the relevant first-order belief. Let us look at pairs of thoughts, first (I, II). Since S has the thought about the question whether *p* in the context of the second-order thought that he believes that *p*, it is plausible also to ascribe a single complex conjunctive thought to S. The following (schema of a) conjunction principle is plausible:[21]

> (Conj-1) If S believes occurrently that he believes that *p*, is aware of the question whether *p*, and takes a positive or negative stance on *p*, then S has an occurrent belief in the conjunction of the contents of the second-order occurrent belief and of his positive or negative stance on *p*.[22]

S thus is in one of the following three situations (brackets indicating the content of the thought): He

[20] Is the indifference stance a second-order attitude or does it involve a second-order belief that one does not have a first-order belief about *p*? One may call such a second-order belief an "indifference stance" but what I have in mind here need not be and usually isn't of the second order, like the state of being epistemically indecisive about whether *p*. Young children or non-human animals might not be able to have higher order attitudes because they lack concepts like *belief* but they can be indecisive about something. Also, one can be indecisive concerning options for choice even if one does not assume a higher-order attitude; how then could there not be a parallel in the case of belief?

[21] See, e.g., Simon Evnine, *Epistemic Dimensions of Personhood* (Oxford: Oxford University Press, 2008), chs. 3-4 for a defense of such principles.

[22] One can argue that (Conj-1) as well as similar conjunction principles hold for dispositional, latent as well as manifest, beliefs, including also unconscious or even "Freudian" beliefs; however, I cannot go into the details of the different cases here. – Other conjunction principles are simpler and more straightforward but false, like the following one: If S believes that *p* and also believes that *q*, then S believes that (*p and q*). More plausible is a principle of conjunction elimination: If S believes that (*p* and *q*), then S believes that *p* and S believes that *q*.

> 1. occurrently believes at t* that (*p* and he believes that *p*),
>
> 2. occurrently believes at t* that (not *p* and he believes that *p*), or
>
> III. occurrently believes at t* that he believes that *p* and leaves it open whether *p* (while being aware of the question whether *p*).

Now, what about III? It makes the antecedent of (Conj-1) false and the principle irrelevant here. But is there, perhaps, another conjunction principle for pairs of beliefs and indifference stances? An indifference stance towards some proposition *p* does, of course, not express itself in the belief that *p* or the belief that not *p*. Hence, there is no direct parallel to 1. and 2. above. It would be nonsense to say something like the following:

> A subject in condition III occurrently believes at t* that (*???* and he believes that *p*).

However, there is a less direct parallel to 1. and 2. above which is direct enough for our purposes here:

> (Conj-2) If S believes occurrently that he believes that *p*, is aware of the question whether *p*, and takes an indifference stance on *p*, then S has a conjunctive occurrent belief of the form "I believe that *p* but the question whether *p* is unsettled for me".[23]

Hence, we also get from III and (Conj-2) to the claim that S

> 3. occurrently believes at t* that (it is unsettled whether *p* and he believes that *p*).

So, from I, II, III plus both conjunction principles we get to the conclusion that our subject can only be in condition 1., 2. or 3. Now, for conceptual reasons S can only be in situation 1. Why? Let us take situation 2 first. Someone who were in that situation would occurrently believe at t* something of the form "I believe that *p*, but not *p*." This is a Moore-paradoxical thought.[24] The problem with that[25] is not

[23] Again: Awareness admits of degrees and something can be more or less in the focus of one's awareness.

[24] See George Edward Moore, "Russell's Theory of Descriptions," in his *Philosophical Papers* (London/ New York: Allen & Unwin/ Macmillan, 1959), 151-195, especially 175-176.

[25] I am only using commissive versions of Moore's paradox here (of the form "I believe that *p* but not *p*"); ommissive versions ("*P* but I don't believe it") are irrelevant to my argument. See also section 5.1 below. – I am leaving aside uses of such phrases by eliminativists about belief: "*p* but since there is no such thing as belief I don't believe it!" (see, e.g., Paul M. Churchland, "Eliminative Materialism and the Propositional Attitudes," *The Journal of Philosophy* 78 (1981): 67-90). Eliminativism about belief is not incoherent or Moore-paradoxical. A Moore-paradoxical thought or utterance presupposes that there are beliefs while eliminativists about belief deny that and are not even in a position to make a truly Moore-paradoxical statement. – Even though

just that it would be an incoherent thought.[26] No, one cannot even have such a thought.[27] Why not? Here we can use an important remark by Wittgenstein:[28] One can mistrust one's senses but one cannot mistrust one's own belief.[29] Having a belief that *p* is incompatible with having a second-order attitude of mistrust towards that belief: for instance, holding that the belief that *p* is false (or suspending judgment on the question whether it is true: see below). However, a subject who is in situation 2 above and thinks and believes something of the form "I believe that *p*, but not *p*" would have such a second-order attitude of mistrust towards his belief. Since the latter is not possible, the former is also not possible (both for conceptual reasons). If someone said or thought something of the form "I believe that *p*, but not *p*," then whatever he was ascribing to himself it couldn't be a belief; hence, he couldn't thereby express or manifest a second-order (manifest) belief; his use of the word "belief" (his attempt at tokening of concept of belief) would show that he hasn't yet mastered the concept of belief. In other words, we can exclude situation 2 as impossible here.[30]

Moore's paradox is often discussed as a problem of assertion, it has long (even before Wittgenstein, *Investigations*) been recognized that the same problem arises for thought not expressed linguistically.

[26] – or an absurd thought; see Uriah Kriegel, "Moore's Paradox and the Structure of Conscious Belief," *Erkenntnis* 61 (2004): 99-121.

[27] Shoemaker, "Moore's Paradox" (1995), sec. IV and Shoemaker, "Moore's Paradox" (1996), sec.IV agree but for different reasons than those presented here.

[28] See Wittgenstein, *Investigations*, 190; more on that also below.

[29] But cf. also Béla Szabados, "Wittgenstein on Mistrusting One's Own Belief," *Canadian Journal of Philosophy* 11 (1981): 603-612.

[30] One could speculate about the possibility of a "division" of the mind and consider cases where one "sub-subject" disagrees with what another "sub-subject" believes. One sub-subject might hold that it believes that *p*, while the other sub-subject might hold that *not p* (compare this to Wittgenstein *Investigations*, 192: "If I listened to the words of my mouth, I might say that someone else was speaking out of my mouth"). However, such deviant cases can be left aside here: They are cases of "split minds" and not ordinary cases of self-attributing beliefs which are topical here. But one might object to this that it makes sense to say something like "Spiders are harmless but when I think about my behavior when I'm near a spider I come to the conclusion that I still believe that spiders are not harmless" or, shorter, "Spiders are harmless but I still believe they aren't"? This makes some sense but it is crucial, again, to acknowledge that such a subject identifies only with part of her mind and treats her behavior as if it were someone else's. Strictly speaking, such a belief is not the subject's belief but the belief (if one may use this word here) of some sub-personal agent or module (see for this also Stephen Stich, "Beliefs and Subdoxastic States," *Philosophy of Science* 45 (1978): 499-518); the fact that some sub-personal agent or module holds a belief (or an attitude like that) does not entail that the person herself holds that belief (compare this with the bad inference from the claim that a particular group

To be sure, this does not mean or imply that one couldn't be less than perfectly confident in one's beliefs (have an intermediate degree of belief), imagine the possibility of being wrong,[31] see one's evidence for one's belief as imperfect, etc. However, all this does not amount to mistrusting one's beliefs.

All this entails that the verb "falsely believe that *p*" has no use in the first-person, present tense.[32] As we will see, only the verb "truly believe that *p*" does. This might explain why we usually skip the qualification "truly" when self-attributing beliefs. False beliefs that oneself might have are "blindspots"[33] in the sense that they are not self-attributable as false beliefs. Having a false belief is an essentially "intransparent" condition insofar as the person cannot know or even believe that he is in this condition while he is in it.

For reasons analogous to the ones above, the situation 3 also turns out to be impossible. Someone who were in that condition would occurrently believe at t* something of the form: "I believe that *p* but it is unsettled whether *p*." This also expresses an attitude of mistrust towards one's own belief that *p* (though a softer one). But one cannot take such an attitude of mistrust towards one's own beliefs. Hence, we can – for similar reasons to the ones concerning situation 2 – exclude situation 3 as impossible (also for conceptual reasons).

But if cases 2 and 3 are excluded as impossible, then only case 1 remains – and there is nothing problematic or incoherent about that one. Hence, we can conclude from the above remarks about cases 1-3 and (8) that

> (9) If S occurrently believes at t* that he believes that *p*, then S occurrently believes at t* that he believes that *p*, and *p* (in the sense of "*p* and I believe that *p*").

There is a plausible principle of distribution of belief over conjunction:[34]

> (Dist) If S believes that (*p and q*), then S believes that *p*.

Given (Dist) we can move from (9) to

> (10) If S occurrently believes at t* that he believes that *p*, then S occurrently believes at t* that *p*.

member holds a certain belief to the claim that the group holds that belief or view).

[31] Concessive self-attributions of beliefs ("I believe it's going to rain but I could, of course, be wrong about that") do not constitute cases of mistrust of one's belief: the confidence that one is right can still be quite firm.

[32] See Wittgenstein, *Investigations*, 190.

[33] See also, more generally, Roy A. Sorensen, *Blindspots* (Oxford: Clarendon, 1988).

[34] See, e.g., John N. Williams, "Wittgenstein, Moorean Absurdity and its Disappearance from Speech," *Synthese* 149 (2006): 225-254, especially sec.7.

3. More on Mistrusting One's Beliefs

Before we move on with the argument, some more remarks about why one cannot mistrust one's own beliefs seem useful. Let us look at the clearest case of a Moore paradoxical belief (similar arguments can be made for other forms of Moore-paradoxicality, like holding that one believes that *p* but suspending judgment about whether *p*): the (alleged) belief that

> (MP) Not *p*, but I believe that *p*.

Why can one not believe something of the form (MP)?

I can mistrust another person's belief. In such a case I hold, so to speak, my belief against another person's belief (or against what I take to be her belief). I compare them and if there is disagreement (given that I can see no reason to revise my own belief), I go with my own belief. Why with mine? Well, that is what it means to have a belief: One goes with it (against alternative beliefs, given that the fact of disagreement or related facts do not itself give one a reason to change one's belief). If I check other people's beliefs, I cannot but use my beliefs as the standard (even if I originally got my beliefs from others and even if I change my beliefs under the influence of other people's beliefs). Sometimes – like in the case we're focusing on here – I have an explicit belief about the relevant subject matter ("He thinks it's raining but it isn't"). But at other times I don't: there might just be a reluctance to judge the whole thing ("He thinks it's raining but that's not clear at all"). This reluctance, however, is also based on certain beliefs ("He doesn't know the weather conditions," "Conditions of perception are much too bad to judge this," etc.). In both cases, I have a belief or a set of beliefs that is the basis for mistrust towards another person's belief.[35]

I cannot do anything like that in my own case. I would have to treat myself as if I were not myself but another person. Given that in (MP) I would have to think of myself as myself ("I"), I would have to think of myself as myself and as another person. However, mastering the notion of oneself as well as the notion of others involves knowing that oneself is not another person different from oneself. Whoever says something to the effect of "I am not myself" is either not sincere or uses language in a special way or just documents that he has not mastered words like "I" and "someone else." Mastering such notions is a precondition of being able

[35] Couldn't a non-propositional mental state be the basis for my mistrust of someone's belief? Suppose sensations, for instance, are such non-propositional states. But in what sense could they be a basis for my mistrust if they don't lead to certain beliefs which then form the more immediate basis for my mistrust? This touches on a whole series of questions which cannot be pursued in further detail here.

to have second-order beliefs. Hence, one cannot have thoughts or beliefs of the form (MP). I cannot hold my own beliefs against my own beliefs as an (allegedly) independent standard;[36] I cannot mistrust my own beliefs. In other words, when I say that S holds something false true I imply (or implicate) that something that is acknowledged by me is not acknowledged by S. I am thus assuming that there is an epistemic asymmetry between me and S which explains why S is wrong and I am right. This only makes sense given the assumption that I am not S. Since I cannot take myself to be S (not me), I cannot apply the above asymmetry to my own case. In other words, I cannot take myself to hold something true that I think is false.[37]

Hence, the subject cannot mistrust his own beliefs and believe something of the form of

> (MP) Not *p*, but I believe that *p*.

Somebody who sincerely claims to believe such a thing only shows that he has not mastered the concept of belief. Even if one were to argue that he expresses some kind of second-order belief, it wouldn't and couldn't be one with the content (MP). Hence, (MP) cannot express a self-ascription of a belief. Not that it constitutes an irrational or defective self-ascriptions of a belief; rather, (MP) does not express any possible self-ascription of a belief at all.

4. The Argument: Second Part

Back to

> (10) If S occurrently believes at t* that he believes that *p*, then S occurrently believes at t* that *p*.

(10) is not our thesis, even though quite close: It does not say that (omitting the reference to a given point in time)

> (2) If S believes that he believes that *p*, then he does believe that *p*

– where "belief" is used in the wide sense including non-manifest, merely dispositional belief as well as occurrent belief. Can we generalize (10) to include situations in which S believes that he believes that *p* but only in a latent and non-manifest way? Can we generalize (10) such that it entails (2)?

First a brief reminder. Suppose that S, at t-2, believes (in a merely dispositional, that is, latent sense) that he believes that *p*. According to (5) as

[36] See Wittgenstein, *Investigations*, 190.

[37] Suppose I have changed my mind: Yesterday I believed that *p* but today I believe that not *p*. Then, I can say today I was wrong yesterday. However, this does, of course, not amount to saying that my present belief is false (see also the remarks above in section 1).

applied to this case and to some earlier time t-1, the following holds

> If S believes latently at t-2 that he believes that *p*, then S manifests that belief as an occurrent belief at some earlier time (which we can call "t-1" here).

Hence, given our assumptions about S here, S manifests the belief that he believes that *p* at t-1. We can now use the relation between dispositional, latent and manifest belief to show that (2) is true if (10) is.

The crucial point here is that if the occurrent belief of S at t-1 that he believes that *p* leads to the latent belief of S at t-2 that he believes that *p*, then the occurrent belief of S at t-1 that *p* – which comes with the corresponding occurrent second-order belief (see (10) above) – will also lead to the latent belief of S at t-2 that *p*. In other words, the same occurrent second-order belief at t-1 leads to both corresponding second- and the first-order latent beliefs at t-2. In a nutshell: The latent second-order belief can only arise from circumstances which also give rise to the corresponding latent first-order belief.[38] The first comes with the second (for a worry, see below).

Here is a different way to put it. A latent belief in some proposition is a disposition to, amongst other things, think that proposition (given certain triggering conditions). The latent belief that one believes that *p*, for instance, is a disposition to think that one believes that *p*. Since one cannot (see (10)) occurrently think that one believes that *p* without also occurrently thinking that *p*, the same disposition (given the relevant circumstances) triggers the thought that one believes that *p* as well as the thought that *p*. Hence, this very disposition is also a disposition to think that *p*. In other words, if S has a latent belief that he believes that *p*, then S also has a latent belief that *p*:

> (11) If S latently believes at t* that he believes that *p*, then S cannot but latently believe at t* that *p*.

Since beliefs are either manifest or latent, we can put (10) and (11) together and thus get our core thesis (again, skipping temporal indices for the sake of simplicity):

> (2) If S believes that he believes that *p*, then he does believe that *p*.

But, one might ask incredulously, isn't it possible that S, after t-1, continues to believe that he believes that *p* (though in a latent way) but loses the belief that *p*? Just stopping to think about *p* would not be sufficient for that: S has at t-1

[38] If the latent first-order belief already exist independently and antecedently, then there is overdetermination and the second-order belief merely "reconfirms" the first-order belief (see also the third-to-last paragraph in this section). This does not constitute a problem here.

acquired (or reconfirmed) the dispositional belief that *p*. As long as S doesn't change his mind about *p*, we can still attribute the belief that *p* to him. But couldn't S change his mind about *p*? And at the same time stick with his latent belief that he believes that *p*?

Not according to the view defended here. Suppose S changes his mind at t-2 about *p*: He now, e.g., comes to believe that not *p*. The critical assumption (for reduction) is that he still has, at t-2, his dispositional belief that he believes that *p*. Now, as we just saw: If at t-2 S has this belief, then he also has the dispositional belief that *p*. So, our situation would be rather one where the subject has inconsistent beliefs: one belief that not *p*, and another belief that *p* (this differs, of course, from the case of holding a belief that *p* and not *p*). It is not clear whether one can describe this as a change of mind but certainly S has thus not lost his belief that *p*.

But couldn't S change his mind at t-2 in a different way: not by acquiring the belief that not *p* but by simply losing the belief that *p*? Again, our assumption is that he still has, at t-2, the dispositional belief that he believes that *p*. Hence, he would (see above, again) also have the dispositional belief that *p*. The assumption that the subject has just dropped a belief thus leads to an inconsistency not of the beliefs of the subject but in the description of the subject's situation: as both having and not having the belief that *p*.

Our subject thus cannot be in a different mind about *p* without changing his mind about whether he believes that *p*. (2) remains standing and we can conclude that

$BBp \rightarrow Bp$.[39]

The argument for (2) has interesting consequences. If I assent (mentally or linguistically) to *"p*, and I believe that *p*", then I cannot detach the second conjunct and leave the first part behind, so to speak. Others can separate the two "parts" when they think or talk about me: They might think that I believe that *p*, but they need not hold that *p*. From the first-person perspective everything is different:

[39] See, though not quite in agreement with the argument above: Christopher Peacocke, *A Study of Concepts* (Cambridge, MA: MIT Press, 1992), 158, Tom Stoneham, "On Believing that I Am Thinking," *Proceedings of the Aristotelian Society* 98 (1998): 125-144, and U.T. Place, "The Infallibility of Our Knowledge of Our Own Beliefs," *Analysis* 31 (1970/71): 197-204; cf. against that Hugh Mellor, "Conscious Belief," *Proceedings of the Aristotelian Society* 78 (1977/78): 88-101, especially 91f. – I do not rule out that one can believe that not all of one's beliefs are true; the Preface Paradox is not Moore-paradoxical (though related). – The above argument works for full belief; I think a similar argument (though much more complicated in detail) can be made for degrees of belief but I will not attempt this here.

From this perspective "*p*, and I believe that *p*" is not a normal conjunction insofar as I cannot infer "I believe that *p*" from it without committing to *p* at the same time.[40]

A second-order belief is "constitutive" of the corresponding first-order belief. "Constitutive" is meant in a conceptual sense here, not in an empirical sense. The second-order belief brings with it, "involves" the first-order belief, and all that for conceptual reasons.[41] If S at t acquires a second-order belief that he believes that *p* and if S did not, before t, believe that *p*, then she acquires the belief that *p* just because she acquires the second-order belief that she believes that *p*. She might, of course, already believe that *p* before her acquisition of the relevant second-order belief, then think about whether or not *p* and about her views on whether or not *p*, and thus finally come to acquire her second-order belief that she believes that *p*. In this case, the second-order belief does not create the first-order belief but rather "reconfirms" it (see fn.38). It is also possible that the acquisition of a second-order belief creates an inconsistent mind set. Suppose S believes that not *p*. Suppose also that she somehow acquires the belief that she believes that *p* (a clever psychiatrist might convince her that he does). Then she thereby also acquires the belief that *p* – which is inconsistent with her belief that not *p*.[42]

I have only argued for a conditional thesis here (2). Hence, insofar as (2) leaves it open whether we do indeed have second-order beliefs, it is also left open whether we have any true beliefs about our own beliefs. It is left open whether we have any self-knowledge about our own beliefs. Now, it might be the case that one cannot have beliefs without having at least *some* second-order beliefs; that we have first-order beliefs would entail that we also have second-order beliefs. However, I want to leave *that* open here.[43] I would rather assume that, as a matter of fact, we often do have second-order beliefs (whatever the explanation of this fact is). Given what I have just said, this would entail that we are indeed right

[40] See André Gallois, *The World without, the Mind within. An Essay on First-Person Authority* (Cambridge: Cambridge University Press, 1996), 5-7, 46, passim who argues that questions about *p* and questions about my beliefs about *p* are not separate when raised from the perspective of the first person.

[41] See Crispin Wright, "Wittgenstein's Later Philosophy of Mind: Sensation, Privacy, and Intention," *The Journal of Philosophy* 86 (1989): 622-634; Jane Heal, "On First-Person Authority," *Proceedings of the Aristotelian Society* 102 (2002): 1-19.

[42] See Derek Bolton, "Self-Knowledge, Error and Disorder," in *Mental Simulation: Evaluations and Applications*, eds. Martin Davies and Tony Stone (Oxford: Blackwell, 1995), 209-234, and Shoemaker, "Moore's Paradox" (1996), 89-91. See also section 5.5. below.

[43] See, e.g., Donald Davidson, "Rational Animals," *Dialectica* 36 (1982): 317-327.

about our own beliefs when we think about it.[44]

No person needs to know or have true beliefs about all her present first-order beliefs, perhaps not even about any of them. Even if error, the presence of a false belief, about one's own beliefs is impossible, ignorance, the absence of a true belief, is still possible. If S has the belief that p he need not have the second-order belief that he believes that p. Not: Bp → BBp.[45]

5. Objections

Finally, I would like to consider and reply to some objections.

5.1. Believing that One Doesn’t Believe

Let’s start with what is perhaps the most serious objection I can think of. As I have already mentioned above (section 1), I am only dealing with beliefs that one has a belief that p (BBp), not with beliefs that one does not have a belief that p (B not Bp). I don’t see any convincing argument similar to the one above that would support the following claim: B (not Bp) → not Bp. Such a claim would easily lead to a contradiction. Suppose the subject has a suppressed and unconscious belief that p (Bp). Somebody (a clever psychoanalyst for example) convinces her that she does not believe that p (B not Bp). If “B (not Bp) → not Bp” were true, a contradiction would follow: Bp & not Bp.

But isn’t there an argument for “B (not Bp) → not Bp” which is parallel to the one above for “BBp → Bp”? And if the former leads to a contradiction, how then can we still hold on to the latter? Here is the idea (see sections 2-4 above for the details of the parallel). If S believes at t-1 that he doesn’t believe that p, then he believes occurrently at some earlier time t* that he doesn’t believe that p, is aware of the question whether p and thus has some thought and stance about whether p. Given that S can only have the three stances towards p mentioned above, S must

[44] How can I move from the claim that our second-order beliefs are true to the claim that they constitute knowledge? Couldn’t some true second-order beliefs fail to be knowledge? I don’t see how this should be possible – given the type of argument above. However, I need not go into this here because the core claim here is about the truth of our second-order beliefs.

[45] This allows for “Freudian” cases of “repressed” and inaccessible beliefs. – Interestingly, in the *Discours de la Méthode* Descartes – who is sometimes taken as defending the very strong thesis that BBp ⌦ Bp – points out that believing one thing is independent from believing that one does believe that thing; hence people can be ignorant about their own beliefs and they can be wrong about their own beliefs (see René Descartes, *Discours de la Méthode*, in René Descartes, *Oeuvres de Descartes*, eds. Charles Adam and Paul Tannery (Paris: Cerf, 1907-1913), vol. VI, 1-78, especially 23).

be in one of the following three situations:

> I*. occurrently believe at t* that he doesn't believe that *p* and occurrently believe at t* that *p*,
>
> II*. occurrently believe at t* that he doesn't believe that *p* and occurrently believe at t* that not *p*, or
>
> III*. occurrently believe at t* that he doesn't believe that *p* and (occurrently) leave it open at t* whether *p*.

And if the above conjunction principles (Conj-1) and (Conj-2) are plausible, then the following principle would seem plausible, too:

> (Conj-3) If S believes occurrently that he doesn't believe that *p*, is aware of the question whether *p*, and takes a positive or negative stance on *p*, then S has an occurrent belief in the conjunction of the contents of the second-order belief and of his stance on *p*.[46]

As applied to I*, it follows that the subject thinks (something of the form) that

> *p* but I don't believe it.

Given the the alleged parallel to the argument from Moore-paradoxes above, we would have to exclude situation I* as impossible. Only II* and III* would remain, both situations where the subject doesn't believe that *p*.[47] Hence, we have to conclude (in parallel to sections 2-4 above) that B(not B*p)* → not B*p*. And this would get us back into the contradiction above – which would be extremely bad. Given that the argument for *BBp → Bp* is strictly parallel, we should also drop the latter.

But this parallel argument for *B(not Bp) → not Bp* does not work. Why not? Why should there be such an asymmetry? The crucial point is that "*p* but I don't believe it" (in the sense of "I lack the belief that *p*," not of "I believe that not *p*") is Moore-paradoxical but it doesn't constitute a case of mistrusting one's beliefs. Not

[46] The indifference stance would require a different conjunction principle (one parallel to Conj-2):
(Conj-4) If S believes occurrently that he doesn't believe that *p*, is aware of the question whether *p*, and takes an indifference stance on *p*, then S has a conjunctive occurrent belief of the form "I don't believe that *p* and the question whether *p* is unsettled for me."

[47] Leaving something open entails the lack of a belief about the matter. So, III* is a case where S doesn't believe that *p*. One might suspect that case II* has two subcases: one in which S believes occurrently that not *p* without believing occurrently that *p* (II*a), and one in which S believes occurrently that not *p* while also (inconsistently) believing occurrently that *p* (II*b). Doesn't case II*b show that the claim in the text above that in both II* and III* the subject doesn't believe that *p*? No: II*b is ruled out as impossible for the same reasons for which I* is ruled out.

all Moore-paradoxical cases are one's of mistrust towards one's beliefs. Only the commissive cases ("I believe that *p* but not *p*") but not the ommissive ones ("*P* but I don't believe it") are cases of mistrust. One can see the thought that "*p* but I don't believe it" as an admission of epistemic imperfection but not as mistrust of a given belief – there is no belief (that *p*) represented by the subject to itself so that it could be the target of mistrust by the subject. And I don't think it is impossible to think something like "I've won the lottery but I don't believe it" (also think, e.g., about eliminativists about "belief;" see fn.25). This is Moore-paradoxical and irrational but still possible to believe.[48]

5.2. Believing and Holding True

Here is a somewhat lighter problem. Haven't I neglected the difference between believing that *p* and holding-true that *p*? The first entails the second – belief being an attitude of holding true – but not vice versa – as the following shows. Suppose I do not understand some particular thing my friend, the quantum theorist, says, expressing her belief; it is some result in recent quantum physics. She believes that *q* but I do not understand what "q" means. Hence, I cannot believe that *q* (since belief presupposes understanding). But I have good evidence that my friend is speaking sincerely and is usually right about such topics in her field; hence, I have good evidence that what my friend believes about quantum theory is true. Hence, it seems that I might well come to hold the belief (their belief) that *q* true without understanding "*q*" and thus, without having the belief that *q*. I hold-true that *q* but I don't believe it (in the full sense of "believe"). Holding-true is a *de re*-attitude towards the relevant proposition, not a *de dicto*-attitude like belief.

Nothing I have said so far seems to exclude the possibility that someone could falsely believe he understands a sentence and grasps the proposition expressed by it. I might have simply forgotten that I don't understand what "q" means and believe that I do understand it when I don't. In such a case, I would not believe that *q* (because belief presupposes understanding) even if I still hold it true.

[48] To be sure, a thought like that is necessarily false: Given a distribution principle for belief like (Dist), a thinker who holds that "*p* but I don't believe it" also believes *p*; hence, the second conjunct ("I don't believe it") is false and thus also the whole conjunction. – If the subject reflects on her epistemic situation, then she will get from believing that she won the lottery to acknowledging and believing that she so believes. Then it would be very hard to see how she could believe both that she does and does not believe that she has won the lottery. But perhaps one can have beliefs with contradictory contents after all (I will leave this open here). And lacking this step of reflection, the subject could be under the illusion that she has no beliefs about the subject matter. This would be irrational or at least show limited rationality but it would certainly not be impossible.

But (falsely believing that I understand *q*) I might, someone could argue, falsely believe that I believe (and not just hold-true) that *q* when I only hold-true that *q*. In other words, it would seem possible that – contrary to (2) –

(1) S believes that he believes that *p*, but he does not believe that *p*.

My argument above would only have shown that a slightly weaker thesis is true:

(12) If S believes that he believes that *p*, then S holds-true that *p*.

How strong is this objection? First of all, one could simply block this objection early on by insisting that being able to grasp a proposition of the form "B*p*" entails being able to grasp the corresponding proposition *p*. Second, even if it should be possible to believe that one believes something one does in fact not grasp, (12) would still be a very interesting conclusion and almost as strong as (2). Finally, cases of holding-true without belief are exceptions and secondary cases. They are only possible because there are many other cases in which we understand what we hold true. It seems impossible not to have any beliefs and only hold things true; the reason is that holding true involves some belief (e.g., that something is true, etc.). Could there be a subject that holds more propositions true (without believing them) than he believes? What would the life of a subject be like who does not understand the majority or even a substantial portion of what he holds true? Even lacking an argument to the effect that this is impossible, such a scenario seems very unrealistic. So, even if we don't block the objection from the start, we can accept the modification but leave it aside as a secondary case and from now on only look at the standard case of holding a belief true while understanding what one believes.

5.3. Belief and Reflection

Consider the following dialogue (assuming sincerity of the utterances):

February

A: What do you think: How many days does a month have?

B: I believe a month can either have 30 or 31 days!

A: What about February?

B: Oh yes, sorry! I do, of course, not really believe that every month has either 30 or 31 days! Sure, February has less. That's what I really believe!

An objector might point out that such a dialogue makes perfect sense. Doesn't it prove that in his first reply B was wrong about his own beliefs? I do not think so. B did, indeed, believe what he said then. This is compatible with what he says in his second reply because "really believing" obviously includes something like giving the first-order matter some amount of reflection – which he did not do in his first, spontaneous reply. He believed that months have either 30 or 31 days but he did not "really believe it" in the sense that he did not “believe it after due reflection.”[49] This leads to a further clarification of the main thesis: What I have said above concerns the unqualified simple sense of "belief" (concerning *p*) – the sense in which B believed what he said in his first reply. I am not claiming that second-order believing entails reflective first-order believing or that whoever believes that *p*, believes it on the basis of reflection. In other words, this kind of objection does no harm to the thesis I have defended here: B$Bp \rightarrow Bp$.[50]

Here is another example which points into the same direction:[51]

Conversion

> Jack was brought up in a very religious family; everyone he knows believes in God and has no doubts about it. Jack then moves to a big city in a different part of the country where he comes into contact with all kinds of people and all kinds of world views. Initially, he is quite shocked but over time gets used to it. In addition, he slowly loses his faith without even noticing it. One day somebody asks him whether he believes in God. Jack replies that he hasn't really thought about if for quite some time but then adds that, sure, he still believes in God. However, after some reflection he denies his first answer: "Sorry, I think I didn't give you a correct answer. I guess I don't believe in God any more."

Prima facie, this seems to be a case in which the person believes that he believes that *p* (that God exists) but does not really believe that *p*. In other words, her second-order belief would be false.

But again, there is a reply like the one to the February-example above. When Jack first answers (sincerely) that he believes in God he does indeed believe in God (given the argument defended here). He might have lost the belief in God before but – given the constitutive nature of second-order beliefs – he “gets it back” (for a very short time) when he acquires (or "reactivates") his second-order belief that he believes in God. As in the February-example above, this belief in God is an ordinary "simple" belief, not a "belief after due reflection (about the subject matter of that first-order belief)" or a "reflective belief" as we might call it. Then,

[49] See, for a related distinction, Dan Sperber, “Intuitive and Reflective Beliefs,” *Mind and Language* 12 (1997): 67-83.

[50] But doesn’t even B’s first reply require some reflection? Sure, but this is no objection. What matters is that there is a difference between more or less reflection. We do draw lines between “spontaneous” and “more reflective” responses.

[51] I owe this example to Hilary Kornblith.

after some thinking, he gives up his second-order belief that he believes in God. Since at this time there is no other basis for his belief in God than the second-order belief he is just giving up, the belief in God also goes over board. On top of that, he acquires the belief that he doesn't believe in God. This can be understood in two ways. First, he might simply acquire a second-order belief that he lacks the belief in God - without now believing something different, namely that there is no God. These second-order beliefs (of the form "B not B*p*") need not be true (see section 5.1) but in this case Jack's new second-order belief is true. Second, Jack might in addition acquire a second-order belief that he believes that God doesn't exist. Given our argument here, this entails that he does indeed believe that God doesn't exist. This first-order belief might be a simple, straightforward belief not based on reflection or a belief based on reflection - depending on whether it is only based on his second-order belief but not on reflection about God's existence or whether it is also based on such reflection.

What if Jack already believed that God doesn't exist when he was asked about it? In that case, he was initially not aware of his belief in God's non-existence. When he answered the question and acquired or re-activated his second-order belief that he believes that God exists he also acquired a second belief (a simple, first-order one) about God's existence: namely that He exists. For a short time, until he gave up this belief, he entertained two mutally contradictory beliefs: the belief that God doesn't exist and the belief that God exists. It is a controversial question whether it is possible to believe a contradiction but it is certainly possible to hold two beliefs which contradict each other. In Jack's case he reflected about things and the reflective belief "won" over the simple belief. The inconstency was only short-lived.

5.4. More on Reflection: Epistemic and Semantic

The distinction between beliefs based on reflection and beliefs not so based is quite important here. So, let me add a few more remarks on it. Take Jack's example, again. Was his answer to the question really about God? Jack might have thought along the following lines after his first answer: "Yes, God exists. Oh, wait a minute – what does "God" mean again? Right, now I remember, the creator of the universe who is maximally benevolent, omnisicent, and omnipotent. No, no, no, I don't think that a being like that exists." If this is what is going on when Jack thinks about the question more closely, then reflection is not just about the reasons he might have for a given belief but also about the concepts involved in that belief (the concept of God) or, in other words, about the meanings of the words expressing that belief (the meaning of "God"). The reflection Jack engages in after his answer might thus either be more of an epistemic nature (only concerned with reasons for a given belief) or more of a semantic nature. Usually, it will be a mixture of both.

One more remark on semantic reflection. Suppose Ernie and Bert are having a chat about math. Bert is asking Ernie whether he believes Goldbach's conjecture is true. Bert has just learned what that is and knows what he is talking about. Ernie first replies, "Sure, haven't you heard about this guy from New Jersey who's proved it?" Then comes the second thought: "Oh, no, wait, that was Fermat's theorem. Gosh, I have no idea. What do you think?" Should we say that Ernie was thinking and talking about Goldbach's conjecture in his first reply? If yes, then we would have to say similar things as in the cases above. If not, then in his first use (in his first reply) of the expression "Goldbach's conjecture" he did not refer to Goldbach's conjecture but rather to, say, Fermat's theorem. In that case, Ernie was not even thinking about and answering the question Bert asked him. In Jack's case semantic reflection led to a fuller understanding of key words whereas in Ernie's case it might have uncovered a simple misunderstanding of core expressions.

One must therefore be quite careful when using examples like *February*, *Conversion* or the Goldbach-example: Insofar as there is simply a change to topic involved between the first and second answer of the subject, nothing at all follows about the possibility of falsely believing one has a certain belief.

5.5. Belief and Behavior

One example that often comes up in discussions about second-order beliefs has to do with psychiatrists:[52]

> Psychiatrists
>
> Suppose Jill does not believe that her parents abandoned her for some time when she was three years old. Her psychiatrist is trying to convince her that she has a "repressed" belief that that was indeed the case. First, Jill rejects the idea: "No, I don't hold that belief." But after the psychiatrist points out some behavioral evidence to the contrary, Jill comes to accept what he says. She acquires the second-order belief that she believes that her parents abandoned her when she was three years old. Isn't this second-order belief just false?

Again, the answer is negative. According to the analysis proposed here, Jill acquires a first-order belief that her parents abandoned her when she was three by accepting the corresponding second-order belief. This is compatible with the fact

[52] Both Susana Nuccetelli and Hilary Kornblith used very similar examples for an objection against my argument.

that she didn't believe that before. The psychiatrist has not only changed her second-order but also her first-order beliefs. What if she not only lacked the relevant belief before her talk to the psychiatrist but, in addition, positively believed that her parents never abandoned her? In that case, we would either have a case of two mutually incompatible and contradictory beliefs or a case in which one belief "wins" over the other and make it disappear (see above). There is nothing problematic with either assumption.

The psychiatrist's case is also interesting because it hints at a difference between two (not the only two!) different ways of acquiring a second-order belief. A person might become convinced by behavioral evidence that she has a certain first-order belief; this kind of evidence is available from a third-person perspective. On the other hand, she might acquire the second-order belief on the basis of reflection about the subject matter of the corresponding first-order belief; this reflection is done from the first-person perspective. In the first case, the resulting first-order belief might rather be a bit more like the acceptance of a theoretical idea whereas in the second case the person might be more wholeheartedly committed to the truth of her first-order belief.[53] This does not entail that one could see one's own beliefs like the beliefs of another person and perhaps even mistrust them. It only means that a person's second-order beliefs can express different kinds of attitudes towards her own beliefs (but always as to her own beliefs). And in both cases, though, the second-order belief entails the first-order belief. [54]

There are more intricacies having to do with this. Consider the case of parachuter P (not necessarily meant as a counter-example):

Parachuting

> P has jumped a couple of times and is convinced that parachuting is not dangerous (and much less risky than driving around in a car which P happily does every day). P is even aware of her belief that parachuting is not dangerous. P is planning to have another jump today. But surprisingly, P just cannot bring herself to jump today. Does this show that P really believed, at least today, that parachuting is dangerous? Does it show that P's second-order belief ("I am not

[53] See L. Jonathan Cohen, *An Essay on Belief and Acceptance* (Oxford: Clarendon, 1992) on this difference.

[54] If the two kinds of attitudes clash, that is, if the person holds from a first-person perspective that she believes that *p* but holds from a third-perspective that she does not believe that *p* (or vice versa), then we have a case of a divided mind if the subject identifies with only "part" of his mind. See fn.30.

> among those who believe that parachuting is dangerous!") was false?[55]

Some propose to distinguish between two kinds of beliefs:[56] avowed beliefs and behavioral beliefs.[57] The first are relatively easily accessible to consciousness but do not necessarily drive our behavior whereas it is the other way around with the second. Should we restrict our thesis that B$Bp \rightarrow Bp$ then to avowed beliefs? I don't think we are forced to multiply kinds of beliefs here. It seems more plausible to say that beliefs have many different properties: They represent reality but also drive behavior. The parachuting case can be handled by our approach even if we don't multiply kinds of beliefs. People can be in two minds about things and hold mutually incompatible and contradictory beliefs (that it is dangerous, that it is not dangerous). Apart from that, beliefs might or might not affect behavior and if they do, then their effects can be of quite different kinds (more or less direct, etc.). If they don't, other mental states might drive our behavior: emotions like fear for instance (in case the person is interpreted as not having a belief that jumping is dangerous).[58] What drives our behavior and which of our beliefs lead to action under what conditions, is an empirical question that can only be attacked case by case. Our account is compatible with the parachuting case even if we assume that the behavior of the person reveals a hidden belief that parachuting is dangerous.

5.6. Crimmins and the Idiot

Mark Crimmins has come up with an example that might look like a counter-example to what I am saying here. Here is his paper (I quote in full):

> "'You have known me for years', explained Gonzales, 'But there is something you have not discovered. You know me under two guises, just as Lois Lane knows

[55] See also Shoemaker, "Moore's Paradox" (1996), 89 for such cases.

[56] See Georges Rey, "Toward a Computational Account of *Akrasia* and Self Deception," in *Perspectives on Self Deception*, eds. Amélie O. Rorty and Brian McLaughlin (Berkeley etc.: University of California Press, 1988), 264-296, especially, 272-277; Herbert Fingarette, *Self-Deception* (London: Routledge, 1969), 70, 88.

[57] One could even add a third kind: apart from those beliefs we identify on the basis of behavioral output or on the basis of avowals there would also be those beliefs we identify on the basis of informational input. I will not pursue this here. – An alternative would be to argue for a difference between belief and another kind of state which cannot be assimilated to belief; Tamar Szabó Gendler, "Alief and Belief," *The Journal of Philosophy* 105 (2008): 634-663 introduces the notion of an "alief." Eric Schwitzgebel, "Acting Contrary to Our Professed Beliefs or the Gulf between Judgment and Dispositional Belief," *Pacific Philosophical Quarterly* 91 (2010): 531-553 analyzes such cases as "in-between" cases of belief.

[58] See, e.g.,Neil Levy, "Have I Turned the Stove off? Explaining Everyday Anxiety," *Philosophers' Imprint* 16.2 (2016).

> Superman. You do not realize that I am the person you know under another guise. On that way of thinking about me, you have quite different opinions of me. In fact, you think me an idiot.'
>
> 'Knowing your cleverness,' I replied, 'I must with some embarrassment accept what you say. Since I do not know what guise you mean, I do not know which belief to revise. Until I find out, it seems, I falsely believe that you are an idiot!'"[59]

This is interesting but misleading in a subtle way. Crimmins believes something like this:

> Gonzales is no idiot.

Crimmins also learns this:

> ∃x (x=Gonzales & I believe of x that he is an idiot).

The only thing of relevance here Crimmins can infer is:

> I falsely believe *of* Gonzales that he is an idiot,

or, in other words:

> ∃x (x=Gonzales & I falsely believe of x that he is an idiot).

However, Crimmins cannot infer:

> I falsely believe *that* Gonzales is an idiot.

Crimmins can only ascribe a certain *de re* belief to himself but not the relevant *de dicto* belief. Since we are only dealing with *de dicto* beliefs here, Crimmins's case does not constitute a counter-example.[60]

So much for some objections to my main claim. It turns out, I think, that they do not work against our constitutive view of knowing one's beliefs.

6. Conclusion

I have argued for the claim that BB*p* → B*p*: If one believes that one believes that *p*, then one believes that *p*. If Mary believes that she believes that justice is the highest virtue, then she does indeed believe that justice is the highest virtue. This is a surprising claim: Sometimes "believing makes it so." It goes against what many people, especially philosophers, psychologists and cognitive scientists, believe. It might seem even more surprising that there are good arguments supporting this

[59] Mark Crimmins, "I Falsely Believe that *P*," *Analysis* 52 (1992): 191.

[60] See also Alan Hajek, Daniel Stoljar, "Crimmins, Gonzales and Moore," *Analysis* 61 (2001): 208-213, and David M. Rosenthal, "Moore's Paradox and Crimmins's Case," *Analysis* 62 (2002): 167-171; Williams, "Wittgenstein," sec.10 is very close to what I am saying here.

constitutive account of knowledge of one's own beliefs. I have used an argument which is based on considerations on Moore's paradox and on the impossibility of mistrusting one's own beliefs. Accepting the claim that B*Bp* → B*p* certainly has farreaching consequences for the way we should think about self-knowledge.[61]

[61] I would like to thank Sven Bernecker, Monika Betzler, Vivienne Brown, Gisela Cramer, Richard Eldridge, David Hemp, Hilary Kornblith, Teresa Marques, Susana Nuccetelli, Neil Roughley, John Williams, Truls Wyller und audiences in Maribor, Seattle, Belfast, and Aberdeen as well as some referees for comments.

CONTEXTUALISM AND CONTEXT INTERNALISM

David COSS

ABSTRACT: Contextualism is the view that the word 'knows' is context sensitive and shifts according to the relevant standards in play. I argue that Contextualism is best paired with internalism about contexts. That is to say, an attributor's context is completely determined by mental facts. Consequently, in the absence of awareness, external facts do not lead to contextual shifts. I support this view by appealing to the typical cases contextualists employ, such as DeRose's Bank Cases and Cohen's Airport Case. I conclude by reflecting on the nature of attributor's themselves, and suggest this also supports the view that Contextualism is internalistic about contextual shifts.

KEYWORDS: contextualism, Bank Cases, pragmatic encroachment

In this paper I argue that Contextualism is best paired with internalism about context. That is to say, I argue that an attributor's context is fixed by the salient contextual standards presently before her mind. I begin by outlining what contextualism is, then present several cases contextualists use to support their view, which also suggests an internalist reading of context. I conclude by providing more fundamental reasons for thinking contextualism is best paired with context internalism.

1. What is Contextualism?

Contextualism is the view that the meaning of the word 'knows' is context sensitive. More specifically, contextualists argue that the truth of knowledge attributions shift with the relevant contextual standards that are in play. For example, contextualists maintain that when one entertains skeptical hypotheses—or even alternate possibilities—the epistemic threshold for knowledge shifts upward, making it more difficult for attributors to have knowledge. However, in ordinary contexts—those that obtain outside of philosophical study, discussion and reflection—the standards of knowledge are usually lower.[1] In this way,

[1] It's worth mentioning that contextualists think ordinary people naturally find themselves in a low standards context. That is to say, low—or moderately low—epistemic standards are the default. However, given the increased popularity of science fiction films ranging from *Inception*, *The Matrix*, *The Thirteenth Floor*, etc. It is no longer clear whether low standards contexts

contextualists deny knowledge invariantism, the view that there's only one standard of knowledge. Contextualists typically adhere to the following thesis about knowledge.

> The Contextualist Thesis
>
> Whether a knowledge attribution, 'S knows that p,' made by an attributor A, is true or false, depends upon whether A's evidence (or, strength of epistemic position) is strong enough for knowledge relative to standards of knowledge in A's context.

A major motivation for Contextualism is the desire to articulate an effective and satisfying response to external world skepticism.[2] The skeptical worry is that it's impossible to have external world knowledge given classical fallibilism.[3] This this is puzzling, however, since ordinary people, as well as philosophers, take themselves to know many things about the external world. The skeptical worry can be formulated as an argument which runs as follows, where 'K' is the knowledge operator and 'BIV' is a brain-in-a-vat hypothesis, according to which all my external world experiences are generated by an evil scientist manipulating my brain, and 'hands' is a generic placeholder for any external world object:

P1. K(hands) → K˜BIV

P2. ˜K˜BIV

C: ˜K(hands)

While Dretske famously denied P1 (the closure principle), maintaining that one can know that one has hands, even if one doesn't know the falsity of BIV hypotheses,[4] Contextualists are reluctant to abandon this principle. Rather, their answer to skepticism is a rejection of P2 for ordinary contexts.

The skeptic defends P2 by claiming we are never in a strong enough epistemic position to deny this premise. Suppose the BIV scenario is true. Skeptics argue that an envated subject S, and a non-envated subject S*, possess the same quality of evidence when considering propositions related to the external world.

should be considered the default epistemic threshold. However, this is a topic for another paper.

[2] I take classical fallibilism to the conjunction of two views: fallibilism and classical epistemology.

[3] Classical fallibilism is the view that knowledge doesn't require truth entailing evidence. In other words, subjects can know propositions even if they are not epistemically certain of its truth. Hence, S could know that p even if logical space affords her the possibility of being mistaken.

[4] Epistemic closure is a principle whereby knowledge is closed under known entailment. The principle is as follows: (sKp & sK(p → q)) → sKq. For more on the denial of closure, see Fred Dretske, "Epistemic Operators," *Journal of Philosophy* 67, 24 (1970): 1007-1023.

Since the quality of evidence is the same for both S and S*, and consequently indistinguishable by perceptual evidence alone, the skeptic claims external world knowledge is impossible.

Contextualists draw attention to a conflict within our belief structure. On the one hand, skepticism seems convincing. The argument for skepticism is valid and appealing to one's epistemic intuitions, it seems sound, although the conclusion strikes many philosophers as unacceptable.

A virtue of the contextualist response to skepticism is twofold. First, viewing 'knows' as context-sensitive allows the contextualist to respond to skeptical worries without abandoning fallibilism.[5] Second, while contextualists accept the conclusion of skeptical arguments in contexts when skeptical possibilities are entertained, they deny that skeptical arguments are applicable in all contexts. In ordinary situations, when skeptical worries and alternative possibilities are not entertained, many 'S knows that p' statements come out true, assuming such true beliefs meet the lower evidential threshold. In other words, contextualism responds to skepticism, while also appreciating the philosophical thrust of the problem.[6]

2. Internalism

Before outlining two ways of viewing contexts, I will explain the internalism/ externalism distinction as it relates to epistemic justification. In their most basic forms, internalists views impose constraints on justification-determining factors that externalists reject. For example, according to internalism, a justified belief must be recognizable on reflection, whereas externalism denies this.[7] According to

[5] One would like to adhere to fallibilism so as to avoid widespread Cartesian skepticism.

[6] One might be inclined to wonder how contextualism differs from an alternative approach called the "ambiguity theory of knowledge." According to this theory, there are multiple senses of the word 'knows.' While contextualism is similar to this view, there are marked differences which delineate the two. Perhaps the most important difference is the way in which each view the role context plays in determining the truth of knowledge attributions. For the ambiguity theory, one can simply stipulate which sense of the word 'knows' one is employing (much the same way as I can stipulate that I am talking about a financial institution when I use the term 'bank'). Context, therefore, plays either no role, or a marginal one, in determining true knowledge attributions. Contextualists, on the other hand, make the knowledge attributors slaves to context. Contextual features determine the evidential threshold, and therefore determine whether a knowledge attribution is true. In other words, the main difference is that for the ambiguity theorist, agents control which sense of 'knows' they employ, while contextualists depend upon context to determine whether a knowledge attribution is true.

[7] Michael Bergmann has argued that internalism doesn't necessarily require awareness. For

internalist epistemologists, the transformation from an unjustified to a justified belief occurs by having the right mental states (usually by possessing and employing evidence in the belief formation process).

While context internalism diverges from justificatory internalism, both in its subject matter and aim, there's nevertheless an important parallel: something mental entirely fixes either justification or contexts.

Context internalism can be understood in several ways, such as the imposition of constraints in terms of awareness, access, mentality, or perception. Perhaps the best way to understand context internalism is through a subject's attitudes, beliefs, desires, intentions etc. in the formation and construction of a context. An implication of this view is that two subjects (or attributors) could be similarly situated in external circumstances, but be in different epistemic states depending on their beliefs.[8]

3. Contextualism and Context Internalism

We can start by making an obvious observation: contexts are fixed by factors that are either entirely internal or partially external. If what fixes an epistemic context is completely internal, only mental factors are relevant in judging what context an attributor or subject is in.

In making the argument that contextualism is best paired with context internalism, we need to further specify how contextual standards of the word 'knows' shift.

Here is my primary reason for thinking that contextualism is best paired with context-internalism. When contextualists evaluate which context an attributor is in, they consider factors that are presently before a subject's mind. External factors, inasmuch as they are not salient, or worse, fail to be cognitively accessible to subjects or attributors, fails to elevate the epistemic threshold for knowledge.

Consider classic cases presented by both Keith DeRose and Stewart Cohen. In reviewing these cases, it's important to keep in mind several questions: do

brevity, I will not engage with his arguments here. For those interested, consult ch. 3 of Michael Bergmann, *Justification without Awareness* (Oxford: Oxford University Press, 2006).

[8] While it is a worthwhile task to evaluate the plausibility of context internalism, I will not pursue this task here. A robust account of context would need to take into consideration arguments and findings from fields like philosophy of language, mind and metaphysics, as well as those from psychology and cognitive science. However, the features of context which need elucidation are only those which relate to the epistemic standards associated with the word 'knows.'

external factors *themselves* determine the attributor's context? Or is it the subject's *awareness* of them? Second, in the absence of such awareness, would contextual shifts occur?

> **Bank Case A.** My wife and I are driving home on a Friday afternoon. We plan to stop at the bank on the way home to deposit out paychecks. But as we drive past the bank, we notice that the lines inside are very long, as they often are on Friday afternoon. Although we generally like to deposit our paychecks as soon as possible it is not especially important in this case that they be deposited right away, so I suggest we drive straight home and deposit our paychecks on Saturday morning. My wife says 'Maybe the bank won't be open tomorrow. Lots of banks are closed on Saturdays.' I reply, 'No, I know it will be open. I was just there two weeks ago on Saturday. It's open until noon.'
>
> **Bank Case B.** My wife and I are driving home on a Friday afternoon, as in Case A, and notice the long lines. I again suggest we deposit our paychecks on Saturday morning, explaining that I was at the bank on Saturday morning only two weeks ago and discovered that it was open until noon. But in this case, we have just written a very large and very important check. If our paychecks are not deposited into our checking account before Monday morning, the important check we wrote will bounce, leaving us in a *very* bad situation. And, of course, the bank will not be open on Sunday. My wife reminds me of these facts. Then she says, 'Banks do change their hours. Do you know the bank will be open tomorrow?' Remaining as confident as I was before that the bank will be open then, still, I reply, 'well, no, I don't know. I'd better go in and make sure.'[9]

Inspecting DeRose's Bank Cases reveals that Keith's context shifts from a low, to a high standards one relative to his wife making salient the possibility of the bank changing its hours. In other words, it's salience of error, not merely the possibility of error, that leads to an upward shift in contextual standards.

We arrive at the same conclusion when considering Cohen's Airport case. Mary and John's context doesn't shift upward until the possibility of error is made salient.

> The Airport Case
>
> Mary and John are at the L.A. airport contemplating taking a certain flight to New York. They want to know whether the flight has a layover in Chicago. They overhear someone ask a passenger Smith if he knows whether the flight stops in Chicago. Smith looks at the flight itinerary he got from the travel agent and respond, 'Yes I know—it does stop in Chicago.' It turns out that Mary and John have a very important business contact they have to make at the Chicago airport. Mary says, 'How reliable is that itinerary? It could contain a misprint. They could have changed the schedule at the last minute.' Mary and John agree that Smith

[9] Keith DeRose, *The Case for Contextualism* (Oxford: Oxford University Press, 2009), 1-2

> doesn't really *know* that the plane will stop in Chicago. They decide to check with the airline agent.[10]

John and Mary start off in a low standards context, and it's only after they are made aware of the potential for a misprint in the itinerary that an upward contextual shift occurs.

Internal, rather than external facts, fix the context in all three of these cases. If Keith's wife hadn't reminded him that banks sometimes change their hours, he would still be in a low standards context. In Cohen's Airport case, Mary and John both start off in a low standards context and it's only when certain error possibilities are entertained that their context becomes more epistemically demanding, consequently elevating the epistemic threshold for knowledge.

Another reason to think contextualists ought to endorse internalism about contexts is the view's inability to handle other bank-style cases. Stanley argues that contextualism gives the wrong answer in cases that lack saliency of error. For example, consider his case.

> Ignorant High Stakes
>
> Hannah and her wife Sarah are driving home on a Friday afternoon. They plan to stop at the bank on the way home to deposit their paychecks. Since they have an impending bill coming due, and very little in their account, it is very important that they deposit their paychecks by Saturday. But neither Hannah nor Sarah is aware of the impending bill, nor the paucity of available funds. Looking at the lines, Hannah says to Sarah, 'I know the bank will be open tomorrow, since I was there just two weeks ago on Saturday morning. So we can deposit out checks tomorrow morning.[11]

Since neither Hannah nor Sarah is aware of the impending bill, Stanley argues that, by contextualisms lights, they are in a low standards context. Consequently, Stanley argues that contextualists must maintain that they know the bank is open.[12]

Finally, there's a more basic reason for thinking contextualists should

[10] Stewart Cohen, "Contextualism, Skepticism and the Structure of Reasons," *Philosophical Perspectives* 13, 13 (1999): 58

[11] Jason Stanley, *Knowledge and Practical Interests* (Oxford: Clarendon Press, 2008), 5

[12] One might worry that Stanley's case can be explained in alternative ways. For example, Hannah and Sarah seem to behave irresponsibly, and perhaps what explains their lack of knowledge is this fact. However, this applies to all high stakes bank cases. If one has an impending bill due, it's irresponsible to put it off even if one knows the bank will be open. For example, even if S knows the bank will be open, S might not know she will get into a car accident on the way there, or perhaps she will misplace the check. While the point about irresponsibility is an important one, I for the sake of brevity, I will not entertain it further.

endorse context internalism: knowledge attributors are the locus of contextual shifts. Broadly speaking, the nature of a knowledge attributor requires awareness of what is being attributed. If there's an upward shift in the contextual standards, an attributor S must, on some level, be aware and sensitive to things like possibilities of error. Given the cases presented above, and the nature of attributors, it's plausible to view contextualism as internalistic.[13]

[13] One might deny that the knowledge attributors needn't be aware of what they attribute. Consider the snarky skeptic who just goes around denying people know anything, but isn't aware of what she's saying. In this sense, one might say that one knowledge attributors—or attributors more generally, don't require awareness. While this is an interesting criticism, and requires a detailed response, I will not pursue it at length here. However, I am inclined to develop an account of authentic versus inauthentic knowledge attributors. Another response is that perhaps ordinary knowledge attributions don't require awareness (after all, people use words like 'know' frequently without fully understanding them). However, in cases where an attributor makes salient skeptical situations or possibility of error scenarios, it seems like they are aware—on some level—of what they are doing. However, since these responses are in an immature state, I will save their development for a different paper.

LIMITATIONS AND THE WORLD BEYOND

Patrick GRIM and Nicholas RESCHER

ABSTRACT: This paper surveys our inescapable limits as cognitive agents with regard to a full world of fact: the well-known metamathematical limits of axiomatic systems, limitations of explanation that doom a principle of sufficient reason, limitations of expression across all possible languages, and a simple but powerful argument regarding the limits of conceivability. In ways demonstrable even from within our limits, the full world of fact is inescapably beyond us. Here we propose that there must nonetheless be a totality of fact, and that despite our limits we can know something of its general character. The world as the totality of fact must form a plenum, with a radically unfamiliar formal structure that contains distinct elements corresponding to each element of its own power set.

KEYWORDS: truth, language, Georg Cantor, Kurt Gödel, totality, plena

1. Introduction

Our topic is that of limits: the metamathematical limits of axiomatic systems, epistemic limits of explanation, linguistic limitations of expression, conceptual limits of conceivability, and ultimately questions of ontological and metaphysical limits as well. The limitations of axiomatic demonstration and of mechanical computation are clear from the Turing and Gödelian traditions. In section 2 we pursue extensions and analogies to limitations intrinsic in the structure of explanation, restrictive on a principle of sufficient reason PSR. In section 3 we consider the limitations on expression entailed by recursive linguistic structure, extending the argument from single languages to sets of possible languages and showing that even the properties of languages inevitably outstrip the properties expressible within those languages. In Section 4 we pause to consider epistemic implications, extending the discussion beyond language to incompleteness of any body of *conceivable* truths in the face of a demonstrably larger realm of fact. We suggest nevertheless that something can be shown of its general character. The world as the totality of fact must form a plenum,[1] with implications we here set out

[1] Nicholas Rescher and Patrick Grim, "Plenum Theory," *Noûs* 42 (2008): 422, *Beyond Sets: A Venture in Collection-Theoretic Revisionism* (Frankfurt: Ontos Verlag, 2011).

to explore.

Plato's *Timaeus* launched the pivotal belief of ancient Neo-Platonism that Reality reflects the operations of Reason and accordingly constitutes a rationally intelligible manifold. In consequence man, the rational animal, is able to get a reason-engendered cognitive grip on Reality's key features. This fundamental idea was to become one of the mainstays of Western philosophy. But no-one, then or since, maintained that human reason's grip on Reality was complete or completable--that human cognition and speculation could exhaust the unbounded vastness of possibility and plumb the bottomless depths of its relationship to the real--a task which, if achievable at all, required an intelligence of supra- and super-human capacity. But just where can we find clear signs of the limits of human intellection and pinpoint some of the issues that lie beyond the horizons of our cognitive reach. No doubt this is a difficult question but there are some things that can plausibly be said on the problem and hopefully some of them will be said here.

The limitations we track are characteristically not some boundary imposed from without but intrinsic limitations of reach from within an entire method of axiomatization, explanation, expression, or comprehension. The problematic clearly traces to Kant for whom human cognition has limits by way of limitations (Grenzen) but not boundaries (Schranken), there being no wall or fence that somehow ontologizes those limits. For us those limits lie not as with Kant, in the faculty structure of the human intellect, but in the nature of the conceptual resources characteristic of our cognition, or perhaps of any cognition.

In sections 5 and 6 we attempt to go farther metaphysically and ontologically, for a glimpse of the world beyond our limits. The attempt itself sounds paradoxical, and it is in fact paradox that we take as the key. The world as the totality of fact lies inevitably beyond our limitations—explanatory, expressive, and conceptual. But we propose we can nonetheless know something of its general character. The world as the totality of fact must form a *plenum*.

2. Limits from Axiomatization to Explanation

The limitations of axiomatization are well known. No formal system adequate for basic arithmetic can be both consistent and complete. No axiomatic system can contain as theorems both all and only the truths expressible in the formal language of the system. We cannot hope to grasp all of mathematical truth—restricted even to the mathematical truth we have the means to express—with the techniques of axiomatization.

It is a short step from Gödel to Turing, from formal systems to mechanical algorithms. By the same token, and in much the same way, no mechanical algorithm can give us all and only correct answers to some easily expressible questions about the function of mechanical algorithms. In both the Gödel and Turing results, it is the system itself—by a particular power of embedding—that reveals its own limitations. It is because a system for number theory can represent (or echo) any mechanism of axiomatic deduction that any axiomatic system will be provably incomplete. It is because Turing machines can echo and embed any algorithmic mechanism that there can be no faultless algorithmic mechanism for any of a range of basic questions regarding them all.

We will return to explanation and the Principle of Sufficient Reason (PSR) at a number of points. Here we start with a particularly simple version:

> (PSR-T) For every truth there is some other, epistemically distinct truth that provides a cogent explanation for it.

If we take 'explanation' to demand a deductively valid accounting, PSR-T will be untenable for precisely Gödelian reasons. Any deductive system adequate for scientific explanation will have to be adequate for arithmetic. But any deductive system adequate for arithmetic, there will be truths expressible in the system which will not be deducible as theorems. Those will be truths in violation of PSR-T.

We can take the result further, and make it more pressing, by *replacing* the concept of deduction in the Gödel result with a concept of explanation instead. Mathematical exploration through the last century, eloquently expressed in Hilbert, was a vision of some distant but reachably complete completed mathematics. That vision died with Gödel's proof. A vision of a completed explanatory *science* has spurred scientific exploration in much the same way. That vision of scientific explanation is as impossible as the correlate vision of mathematical explanation, and for precisely the same reasons.

Suppose a science which contain (a) a complete set of basic facts, and (b) a complete set of 'explanatory consequence' principles whereby further facts follow from others. It is clear that any such system must also contain the mechanisms of any system adequate for arithmetic. Among its 'basic facts' must be the axioms and among its 'explanatory consequences' principles must be the rules of inference which are required for basic arithmetic. It then follows that there will be true statements in the language of such a science for which our 'completed science' will be unable to offer a scientific explanation.

There is an older and simpler problem with PSR-T, of course. The explanatory project confronts us with the prospect of basic explanatory elements analogous to axioms which by hypothesis cannot be derived from anything else. Further forms of the principle of sufficient reason, correlate to even wider limitations on explanation, reappear later in our discussion.

3. Intrinsic Limits of Language and Truth

We humans conduct our cognitive business by means of language, broadly conceived to include all processes of symbolic communication. Linguistic articulation, both in human communicative reality and in its formal representation, is fundamentally recursive. Beginning with a finite vocabulary it exfoliates meaningful statements by means of a finite number of grammatical rules of combination. The result is a potentially infinite number of meaningful statements in in any such language, but those statements will be enumerable and thereby denumerable in number. And of course if the meaningful statements (the well-formed formulas as a whole) can be enumerated (and thus be denumerable in number) this will also have to hold for the subset of them that are true. The truths expressible in any language, in sum, form a denumerable manifold.

At this point a distinction between truths and facts becomes critical. We take truths to be linguistically articulated claims—specifically those that are correct. We take facts to be something else again: states of affairs that obtain and do so independently of any articulation by linguistic means.

A. We begin with the simplest formal case, which is also closest to the reality of human languages. Consider a language with a finite number of basic symbols and a finite number of recursive rules for combination. Such a language will afford us with a countably infinite number of formulae. At best, the expressible truths for such a language will be countably infinite.

It's clear that there will be more than countably infinite facts, a point provable using the example of this language alone. The formulae of any such language *L* form a countably infinite set. But by the basic mechanisms of Cantor's Theorem, there will be more elements of the power set of any set than elements of that set itself. Consider then the power set *PL* of the set of formulae of this initial language. For each set element of *PL* there will be a distinct fact: the fact that a specific formulae does or does not belong to that set, for example. Even this small corner of a world of fact—facts about the language *L*—will have facts inexpressible in *L* itself. The facts *about* such a language inevitably outstrip the truths it can express.

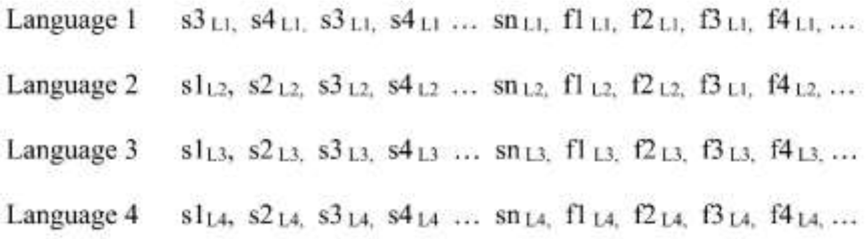
What are we to make of there being infinitely more actual facts than articulable truths? With human knowledge functioning linguistically by way of a recognition and acknowledgement of truths, does this disparity between facts and truths not entail the existence of an unknowable truth?

Here it is instructive to begin with a simple analogy: that of Musical Chairs.

Where there are more players than chairs it is inevitable that some will be left unseated when the music stops. So the existence of *unseated* players is inescapable. But this of course does not itself mean that any players are *unseatable* so that it is in principle impossible for such a player to be seated. The prospect of seating cannot be denied to any of them.[2] When this situation is analogized to the truth/fact situation, we will have it that the inevitability of *unknown* facts does not of itself establish the existence of *unknowable* ones. All we can maintain at this point is that there are bound to be *unknown* facts: that there are unknowable ones does not follow. That *not every fact can be known* does not of itself enjoin that *some fact cannot possibly be known*. The quantitative disparity between formulable truths and objective facts does not immediately establish the existence of unknowable facts.

B. What of the truths expressible by any *possible* language of this simple formal and very human form, involving finitely many basic symbols and finitely many recursive rules of combination? We begin by supposing that each possible language takes its basic symbols from some zingle but countably infinite reservoir of possible symbols, awash with as many basic symbols as there are numbers 1, 2, 3… On that assumption, the basic symbol sets of the full set of our possible languages will be enumerable: there will be only a countably infinite number of basic symbol sets.

Because those finite sets of symbols can simply be appended as the first of the countably infinite formulae generable using them, within our basic assumptions we can envisage an enumeration of all formulae of all possible languages of this form as an infinite series of infinite arrays. Using $s1_{L1}$ through sn_{L1} to represent the finitely many basic symbols of language 1 and $f1_{L1}$, $f2_{L1}$, $f3_{L1}$… to represent its infinitely many compound formulae, such an array might take this form:



Language 1 $s3_{L1}$, $s4_{L1}$, $s3_{L1}$, $s4_{L1}$ … sn_{L1}, $f1_{L1}$, $f2_{L1}$, $f3_{L1}$, $f4_{L1}$, …

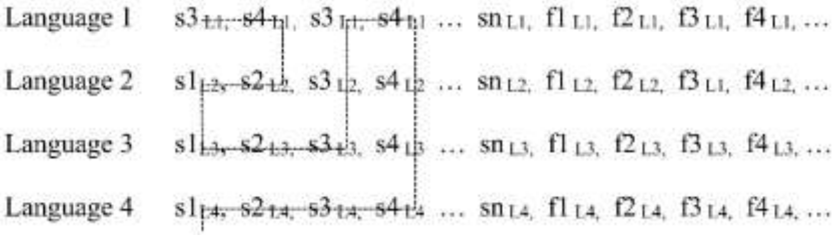
Language 2 $s1_{L2}$, $s2_{L2}$, $s3_{L2}$, $s4_{L2}$ … sn_{L2}, $f1_{L2}$, $f2_{L2}$, $f3_{L1}$, $f4_{L2}$, …

Language 3 $s1_{L3}$, $s2_{L3}$, $s3_{L3}$, $s4_{L3}$ … sn_{L3}, $f1_{L3}$, $f2_{L3}$, $f3_{L3}$, $f4_{L3}$, …

Language 4 $s1_{L4}$, $s2_{L4}$, $s3_{L4}$, $s4_{L4}$ … sn_{L4}, $f1_{L4}$, $f2_{L4}$, $f3_{L4}$, $f4_{L4}$, …

As in Cantor's proof for the countability of the rationals, however, we can introduce a circuitous but systematic enumeration of every item in that array as well:

[2] In logical notation, the different at issue is that between ☒$(\exists x){\sim}Sx$ and $(\exists x)$☒${\sim}Sx$.



Language 1 s3$_{L1}$, s4$_{L1}$, s3$_{L1}$, s4$_{L1}$ … sn$_{L1}$, f1$_{L1}$, f2$_{L1}$, f3$_{L1}$, f4$_{L1}$, …

Language 2 s1$_{L2}$, s2$_{L2}$, s3$_{L2}$, s4$_{L2}$ … sn$_{L2}$, f1$_{L2}$, f2$_{L2}$, f3$_{L1}$, f4$_{L2}$, …

Language 3 s1$_{L3}$, s2$_{L3}$, s3$_{L3}$, s4$_{L3}$ … sn$_{L3}$, f1$_{L3}$, f2$_{L3}$, f3$_{L3}$, f4$_{L3}$, …

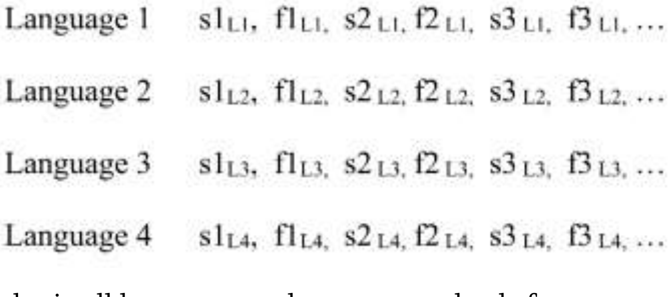
Language 4 s1$_{L4}$, s2$_{L4}$, s3$_{L4}$, s4$_{L4}$ … sn$_{L4}$, f1$_{L4}$, f2$_{L4}$, f3$_{L4}$, f4$_{L4}$, …

On the assumption of a countable reservoir of basic symbols, then, there will be only countably many truths expressible in all *possible* languages of this basic form. We know the facts of even one of those languages form more than a countable set, and thus the facts regarding even one of these possible languages outstrip the truths expressible in all such possible languages.

C. But perhaps we've sold linguistic possibilities short. We can expand our conception of formal languages, recognizing as we do so that we are leaving the limitations of human languages behind.

Limitations like those above are demonstrable for even some superhuman languages. Let us start by allowing a language to contain more than a finite number of basic symbols. It is indeed standard in outlining formal systems to envisage a countably infinite number of basic formulae p1, p2, p3…. That change alone won't alter the results for single languages. The countably infinite basic symbols of such a language can be interwoven with the countably infinite formulae that can be recursively generated from those formulae, giving us no more than countably infinite formulae over all. The cardinality of our formulae, the factual limitations of truths, will remain.

As long as our basic symbols are drawn from a countably infinite pool, the same will hold for all *possible* languages of such a form. For each language we can envisage an enumeration that interweaves the countable series of basic symbols with the countable series of recursively combinatorial formulae:



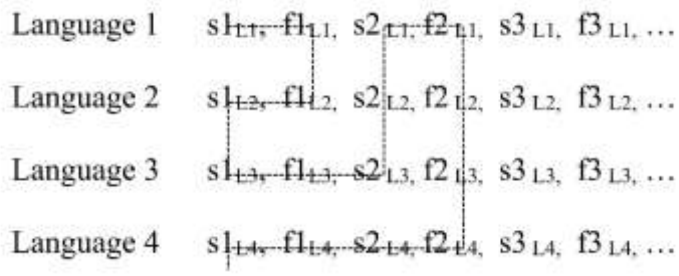
Language 1 s1$_{L1}$, f1$_{L1}$, s2$_{L1}$, f2$_{L1}$, s3$_{L1}$, f3$_{L1}$, …

Language 2 s1$_{L2}$, f1$_{L2}$, s2$_{L2}$, f2$_{L2}$, s3$_{L2}$, f3$_{L2}$, …

Language 3 s1$_{L3}$, f1$_{L3}$, s2$_{L3}$, f2$_{L3}$, s3$_{L3}$, f3$_{L3}$, …

Language 4 s1$_{L4}$, f1$_{L4}$, s2$_{L4}$, f2$_{L4}$, s3$_{L4}$, f3$_{L4}$, …

All formulae in all languages can be enumerated as before:



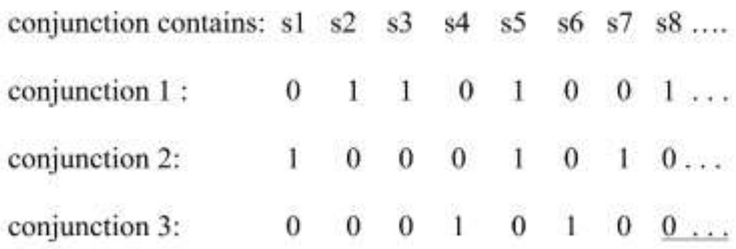
Language 1 $s1_{L1}$, $f1_{L1}$, $s2_{L1}$, $f2_{L1}$, $s3_{L1}$, $f3_{L1}$, …
Language 2 $s1_{L2}$, $f1_{L2}$, $s2_{L2}$, $f2_{L2}$, $s3_{L2}$, $f3_{L2}$, …
Language 3 $s1_{L3}$, $f1_{L3}$, $s2_{L3}$, $f2_{L3}$, $s3_{L3}$, $f3_{L3}$, …
Language 4 $s1_{L4}$, $f1_{L4}$, $s2_{L4}$, $f2_{L4}$, $s3_{L4}$, $f3_{L4}$, …

The formulae of all possible languages based on countably infinite symbols from a countably infinite pool will still form merely a countable set. The truths expressible in all possible languages of such a form will be merely countable.

D. The situation changes if we further broaden assumptions, leaving human capabilities even farther behind. Consider the possibility of a larger reservoir from which a language might draw its basic symbols: a reservoir that has as many basic symbols not merely as the rationals, for example, but as many as the reals.

Any language that has either a finite number of basic symbols drawn from such a pool or a countably infinite number of such symbols will be limited, as above, to a countably infinite number of formulae. But the conclusions drawn so far will not hold for all *possible* languages of this expanded form. A very simple way of seeing this is to envisage those languages that have merely one basic symbol. Since that symbol can be any of a collection as large as the reals, we will not be able to enumerate all of those languages, prohibiting the countable list of languages used on the left axis in the arrays above. For languages with basic symbols drawn from a set the size of the reals, then, formulae of *each* language will be countable but formulae of all *possible* such languages will not.

Limitations of countably many formulae are obviously lifted for even *single* languages if we allow a language to have as many simple formulae as the reals. Somewhat less obviously, limitation to the countably infinite is lifted for a single language with countably many basic formulae and infinite combinations: infinite conjunctions or disjunctions, for example. We might list conjunctions in such a language by using 0 or 1 to indicate whether they include symbol 1, symbol 2, symbol 3, and so on:



conjunction contains:	s1	s2	s3	s4	s5	s6	s7	s8 ....
conjunction 1 :	0	1	1	0	1	0	0	1 . . .
conjunction 2:	1	0	0	0	1	0	1	0 . . .
conjunction 3:	0	0	0	1	0	1	0	0 . . .

It is clear that every infinite series of 1's and 0's will be represented by some conjunction in such a system. But these correspond to the infinite decimals between 0 and 1, which correspond to the reals. Cantor's proof that there are non-denumerably many reals may be performed quite directly on any proposed enumeration of these conjunctions. We can produce a conjunction *not* on the list by exchanging 1's and 0's on the diagonal.

E. If we weaken assumptions and stretch possibilities for languages far enough, then, we can have sets of possible languages and even single languages that transcend the limits of a countable infinity of expressible truths. In a very real sense, however, such languages bring us no closer to the world of facts.

No matter how large the set of formulae expressible in any of these languages, the power set of that set will be larger than the set itself. For every set element of that power set there will be a fact: the fact that a given formulae is or is not an element of that set, for example. There will still be more facts expressible in any given language.

Given any set of specifications for a form of language, there will be a set of formulae and thus a set of truths expressible in all *possible* languages of that form. The power set of *that* set of all possible formulae or expressible truths will be larger still, and thus the facts even about sets of truths expressible in all possible languages of a specific form will transcend the truths so expressible. Like the individual languages within them, ranges of possible languages embody more facts than they can possibly express.

F. All of the arguments presented to this point have been written in terms of syntax: numbers of formulae generable within a given language. But languages in the sense we are after are perhaps better conceived of semantically, such that formulae are *about* certain things, using predicates to express properties *of* certain things. A more semantic and in that sense more philosophical form of the argument makes the point in its most general form.

With any language there will be those things that it can say things about: what we might term the *linguistic objects* of a language. Things in general, linguistically reachable within such a language or not, we can term *factual objects*. Any language will also have those things that it can say *about* things: its *predicates*. Factual properties that actually hold of things, linguistically bound or not, we will simply term *properties*.

On this simple outline, it's clear that the predicates of any language will themselves be factual objects. By an analog of Cantor's Theorem, we know that the sets of those objects outnumber the objects themselves. But for each such set, there is a unique property, indeed an extensional property: the property of belonging to such a set, for example. There are therefore more properties of factual objects than there are predicates available in any language to express those properties. Indeed there are more properties of the predicates of any language than there are predicates in the language. The facts of properties inevitably outstrip truths expressible by predicates.

What holds for a single language holds for all possible languages. If we consider the predicates applicable in any possible language, of whatever form, we are considering a set of factual objects. But there will be more sets of such objects, and thus there will be more factual properties, than there are predicates applicable in any possible language.

G. Does this entail that there is any specific inexpressible truth? One can hardly ask for an example. To this point, considering languages both syntactically and semantically, the image of musical chairs still holds: each language will leave out some fact, but nothing yet identifies a specific fact that will be left out.

Languages are more than syntactic structures, more even than syntactic structures with correspondences to objects and properties. Languages are means of managing information. Information is packaged in the form of expressions, unpacked by means of derivation. It is in terms of information that we can begin to see some specifics regarding linguistic limits: for any language L, a *specific* body of information beyond it.

We have termed truths those linguistic elements that correspond to facts. For any language there will be those truths expressible in the language. Each truth will embody some information, reflecting some fact. But there is one body of information that will inevitably escape a language, in one way or another: that body of information that is represented in *all* of its truths combined. For any language L, we will term that megafact M_L. There is no single truth in L that can capture this megafact: totalistic self-representation cannot be internalized declaratively.

Suppose any language L, and all truths expressible in L. Consider moreover a truth-preserving set of rules of derivation R employed in L which allows one to squeeze out as consequent truths the information contained in a given truth. Finally, consider ML, the information contained in all the truths of L. ML will either be L-inexpressible or R-inaccessible at least in part. ML will be inexpressible in L in any way in which all the information in ML will be derivable by R.

Our discussion started with Gödel and Turing as a foundation, moving from there to considerations that were largely Cantorian. Here the argument turns on Gödel once again. Were ML both L-expressible and fully R-accessible, there would be an axiomatic system with R as its rules and ML as an axiom from which all truths expressible in L were derivable. By Gödel, there can be no such axiom system.

The result will clearly hold for all the kinds of languages to which Gödel applies: all those satisfying the minimal requirements of an L and R adequate for arithmetic. It is also possible to generalize the result beyond those specific requirements.[3] Given any rules of derivation R, a language that can represent R-derivability we will call R-expressive. A language that can take any of its own expressions as object we will term expressibility-reflective. For any expressibility-reflective language L that is R-expressive, for any truth-preserving R, the megafact ML for that language will either be inexpressible in L or R-inaccessible at least in part: either ML will be inexpressible in the language of L, or there will be information in ML that will be underivable by R.

For any language within these minimal constraints there will be a particular fact that proves inaccessible for it: the megafact ML that represents the totality of information in the facts that it does represent. Note that ML doesn't have to extend to all facts. It is specified relative to a language and encapsulates merely the information expressible in the facts captured in that language. Even that smaller language-relative totality of facts escapes the nets of language and derivability.

H. Here again our reflections impact the principle of sufficient reason.

Anything rationale offered as an explanation, in any language, will be a set of expressions within that language. The available rationales for any language will therefore be limited by the available expressions. For a standard language L with countably many expressions, for example, there will be only countably many possible finite rationales.

It's clear from the pattern of argument above that for any L there will be not only more facts than linguistically expressible truths, but more facts than there are available rationales. Using 'explanationL' to indicate rationales in language L, then, the following version of the PSR will fail for any L:

[3] Here the generalization of Gödel follows roughly the lines of chapter 3 of Patrick Grim, *The Incomplete Universe* (Cambridge, MA: MIT Press, 1991).

(PSR-F) Every distinct fact has a distinct explanationL.

The lesson will extend to the languages of non-standard forms considered above. It will also extend to explanation in any or all possible languages. If we consider the rationales expressible in any possible language, of whatever form, we are considering a set of factual objects. But there will be more sets of rationales than rationales themselves. For each of those sets there will be a distinct fact. There will therefore be more distinct facts than distinct rationales in any possible language. Generalizing 'explanation' from 'explanation in L' to 'explanation in *any* possible language,' this more encompassing version of the PSR will fail as well:

(PSR'-F) Every distinct fact has a distinct explanation.

4. Epistemic Reflections and Conceivability

What does this disparity between linguistic truth and trans-linguistic fact mean for our knowledge? To what extent do the limitations of language extend to limits of conceivability?

A. At first glance, axiomatization as a model of a distinction a distinction between explicit and implicit knowledge might seem to offer some hope.

By the Cantorian argument, the expressible truths of any language will be outnumbered by the facts. But there are two ways of affirming or claiming a fact. One is to state it explicitly and specifically, in the form for example of a corresponding truth. Another is to affirm it obliquely and implicitly by stating other facts from which it follows. In that sense a single statement---the conjunction of the axioms of a system, for example—can be seen as implicitly containing the full information of all theorems of the system. It lies in the logic of things that one truth can informatively encompass a vast—indeed a potentially infinite—multitude of other distinct claims.

One true claim, such as a conjunction of the axioms of plane geometry, can informationally encompass the entire field. Finite access to claims does not itself therefore entail finitude in knowledge. Given the distinction between explicit expression and implicit deducibility on the model of axioms, the quantitative disparity between truth and fact might not seem all that portentious.

We might, then, distinguish two basic questions:

Q1. Can the totality of the facts in the domain at issue be stated and acknowledged explicitly in terms of coordinate truths?

Q2. Can the totality of fact of the domain at issue be substantiated at least obliquely and implicitly by way of inferential axiomatization?

The force of the Cantorian argument—there are more facts than truths with which to express them—is that the answer to Q1 is a clear 'No.' But for standard systems, at least, a Cantorian argument shows that the answer to Q2 must be 'No' as well.

Standard systems will have only a countable number of theorems. Even implicitly, therefore, their axioms will contain only a countable number of truths. Implicit knowledge amounts to deductive closure: we implicitly know whatever can be derived from what we explicitly know. Derivation is a recursive process. It begins with premisses and applies stepwise any of a finite register of inferential rules. A body of explicit axioms, then, be it finite or countably infinite, can never represent more than a countable body of implicit knowledge. In the previous section we envisaged systems beyond standard systems. But even these will have only some limited cardinality of implicit theorems—a cardinality that will be provably exceeded by the range of fact...even the range of fact about those theorems.

If our model of implicit knowledge is axiomatic, it must be recognized that the power of an axiomatic system cannot exceed that of the language in which it is expressed. Our results above hold for all languages, and thus for the implicit knowledge contained in any axioms written within those languages as well. Any hope for conceivability beyond linguistic limits must appeal to something beyond implicit knowledge, at least implicit knowledge conceived on the model of axiomatization.

B. Given the distinction between facts and linguistic truths employed throughout, there is another question close to that above. Here the question is again one of implicit as opposed to explicit knowledge, but limited merely to the facts expressible in a language:

> Q3. Can the totality of *truth* in the domain at issue be claimed and affirmed at least obliquely and implicitly on the model of inference from axioms?

This question demands something more like a Gödelian than a Cantorian analysis. Here again, in ways allied with considerations above, the answer will be 'no.'

For any system adequate for arithmetic, and therefore of course for realms of truth and fact at large, there will be *truths* expressible in the language that are not deducible from the axioms. If even expressible truths within a language outstrip the implicit information of any axiom set, the implicit knowledge contained in axioms does not seem to offer an escape.

The question of implicitly knowing ML is particularly instructive. ML is too 'large' to be seen as a consequence of some other truth in the system: it contains by definition all information of all truths in the system. Nor can it function as an axiom which implicitly contains all other information, as long as 'implicitly' is taken on the model of inference from a consequence function R. By the results above, no system can express an ML from which all information is recoverable by inference.

Any appeal to implicit knowledge in the hopes of overcoming the limits we've documented above must appeal to implicit knowledge conceived on some model other than that of axiomatic containment or logical inference. The distinction between implicit and explicit knowledge remains an intriguing one, one that will reoccur in thinking about conceivability and reference to a world beyond.

C. Does essential limitation of knowledge doom us to error?

The numerical discrepancy between truth and fact means that our knowledge of a world of fact is bound to be imperfect. Specifically it means this knowledge is incomplete. Does it also mean that it is incorrect—that it contains not only gaps but errors? After all, suppose that you are otherwise fully informed about swans in general but totally unaware the some Australian Swans are black. One is then bound to arrive at the erroneous conclusion that all swans are white.

The incompleteness of our knowledge does not, of course, *ensure* its incorrectness—after all, even a single isolated belief can represent a truth. But it does strongly *invite* it. For if our information about some object is incomplete then it is bound to be unrepresentative of the objective make up-as-a-whole so that a judgment regarding that object is liable to be false. The situation is akin to that depicted in John Godfrey Saxe's "The Blind Men and the Elephant" which tells the story of certain blind sages who variously read incomplete evidence as indicating a creature like a wall, like a spear, a snake, a fan, or a rope. "Each was partly right," Saxe concludes, "And all were in the wrong."

The lesson is clear. The incompleteness of object-descriptive statements certainly does not entail their incorrectness: incomplete information does not ensure false belief with categorical necessity. But it does ensure inadequate understanding since at the level of generality there will be too many gaps that need filling in. There are just too many alternative ways in which reality can round out an incomplete account to warrant confidence in the exclusion of error.

This vulnerability of our putative knowledge of the world in the face of potential error is rather *exhibited* than *refuted* in our scientific knowledge. For this is by no means as secure and absolute as we like to think. We cannot but recognize in our heart of hearts that our putative truth in fact incorporates a great deal of error. There is every reason to believe that where scientific knowledge is concerned further knowledge does not just supplement but generally corrects our knowledge-in-hand, so that the incompleteness of our information implies its presumptive incorrectness as well.

D. To this point we have concentrated on the disparity between the limited world of linguistic truth and the larger world of fact beyond, but the range of these deliberations can be extended yet further. It is not merely in language that we manage our attempts at grasping facts, but in conceptualization and thought. Although neither speculation nor conceptualization need be recursively conceived or recursively limited, the same quantitative disparity between epistemic thinkability and ontological actuality will obtain in these contexts as well.

Is there reason to think that the realm of fact must outstrip pure conceivability? We have seen the limitations of language, and a long philosophical tradition insists that the limitations of language are necessarily the limitations of conceivability and therefore of knowledge as well. If we conduct the business of conception and knowledge via language, the limitations we've already noted, essential to any language, will be limitations of conceivability and knowability as well.

But limitations will still face us even if we abandon the assumption that conceivability and knowledge are tied to language. Let us assume a notion of conceivable propositions beyond the limits of linguistic expression: the conceptual parallel to facts rather than truths, perhaps. Consider all the propositions you have entertained in the course of reading this article, or all the propositions that have come to mind throughout the day. Consider all the propositions you have ever entertained, or all the propositions which you will in fact entertain throughout your lifetime.

The world of fact will necessarily outstrip any such set of propositions. There will be more subsets of propositions than there are propositions themselves. For each of these, there will be a specific fact: that a given proposition P is or is not a member of that set, for example. There will then be more factual propositions than those that you conceive in a day or indeed that all humans conceive in the course of human history. The world of fact will necessarily outstrip the realm of propositions conceived, and thus of course of things known.

The argument takes us even further. For consider not merely the propositions that have or will be conceived, but the propositions it is in any way possible to conceive: not merely the conceived but the *conceivable* propositions. For even these a numerical argument will apply: there will be more subsets of propositions, and thus more facts, than there are *conceivable* propositions.

The implication is that there are facts that are not even *conceivable*. That conclusion, of course, is one that holds on the level of generality. We cannot meaningfully claim to know—or even conceive—of any of them. The claim that there are inconceivable facts is in that regard like the claim that there are facts that I do not in fact know. I can conceive of there being inconceivable facts, of course without being able to conceive of any of the specifics, just as I can know there are facts I don't in fact know, of course without knowing any of those specific facts.

Unlike the image of musical chairs, the inconceivable facts would have to be specific inconceivable propositions. The realm of what is actually conceived, by a person on a day, in a lifetime, across all human history or by all creatures capable of entertaining propositions might have been different. But the realm of what is conceiv*able* in any of these categories would seem to be metaphysically fixed. If there are more facts than there are conceivable propositions, there must be *specific* facts beyond the range of propositional conceivability.

E. There is an air of paradox at this point: in conceiving of inconceivable facts, have we not somehow made them conceivable after all? Hints of paradox do mark any attempt to glimpse the world beyond, but there are several relevant considerations here.

Here as before we might appeal to a distinction between explicit and implicit conception, direct or indirect, full or weakly oblique. In a full sense a propositions is conceived in a full sense only when it is entertained in full content and with genuine understanding. In a far weaker sense, a proposition may be conceived *of* in any of a number of indirect ways—as the core propositions that a speaker will be arguing for, for example, but that I have not yet heard. We can thus think of the numerical argument as leading us to the weaker conception of propositions that are beyond conceivability in the full sense.

We can perhaps press the paradoxical character of the argument, however, by explicitly considering all facts that might be conceived *of*. In a similar fashion, we might consider all the facts that might be referred to in any way, either directly or obliquely. Given the basic Cantorian argument, there will be more facts than can be conceived of, and more facts that can be referrred to in any way. If the 'however possible' defines a fixed set, there will be specific facts that cannot even be conceived *of*, and which cannot be referred to in any way. But have we not just conceived of those? Have we not just referred to them?

There is an escape clause here that we will return to below and that we will in fact use as a window to the world beyond. For now let us note that the core argument, like its predecessors, relies on essential assumptions of number: the assumption of fixed collectivities with a given cardinality.

Applied to conceivable or referrable facts, the argument take the same form as that used earlier to show that the truths within any language will be outstripped by the facts of a world beyond. In that case the character of languages does indeed commit us to a fixed collectivity of expressions and thus of expressible truths that has a specific cardinality. When it comes to facts conceivable in any sense, to facts referrable in any sense, or to all the facts themselves, it will be these assumptions of collectivities bound by familiar principle of number that we will have to leave behind.

Our proposal is that we not treat the reasoning that leads us to such a point as somehow illegitimate, in need of a 'solution.' Our proposal is that we let the logic lead us to a genuine though radically unfamiliar realm beyond.

5. Facing Facts

Any world of fact must extend beyond language and beyond explanation. In at least some sense, it must extend beyond conceivability as well. Is any glimpse of the character of such a world simply impossible?

We think not. Our goal is to offer a glimpse of that world beyond.

The results that have led us here should warn us that the full world of fact will not be conceived in standard terms. Some of our familiar ways of approaching things must be compromised. Interestingly, they may be compromised in any of several ways.

If there is a world of fact, we will propose, its collectivity must be conceived as a *plenum*. Plena are supra-numerical collectivities that violate at least one of several standard logical assumptions. Among such supra-numerical collectivities are the totality of all things, of all abstract objects, of all propositions. Like these, we propose, the world of fact constitutes a plenum.

A. Consider a Cantorian argument applied directly to the totality of facts. Given any such totality, there will be more sub-collections of the totality than there are members. But for each of those sub-collections there will be a distinct fact: that a given fact f is or is nor a member of that sub-collection, for example. There will then be more facts than contained in the totality of facts.

Something has to give. The argument can be perspicuously rendered as an aporetic triad:

> 1. The Cantorian assumption: There will be more sub-collectivities of any collectivity than there are members of that collectivity.
>
> 2. The Factual assumption: For any sub-collectivity of any collectivity there will be a distinct fact.

3. The Totality assumption: there is a collectivity that contains all facts.

Given (1), there will be more sub-collectivities of the collectivity assumed in (3) than there are members of (3). Given (2), there will be more facts than there are members of (3). Given (3), there will be more facts than are contained in a collectivity that contains all facts.

In this form the aporia is clearly one of number: any supposed totality of facts will have more members than it has members. Whatever *number* it contains, it must contain more than that number. Our exploration will involve digging beneath that concept of number. We begin, however, by surveying possible options.

One option is to deny (3). Despite appearances, despite deep intuitions, and perhaps despite our apparent ability to quantify over facts in general, there simply is no totality of facts. The world of facts is essentially incomplete: facts refuse to form a whole. The universe, on such an approach, is incomplete. It is this option that one of us has argued for in earlier work.[4] Aristotle, Kant, and Russell can be seen as precursors.[5] 'Indefinite extensibility' approaches, in denying a completed totality, can also be seen in this tradition.[6]

Another option is to deny (2). Despite appearances and despite deep intuitions there are things regarding which there are no facts. The things are there, they are what they are, but there is no fact regarding them. However difficult to believe, such an approach has also been attempted.[7]

The third option, which we will pursue, is to deny (1). There are collectivities for which Cantorian assumptions do not hold: collectivities beyond standard principles of number.

These collectivities will in fact be *defined* as having a unique member for each of their sub-collectivities. For any conception of their contents at any moment of thought—for any snapshot of membership at any conceptual moment—these collectivities will contain more. These collectivities, beyond standard assumptions of either sets or any collectivities like them, are *plena*.

These collectivities will in fact be *defined* as having a unique member for

[4] Grim, *The Incomplete Universe*.

[5] Graham Priest, *Beyond the Limits of Though* (New York: Oxford University Press 2002), 229.

[6] See Stewart Shapiro and Crispin Wright, "All Things Indefinitely Extensible," in *Absolute Generality*, eds. Agustín Rayo and Gabriel Uzquiano (Oxford: Oxford University Press, 2006), 255.

[7] Keith Simmons, "On An Argument Against Omniscience," American Philosophical Association, New Orleans, April 1989.

each of their sub-collectivities. For any conception of their contents at any moment of thought—for any snapshot of membership at any conceptual moment—these collectivities will contain more. These collectivities, beyond standard assumptions of either sets or any collectivities like them, are *plena*.

We can construct a graphic example if we think of patterns of one or more patches on a two-dimensional plane, where each patch of a pattern must have an area. A pattern in our sense consists of a collection of patches that need not be contiguous, and indeed that might overlap. Graphically portrayed, one might think of a patch within another patch distinguished by a different color. For completeness, we include a completely blank plane as a pattern as a well.

Given this concept of patterns, it is clear that both any sub-pattern of a pattern and any collectivity of patterns will themselves constitute a pattern. The totality of all patterns will constitute a plenum, since every collectivity of elements of that totality—analogous to the elements of the power set of a set—will also constitute an element of the totality.

If propositions are understood as claims to facticity in the abstract, beyond any linguistic limits of mere statements, the totality of all propositions will constitute a plenum. For every collectivity of propositions there will be a distinct proposition—that a favored proposition p is included in that collectivity, for example (whether true or not)—and thus the totality of propositions will contain as many propositions as there are collections of propositions. The totality of things will constitute a plenum, if 'things' is broad enough to include collections. Every collectivity of things will constitute a thing in its own right. The totality of abstract objects will constitute a plenum for similar reasons.

Moreover, facts taken as a whole will form a plenum as well. There indeed is a world beyond language, sets, and systems. This, to be specific, is the plenum constituted by the world of facts.

There are several approaches to the aporetic triad that have points in common with the approach we take here, though we regard these as mere points of contact, short of the full metaphysical vision of a trans-numeric world of fact that we propose. In an attempt to understand truth, Hans Herzberger, Anil Gupta, and Nuel Belnap envisage truth as a concept that forces its own revision, much in the way that any attempt to conceive of the contents of a plenum as a fixed collectivity forces a revised vision of its further extent.[8] In the same light, an approach in terms of 'indefinite extensibility' has points of contact with our own. Graham Priest urges us to welcome any inconsistency in the aporetic triad for its

[8] Hans Herzberger, "Notes on Naïve Semantics," *Journal of Philosophical Logic* 11 (1982): 61, Anil Gupta and Nuel Belnap, *The Revsion Theory of Truth* (Cambridge, MA: MIT Press, 1993).

own sake, opening dialethic arms to 'true contradictions.'[9] We will not knowingly embrace contradiction. There is nonetheless a way of reading some of Priest's conclusions—that totalities at issue are both complete and not—that does resonate to some extent with the vision of plena we wish to present.

B. As expressed above, the aporetic triad turns on a concept of number that is buried within the Cantorian assumption. 'There will be *more* sub-collectivities of any collectivity than there are members of that collectivity.' On a Cantorian conception of number, the claim that a collectivity Y contains more than another collectivity X means simply that any line-up of the two such that every member of X is assigned a distinct member of Y will leave out some member of Y: the 'more' that Y contains.

Cantor's theorem is that the subsets of any set S—elements of its power set **P**S—will necessarily outnumber the elements of S. The proof is a proof that there can be no mapping M of elements of S onto distinct elements of **P**S that doesn't leave some element of **P**S out. For any proposed M, the proof offers a specific element of **P**S that must be left out. Here two points are of particular note. The first is that the 'specific element of **P**S' or subset of S that is necessarily excluded from the mapping M is itself specified in terms of M and a specific relation R. The second is that the 'necessary exclusion' of that element is exclusion on pain of contradiction. Derivation of the contradiction demands exclusive and exhaustive alternatives regarding an element of **P**S and that element of S mapped to it by M. S must either stand in relation R to its corresponding element or not. At its foundations, then, the 'more' of our aporetic triad is a matter of contradiction given exclusive alternatives and a peculiar reflexivity involving a mapping M and relation R.[10]

Although our target is collectivities well beyond mere sets, it is worthwhile to review the general mechanisms of the familiar set-theoretic proof. We assume any mapping M designed to assign each member of S to a unique member of its power set **P**S. The relationship R is set-membership, a crisp binary relationships fully obtaining or failing to obtain between any two candidates. We then consider a particular subset of our original set, specified in terms of M and R: the set D (for diagonal) of precisely those members of S which are not members of the subset to which they are assigned by our mapping M. If M fulfilled the conditions of 'same number,' giving us a one-to-one correspondence onto all elements of **P**S, it would

[9] Priest, *Beyond the Limits of Thought*.

[10] Patrick Grim and Nicholas Rescher, *Reflexivity: From Paradox to Consciousness* (Frankfurt: Ontos Verlag 2012).

assign some member s of our original set to D. But given either of two exclusive and exhaustive alternatives regarding membership, any such assignment leads to contradiction. If s of S is a member of D, it will by specification of D *not* be a member: D is to contain *only* those elements of S that are not members of the subset assigned by M. If s is not a member of D, it will by specification of D be a member of D: D is to contain *all* those elements of S that are not members of their corresponding subset.

The power set of any set must be larger than the set itself. In the context of classical set theory, the obvious next question has always been 'and what of the set of all sets?' By virtue of containing all sets, it must contain the elements of its own power set. But won't we then be forced to conclude that it is larger than itself?

With an eye to possible exportation to the aporia regarding all facts, consider standard responses to the strictly set-theoretic issue of a set of all sets. The standard line, despite appearances, despite intuitions, and perhaps despite our apparent ability to quantify over sets in general, is to deny the existence of a set of all sets. One move here, kicking the problem upstairs, is to create a new department of 'classes,' to one of which all sets (but of course not all classes) are assigned.[11] Another move is to deny or restrict the power set axiom, required in standard axiomatization to give us **P**S for arbitrary sets S to begin with.[12] A third move, echoing a theory of types, is to attempt to restrict the specifications of subsets so as to exclude the specification required to give us D.

C. None of the standard options for dealing with a set of all sets can be said to be intuitive. All look like cheating. All carry an atmosphere of the ad hoc. Parallels to those options become even less intuitive when we attempt to export them to the issue of a totality of facts.

For every collectivity of facts there will be a distinct fact: that a chosen fact is an element of that collectivity, for example, or that it is not. That a chosen fact is entailed by the collectivity, or that it is not. That the collectivity is finite, for example, or that it is not. That some of its elements entail other elements, or that all elements of that collectivity are logically distinct. Consider any of these 'collectivity facts' regarding the facts of a specific collectivity.

Consider now (a) the elements of a collectivity of all facts and (b) facts regarding collectivities of these, of any of the forms above: facts as to the facts they contain, facts regarding the facts they entail, the finitude or infinitude of the

[11] The further sorrows of class theory are documented in Grim, *The Incomplete Universe* and Priest, *Beyond the Limits of Thought*.

[12] Christopher Menzel, "On Set Theoretic Possible Worlds," *Analysis* 46, no 2 (1986): 68. See also Menzel, "Sets and Worlds Again," *Analysis* 72, 2 (2012): 304.

collectivities at issue, or the like. We can think of the facts falling within the collectivity of a collectivity fact (b) as facts within its domain. Somewhat more informally, but to the same point, we might think of the facts within the domain of a collectivity fact as facts it is *about*.[13]

Take any one-to-one mapping M from the facts of (a) to the collectivity facts of (b). Any such mapping must leave some element of (b) out. Consider in particular all those facts on the left that do not fall within the domain of their associated collectivity fact. There will be a fact df about precisely that collectivity: that it entails a chosen fact f or that it does not, that it is finite or infinite, and the like. But there can be no element f* of (a) mapped to fact df. If f* falls within the domain of df, it cannot, by specification of df in terms of our mapping M. If f* does not fall within the domain of df, it must, again by specification of df.

In the context of the argument targeted to facts, the option of denying the existence of a set of all sets would be paralleled by a denial of any totality of all facts: denial of (3) in our aporetic triad above. On that line there is no world of all facts: the factual world refuses to form a coherent whole.[14] Such a route seems to violate the concept of a world.

The option of avoiding a set-theoretic diagonal set D by denying all sets within a power set **PS** is can be paralleled here by avoiding df, denying that any collectivity of facts is something about which there will be a fact. This amounts to a denial of (2) above. This route seems to violate the very concept of facts.

Neither of these options allows us a world of facts. One offers us a totality of something short of the ubiquity of facts. One offers us facts without a totality. On either approach, on pain of contradiction, we are again forced to conclude that there are too 'many' Cantorian facts to form a world.

One might choose simply to revel in contradiction. We take the result more seriously than that, as an invitation to explore a realm beyond. In the present line of inquiry we assume a genuine world of fact. We ask what results such as these have to show us about the possible character of that world, however strange.

What we explore is what must follow if we deny (1) of the aporetic triad. The world of facts, we propose, lies beyond a number of the Cantorian

[13] The difficulties of pinning down the concept of aboutness in even the context of linguistic statements, making free use of the concept of designating expressions, became evident long ago in an exchange between Rescher and Goodman (Goodman, "About," *Mind* 70 (1961): 1, Rescher, "A Note on 'About'," *Mind* 72 (1963): 268). The current deliberations extend beyond language, targeting a relation of aboutness between facts and facts. In the context of facts, we'll argue, the concept of aboutness is not merely difficult to define but indeterminate in application.

[14] As in Grim, *The Incomplete Universe*.

assumptions. The world of facts forms a plenum.

6. The World of Fact as Plenum

We define a *plenum* as a collectivity that contains distinct elements corresponding to each of its sub-collectivities, where sub-collectivities follow the same pattern as subsets: something qualities as a sub-collectivity of a collectivity C just in case each of its members is a member of C.

In a membership plenum, such as a collectivity of all collectivities A, each sub-collectivity is itself a member of A. In other plena, such as the collectivity of all facts F, there is a fact regarding each sub-collectivity of F that is itself a member of F. Membership plena contain their own power collectivities. Other forms of plena contain members that map onto their power collectivities.

We take such plena to exist, with the world of fact as an example so intuitive as to be undeniable. The question for us, then, is not whether there is a world of fact but what such a world must be like.

A. We assume both (2) and (3) of the aporetic triad above. For anything that exists—and thus for any sub-collectivity of any collectivity—there will be a distinct fact. There is moreover a world of all facts. What we must deny, then, is the Cantorian core in (1): the claim that there will be more sub-collectivities of any collectivity than there are members of that collectivity.

The key to the Cantorian argument is that crucial concept of number: the claim that there will be *more* sub-collectivities of any collectivity than there are members of that collectivity. That 'more' amounts to the thesis that there can be no one-to-one mapping M from elements of a collectivity C to elements of its power-collectivity **P**C or some collectivity F**P**C which contains distinct members for each element of **P**C.

If we are to embrace plena as collectivities with members for each sub-collectivity, we must deny that there will be 'more' of the latter. We must hold that there **P**C be a mapping M from C to **P**C or F**P**C which leaves no element of the latter out.

In doing so we have to find the loophole in the Cantorian argument that attempts to show there can be no such M. That argument rests on specification of a particular element D of **P**C or F**P**C which stands in relation R to all and only those elements of C to which their corresponding M-correlate does *not* stand in relation R. Our assumed mapping, in assigning an element of C to every element of **P**C or F**P**C, must assign an element d to D.

B. The crucial step in the argument is the dilemma step. Does d stand in relation R to D, or not? If not, by specification of D in terms of M, d must stand in relation R to D. But if it does, again by specification of D, it cannot.

The lesson, we believe, is that for any plenum there will be inherent indeterminacy in R. For any M, any R, and any D definable in terms of M and R, the M-correlate to that D neither will nor will not stand in relation R to D. In the case of a simple membership plenum C, for every way M of assigning elements of C to elements **P**C one-to-one, the element d of the plenum assigned to that D by M neither will nor will not be a member of D. In at least some cases, the Law of Excluded Middle LEM will fail for the membership relation within plena. For some items x within a plenum P, it will be neither the case that $x \in P$ nor $x \notin P$. In that sense, some of the borders of plena will be imperfect, imprecise, or indeterminate.[15]

The lesson regarding a world of all facts is clear as well. The Cantorian argument regarding facts relies on 'collectivity facts': facts regarding whether a specific collectivity of facts contains or entails a specific fact, for example, or is finite or infinite. The crucial question of that argument is whether a specific fact lies within the domain of such a fact: somewhat informally, whether it is one of the facts that collectivity fact is about. Because the world of facts is a plenum, the relevant relationship—that a fact lies within the domain of another, is one of the facts it is about, or is one of the facts for which the collectivity fact holds—must in at least some cases be indeterminate. It is not always the case that a fact is either an element of a specified collectivity of facts or is not. It is not always the case that a fact is either one of the facts another fact is true of or is not. It is not always the case that one fact subsumes another, or is about another, or is not.

That, we suggest, is the lesson to be drawn from the clear existence of a world of fact. Given a total world of fact, various facts about the world will have to be indefinite, indeterminist, or undefined. Corresponding to a multitude of collectivity-defining characteristics Y there will be a multitude of factual theses of the form 'It is not always the case—it is not always itself a fact—that a particular fact f is either Y or not Y. What might be called alethic indeterminacy—indeterminacy of fact—will pervade the world of fact.

Our reflections have brought us to alethic indeterminacy from consideration of a fact's membership in a given collectivity of facts, or having a characteristic shared by certain facts. In that train of thought, it appears to be on the meta-level of facts about facts that is crucial. At this point both the substance and form of the result are reminiscent of Gödel, though with an enlarged perspective. Gödel

[15] This indeterminism bespeaks a curious parallelism between the ream of the theoretically very large—plena—and the physically very small—quanta.

showed that any consistent systematization of arithmetic will be incomplete, leaving the provable truth or falsity of certain arithmetical truths undetermined. The proof involves the technique of Gödel numbering, allowing statements of the base language to correspond to or 'encode' second-order statements regarding theoremhood within the system. Our conclusion also involves reflexivity, though it applies in the metaphysical realm well beyond logical systems: any totalization of fact is going to leave the status of certain factuality-claims indeterminate. Given the structural similarities, resonant results in this enlargement of perspective should perhaps not be entirely surprising.[16]

It should be emphasized that the denial of LEM at issue throughout is a *strong* denial, rather than invocation of either a third alternative or any number of additional alternatives. Were we to think in terms of three exhaustive categories—that a fact (i) falls within the domain of another, (ii) oes not, or (iii) neither does nor does not—we could construct a relation R in terms of teh second two that would be sufficient for resurrection of the basic argument. Were there *any* totality of exhaustive categories, we could do precisely the same. The strong denial of LEM is a denial that there is *any* set of exhaustive categories regarding the relationships between facts and collectivity facts at issue.[17] The lesson to be drawn from the clear existence of a world of fact is that a prime characteristic of some facts—that they take other as part of their subject collectivity—does not hold in terms of any set of exhaustive categories regarding all pairs of facts. In that sense, the lesson of a world of fact is that certain characteristics of facts themselves are not what we might have taken them to be.

The argument may well generalize to other characteristics of facts. It is worthy of note, however that it will not generalize to all. The Cantorian argument cannot be plausibly constructed in terms of just any relation R.

Consider an attempt to construct the argument in terms of logical entailment, for example. Some facts and some sets of facts logically entail others. For any M from facts to elements of the power set of a set of all facts, we might then envisage D as all those facts which are not entailed by the elements of the power set to which M assigns them. M must assign a fact d to that D.

But what then is the crucial question required for a Cantorian dilemma? We might first phrase the question as one of membership: Will d be a member of D or not? If it *is* a member, it will not be entailed by its corresponding set D. Interestingly, we cannot maintain that option: if d is a member of D, D certainly will entail d. But we can maintain that d is *not* a member of D. It follows that D

[16] See also Grim and Rescher, *Reflexivity*.

[17] See Rescher and Grim, *Beyond Sets*, chapter 6.

will entail d without containing it, but that does not give us contradiction. A set of propositions may entail many that it does not strictly contain.

We might alternatively ask whether d will be logically entailed by D. If it is not, it is an element of C not entailed by its M-correlate, and so will be a member of D. But as a member of D, of course, it will be logically entailed by D. The hypothesis that d will not be logically entailed by D is inconsistent. But the hypothesis that d *will* be logically entailed by D is not. In that case d, though not a member of D, will be entailed by D. Once again, a set of propositions may entail many that it does not strictly contain.

Given a world of facts, some relations—whether one fact falls within the collectivity addressed by another, for example—must be indeterminate. Logical entailment, on the other hand, need not be.

Though short of contradiction, there is a strange consequence of the argument phrased in terms of logical entailment. Because it can be run for any proposed one-to-one correspondence M from facts to collectivities of facts, the Diagonal construction D for *every* such M will entail whatever d is assigned to it.

C. We have defined plena as collectivities which take as members either their own subsets or elements such as facts mapped onto their subsets. Any world of fact would necessarily meet that criterion.

There are, we think, four options regarding plena:

> 1. Using standard logical principles, we might insist on Cantorian grounds that plena do not and cannot exist.
>
> 2. We might hold that plena do exist, but that the law of excluded middle fails to hold for all cases membership and crucial relations R.
>
> 3. We might hold that they do exist, but that the law of non-contradiction NC fails to hold in all cases for membership and crucial relations R.
>
> 4. We might hold that plena do exist, with every element of their power set as or corresponding to a member, and with power sets that are indeed larger than they are.

On the assumption of a world of all facts, (1) must be rejected. We have outlined (2) as a favored option, tracking some of its implications for the nature of facts. We consider (3) and (4) more radical options, but include consideration of these as well.

D. The dilemma at the core of the Cantorian argument takes the form 'Does d stand in relation R to D or not?' That dilemma assumes that its options are exhaustive—precisely the assumption denied in putting aside the law of excluded middle for such a case. That dilemma also assumes, however, that its options and their consequences are exclusive: that something cannot both stand in relation R to D *and* not. The force of the argument can be broken at that point if we simply shrug and accept both options.

The implication would be that for plena, issues of membership can be both 'yes' and 'no': in some cases collectivity c can both be a member of another collectivity c' and not be a member. In some cases a fact f can both fall within the domain of another fact f' and not fall within that domain.

Here consequences are roughly the dual of those outlined above. On denial of LEM, membership and whether a fact is among those another fact applies to are indeterminate in some cases. On a denial of the law of non-contradiction, these will be overdeterminate in some cases. In one case it is exhaustiveness of alternatives that is denied—that a fact is either among the collectivity to which another applies or that it is not. In another case it is exclusiveness of alternatives that is denied—that a fact cannot be both.

Our tendency, as noted, is to go for indeterminacy and the LEM. Another tack, however, would be to derive a disjunctive lesson. For plena, membership must either be indeterminate or overdeterminate in some cases. For facts, whether one fact falls within the domain of another must be either indeterminate or overdeterminate in some cases.

E. A last option, though the most radical, also has its attractions. Could there be a one-to-one mapping from a plenum to its subsets? From facts to sub-collectivities of facts? The answer from (2) and (3) is that there could be such a mapping. Plena need not be larger than themselves.

The last option is to accept the conclusion of the Cantorian argument. There can be no exhaustive mapping from a plenum to its subsets. Its power set is larger than it is, in that sense. But every one of its subsets appears as a member. It is therefore larger than itself. On this approach we maintain both the law of non-contradiction and the law of excluded middle. All the assumptions of the Cantorian argument stand, as does its conclusion.

Such an approach has some aesthetically pleasing elements. The idea that plena will be larger than themselves has an intuitive resonance with feelings one gets when thinking about a totality of fact, for example: having thought one had them all, one finds they are more. Plena seem to expand under our gaze.

There is also something pleasing in thinking of plena as the third step in size conception of collectivities. Finite sets are collectivities such that all proper sub-collectivities are smaller than the collectivity itself. Infinite collectivities are those such that some proper sub-collectivities are as large as the collectivity itself. Plena are collectivities such that some proper sub-collectivities are larger than the collectivity itself.

There is however a major sacrifice here as well. On such an approach there will be no one-to-one mapping from plena onto themselves. If there were, there would be a mapping onto their power set, violating the conclusion of the Cantorian argument.

The non-existence of a one-to-one mapping for plena would mean that there is no relation that holds one-to-one between members of a plenum. That would seem to force us to the most radically contentious option of all: to hold that items of a plenum will even fail to map onto themselves by way of a relation of identity. Even self-identity will fail for at least some items of a plenum.

For collectivities, this would appear to mean that whether something is identical to another—is the same collectivity as another—would in some cases be indeterminate. For facts, this would mean that whether something is the same fact as another would be indeterminate. On such a view we would have individual facts, we would have a totality of all facts as a plenum, but the concept of 'the same fact' would lose its grip. Here perhaps is the most complete sense in which we would lose the concept of number: we would lose the concept of distinct entities involved in the counting.

We cannot say that we recommend such a route: after all, "everything is what it is, and not another thing." Were one to take such an approach, however, we think the appropriate route would be to emphasize the extent to which the concept of identity in general becomes problematic at this juncture. Classically, identity is detailed in terms of features or properties: $x = y$ for $(\forall F)(Fx \equiv Fy)$. If having certain properties itself becomes problematic for elements of plena, the applicability of identity so understood may become problematic as well. It should also be noted that such a route, however radically contentious, is not without precedent: Peirce denies identity for elements of a continuum, which has a number of points of contact with plena as considered here.[18]

F. With the concept of plena in hand, we can return to some of the issues raised in previous sections.

[18] See Wayne C. Myrvold, "Peirce on Cantor's Paradox and the Continuum," *Transactions of the Charles S. Peirce Society* 30, 3 (1995): 508, Fernando Zalamea, *Peirce's Logic of Continuity* (Boston, MA: Docent Press, 2012), Benjamin Lee Buckley*, The Continuity Debate: Dedekind, Cantor, du Bois-Raymond, and Peirce on Continuity and Infinitesimals* (Boston, MA: Docent Press, 2012).

It is clear by Cantorian argument that there will be more facts than there are propositions conceived of in the course of human history. There are more sets of those propositions than there are those propositions themselves. But for each such set there will be a distinct fact. The world of fact will outstrip the world of human conception.

That alone may not seem surprising. Extending the argument in section III, however, seemed to lead us into paradox. A Cantorian argument can be run not merely on all proposition that have or will be conceived, but the propositions it is in any way possible to conceive: a collectivity of all *conceivable* propositions. On such an argument it appears that there will be propositions that cannot in any way be conceived. But does not our grasp of the argument itself demonstrate that we have in some way conceived of them? Similar paradoxes accompany Cantorian arguments regarding all facts that might be referred to in any way, either explicitly or indirectly. There will be more facts than these; but are we not at this point referring to those facts supposedly beyond reference?

In section III we alluded to an escape clause, pointing out that each of these relies on the essential assumption of fixed collectivities with a given cardinality. That assumption is the Cantorian assumpton (1) that we have abandoned in favor of plena in exploring a world of facts.

An escape from paradox by way of a similar denial seems called for in these cases as well.

These apparent paradoxes, we propose, like the question of a totality of facts, point to the existence of plena. The realm of conceivable propositions, conceivable facts, and facts to which we might at least obliquely refer may all form plena: collectivities for which every sub-collectivity corresponds to a member. If all of this holds for actual facts, it will clearly hold for the still richer realm of possibilities: these will all the more emphatically constitute a plenum.

On the assumption of such plena, cashed out in any of the ways we've outlined—by strong denial of the law of excluded middle, exceptions to the law of non-contradiction, or a vagueness of identity—the Cantorian argument falls short. When broadly construed so as to include oblique conception and reference, we can see the realm of possible reference and conception—like the world of fact itself—as forming a plenum.

We should remind ourselves that in a familiar range of more restricted considerations all the classical principles can still be maintained. It is only when we reach for a grasp of totalities such as the world of all facts that we turn the page, forcing us to resort to new devices. Are compromises in familiar principles such as the law of excluded middle too high a price to pay for recognizing the existence of plena? Here the simplest answer, we think, is that we have no choice: it seems inescapable that there must be a world of fact as a whole. If so, here as elsewhere, it is our thinking we must mold to the world rather than the other way around.

Newtonian physicists confronted modern science with a physically infinite astronomical cosmos the contemplation of whose vastness filled Pascal with vertiginous fright. Cantorian set theory confronted modern mathematics with a qualitatively infinite numerical realm of numberless quantities.

The present deliberations confront modern philosophy with an epistemically infinite manifold of fact. Modernity is replete with challenges of coming to terms with the many guises of infinitude. Our discussion here is simply another instance of this larger phenomenon.

7. Conclusion

It is clearly demonstrable, from a number of sources and in a number of ways, that we face major limitations in the face of a world beyond the accustomed horizons of thought.[19] Our axiomatics imposes limits on formalization, with corresponding limits on explanation and the principle of sufficient reason. Godelian arguments show that demonstrable fact cannot exhaust fact.

Our language imposes limits on expressibility, limits that extend even to all possible languages. We argue that even expressible fact cannot exhaust fact. Beyond these, even conceivability faces inherent limits: the world of facts necessarily outstrips the world as we conceive it.

Despite those limitations, we propose that we can get a glimpse of the world of fact beyond. We can limn its general shape as that of a plenum: a collectivity that includes elements corresponding to all sub-collectivities.

Recognition of that fact, however, also forces us to recognize that such a world is unfamiliar in at least one of several ways. There is indeed a world of fact. But certain relations of facts to facts that might be assumed unproblematic—such as the question of whether one fact falls in the subject domain of another--will have weaker logical properties than we might have assumed. We have to conclude that whether one fact is about another may be indeterminate, in the sense of a strong denial of the law of excluded middle, or over-determinate, in the sense of a violation of the law of non-contradiction. A third alternative is that both of these hold, but hold for a range of things that are themselves less determinate that we

[19] Although such a phrase and much of the spirit of our piece echo Graham Priest's title for *Beyond the Limits of Thought*, it should be clear that his acceptance of contradictions is just one of the approaches we've outlined.

might have taken them to be. On the third alternative, it is a principle of identity that fails to hold in all cases: ‘the same fact’ loses its grip.

Language is a purposive instrument. Ordinary language has evolved for everyday use. Logico-Mathematical language primarily for logico-mathematical purposes. But beyond those familiar purposive horizons there lies the realm of abstract deliberation—a conceptual Wild West outside the pale of familiar logical law. Here the very questions one asks tend to be nonstandard. When you ask extra-ordinary questions, we propose, you must expect extra-ordinary answers. The reality beyond our conceptual horizons is a world about whose *being* we can reasonably say something but regarding whose n*ature* we do and can know effectively nothing. Our acknowledgment of this world is a constructive reminder to being honest and humble. It is the epistemic equivalent of the Roman functionary whose task was to give the emperor an ongoing reminder: “Remember that thou are but mortal.”

EPISTEMIC UTILITY AND THE NORMATIVITY OF LOGIC

Richard PETTIGREW

ABSTRACT: How does logic relate to rational belief? Is logic normative for belief, as some say? What, if anything, do facts about logical consequence tell us about norms of doxastic rationality? In this paper, we consider a range of putative *logic-rationality bridge principles*. These purport to relate facts about logical consequence to norms that govern the rationality of our beliefs and credences. To investigate these principles, we deploy a novel approach, namely, epistemic utility theory. That is, we assume that doxastic attitudes have different epistemic value depending on how accurately they represent the world. We then use the principles of decision theory to determine which of the putative logic-rationality bridge principles we can derive from considerations of epistemic utility.

KEYWORDS: normativity of logic, epistemic value, accuracy-first epistemology

How does logic relate to rational belief? Is logic normative for belief, as some say? What, if anything, do facts about logical consequence tell us about norms of doxastic rationality? Here are some putative norms that seek to connect logic and rational belief:

> (BP1) If Priest's Logic of Paradox governs propositions A and B, and B is strictly stronger than A in that logic, then, if you believe A, then you ought to believe B.
>
> (BP2) If classical logic governs $A_1, \dots, A_n, B$, and A_1, …, A_n together entail B in that logic, then you ought not to believe each of A_1, …, A_n while disbelieving B.
>
> (BP3) If you know that strong Kleene logic governs A and B, and you know that A entails B in that logic, then you have reason to see to it that your credence in A is at most your credence in B.

These illustrate something of the variety of claims that we might make in this area. Following John MacFarlane, we call such claims *bridge principles*—in particular, they are *logic-rationality bridge principles*.[1] Below, I will extend MacFarlane's taxonomy of such bridge principles to bring some order to this variety. Having done that, I wish to explore a novel way of adjudicating between them. In the existing literature, the following sorts of reasons are used to justify rejecting a given proposal of this sort:

[1] John MacFarlane, "In What Sense (If Any) Is Logic Normative for Thought?" (unpublished manuscript).

> *Conflicts with intuition.* For instance, we might reject (BP2) by appealing to our intuitive reaction to cases like Makinson's Preface Paradox.[2] Suppose A_1, …, A_n enumerate all of my beliefs about British birdlife. So, for each A_i, I believe it. But I also realise that I am fallible on this topic. And thus, I disbelieve B, the proposition that all of my beliefs are true—that is, I disbelieve $B = A_1$ & … & A_n. Nonetheless, A_1, …, A_n together entail B. So I violate (BP2). Yet intuitively, we judge that I am perfectly rational. For this reason, some argue, we should reject (BP2).
>
> *Conflicts with ought-can.* It is often noted that principles like (BP1) are extremely demanding, partly because we are not in a position to discover *all* the logical consequences of our beliefs, but also because, even if we could, we would be unable to store beliefs in all of them.[3] Suppose, for instance, that A is the conjunction of the second-order Dedekind-Peano axioms for arithmetic. Then presumably we cannot discover all of the consequences of A; and even if we could, we could not store them.[4] Thus, we might take (BP1) to fail on the grounds that it conflicts with an ought-can principle.

Also, in recent unpublished work, Claire Field and Bruno Jacinto have tried to justify bridge principles in the following way:[5]

> *Justification on the basis of norms.* They consider various norms that govern our beliefs. They consider the Truth Norm of Belief and the Knowledge Norm of Belief. And they ask which bridge principles follow from those norms. When considering the consequence of the Knowledge Norm for Belief, they consider the effects of assuming different frame conditions on the accessibility relation in the epistemic logic.

I wish to explore an alternative approach: sometimes this approach supplies a reason for rejecting a putative logic-rationality bridge principle, and sometimes it supplies a justification for accepting such a principle.

> *Justification by appeal to epistemic utility.* In recent years, a number of philosophers have appealed to considerations of epistemic utility in order to justify various epistemic norms. Kenny Easwaran, Branden Fitelson, and Kevin Dorst have sought to establish the Lockean thesis concerning the normative link between credences and full beliefs, while Ted Shear, Branden Fitelson, and

[2] David Makinson, "The Paradox of the Preface," *Analysis* 25, 6 (1965): 205-207.

[3] Gilbert Harman, *Change in View* (Cambridge, MA: MIT Press, 1986)

[4] And, even if we could store them, surely that would not be a good use of our storage facilities. Note, however, that this last point does not turn on a conflict with an ought-can principle, but rather a conflict with a plausible principle governing how we should sensibly use our limited storage capacities.

[5] Field and Jacinto presented this work at a conference, *The Normativity of Logic*, held at the University of Bergen, 14-16 June 2017.

> Jonathan Weisberg have offered justifications of some of the principles of belief revision.[6] On the credal side of epistemology, Jim Joyce and I have offered very closely related epistemic utility arguments for Probabilism,[7] Together, Hilary Greaves and David Wallace have argued for Conditionalization on this basis, and R. A. Briggs and I have recently offered an alternative justification of that updating rule;[8] Jason Konek and I have both sought to justify the Principal Principle;[9] I have provided a rationalisation of the Principle of Indifference;[10] Sarah Moss and Ben Levinstein have both sought norms that govern peer disagreement situations;[11] and Miriam Schoenfield has appealed to accuracy considerations to motivate a particular solution to the problem of higher-order evidence.[12]

We will spell out the idea behind these arguments in detail below, but roughly it is this. Our actions have different pragmatic value given different ways the world might be. We call this their *utility*. For instance, my action of betting that Labour will win the next UK General Election has high utility in worlds where they win and low utility in worlds where they lose. Similarly, our doxastic states—either our full beliefs, disbeliefs and suspensions of judgements, or our

[6] Kevin Dorst, "Lockeans Maximize Expected Accuracy," *Mind* (forthcoming), Kenny Easwaran "Dr Truthlove, Or: How I Learned to Stop Worrying and Love Bayesian Probabilities," *Noûs* 50, 4 (2016): 816–853, Kenny Easwaran and Branden Fitelson "Accuracy, Coherence, and Evidence," in *Oxford Studies in Epistemology*, volume 5, eds. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2015), Ted Shear, Branden Fitelson, and Jonathan Weisberg, "Two Approaches to Belief Revision" (unpublished manuscript).

[7] James M. Joyce, "A Nonpragmatic Vindication of Probabilism," Philosophy of Science 65, 4 (1998): 575–603, James M. Joyce. "Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief' in Franz Huber, & Christoph Schmidt-Petri (eds.) Degrees of Belief. (Springer, 2009), Richard Pettigrew, *Accuracy and the Laws of Credence* (Oxford: Oxford University Press, 2016).

[8] Hilary Greaves and David Wallace, "Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility," *Mind* 115, 459 (2006): 607–632, R. A. Briggs and Richard Pettigrew, "An Accuracy-Dominance Argument for Conditionalization" (unpublished manuscript).

[9] Jason Konek, "The Simplest Possible Accuracy Argument for the Principal Principle" (unpublished manuscript), Richard Pettigrew, "A New Epistemic Utility Argument for the Principal Principle," *Episteme* 10, 1 (2013): 19–35.

[10] Richard Pettigrew, "Accuracy, Risk, and the Principle of Indifference," *Philosophy and Phenomenological Research* 92, 1 (2016): 35–59.

[11] Sarah Moss, "Scoring Rules and Epistemic Compromise," *Mind* 120, 480 (2011): 1053–1069. Benjamin A. Levinstein, "With All Due Respect: The Macro-Epistemology of Disagreement," *Philosophers' Imprint* 15, 3 (2015): 1–20.

[12] Miriam Schoenfield, "An Accuracy-Based Approach to Higher-Order Evidence," *Philosophy and Phenomenological Research* (forthcoming).

credences—have different epistemic value given different ways the world might be. We will call this their *epistemic utility*. For instance, we might say that a true belief is more valuable than a false one, and a high credence in a true proposition is more valuable than that same high credence in a false proposition. Just as we choose between our actions using the principles of decision theory, so we might pick between different doxastic states using those same principles. After all, these decision-theoretic principles are simply claims about how facts about rationality are determined by facts about value—they govern how epistemic utility determines epistemic rationality just as much as they govern how pragmatic utility determines pragmatic rationality. Thus, just as previous authors have tested norms like Probabilism, Conditionalization, etc. by asking whether they follow from the principles of decision theory together with a particular account of epistemic utility, so we might test principles like (BP1), (BP2), (BP3), and their ilk, which claim to connect logic and rationality, in the same way.

It's worth emphasising here that proceeding in this way seems natural—more natural, perhaps, than appealing to intuition or to an ought-can principle. Presumably a large part of the reason why we think that logic might be normative for belief is that we think that beliefs aim at the truth; that is, we think that beliefs are better when true and worse when false. And presumably we also recognise that logic is the study of the relationships between the truth values of different propositions. If that's right, you should expect logic to tell you something about how best to obtain the aim of belief. Epistemic utility theory allows us to explore exactly how this might work. That's not to say that this is the only framework in which to explore this: seeking out the consequences of the Truth Norm for Belief, as Field and Jacinto do, is an alternative approach. But I hope to convince you that it is a fruitful way to do so.

A Taxonomy of Bridge Principles

Each principle that purports to connect logic and rationality shares the same form. It is a conditional. Its antecedent is a proposition $A(T \,\&\, L)$. T is a claim about the logic that governs some set of propositions; L is a claim about the consequence relation of that logic; A is a propositional operator that acts on the conjunction, T & L. The consequent of a bridge principle is a normative claim C concerning an agent's beliefs or credences. Thus, our logic-rationality bridge principles have the form $A(T \,\&\, L) \rightarrow C$.

In (BP1), T is the claim that the Logic of Paradox governs A and B, and L is the claim that A entails B in that logic, but B does not entail A. In (BP3), T is the claim that strong Kleene logic governs A and B, while L is the claim that A entails

B. In (BP1) and (BP2), $A(T \,\&\, L)$ is just $T \,\&\, L$, so A is the identity operator in this case, whereas in (BP3), $A(T \,\&\, L)$ is the proposition that you know $T \,\&\, L$, so that A is the knowledge operator in this case. In (BP1), C is the conditional: if you believe A, then you ought to believe B. That is, C is a *narrow scope norm*. In (BP2), C is a *wide scope norm*: it ought not to be that you believe each of A_1, …, A_n and you disbelieve B. In (BP3), the normative claim in the consequent is not stated in terms of ought at all; it is stated in terms of reasons, so it is weaker.

We now expand a little on John MacFarlane's taxonomy for bridge principles. MacFarlane lists a number of dimensions along which bridge principles can differ, and he lists the ways in which they might differ along these different dimensions. I simply add a couple of further dimensions to his list.

Grain What sort of doxastic states does the norm govern?

Credences The norm governs credences or degrees of belief.

Full beliefs The norm governs full beliefs, full disbeliefs, and suspensions of judgment.

Normativity What sort of norm is $A(T \,\&\, L) \rightarrow C$?[13]

Evaluation It is used only to evaluate an agent's doxastic state.

Appraisal It is used to apportion epistemic blame and fault to the agent.

Directive It is used to direct the agent's doxastic life.

Governing logic Which logic governs the propositions in question, according to T?

Classical We denote the consequence relation of this logic $\vDash_{\text{cl}}$

Strong Kleene logic We denote the consequence relation of this logic $\vDash_{\text{skl}}$

Logic of Paradox We denote the consequence relation of this logic $\vDash_{\text{lp}}$

and so on...

Strength of logical claim What is the strength of the claim L about logical consequence that occurs in the antecedent? Weak or strong?

Weak $A \vDash B$.

Strong $A \vDash B$ and $B \nvDash A$.

Antecedent operator What is the operator A in the antecedent $A(T \,\&\, L)$? That is, under what conditions on $T \,\&\, L$ does the bridge principle get triggered?

Identity $A(T \,\&\, L) = T \,\&\, L$.

[13] Cf. Florian Steinberger, "Three ways in which logic might be normative" (unpublished manuscript).

Obvious $A(T \,\&\, L) = T \,\&\, L$ is obvious.

Knowledge $A(T \,\&\, L) =$ You know $T \,\&\, L$.

Belief $A(T \,\&\, L) =$ You believe $T \,\&\, L$.

Number of premises L is a fact about logical consequence. That is, it is a proposition of the form $A_1, \ldots, A_n \vDash B$ for some propositions $A_1, \ldots, A_n, B$. What is n?

Consequent operator What is the operator in the consequent C? Is it an ought operator, a reasons operator, or a permission operator?

Ought C states a norm in terms of *ought*.

Reasons C states a norm in terms of *reasons*.

Permission C states a norm in terms of *permission*.

Consequent scope What is the scope of the operator found in the consequent C? Does it apply to the consequent of the conditional only, both antecedent and consequent separately, or the whole conditional together?

Consequent C takes the form $X \rightarrow N(Y)$, where N is the normative operator identified in the previous condition. Thus, C is a narrow scope norm.

Whole C takes the form $N(X \rightarrow Y)$. Thus, C is a wide scope norm.

Both C takes the form $N(X) \rightarrow N(Y)$.

Polarity What is the strength of the claim in the consequent of the conditional in C?

Positive The consequent of the conditional in C is a positive demand that the agent has a particular attitude.

Negative The consequent of the conditional in C is a negative demand that the agent does not have a particular attitude.

Picking a different answer for each of these gives a different putative normative claim about the connection between logic and rationality. Thus, for instance, (BP1) arises from the following choices: it governs full beliefs; the logic is Logic of Paradox; the logical claim is weak; the operator in the antecedent is the identity operator; the claim about logical consequent involves just a single premise; the operator in the consequent is the ought operator and that operator takes narrow scope in the consequent; and the polarity of the consequent is positive.

In what follows, we'll use epistemic utility to adjudicate between these different bridge principles. We'll divide our treatment into two parts: first, we'll treat full beliefs; second, we'll treat credences.

Bridge Principles for Full Beliefs

We begin by considering epistemic utility for full beliefs. I will present the now-standard veritist story for the classical case. This originates with Carl Hempel, but in its current form it is due to work by Kevin Dorst, Kenny Easwaran, and Branden Fitelson.[14] After that, I will extend it to the non-classical case.

Suppose you entertain a particular proposition; it is there before your mind. Then there are three categorical doxastic attitudes that you might adopt towards it: you can believe it (**B**), disbelieve it (**D**), or suspend judgment on it (**S**). Suppose $\mathcal{F}$ is the set of propositions that you entertain. We can represent your doxastic state by a function $b : \mathcal{F} \to \{\mathbf{B}, \mathbf{S}, \mathbf{D}\}$. We call this your *belief function*. Our first order of business is to describe an epistemic utility function for doxastic states represented in this way. An epistemic utility function takes a doxastic state and a possible world and returns a measure of how much epistemic utility that state has at that possible world. Here and throughout, we will assume a *veritist* account. That is, we will assume that the sole fundamental source of epistemic value for doxastic states is their accuracy; a doxastic state has greater epistemic value the more accurately it represents the world. One consequence of this is that the epistemic utility of your doxastic state at a possible world depends only on the truth values at that world of the propositions that you entertain. So, just as we can represent a doxastic state as a function from $\mathcal{F}$ to the set of possible doxastic attitudes, so we can represent a possible world as a consistent valuation function from $\mathcal{F}$ to the set of possible truth values. Since we are currently presenting the classical case, the set of truth values is $\{\mathbf{t}, \mathbf{f}\}$, and the consistency in question is classical consistency.

Now, we wish to define a function EU such that, if $b : \mathcal{F} \to \{\mathbf{B}, \mathbf{S}, \mathbf{D}\}$ is a belief function on $\mathcal{F}$ and $w : \mathcal{F} \to \{\mathbf{t}, \mathbf{f}\}$ is a classical valuation function on $\mathcal{F}$, then $\mathrm{EU}(b, w)$ is the epistemic utility of the doxastic state represented by b at the possible world represented by w. First, we assume that EU is *additive*: that is, $\mathrm{EU}(b, w)$ is the sum of the epistemic utilities at w of the different doxastic attitudes that b comprises. That is, there is a *local epistemic utility function* $\mathrm{eu} : \{\mathbf{t}, \mathbf{f}\} \times \{\mathbf{B}, \mathbf{S}, \mathbf{D}\} \to [-\infty, \infty]$ such that

$$\mathrm{EU}(b, w) = \sum_{X \in F} \mathrm{eu}(w(X), b(X))$$

[14] Carl Hempel, "Deductive-Nomological vs. Statistical Explanation," in *Minnesota Studies in the Philosophy of Science*, vol. III, eds. Herbert Feigl and Grover Maxwell (Minneapolis: University of Minnesota Press, 1962), 98–169, Dorst, "Lockeans Maximize Expected Accuracy," Easwaran "Dr Truthlove," Easwaran and Fitelson, "Accuracy, Coherence, and Evidence."

Thus, $\text{eu}(\mathbf{t}, \mathbf{B})$ is the epistemic utility of believing a proposition when it is true, while $\text{eu}(\mathbf{t}, \mathbf{D})$ is the epistemic utility of disbelieving a proposition when it is true, and so on. And $\text{eu}(\mathbf{t}, \mathbf{B}) + \text{eu}(\mathbf{f}, \mathbf{B})$ is the epistemic utility of an agent with one belief in a truth and one belief in a falsehood and no other doxastic attitudes. Next, we identify our proposed local epistemic utility function. It is this:

$\text{eu}(\mathbf{t}, \mathbf{B}) = \text{eu}(\mathbf{f}, \mathbf{D}) = R$ (for getting it Right)

$\text{eu}(\mathbf{t}, \mathbf{S}) = \text{eu}(\mathbf{f}, \mathbf{S}) = 0$

$\text{eu}(\mathbf{t}, \mathbf{D}) = \text{eu}(\mathbf{f}, \mathbf{B}) = -W$ (for getting it Wrong)

where $R, W > 0$. Thus, true beliefs and false disbeliefs are equally valuable, with epistemic utility R; and they are more valuable than suspensions, which are equally valuable whatever the outcome, with epistemic utility 0; and they, in turn, are more valuable than false beliefs and true disbeliefs, which are equally valuable, with epistemic utility $-W$. According to William James, two principles guide our epistemic life: *Believe truth! Shun error!*.[15] If you agree, you might take R and W to measure the strength of those two exhortations, respectively. The higher R, the more you care about getting things right; the higher W, the more you care about not getting things wrong. Thus, if $R > W$, you might call yourself an epistemic radical; if $R = W$, you are an epistemic centrist; and if $W > R$, you are an epistemic conservative.

Now, let's see what these different positions have to say about the logic-rationality bridge principles that we categorized at the beginning of the paper. Throughout, we will have cause to refer to five different belief functions defined on A and B. We define them here for ease of reference:

	A	B
b_1	B	D
b_2	B	S
b^*	S	S
$b_1^\dagger$	D	B
$b_2^\dagger$	S	B

Thus, for instance, we might take A to be *Labour will win* and B to be *Labour or the Greens will win*. Thus, b_1 believes that Labour will win, but disbelieves that Labour or the Greens will win, while $b_1^\dagger$ switches those attitudes, disbelieving that Labour will win, but believing that Labour or the Greens will

[15] William James, "The Will to Believe," in *The Will to Believe, and Other Essays in Popular Philosophy* (New York: Longmans Green, 1905).

win. Similarly, b_2 believes Labour will win, but suspends on Labour or the Greens winning, while $b_2^\dagger$ switches those attitudes. And b^* suspends on both propositions.

Let's start with the epistemic conservative; for them, recall, $W > R$. Then we have an epistemic utility argument for the following logic-rationality bridge principle:[16]

> (BP4) If $\vDash_{\text{cl}}$ governs A and B, and $A \vDash_{\text{cl}} B$, then you ought not to believe A while disbelieving B.

That is, when A and B are classical propositions, and A classically entails B, you ought not to have the belief function b_1. This is the single-premise version of the bridge principle that MacFarlane calls (Wo-).

Here's the argument for (BP4). Suppose A classically entails B. Then consider b_1 and b^*. While b_1 believes A and disbelieves B, b^* suspends judgment on both. Now consider the different ways the world might be and the epistemic utility of the two belief functions at those different worlds:

	A	B	$\text{EU}(b_1, w)$	$\text{EU}(b^*, w)$
w_1	t	t	$R - W$	$0 + 0$
w_2	t	f	$R + R$	$0 + 0$
w_3	f	t	$-W - W$	$0 + 0$
w_4	f	f	$-W + R$	$0 + 0$

Since $W > R$, it follows that $R - W < 0$. Thus, at all worlds except the one at which b_1 gets everything right—the world at which A is true and B is false—the epistemic utility of b_1 is negative; and the epistemic utility of b^* is always 0. However, given that A entails B, there is no world at which A is true and B is false—that is, w_2 is not a classically consistent valuation and thus does not represent a genuine possibility. So, at all logically possible worlds—that is, at w_1, w_3, and w_4—b^* has greater epistemic utility than b_1. That is, as a matter of logical necessity, it is epistemically better to have belief function b^* than b_1. That is, $\text{EU}(b_1, w) < \text{EU}(b^*, w)$ for all logically possible worlds w. In such cases, we say that b^* *strictly logically dominates* b_1 *relative to* EU.

Now, in decision theory, strict logical dominance is often taken to be a sign of irrationality. That is, the following is taken to be a principle of rationality:

> **Strict Logical Dominance** If option o^* has greater utility than option o at every logically possible world, then o is irrational.

[16] See Easwaran's "Dr Truthlove," for very closely related results in which the only categorical doxastic attitudes are belief and suspension.

Thus, we have our first epistemic utility argument for a logic-rationality bridge principle:

(EU1) Epistemic Conservatism + Strict Logical Dominance ⇒ (BP4).

Before we move on, it helps to see this argument in a particular case. Consider, then, the person who believes that Labour will win, and disbelieves that Labour or the Greens will win. Such a person would do better for sure if they were to suspend judgment on both propositions. If Labour do win, then their belief is true but their disbelief false, and that means that they have negative epistemic utility; if Labour don't win but the Greens do, then they do maximally badly, since both attitudes are wrong, so they have negative epistemic utility; and if neither Labour nor the Greens win, then they are in the same situation as when Labour wins, namely, that one attitude is right and the other wrong, and that means that they have negative epistemic utility. Thus, they are guaranteed to have negative epistemic utility. On the other hand, if they were to suspend on both propositions, they would be guaranteed to have a neutral epistemic utility of 0. Thus, suspending dominates.

Hopefully, this gives a taste of the sort of epistemic utility argument we will pursue in this paper. Each argument consists of the components: (i) an account of epistemic utility—in this case, we assumed Epistemic Conservativism; (ii) a decision-theoretic principle—in this case, Strict Logical Dominance; (iii) a mathematical fact that shows that, if you apply the decision-theoretic principle using the account of epistemic utility, you obtain the epistemic norm that you seek, such as (BP4)—in this case, we demonstrated the mathematical result using the truth table above.

There are a number of ways in which we might try to adapt this argument. We might ask what happens when we switch Epistemic Conservatism for Epistemic Centrism or Epistemic Radicalism; or when we include more than one premise in the fact about logical consequence in the antecendent; or when we replace classical logic with some non-classical alternative; or when we consider the possibility of rational ignorance of logical truths.

Epistemic Conservatism, Centrism, and Radicalism

First, let's see what happens when we move from Epistemic Conservatism to Epistemic Centrism or Epistemic Radicalism. Recall: according to Epistemic Centrism, $R = W$—getting things right is exactly as good as getting things wrong is bad. Now, we can see from the table above that, for the Epistemic Centrist, b^* does not strictly logically dominate b_1. After all, at worlds at which A and B are both true or both false, the epistemic utility of b_1 (namely, $R - W$) is the same as the epistemic utility of b^* (namely, 0); it does not exceed it. And indeed it is straightforward to see that no alternative belief function strictly dominates b_1.[17] Nonetheless, note that b^* is at least as good, epistemically speaking, as b_1 at all logically possible worlds, and strictly better at some. In such cases, we say that b^* *weakly logically dominates* b_1 *relative to* EU. In decision theory, weak logical dominance is often taken to be a sign of rationality in the same way that strict logical dominance is. That is, the following is taken to be a principle:

> **Weak Logical Dominance** If option o^* has at least as great utility as option o at every logically possible world, and greater utility at some, then o is irrational.

Now, note that, in order to apply this to the choice of belief functions on A and B, we must ensure that it is genuinely logically possible that A is false and B true. That is, Weak Logical Dominance will tell us nothing when B entails A. In that situation, b_1 and b^* are equally good in every logically possible world, and there's nothing irrational about picking an option with that feature. This gives us an epistemic utility argument for a slightly weaker version of (BP4):

> (BP5) If $\vDash_{\text{cl}}$ governs A and B, and $A \vDash_{\text{cl}} B$, and $B \nvDash_{\text{cl}} A$, then you ought not to believe A while disbelieving B.

Here's the argument:

> (EU2) Epistemic Centrism + Weak Logical Dominance ⇒ (BP5).

Weak Logical Dominance also proves crucial when we move to Epistemic Radicalism—that is, the claim that $R > W$. It is clear from the table above that b^* neither strictly nor weakly logically dominates b_1 when $R > W$. After all, in this situation, b_1 outperforms b^* when A and B have the same truth value, since $R - W > 0$. As above, it is straightforward to see that there is no alternative belief function that strictly dominates b_1 for the Epistemic Radicalist. But there is an alternative that weakly dominates it, namely, $b_1^\dagger$ from above. Recall: $b_1^\dagger$ disbelieves A and believes B.

	A	B	$\text{EU}(b_1, w)$	$\text{EU}(b_1^\dagger, w)$
w_1	t	t	$R - W$	$-W + R$
w_2	t	f	$R + R$	$-W - W$
w_3	f	t	$-W - W$	$R + R$
w_4	f	f	$-W + R$	$R - W$

[17] By checking cases, we can see that, if a belief function b strictly outperforms b_1 when A and B are both true, then b_1 strictly outperforms b when A and B are both false.

Thus, b_1 and $b_1^\dagger$ have the same epistemic value when A and B are both true or both false, and $b_1^\dagger$ is strictly better than b_1 when A is false and B is true. Thus, we have another epistemic utility argument for (BP5):

> (EU3) Epistemic Radicalism + Weak Logical Dominance ⇒ (BP5).

Thus, in sum: for every point on the scale between Epistemic Conservatism and Epistemic Radicalism, there is an epistemic utility argument for (BP5). Indeed, for every point on that scale, $b_1^\dagger$ weakly logically dominates b_1. Thus, we have:

> (EU4) Weak Logical Dominance ⇒ (BP5).

And, for Epistemic Conservatism, there is an epistemic utility argument for (BP4), namely, (EU1).

Moreover, note that a similar trick can be used to establish the following bridge principle:

> (BP6) If $\vDash_{\mathrm{cl}}$ governs A and B, and $A \vDash_{\mathrm{cl}} B$, and $B \nvDash_{\mathrm{cl}} A$, then you ought not to believe A while suspending judgment in B.

That is, if A is strictly stronger than B, then you ought not to have the belief function b_2 defined above. We can justify this by noting that b_2 is weakly logically dominated by $b_2^\dagger$, which we defined above. Recall: $b_2^\dagger$ suspends on A and believes B.

	A	B	$\mathrm{EU}(b_2, w)$	$\mathrm{EU}(b_2^\dagger, w)$
w_1	t	t	$R - 0$	$0 + R$
w_2	t	f	$R + 0$	$0 - W$
w_3	f	t	$-W - 0$	$0 + R$
w_4	f	f	$-W + 0$	$0 - W$

Thus,

> (EU5) Weak Logical Dominance ⇒ (BP6)

Thus, if $\vDash_{\mathrm{cl}}$ governs A and B, and $A \vDash_{\mathrm{cl}} B$, and $B \nvDash_{\mathrm{cl}} A$, then we have an argument against believing A and disbelieving B, and an argument against believing A and suspending on B. Since belief, disbelief, and suspension are the only available categorical doxastic attitudes to a proposition, it seems at first sight that these two arguments then furnish us with a further argument that, if you believe A, and you adopt any categorical doxastic attitude towards B, then you ought to believe B. But that's not quite right.[18] The problem is that, while arguments (EU4) and (EU5) establish flaws in believing A and either suspending on B or disbelieving B, they do not rule out the possibility that believing A and believing B is also flawed in the same way. Now it turns out that, if A is a contradiction then that is indeed the case. If A is a contradiction, believing A and believing B is strictly logically dominated by disbelieving A and believing B. However, if A is not a contradiction, then believing A and believing B is not even weakly logically dominated.[19] Thus, we have an argument for:

> (BP7) If $\vDash_{\mathrm{cl}}$ governs A and B, and $A \vDash_{\mathrm{cl}} B$, and $B \nvDash_{\mathrm{cl}} A$, and A is not a classical contradiction, then you ought to see to it that, if you believe A, and you entertain B, then you believe B.

The argument:

> (EU6) Weak Logical Dominance ⇒ (BP7)

(BP7) doesn't say that you ought to see to it that you believe B if you believe A. Rather, it says that you ought to see to it that, *if B is a proposition that you entertain and to which you assign an attitude at all*, then you believe B if you believe A. It does this by showing that it would be irrational to assign either of the alternative attitudes to B in a way that it wouldn't be irrational to believe B. This is as close as we can get to the principle that MacFarlane calls (Wo+).

The upshot of this section is that, when the logic is classical and the entailment is a strict single premise entailment, epistemic utility considerations vindicate the logic-rationality bridge principles that seem most natural. They justify the wide scope versions of the norms, and they justify the versions with negative polarity. So they support bridge principles that are not vulnerable to some of Harman's main criticisms. They are not excessively demanding, since they do not demand that an agent have any attitude at all towards B; rather, they only say what she should do if she does have an attitude towards B. And they posit wide scope norms, so they are not vulnerable to Harman's objection that narrow scope norms arbitrarily favour one way of resolving an inconsistency in your beliefs.

Multi-Premise Entailments

Next, let's see what happens when we include more premises. It turns out that the answer depends on the values of R and W. In this section, we'll have cause to refer to three different belief functions. Again, we define them now for ease of

[18] Thanks to Anandi Hattiangadi on this point.

[19] No alternative does as well when A and B are both true.

reference:

	A_1	…	A_n	B
b_3	B	…	B	D
b_4	B	…	B	S
$b^\diamond$	S	…	S	S

Suppose $\vDash_{\text{cl}}$ governs A_1, …, A_n, B, and $A_1, \dots, A_n \vDash_{\text{cl}} B$. Now, we might say that an assignment of epistemic utility is extremely conservative if $nR < W$. Then, assuming Extreme Epistemic Conservatism, b_3 is strictly logically dominated by $b^\diamond$ relative to EU. After all, $b^\diamond$ has exactly the same epistemic utility at every world—namely, 0—while b_3 performs best when A_1, …, A_n are all true, but B is false, and in that situation b_3 has epistemic utility $nR - W$, which by hypothesis is less than 0. Thus, we have an epistemic utility argument for the following bridge principle:

(BP8) If $\vDash_{\text{cl}}$ governs A_1, …, A_n, B, and $A_1, \dots, A_n \vDash_{\text{cl}} B$, then you ought not to believe each of A_1, …, A_n while disbelieving B.

This is the multi-premise analogue of (BP4) and MacFarlane's (Wo-). Here's the argument:

(EU7) Extreme Epistemic Conservatism + Strict Logical Dominance ⇒ (BP8)

(EU7) rules out b_3 as irrational. Interestingly, we cannot strengthen this argument to rule out b_4 as irrational as well. That is, we cannot give an argument for

(BP9) If $\vDash_{\text{cl}}$ governs A_1, …, A_n, B, and $A_1, \dots, A_n \vDash_{\text{cl}} B$, then you ought not to believe each of A_1, …, A_n while suspending judgment in B.

Indeed, for any number of premises n, and any values R, W for the goodness of getting things right and the badness of getting things wrong, respectively, there are classical propositions A_1, …, A_n, B such that $A_1, \dots, A_n \vDash_{\text{cl}} B$, and such that the belief function b_4 is not dominated. Indeed, there is always a regular probability function p that expects the belief function b_4 to have the highest epistemic utility of all possible belief functions defined on A_1, …, A_n, B.[20] And this is sufficient to show that b_4 is not weakly logically dominated.[21]

[20] A probability function is regular if it assigns strictly positive credence to every possibility.

[21] Let's see why this is so. Suppose that one option o^* strictly logically dominates another o. Then o^* has strictly greater expected utility than o by the lights of *any* probability function. After all, the utility of o^* is greater than the utility of o at every world. So any weighted sum of the utilities of o^* will be greater than the corresponding weighted sum of the utilities of o. And of

Given how useful it is to know that a belief function maximises expected epistemic utility relative to a probability function, let's spell out exactly how this works. The *expected epistemic utility of a belief function b by the lights of probability function p* is defined as follows:

$$\mathrm{Exp}_{\mathrm{EU}}(b \mid p) = \sum_{w} p\,(w)\mathrm{EU}(b, w)$$

It is straightforward to see that:

$$\begin{aligned}\mathrm{Exp}_{\mathrm{EU}}(b \mid p) &= \sum_{w} p\,(w)\mathrm{EU}(b, w)\\ &= \sum_{w} p\,(w) \sum_{X \in \mathcal{F}} \mathrm{eu}\,(w(X), b(X))\\ &= \sum_{X \in \mathcal{F}} \sum_{w} p\,(w)\mathrm{eu}(w(X), b(X))\\ &= \sum_{X \in \mathcal{F}} p\,(X)\mathrm{eu}(\mathbf{t}, b(X)) + p(\overline{X})\mathrm{eu}(\mathbf{f}, b(X))\end{aligned}$$

That is, the expected utility of b is the sum of the expected utilities of the individual attitudes it assigns. Thus, b has maximal expected epistemic utility by the lights of p iff each attitude that b assigns has maximal expected epistemic utility by the lights of p.

Now, note the following fact:

Theorem 1 (Hempel-Easwaran-Dorst)

If $W \geq R$, then

i. belief in X has maximal expected EU by the lights of p if $1 \geq p(X) \geq \frac{W}{W+R}$

course the expected utilities of o and o^* are just such weighted sums. Thus, if o maximises expected utility by the lights of some probability function, there is no option that strictly logically dominates o, for such an option would have strictly greater expected utility by the lights of that probability function. Next, suppose that o^* weakly logically dominates o. Then the utility of o^* is at least the utility of o at every world and strictly greater at some. So any weighted sum of the utilities of o^* that assigns strictly positive weight to each will be greater than the corresponding weighted sum of the utilities of o. And again the expected utilities of o and o^* by the lights of a regular probability function are just such weighted sums. Thus, if o maximises expected utility by the lights of some regular probability function, there is no option that weakly logically dominates o, for such an option would have strictly greater expected utility by the lights of that probability function.

ii. suspension in X has maximal expected EU by the lights of p if $\frac{W}{W+R} \geq p(X) \geq \frac{R}{W+R}$

iii. disbelief in X has maximal expected EU by the lights of p if $\frac{R}{W+R} \geq p(X) \geq 0$.

If $W < R$, then

i. belief in X has maximal expected EU by the lights of p if $1 \geq p(X) \geq \frac{1}{2}$

ii. suspension in X never maximises expected EU by the lights of p

iii. disbelief in X has maximal expected EU by the lights of p if $\frac{1}{2} \geq p(X) \geq 0$.

It is in this sense that epistemic utility theory vindicates a normative reading of the Lockean thesis. Suppose $W \geq R$ — that is, you are an epistemic conservative or centrist. Then there is a threshold $t = \frac{W}{W+R}$ such that you are rationally required to believe a proposition if your credence in that proposition exceeds t, you are rationally required to suspend on that proposition if your credence lies strictly between $1 - t$ and t, and you are rationally required to disbelief it if you credence lies below $1 - t$. If your credence is exactly t, then believing and suspending both maximise expected EU; if your credence is exactly $1 - t$, then disbelieving and suspending both maximise expected EU. Next, suppose $W < R$ — that is, you are an epistemic radical. Then there is a threshold $t = \frac{1}{2} = 1 - t$ such that you are rationally required to believe a proposition if your credence in that proposition exceeds t, and you are rationally required to disbelief it if you credence lies below t. If your credence is exactly t, then believing and disbelieving both maximise expected EU.

Thus, given $W \geq R$, and n, in order to find propositions A_1, …, A_n, B such that $A_1, \dots, A_n \vDash_{\text{cl}} B$ and a probability function by the lights of which $b_4(A_1) = \cdots = b_4(A_n) = \mathbf{B}$ and $b_4(B) = \mathbf{S}$ has maximal expected epistemic utility, we need only find p such that $p(A_1) = \cdots = p(A_n) \geq \frac{W}{W+R}$ and $p(B) \leq \frac{W}{W+R}$. And that is straightforward to do, since conjunctions typically have lower probability than their conjuncts.[22] If $W < R$, the situation is a little more complicated, since there is no probability function by the lights of which b_4 has maximal expected utility, since there is no probability function for which suspending judgment has maximal

[22] Let X and Y be logically independent propositions. Let $A_1 = X$ and $A_2 = \cdots = A_n = Y$, and $B = X \,\&\, Y$. So $A_1, \dots, A_n \vDash_{\text{cl}} B$. Then let $p(X), p(Y) = \frac{W}{W+R}$ and $p(X\overline{Y} \vee \overline{X}Y) > 0$. So $\frac{R}{W+R} < p(X \,\&\, Y) < \frac{W}{W+R}$. Thus, $p(A_1) = \cdots = p(A_n) = t$, while $1 - t < p(B) < t$.

expected utility. But there are regular probability functions such that the only belief function that has higher expected epistemic utility than b_4 assigns belief to each A_i and assigns belief or disbelief to B; and it is easy to see that neither of those strictly or weakly dominates b_4; so nothing does. So, unlike in the single-premise case, we cannot give a dominance argument for (BP9).

Similar reasoning shows that, if we do not assume Extreme Epistemic Conservatism, then there is no guarantee even that b_3 is weakly or strictly logically dominated. That is, we can find propositions A_1, …, A_n, B, and a probability function p such that b_3 has maximal expected epistemic utility by the lights of p. Here's an example. Suppose there is a fair lottery with $n+1$ tickets. Let A_i be proposition *Ticket i does not win*, for $1 \leq i \leq n$, and let B be the proposition *Ticket $n+1$ wins*. Then $A_1, \dots, A_n$ entail B. However, if we suppose that each ticket has the same chance of winning, then $p(A_i) = \frac{n}{n+1} \geq \frac{W}{W+R}$, for each $1 \leq i \leq n$, while $p(B) = \frac{1}{n+1} \leq \frac{R}{W+R}$.[23] Thus, believing that each of the first n tickets does not win whilst disbelieving that the final ticket will win is not strictly or weakly logically dominated. Indeed, not only is it not dominated, it is in fact the belief assignment recommended by the objective chance function in this context.

The upshot of this section is that epistemic utility considerations vindicate intuitions such as the Preface Paradox, which entail that logical consistency is not a rational requirement on beliefs. If we care so much more about avoiding error than about believing truths, then we can recover the bridge principle that prohibits believing each of a set of propositions whilst disbelieving one of their classical logic consequences. But if we do not, we cannot. There will be situations, such as lottery or preface cases, in which such doxastic attitudes will be rationally required by the natural probability function that governs them.

Non-Classical Logics

Next, let's look at what happens when we move from classical logic to a non-classical alternative. We'll focus on two particular non-classical logics: Kleene's strong logic of indeterminacy (skl) and Priest's Logic of Paradox (lp) . Both have three truth values: $\{\mathbf{t}, \mathbf{u}, \mathbf{f}\}$. And both specify the same truth-functional definitions for the connectives, namely,

[23] Note: $\frac{n}{n+1} \geq \frac{W}{W+R}$ iff $nR \leq W$ iff $\frac{1}{n+1} \leq \frac{R}{W+R}$.

X	$\overline{X}$
t	f
u	u
f	t

&	t	u	f
t	t	u	f
u	u	u	f
f	f	f	f

$\vee$	t	u	f
t	t	t	t
u	t	u	u
f	t	u	f

They differ in the interpretation of the third truth value **u**. In strong Kleene logic it is taken to mean *neither true nor false*, while in Logic of Paradox, it is taken to mean *both true and false*. And they differ in the role that those truth values play in the definition of logical consequence for the two logics. In both logics, A_1, …, A_n entails B iff whenever each of A_1, …, A_n takes one of the designated truth values, B does as well. But they differ in the specification of the designated truth values. For Kleene's logic, **t** is the only designated truth value. For the Logic of Paradox, **t** and **u** are both designated. Thus, $A \vee \overline{A}$ is not a tautology in strong Kleene logic, since it has truth value **u** when A does, and **u** is not designated; but it is a tautology in Logic of Paradox, since it has value **t** or **u** regardless of the truth value of A, and both are designated.

Having introduced strong Kleene logic and Logic of Paradox, how might we define epistemic utility for belief functions when one of those logics governs the propositions that our agent entertains? It is easy to see what the possible worlds are in such a situation. They are the logically consistent valuation functions $w\colon \mathcal{F} \to \{\mathbf{t}, \mathbf{u}, \mathbf{f}\}$. And a local epistemic utility function is a function $\mathrm{eu}_{\mathrm{skl}}/\mathrm{eu}_{\mathrm{lp}}\colon \{\mathbf{B}, \mathbf{S}, \mathbf{D}\} \times \{\mathbf{t}, \mathbf{u}, \mathbf{f}\} \to [-\infty, \infty]$. We assume that the local epistemic utility functions in these situations extend the classical ones, which specify the epistemic utility for **B**, **S**, and **D** when the proposition towards which the attitude is directed takes truth value **t** or **f**. Thus, we need only specify the epistemic value of each of these attitudes when directed towards a proposition with truth value **u**.

Take strong Kleene logic first. Here, **u** is interpreted to mean *neither true nor false*. We must define $\mathrm{eu}_{\mathrm{skl}}(\mathbf{u}, \mathbf{B})$, $\mathrm{eu}_{\mathrm{skl}}(\mathbf{u}, \mathbf{S})$, and $\mathrm{eu}_{\mathrm{skl}}(\mathbf{u}, \mathbf{D})$. There are a number of views one might take on these values. These will depend in part on the use to which the logic is being put, but there will also be disagreements once we have fixed the use of the logic. I do not seek to adjudicate these disagreements here, but rather to spell out their consequences for the logic-rationality bridge principles.

For instance, Hartry Field, following Kripke, claims that strong Kleene logic is the logic that governs the liar sentence.[24] That is, the liar sentence takes truth value **u**; it is neither true nor false. What's more, he takes the ideal attitude to

[24] Hartry Field, *Saving Truth from Paradox* (New York: Oxford University Press, 2008), Saul Kripke, "Outline of a Theory of Truth," *Journal of Philosophy* 72, 19 (1975): 690–716.

propositions with truth value $\mathbf{u}$ to be disbelief (or rejection). Indeed, Michael Caie notes that this is the consensus amongst those who take a paracomplete approach to semantic paradoxes.[25] Thus, for Field, the following is the natural assignment of epistemic value to the various categorical doxastic attitudes to a proposition with truth value $\mathbf{u}$:

$$\begin{array}{ccccccc} \mathrm{eu}(\mathbf{t},\mathbf{B}) & = & \mathrm{eu}(\mathbf{f},\mathbf{D}) & = & \mathrm{eu}(\mathbf{u},\mathbf{D}) & = & R \\ \mathrm{eu}(\mathbf{t},\mathbf{S}) & = & \mathrm{eu}(\mathbf{f},\mathbf{S}) & = & \mathrm{eu}(\mathbf{u},\mathbf{S}) & = & 0 \\ \mathrm{eu}(\mathbf{t},\mathbf{D}) & = & \mathrm{eu}(\mathbf{f},\mathbf{B}) & = & \mathrm{eu}(\mathbf{u},\mathbf{B}) & = & -W \end{array}$$

Given this, we have a strict dominance argument for the strong Kleene version of (BP4) and weak dominance arguments for the strong Kleene versions of (BP5) and (BP6), which are obtained from those principles by replacing $\vDash_{\mathrm{cl}}$ with $\vDash_{\mathrm{skl}}$. In fact, this follows from a more general fact, which also covers the original, classical versions of (BP4-6):[26]

Theorem 2

Suppose:

i. The logical consequence relation for a many-valued logic k is defined in terms of the preservation of designated truth values.

ii. That is, $A_1, \dots, A_n \vDash_{\mathrm{k}} B$ iff, for all worlds w, if $w(A_i)$ is a designated truth value in k, for each $1 \leq i \leq n$, then $w(B)$ is a designated truth value in k.

iii. $A \vDash_{\mathrm{k}} B$

iv. If $\mathbf{i}$ is a designated truth value, then $\mathrm{eu}(\mathbf{i},\mathbf{B}) = R$, $\mathrm{eu}(\mathbf{i},\mathbf{S}) = 0, \mathrm{eu}(\mathbf{i},\mathbf{D}) = -W$.

v. If $\mathbf{i}$ is not a designated truth value, then $\mathrm{eu}(\mathbf{i},\mathbf{B}) = -W$, $\mathrm{eu}(\mathbf{i},\mathbf{S}) = 0, \mathrm{eu}(\mathbf{i},\mathbf{D}) = R$.

Then:

a. If Epistemic Conservatism holds, then believing A and disbelieving B is strictly logically dominated by suspending on A and suspending on B.

[25] Michael Caie, "Belief and Indeterminacy," *Philosophical Review* 121, 1 (2012): 1–54.

[26] *Proof.* The proof is easily adapted from the classical case. In that case, there were three possibilities: worlds at which A and B both take value $\mathbf{t}$, worlds at which A takes $\mathbf{f}$ while B takes $\mathbf{t}$, and worlds at which A and B both take value $\mathbf{f}$. In the present case, the worlds can also be divided into three groups: worlds at which A and B both take a designated value, worlds at which A takes an undesignated value and B takes a designated value, and worlds at which A and B both take an undesignated truth value. Because of (iii) and (iv), b_1, $b_1^\dagger$, etc. have the same epistemic values at each of these three possibilities as they have at the corresponding possibility in the classical case. Thus, the reasoning in the classical case transfers to this case. QED.

b. That is, b_1 is strictly logically dominated by b^*

c. Believing A and disbelieving B is weakly logically dominated by disbelieving A and believing B. That is, b_1 is weakly logically dominated by $b_1^\dagger$.

Believing A and suspending on B is weakly logically dominated by disbelieving A and suspending on B.

That is, b_2 is weakly logically dominated by $b_2^\dagger$.

We can also apply this in the case of the Logic of Paradox, if we follow Priest's claim that belief (or acceptance) is the correct attitude to a proposition that is assigned truth value **u**, which he interprets as *both true and false*.[27] In that case, the natural account of local epistemic utility is this:

$$\begin{array}{ccccccc} \mathrm{eu}(\mathbf{t},\mathbf{B}) & = & \mathrm{eu}(\mathbf{f},\mathbf{D}) & = & \mathrm{eu}(\mathbf{u},\mathbf{B}) & = & R \\ \mathrm{eu}(\mathbf{t},\mathbf{S}) & = & \mathrm{eu}(\mathbf{f},\mathbf{S}) & = & \mathrm{eu}(\mathbf{u},\mathbf{S}) & = & 0 \\ \mathrm{eu}(\mathbf{t},\mathbf{D}) & = & \mathrm{eu}(\mathbf{f},\mathbf{B}) & = & \mathrm{eu}(\mathbf{u},\mathbf{D}) & = & -W \end{array}$$

And, since **u** is a designated truth value in Logic of Paradox, this account satisfies the hypotheses of Theorem 2. Thus, we have a strict dominance argument for the Logic of Paradox version of (BP4) and weak dominance arguments for the Logic of Paradox versions of (BP5) and (BP6), which are obtained from those principles by replacing $\vDash_{\mathrm{cl}}$ with $\vDash_{\mathrm{lp}}$.

However, there are other ways in which strong Kleene logic and Logic of Paradox may be applied for which the local epistemic utility functions described so far are not appropriate.[28] Suppose, for instance, that strong Kleene logic governs propositions that involve vague predicates.[29] Then the appropriate doxastic attitude to a proposition with truth value **u** is surely suspension, not disbelief. If the colour of my socks lies in the borderline region between determinately red and determinately orange, it seems better to suspend judgment on the proposition *My handkerchief is red* than to believe or disbelieve it. Thus, we might think that the local epistemic utilities are assigned as follows:

[27] Graham Priest, *Doubt Truth to be a Liar* (Oxford: Oxford University Press, 2005)

[28] Thanks to Hartry Field, Patrick Greenough, and Ole Thomassen Hjortland for helpful discussion on this point.

[29] See, for instance: Michael Tye, "Sorites Paradoxes and the Semantics of Vagueness," in *Philosophical Perspectives: Logic and Language, vol. 8*, ed. James Tomberlin (Atascadero: Ridgeview Press, 2008), 189–208, Hartry Field, "No Fact of the Matter," *Australasian Journal of Philosophy* 81 (2003): 457–480.

$$\begin{array}{rcrcl}
\mathrm{eu}(\mathbf{t},\mathbf{B}) & = & \mathrm{eu}(\mathbf{f},\mathbf{D}) & = & R \\
 & & \mathrm{eu}(\mathbf{u},\mathbf{S}) & = & N \\
\mathrm{eu}(\mathbf{t},\mathbf{S}) & = & \mathrm{eu}(\mathbf{f},\mathbf{S}) & = & 0 \\
\mathrm{eu}(\mathbf{u},\mathbf{B}) & = & \mathrm{eu}(\mathbf{u},\mathbf{D}) & = & -Z \\
\mathrm{eu}(\mathbf{t},\mathbf{D}) & = & \mathrm{eu}(\mathbf{f},\mathbf{B}) & = & -W
\end{array}$$

where $-W \leq -Z < 0 < N \leq R$. In this case, under certain assumptions, we can again argue for strong Kleene versions of (BP4-6). After all, consider the truth table:

	A	B	$\mathrm{EU}(b_1,w)$	$\mathrm{EU}(b^*,w)$	$\mathrm{EU}(b_1^\dagger,w)$	$\mathrm{EU}(b_2,w)$	$\mathrm{EU}(b_2^\dagger,w)$
w_1	**t**	**t**	$R-W$	$0+0$	$-W+R$	$R+0$	$0+R$
w_2	**t**	**u**	$R-Z$	$0+N$	$-W-Z$	$R+N$	$0-Z$
w_3	**t**	**f**	$R+R$	$0+0$	$-W-W$	$R+0$	$0-W$
w_4	**u**	**t**	$-Z-W$	$N+0$	$-Z+R$	$-Z+0$	$N+R$
w_5	**u**	**u**	$-Z-Z$	$N+N$	$-Z-Z$	$-Z+N$	$N-Z$
w_6	**u**	**f**	$-Z+R$	$N+0$	$-Z-W$	$-Z+0$	$N-W$
w_7	**f**	**t**	$-W-W$	$0+0$	$R+R$	$-W+0$	$0+R$
w_8	**f**	**u**	$-W-Z$	$0+N$	$R-Z$	$-W+N$	$0-Z$
w_9	**f**	**f**	$-W+R$	$0+0$	$R-W$	$-W+0$	$0-W$

If $A \vDash_{\mathrm{skl}} B$, then w_2 and w_3 do not represent logical possibilities, but the rest do. Thus:

- If $N+Z > R$ and $W \geq R$, b_1 is weakly logically dominated by b^*.
- Even if $N+Z > R$ and $W \geq R$, b_1 is not even weakly logically dominated by $b_1^\dagger$.
- After all, b_1 outperforms $b_1^\dagger$ when A has truth value **u** and B has truth value **f**. And indeed there are values of $W \geq Z$ and $N \leq R$ such that nothing even weakly logically dominates b_1 for those values.
- If $W-Z=N$, b_2 is weakly logically dominated by $b_2^\dagger$.

Thus, we have arguments for the strong Kleene versions of (BP4-6), which are obtained from those principles by replacing $\vDash_{\mathrm{cl}}$ with $\vDash_{\mathrm{skl}}$. But those arguments are weaker than the corresponding arguments for the original, classical versions of (BP4-6), since they make stronger assumptions about local epistemic utilities:

(EU8) $(N+Z>R)$ + $(W \geq R)$ + Weak Logical Dominance ⇒ (BP4 $_{\mathrm{skl}}$) and (BP5 $_{\mathrm{skl}}$).

(EU9) $(W+Z=N)$ + Weak Logical Dominance ⇒ (BP6 $_{\mathrm{skl}}$).

Next, suppose that the Logic of Paradox governs future contingents.[30] Thus, the truth value of a future contingent X is: (i) **t** if X is true in all possible futures; (ii) **f** is X is false in all possible futures; and (iii) **u** if X is true in some futures and false in others. The idea is that, in the latter case, the proposition is both true at some point in the future and false at some point in the future, and thus both true and false now. This is a paraconsistent approach to the logic of future contingents. In this situation, we might think it natural to order the local epistemic utilities of the various categorical doxastic attitudes as follows:

$$\begin{array}{rcccc}
\mathrm{eu}(\mathbf{t},\mathbf{B}) & = & \mathrm{eu}(\mathbf{f},\mathbf{D}) & = & R \\
\mathrm{eu}(\mathbf{u},\mathbf{B}) & = & \mathrm{eu}(\mathbf{u},\mathbf{D}) & = & N \\
\mathrm{eu}(\mathbf{t},\mathbf{S}) & = & \mathrm{eu}(\mathbf{f},\mathbf{S}) & = & 0 \\
 & & \mathrm{eu}(\mathbf{u},\mathbf{S}) & = & -Z \\
\mathrm{eu}(\mathbf{t},\mathbf{D}) & = & \mathrm{eu}(\mathbf{f},\mathbf{B}) & = & -W
\end{array}$$

where $-W \leq -Z < 0 < N \leq R$. If we do this, here's the truth table:

	A	B	$\mathrm{EU}(b_1, w)$	$\mathrm{EU}(b^*, w)$	$\mathrm{EU}(b_1^\dagger, w)$	$\mathrm{EU}(b_2, w)$	$\mathrm{EU}(b_2^\dagger, w)$
w_1	**t**	**t**	$R - W$	$0 + 0$	$-W + R$	$R + 0$	$0 + R$
w_2	**t**	**u**	$R + N$	$0 - Z$	$-W + N$	$R - Z$	$0 + N$
w_3	**t**	**f**	$R + R$	$0 + 0$	$-W - W$	$R + 0$	$0 - W$
w_4	**u**	**t**	$N - W$	$-Z + 0$	$N + R$	$N + 0$	$-Z + R$
w_5	**u**	**u**	$N + N$	$-Z - Z$	$N + N$	$N - Z$	$-Z + N$
w_6	**u**	**f**	$N + R$	$-Z + 0$	$N - W$	$N + 0$	$-Z - W$
w_7	**f**	**t**	$-W - W$	$0 + 0$	$R + R$	$-W + 0$	$0 + R$
w_8	**f**	**u**	$-W + N$	$0 - Z$	$R + N$	$-W - Z$	$0 + N$
w_9	**f**	**f**	$-W + R$	$0 + 0$	$R - W$	$-W + 0$	$0 - W$

If $A \vDash_{\mathrm{lp}} B$, then worlds w_3 and w_6 fail to represent logical possibilities. But given this, we can see that no alternative weakly or strictly dominates b_1. The reason is that no alternative belief function performs as well as b_1 at world w_2. Similarly, no alternative either weakly or strictly dominates b_2. The only alternatives that perform as well as b_2 at w_2 assign **B** to A and either **B** or **D** to B; but these perform worse than b_2 at worlds w_9 and w_7, respectively. So we do not have an Logic of Paradox versions of (BP4-6) in this case.

Indeed, this all follows from a more general fact:[31]

[30] Roberto Ciuni and Carlo Proietti, "The Abundance of the Future: A Paraconsistent Approach to Future Contingents," *Logic and Logical Philosophy* 22, 1 (2013): 21–43.

[31] *Proof.* The truth value **i** from (4) is either designated or undesignated. Suppose first that it is designated. Then the only alternative belief function that performs at least as well as b_1 at the logical possibility where A takes **t** and B takes **i** is the belief function that assigns belief to A and belief to B. But that performs worse than b_1 when A and B both take **f**. Next, suppose that it is

Theorem 3

Suppose:

i. The logical consequence relation for a many-valued logic k is defined in terms of the preservation of designated truth values.

ii. $A \vDash_{\mathrm{k}} B$

iii. eu_{k} extends $\mathrm{eu}_{\mathrm{cl}}$.

iv. There is a truth value **i** such that

v. $-W \leq \mathrm{eu}_{\mathrm{k}}(\mathbf{i}, \mathbf{S}) < \mathrm{eu}_{\mathrm{k}}(\mathbf{i}, \mathbf{B}) = \mathrm{eu}_{\mathrm{k}}(\mathbf{i}, \mathbf{D}) \leq R$

Then:

a. No alternative even weakly logically dominates believing A and disbelieving B.

That is, nothing weakly dominates b_1.

The upshot of this section is that the fate of logic-rationality bridge principles is sensitive to the logic that governs the propositions in question, the interpretation of the truth values in that logic, and the resulting assignments of epistemic value to beliefs, disbeliefs, and suspensions in propositions that take truth-values other than **t** or **f**. This explains why we have been careful throughout to specify in the antecedent of those principles which logic governs the propositions in question. There are bridge principles that hold when the logic is classical that do not hold for alternative logics.

Logical, Doxastic, and Epistemic Possibilities

In the preceding sections, we have offered epistemic utility arguments in favour of certain logic-rationality bridge principles, and we have given epistemic utility-based reasons for doubting that others can be justified. For each of the bridge principles we have considered, its antecedent is a plain fact about logical consequence—something of the form $\vDash_{\mathrm{k}}$ *governs* $A_1, \ldots, A_n, B$, *and* $A_1, \ldots, A_n \vDash_{\mathrm{k}} B$ for some logic k and some $n \geq 1$. As a result, the principles are quite demanding. They require that you manage your beliefs in line with a logical fact that you might not know or even believe. We have been able to justify these demanding principles only because we've assumed similarly demanding principles of decision theory. For instance, Strict Logical Dominance says that an option is irrational if

undesignated. Then the only belief function that performs at least as well as b_1 at the logical possibility where A takes **i** and B takes **f** is the belief function that assigns disbelief to A and disbelief to B. But that performs worse than b_1 when A and B both take **t**. QED.

there is an alternative that is better at all *logically possible worlds*; Weak Logical Dominance says that an option is irrational if there is an alternative that is at least as good at all *logically possible worlds*, and better at some. And you might think that these are too strong, even in the practical case. For instance, Strict and Weak Logical Dominance render it irrational for my nine year old niece to pay any positive amount for a bet against Fermat's Last Theorem, even if she has never heard of it until I describe it to her, and even if I tell her nothing about its proof status. If we weaken these decision-theoretic principles, we obtain epistemic utility arguments for the correspondingly weakened logic-rational bridge principles.

Here are general versions of our dominance norms, where $\mathcal{C}$ is a set of worlds.

> **Strict $\mathcal{C}$ Dominance** If option o^* has greater utility than option o at every world in $\mathcal{C}$, then o is irrational.

> **Weak $\mathcal{C}$ Dominance** If option o^* has at least as great utility as option o at every world in $\mathcal{C}$, and greater utility at some, then o is irrational.

We obtain Strict/Weak Logical Dominance if $\mathcal{C}$ is the set of logically possible worlds. We obtain Strict/Weak Doxastic Dominance if $\mathcal{C}$ is the set of doxastically possible worlds—that is, the worlds at which everything she believes is true and everything she disbelieves is false. And we obtain Strict/Weak Epistemic Dominance if $\mathcal{C}$ is the set of epistemically possible worlds—that is, the worlds compatible with what the agent knows. And so on.

Now, suppose we replace Strict/Weak Logical Dominance with Strict/Weak Doxastic or Epistemic Dominance in our arguments for logic-rationality bridge principles. Then surely we obtain arguments for the corresponding bridge principles in which the antecedent is no longer just a proposition about logical consequence, but is rather the proposition that the agent believes or knows that proposition about logical consequence.[32]

Thus, for instance, let's assume Weak Doxastic Dominance:

> **Weak Doxastic Dominance** If option o^* has at least as great utility as option o at every doxastically possible world, and greater utility at some, then o is irrational.

Then, in order to adapt argument (EU2), we need to assume that the agent believes that classical logic is the correct logic and that A is strictly stronger than B in that logic. By believing that classical logic is the correct logic, our agent narrows down the set of doxastically possible worlds to the four—w_1, w_2, w_3, w_4—represented in the relevant table above; by also believing that A is strictly stronger

[32] J. R. G. Williams, "Rational Illogicality" *Australasian Journal of Philosophy* (forthcoming).

than B, our agent narrows the field further by ruling out world w_2 at which A is true and B is false, but retains world w_3 at which A is false and B is true—that is, she narrows the field to w_1, w_3, w_4. We thus obtain:

> (BP10) If you believe that $\vDash_{\mathrm{cl}}$ governs A and B, and $A \vDash_{\mathrm{cl}} B$, and $B \nvDash_{\mathrm{cl}} A$, then you ought not to believe A while disbelieving in B.

Or so it seems. The problem with using doxastic possibility in this context is that we are using facts about what we believe to delimit the set of worlds that feature in a dominance principle that we then use to choose our beliefs! Why might this be problematic? Initially, you might think that it could give rise to a sort of instability in the beliefs it is rational for you to have. You start with a set of beliefs. They determine the worlds that are doxastically possible for you. Determined in this way, it turns out that epistemic utility theory rules out your beliefs as irrational. So you pick another set of beliefs. They determine the worlds that are doxastically possible for you. Determined in this way, it turns out that epistemic utility theory rules out those beliefs as irrational. And so on. At first sight, this seems a possibility. But, in fact, it is the opposite that happens. Pick a set of consistent beliefs and disbeliefs. These then determine a set of doxastically possible worlds. At each of these worlds, each of the beliefs you picked is true and each of the disbeliefs you picked is false. Thus, you have maximal epistemic utility at each of these worlds. Any alternative assignment of beliefs, suspensions, and disbeliefs to the same propositions will be weakly doxastically dominated. Thus, any consistent set of beliefs renders itself the only rational option.

Here's another way in which it might be problematic to use beliefs to determine the doxastic possibilities, and then use those possibilities to pick the beliefs. Suppose I believe that A entails B, and I believe A, but I disbelieve B. Then there are no worlds at which all of my beliefs are true and all my disbeliefs false. Thus, there are no worlds that are doxastically possible for me. One consequence of that is that every belief function is strictly doxastically dominated by every other one — for every pair of belief functions b and b' on the same set of propositions, it is vacuously true that b is strictly better than b' at all doxastically possible worlds, for there are no doxastically possible worlds. Thus, unless we restrict Strict Doxastic Dominance, every belief function is irrational. In fact, we're best to restrict Strict Doxastic Dominance in this case, and say that dominance principles only apply when the relevant set of worlds is non-empty. But if we do this, nothing is ruled irrational for the agent who believes that A entails B, believes A, and disbelieves B. And that looks troubling too!

A final worry about moving to Strict/Weak Doxastic Dominance principles. The norms that result from the epistemic utility arguments that appeal to those principles are narrow scope norms of the sort that we typically reject in this area. Consider the standard narrow scope norm in this area: if you believe *If A, then B*, and you believe that *A*, then you ought to believe that *B*. As Harman noted in *Change in View*, such a norm cannot be correct, since it is just as legitimate to respond by dropping your belief that *If A, then B* or by dropping your belief that *A* as it is to respond by keeping both of those beliefs and further adopting a belief that *B*. Similarly, surely the logic-rationality bridge principles that follow from the doxastic dominance arguments cannot be correct either. Surely it is just as legitimate to respond to your belief that *A and B are governed by* $\vDash_{\mathrm{cl}}$ and your belief that $A \vDash_{\mathrm{cl}} B$ by dropping one or other or both of those beliefs as it is to retain both beliefs and then ensure that you do not believe *A* and disbelieve *B*.

I offer two different solutions to these problems. First, the *Two-Tier Solution*. We might save our new doxastic dominance arguments for the doxastic versions of the bridge principles if we take our beliefs about logical consequence to be of a rather different sort from our beliefs about other matters. We might take the beliefs about logical consequence to delimit the doxastically possible worlds, perhaps, and then use those in our dominance principles to assess the different possible sets of beliefs we might have towards other propositions. If you opt for this solution, you owe an account of why there are two sorts of beliefs, ones that get to delimit doxastic possibilities and ones that don't. And you have to say, in particular, why logical beliefs—beliefs about the logic that governs a class of propositions, and beliefs about the consequence relation of that logic—are of the former sort. You might appeal, for instance, to Quine's notion of a web of belief.[33] It is perhaps the propositions sufficiently close to the centre of our web of belief—that is, those least vulnerable to revision—that delimit the set of doxastic possibilities. It is then those further out—those more vulnerable to revision—that are governed by Strict/Weak Doxastic Dominance. This would provide a principled distinction between the two sorts of belief, and it would also explain why the narrow scope norm is appropriate. It explains why it is legitimate to demand that you respond to your logical belief that *A and B are governed by* $\vDash_{\mathrm{cl}}$ and your logical belief that $A \vDash_{\mathrm{cl}} B$ by ensuring that you do not believe *A* and disbelieve *B*, rather than by dropping one or other or both of your logical beliefs. The explanation is that the logical beliefs lie closer to the centre of the web of belief—when something's got to give, it shouldn't be them.

Here's the second solution to the problems raised above for the doxastic dominance arguments for the doxastic versions of the logic-rationality bridge

[33] W. V. O. Quine, "Two Dogmas of Empiricism," *Philosophical Review* 60, 1 (1951): 20–43. W. V. O. Quine and Joe Ullian, *The Web of Belief* (New York: Random House, 1970).

principles we've been considering. We might call it the *Wide Scope solution*. It proceeds by analogy with the standard retreat from narrow to wide scope norms in the face of Harman's criticism. Thus, instead of trying to justify the narrow scope norm (BP10), we might try to justify the wide scope version of it:

> (BP11) You ought to see to it that you don't believe that $\vDash_{\mathrm{cl}}$ governs A and B, believe $A \vDash_{\mathrm{cl}} B$, believe $B \nvDash A$, believe A, but disbelieve B.

Can we offer an epistemic utility argument for (BP11)? We have five propositions in play: (i) $\vDash_{\mathrm{cl}}$ *governs* A *and* B; (ii) $A \vDash_{\mathrm{cl}} B$; (iii) $B \nvDash_{\mathrm{cl}} A$; (iv) A; and (v) B. (BP11) says that you ought not to believe (i)-(iv) whilst disbelieving (v). But now notice that (i)-(iv) entail (v). So, we have a multi-premise entailment and we wish to justify a norm that prohibits believing the premises and disbelieving the conclusion. Thus, if we simply treat each of these propositions as a normal proposition, and if it is legitimate to assume that any agent can see the entailment from (i)-(iv) to (v), and that is something they should never give up, then we can simply turn our Strict Logical Dominance argument for (BP8) into a Strict Doxastic Dominance argument for (BP11). Now, notice that the multi-premise entailment in question is a four-premise entailment. So in order to run our dominance argument for it, we need to assume a version of Extreme Epistemic Conservatism, namely, that $4R < W$. But if we have that, then we can conclude (BP11).

Thus, we have two putative solutions to our problems. On the Two-Tier solution, we retain the narrow scope norm (BP10) by saying that some propositions—including those that pertain to the correct logic and the consequence relation of that logic—fix the doxastic possibilities, while others are determined after those possibilities have been fixed by considerations of epistemic utility. On the Wide Scope solution, we do not assign the logical propositions any special role, and instead treat them just like other propositions, giving us the wide scope version (BP11).

So much for our solutions to the problems raised above. In the remainder of this section, we make a handful of further observations on the move from logic-rationality bridge principles with purely logical antecedents to the versions with doxastic antecedents. First, we note that there are two ways in which you might not know or believe all the logical truths. You might know what the correct logic is—for instance, you might know that classical logic governs A and B—but you might not know that A entails B within that logic. But you might not even know what the correct logic is—you might not know whether strong Kleene logic, classical logic, or Logic of Paradox governs A and B. Above, we focussed on the first sort of case, assuming that there was some particular logic that our agents believed to be the correct one. But there are some things to say about the second case as well.

Suppose, for instance, that our agent believes that the correct logic is either strong Kleene logic or classical logic; suppose she knows that A entails B; and suppose $N + Z > R$ and $W \geq R$; then it would be irrational for her to believe A and disbelieve B—that is, irrational for her to have belief function b_1. After all, if we pool all of the worlds that are possible relative to classical logic and all of the worlds that are possible relative to strong Kleene logic, then b^* dominates b_1. The reason is that every classically possible world is also logically possible from the point of view of strong Kleene logic. Thus, since b^* dominates b_1 relative to the strong Kleene worlds, it dominates b_1 relative to all the strong Kleene *and* classical worlds.

That might tempt us to think that if a logic-rationality bridge principle can be justified by logical dominance reasoning relative to one logic and justified by logical dominance reasoning also relative to another logic, then it can be justified by doxastic dominance reasoning for someone who believes that one or other of these logics is correct, but isn't certain which. But that is not the case. The reason: it might be that a belief function b is dominated by b' and not by b'' relative to the first logic, while it is dominated by b'' and not by b' relative to the second. In that case, neither b' nor b'' dominates b relative to the disjunction of the logics. In this case, we have a situation akin to the Miners Paradox.[34] If the first logic is actual, the agent ought not to choose b (since it is dominated by b'); if the second logic is actual, the agent ought not to choose b (since it is dominated by b''); the first logic is actual or the second is; but it does not follow that the agent ought not to choose b.[35]

My next observation on logic-rationality bridge principles with doxastic antecedents concerns the argument due to MacFarlane that we should not be satisfied with them. The problem with these principles, MacFarlane argues, is this: if they are the strongest norms in the vicinity, it seems that the less you know, logically speaking, the less restricted are your beliefs; by remaining ignorant of logical facts, you are less likely to be irrational, since less stringent restrictions are placed upon you. And this seems counterintuitive. It seems to give an incentive to remain logically uninformed. But this should be nothing new. For many philosophers—subjectivist Bayesian epistemologists, for instance—the less evidence you have, whether logical or not, the fewer restrictions are placed upon you. But this only gives an incentive to remain uninformed if avoiding irrationality

[34] Derek Parfit, "What We Together Do" (unpublished manuscript).

[35] For a related discussion, see J. R. G. Williams, "Rational Illogicality."

is the only thing you care about. And of course it is not. You also care about making good decisions and having accurate beliefs. In the case of non-logical facts, the *value of learning theorem* due to I. J. Good and Frank Ramsey shows that you can expect to make better decisions after you have learned those facts than before;[36] and it is straightforward to adapt the epistemic utility-based argument for Conditionalization to show that you can expect to have more accurate beliefs after you learn non-logical facts than before.[37] Now, there is no reason why these arguments shouldn't apply to learning logical facts as well. Thus, we can take the doxastic versions of the logic-rationality bridge principles to be the strongest principles in the vicinity, whilst also thinking that agents should try to know as many logical facts as possible, and should then manage their beliefs in line with the logical facts that they believe. However, their reasons for doing so are just their usual reasons for learning, and then managing their beliefs in line with what they've learned.

This leads us to our final point in this section. Given a particular logic-rationality bridge principle, we can ask what role is played by the logicality of the fact about logical consequence that appears in the principle's antecedent.[38] Are there principles of rationality of the same form that feature a non-logical fact in the antecedent? Does the argument for the bridge principle pay any special attention to the logicality of the fact about logical consequence? The structure of the answer depends, I think, on whether you embrace the Two-Tier solution or the Wide Scope solution above; but the conclusion doesn't. Whichever of those solutions you choose, the logicality of the logical facts plays no special role. Let's see why. First, suppose you opt for the Two-Tier solution. That is, you say that there are two different roles that beliefs can play: they can circumscribe the set of doxastic possible worlds, and they can be evaluated for their epistemic utility. What's more, you say that all logical beliefs play the first role, while some non-logical beliefs play the second. But there is no reason to suppose that it is *only* the logical beliefs that can delimit the doxastically possible worlds. And, of course, if we spell out this solution by saying that it is beliefs near to the centre of the web of belief that play the delimiting role, then presumably beliefs concerning analytic or conceptual truths, mathematical truths, or metaphysical necessities will fit the bill

[36] I. J. Good, "On the Principle of Total Evidence," *The British Journal for the Philosophy of Science* 17 (1967): 319–322, Frank P. Ramsey, "Weight or the Value of Knowledge," *The British Journal for the Philosophy of Science* 41 (1990): 1–4.

[37] Hilary Greaves and David Wallace, "Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility," *Mind* 115, 459 (2006): 607–632.

[38]This question also runs through Harman, *Change in View*.

just as well as logical beliefs do. Thus, the logicality of the logical beliefs plays no special role—what is relevant is their location in the web of belief, and plenty of non-logical beliefs occupy nearby locales. Next, suppose you opt for the second solution to the problems outlined at the beginning of this section. Then the irrelevance of the logicality of those beliefs is even more stark. After all, on that solution, we justify the wide scope norm (BP11). But our argument for that appeals only to the obvious entailment from (i)-(iv) to (v) above. It does not appeal at any point to the logicality of (i)-(iii). The argument would run just as well if (i)-(iii) were any premises for which the entailment from those, together with (iv), to (v) is sufficiently obvious. Thus, on both solutions there is nothing about the logicality of the logic-rationality bridge principles that plays a role in our arguments.

Bridge Principles for Partial Beliefs

In the previous section, we asked what we can learn about logic-rationality bridge principles for categorical doxastic states, such as full belief, full disbelief, and suspension of judgment, by looking at the epistemic utility of those doxastic attitudes. In this section, we turn our attention to partial beliefs, or credences as we will call them.

As before, we begin with the now-standard story about the classical case.[39] Suppose our agent has a credence function c defined on a set of propositions $\mathcal{F}$. For each proposition A in $\mathcal{F}$, $c(A)$ gives the agent's credence in A. By convention, we take maximal credence to be 1 and minimal credence to be 0. Thus, $c: \mathcal{F} \to [0,1]$. Now, as above, we must define an epistemic utility function EU that takes a credence function c and a possible world w and returns $\mathrm{EU}(c, w)$, a measure of the epistemic utility of having credence function c at world w. As above, we take it to be additive. That is, we assume that there is a local epistemic utility function $\mathrm{eu}: \{\mathbf{t}, \mathbf{f}\} \times [0,1] \to [-\infty, \infty]$ such that

$$\mathrm{EU}(c, w) = \sum_{X \in \mathcal{F}} \mathrm{eu}\,(w(X), c(X))$$

How do we define eu? We do so in two steps. First, for each proposition A and each possible world w, we take there to be an *ideal* or *perfect* or *vindicated* credence in A at w—we call this $v_w(A)$. Now, as in the case of categorical doxastic attitudes, the standard story in the credal case takes a veritist approach—that is, it assumes that the sole fundamental source of epistemic value for doxastic states is

[39] Joyce, "A Nonpragmatic Vindication," Joyce, "Accuracy and Coherence," Pettigrew, *Accuracy and the Laws of Credence*.

the accuracy with which they represent the world. In the classical case, this suggests:

$$v_w(A) = \begin{cases} 1 & \text{if } v_w(A) = \mathbf{t} \\ 0 & \text{if } v_w(A) = \mathbf{f} \end{cases}$$

We might call this assumption about the ideal credences *Vindicated is Omniscient*.

Second, having defined the ideal credence in a given proposition at a given possible world, we can then define the epistemic utility of credence $c(A)$ at world w to be its proximity to $v_w(A)$. That is, the epistemic *dis*utility of $c(A)$ at w is the distance from $v_w(A)$ to $c(A)$. How are we to measure distance between credence functions? There are various arguments for measuring such distances using the so-called *Bregman divergences*.[40] A divergence is a function $\mathfrak{d}$ that takes a pair of real numbers x and y and returns $\mathfrak{d}(x, y)$, a non-negative real number or ∞. We say that $\mathfrak{d}$ is a divergence iff $\mathfrak{d}(x, y) \geq 0$ with equality iff $x = y$. And we say that $\mathfrak{d}$ *is a Bregman divergence* if there is a strictly convex, continuously differentiable function $\varphi: [0,1] \to [0, \infty)$ such that:

$$\mathfrak{d}(x, y) = \varphi(x) - \varphi(y) - \varphi'(y)(x - y)$$

where φ' is the derivative of φ. That is, the divergence from x to y is the difference between the value at x of φ and the value at x of the tangent to φ taken at y. If eu is a local epistemic utility function, we demand that it is generated by a Bregman divergence $\mathfrak{d}$ as follows:

$$\mathrm{eu}(w(A), c(A)) = -\mathfrak{d}(v_w(A), c(A))$$

That is, the epistemic utility of a credence $c(A)$ in proposition A at world w is the negative of the divergence from $v_w(A)$, the ideal credence in A at w, to $c(A)$. Putting all of this together, we have that

$$\mathrm{EU}(c, w) = \sum_{X \in \mathcal{F}} \mathrm{eu}\,(w(X), c(X)) = -\sum_{X \in \mathcal{F}} \mathfrak{d}\,(v_w(X), c(X))$$

A well known result shows that the epistemic utility functions defined in this way are precisely the so-called *additive and continuous strictly proper inaccuracy measures*. We might call this assumption about epistemic utility *Bregman Divergence*.

[40] Joyce, "Accuracy and Coherence," Pettigrew, *Accuracy and the Laws of Credence*, Benjamin Levinstein, "A Pragmatist's Guide to Epistemic Utility," *Philosophy of Science* (forthcoming), Sophie Horowitz, "Accuracy and Educated Guesses," *Oxford Studies in Epistemology* (forthcoming).

Now, it is natural to ask what logic-rationality bridge principles follow from this account of the epistemic utility of credences when we apply the decision-theoretic principles that we considered above. The following is a well-known result:[41]

Theorem 4

Suppose:

i. EU is an additive and continuous strictly proper inaccuracy measure.

That is, $\mathrm{EU} = \sum_{X \in \mathcal{F}} \mathrm{eu}\,(w(X), c(X)) = -\sum_{X \in \mathcal{F}} \mathfrak{d}\,(v_w(X), c(X))$, where $\mathfrak{d}$ is a Bregman divergence.

Then:

a. If c is not a probability function on $\mathcal{F}$, then there is a probability function c^* on $\mathcal{F}$ such that c^* strictly logically dominates c relative to EU—that is, $\mathrm{EU}(c, w) < \mathrm{EU}(c^*, w)$, for all logically possible worlds w.

Now, a probability function on $\mathcal{F}$ is a credence function c that satisfies the following conditions:

(BP11a) If $\vDash_{\mathrm{cl}}$ governs A, and $\vDash_{\mathrm{cl}} A$, then $c(A) = 1$

(BP11b) If $\vDash_{\mathrm{cl}}$ governs A, and $A \vDash_{\mathrm{cl}}$, then $c(A) = 0$

(BP12) If $\vDash_{\mathrm{cl}}$ governs A, B, and $A \vDash_{\mathrm{cl}} B$, then $c(A) \leq c(B)$

(BP13) If $\vDash_{\mathrm{cl}}$ governs A, B, then $c(A \,\&\, B) + c(A \vee B) = c(A) + c(B)$

The first three are logic-rationality bridge principles: they concern how credences should behave given facts about the consequence relation. The fourth is not: it concerns the interaction between credences in propositions of different logical forms. Williams calls (BP12) the *No Drop principle*. It is the credal analogue to principles like (Wo-) from MacFarlane[42] and (BP4-7) from above. Thus, we have the following epistemic utility argument:

(EU9) Bregman Divergence + Vindicated is Omniscient + Strict Logical Dominance ⇒ (BP11-13).

Before we leave the classical case, it is worth sketching the proof of Theorem 4, since that will show us how that proof might be adapted to the non-classical case. The proof is based on two lemmas. First:

[41] Joel Predd, Robert Seiringer, Elliott Lieb, Daniel Osherson, Vincent Poor, and Sanjeev Kulkarni, "Probabilistic Coherence and Proper Scoring Rules," *IEEE Transactions of Information Theory* 55, 10 (2009): 4786–4792.

[42] MacFarlane, "In What Sense (If Any)."

> **Lemma 5** The set of probability functions is precisely the *closed convex hull* of the set of vindicated credence functions, v_w, for possible worlds w.[43]

Second:

> **Lemma 6** Suppose $\mathfrak{d}$ is a Bregman divergence, and $\mathcal{X} \subseteq [0,1]^n$ is a set of n-dimensional vectors. Then, if z is a point in $[0,1]^n$ that lies outside the closed convex hull of $\mathcal{X}$, then there is a point z^* inside the convex hull of $\mathcal{X}$ such $\sum_{i=1}^{n} \mathfrak{d}\,(x_i, z_i^*) < \mathfrak{d}(x_i, z_i)$ for all x in $\mathcal{X}$.

Thus, suppose c is a credence function that is not a probability function. Then, by Lemma 5, c lies outside the closed convex hull of the vindicated credence functions. Then, by Lemma 6, there is c^* in the convex hull of the vindicated credence functions such that $\mathfrak{d}(v_w(X), c^*(X)) < \mathfrak{d}(v_w(X), c(X))$. And thus, $\mathrm{EU}(c, w) < \mathrm{EU}(c^*, w)$, for all w.

Breaking down the result into these two component parts allows us to see how it might be generalised. Suppose we move to a different logic. And, as a result, we take different credences to be vindicated—that is, we define $v_w(A)$ differently from how we defined it above. Suppose further that we continue to measure the local epistemic utility of a credence $c(A)$ as its proximity to the vindicated credence: that is, $\mathrm{eu}(w(A), c(A)) = -\mathfrak{d}(v_w(A), c(A))$. Then we can find the bridge principle for which strict dominance provides an epistemic utility argument as follows:

First, we characterise the closed convex hull of those new vindicated credence functions—that is, we provide an analogue of Lemma 5.

Second, we note that, if a credence function c lies outside this closed convex hull, then there is an alternative c^* that is closer to each of the vindicated credence functions than c, and thus epistemically better than c at all logically possible worlds—that is, we deploy Lemma 6.

We will see exactly this strategy in action below.

Before we see it in action in the non-classical case, we first observe it in the classical case. We can use Lemma 6 to justify the following bridge principle, (BP14). (BP14) is the general bridge principle for credences that Field[44] defends, drawing on Adams and Edgington[45].

[43] If $\mathcal{X}$ is a set of credence functions, then its convex hull is the smallest convex set that contains $\mathcal{X}$; that is, the smallest set that contains $\mathcal{X}$ and contains every mixture of two credence functions whenever it contains those credence functions. We denote the convex hull $\mathcal{X}^+$. The closure of a set is the union of that set with the set of its limit points.

[44] Hartry Field, "What Is the Normative Role of Logic?" *Proceedings of the Aristotelian Society* (Supplementary Volumes) 83 (2009): 251–268.

[45] Ernest W. Adams, *The Logic of Conditionals* (Dordrecht: Reidel, 1975), Dorothy Edgington,

(BP14) If $\vDash_{\text{cl}}$ governs $A_1, \dots, A_n, B$, and $A_1, \dots, A_n \vDash_{\text{cl}} B$, then $c(A_1) + \cdots + c(A_n) - (n-1) \leq c(B)$

Or, equivalently and more intuitively:

(BP14) If $\vDash_{\text{cl}}$ governs $A_1, \dots, A_n, B$, and $A_1, \dots, A_n \vDash_{\text{cl}} B$, then $\overline{c}(B) \leq \overline{c}(A_1) + \cdots + \overline{c}(A_n)$

where $\overline{c}(X) := 1 - c(X)$ measures an agent's degree of disbelief in X when $c(X)$ measures her degree of belief in X.

Here's the argument: it is easy to see that, if $\vDash_{\text{cl}}$ governs $A_1, \dots, A_n, B$ and $A_1, \dots, A_n \vDash_{\text{cl}} B$, then, for each logically possible world w, the ideal credence function v_w satisfies (BP14). That is,[46]

$$v_w(A_1) + \cdots + v_w(A_n) - (n-1) \leq v_w(B)$$

What's more, whenever two credence functions satisfy (BP14), so does every convex combination of them. And whenever each credence function in an infinite sequence satisfies (BP14), so does the limit of that sequence. Thus, every credence function in the closed convex hull of the ideal credence functions satisfies (BP14). And thus, by Lemma 6, any credence function that violates (BP14) is strictly logically dominated. This establishes (BP14).

Non-Classical Logics

What happens when we move from classical logic to a non-classical alternative? The key issue here is to determine, for each possible world w, what the vindicated credence function v_w is at that world. In the classical case,

$$v_w(A) = \begin{cases} 1 & \text{if } w(A) = \mathbf{t} \\ 0 & \text{if } w(A) = \mathbf{f} \end{cases}$$

But what about the non-classical case? Here is one suggestion:[47]

$$v_w(A) = \begin{cases} 1 & \text{if } w(A) \text{ is designated} \\ 0 & \text{if } w(A) \text{ is not designated} \end{cases}$$

We might call this the *Vindicated is Designated* condition on epistemic utility. Notice that this is analogous to the suggestion in the full belief case that, if a proposition has designated truth value, then belief is the ideal categorical doxastic

"On conditionals," *Mind* 104 (1995): 235–329.

[46] *Proof.* There are two cases. First, if there is A_i such that $v_w(A_i) = 0$, then, since $v_w(A_j) \leq 1$ for all $1 \leq j \leq n$, $v_w(A_1) + \cdots + v_w(A_n) - (n-1) \leq 0 \leq v_w(B)$. Second, if $v_w(A_i) = 1$, for all A_i, then $v_w(B) = 1$ and $v_w(A_1) + \cdots + v_w(A_n) - (n-1) = 1 = v_w(B)$. QED.

[47] J. R. G. Williams, "Gradational Accuracy and Non-Classical Semantics," *Review of Symbolic Logic* 5, 4 (2012): 513–537.

attitude, with value R, while disbelief takes value $-W$, and if that proposition has a non-designated truth value, then disbelief is the ideal attitude, with value R, while belief takes value $-W$. In the credal case, we can then appeal to a result due to Jeff Paris to provide epistemic utility arguments for various bridge principles for a wide variety of non-classical logics.[48]

Theorem 7

Suppose:

i. The logical consequence relation for a many-valued logic k is defined in terms of the preservation of designated truth values.

ii. $w(X \,\&\, Y)$ is designated iff $w(X)$ and $w(Y)$ are both designated.

iii. $w(X \vee Y)$ is designated iff $w(X)$ or $w(Y)$ is designated.

iv. $v_w(A) = \begin{cases} 1 & \text{if } w(A) \text{ is designated} \\ 0 & \text{if } w(A)\text{is not designated} \end{cases}$

v. $\mathrm{EU}(c, w) = \sum_{X \in \mathcal{F}} \mathrm{eu}\,(w(X), c(X)) = -\sum_{X \in \mathcal{F}} \mathfrak{d}\,(v_w(X), c(X))$

Then:

a. c is strictly logically dominated if it is not a *generalized probability function for logic* k.

That is, c is strictly logically dominated if it fails to satisfy any of the following bridge principles:

(BP14a) If $\vDash_{\mathrm{k}}$ governs A, then $\vDash_{\mathrm{k}} A$, then $c(A) = 1$

(BP14b) If $\vDash_{\mathrm{k}}$ governs A, then $A \vDash_{\mathrm{k}}$, then $c(A) = 0$

(BP15) If $\vDash_{\mathrm{k}}$ governs A, B, and $A \vDash_{\mathrm{k}} B$, then $c(A) \leq c(B)$

(BP16) If $\vDash_{\mathrm{k}}$ governs A, B, then $c(A \,\&\, B) + c(A \vee B) = c(A) + c(B)$

In fact, we only require (ii) and (iii) in order to infer (BP16), which is not a logic-rationality bridge principle. Thus, if we are interested only in the bridge principles, we can prove a more general theorem. This is the credal analogue to Theorem 2:[49]

[48] *Proof.* Paris proves that, if logic k satisfies (i), (ii), and (iii), and if v_w is defined as in (iv), then the closed convex hull of the set of vindicated credence functions is precisely the set of credence functions that satisfy (BP14-16). We then simply apply Lemma 6 to obtain the theorem. QED.

[49] *Proof.* Note that, if (i) and (ii) hold, then (BP14a), (BP14b), and (BP15) are all satisfied by each of the vindicated credence functions. What's more, when those conditions are satisfied by two credence functions, they are also satisfied by any convex combination of them; and when they are satisfied by each credence function in a sequence, they are also satisfied by the limit, if such exists. Thus, they are satisfied by everything in the closed convex hull of the vindicated credence

Theorem 8

Suppose:

i. The logical consequence relation for a many-valued logic k is defined in terms of the preservation of designated truth values.

ii. $v_w(A) = \begin{cases} 1 & \text{if } w(A) \text{ is designated} \\ 0 & \text{if } w(A) \text{is not designated} \end{cases}$

iii. $\mathrm{EU}(c, w) = \sum_{X \in \mathcal{F}} \mathrm{eu}\,(w(X), c(X)) = -\sum_{X \in \mathcal{F}} \mathfrak{d}\,(v_w(X), c(X))$

Then:

a. c is weakly dominated if it fails to satisfy any of the following bridge principles:

(BP14a) If $\vDash_{\mathrm{k}}$ governs A, then $\vDash_{\mathrm{k}} A$, then $c(A) = 1$

(BP14b) If $\vDash_{\mathrm{k}}$ governs A, then $A \vDash_{\mathrm{k}}$, then $c(A) = 0$

(BP15) If $\vDash_{\mathrm{k}}$ governs A, B, and $A \vDash_{\mathrm{k}} B$, then $c(A) \leq c(B)$

Thus, we have the following epistemic utility argument for the logics in question:

(EU10) Bregman Divergence + Vindicated is Designated + Strict Logical Dominance ⇒ (BP14-15).

And, as in the classical case, we can also establish

(BP17) If $\vDash_{\mathrm{k}}$ governs $A_1, \dots, A_n, B$, and $A_1 \dots, A_n \vDash_{\mathrm{k}} B$, then $c(A_1) + \cdots + c(A_n) - (n-1) \leq c(B)$

That is,

(BP17) If $\vDash_{\mathrm{k}}$ governs $A_1, \dots, A_n, B$, and $A_1 \dots, A_n \vDash_{\mathrm{k}} B$, then $\overline{c}(B) \leq \overline{c}(A_1) + \cdots + \overline{c}(A_n)$

So we obtain Field's logic-rationality bridge principle for all such logics.

However, as in the full belief case, while this may be the correct account of epistemic value for Field's use of strong Kleene logic or Priest's use of Logic of Paradox, it is not obviously the correct account for the application of strong Kleene logic to vague propositions nor the application of Logic of Paradox to propositions concerning future contingents. But, as the following theorem shows, as soon as we abandon this account of epistemic value, we lose the No Drop principle, (BP15), and with it the logic-rationality bridge principle that Field endorses, namely, (BP17). This is the analogue of Theorem 3:[50]

functions. QED.

[50] *Proof.* The truth value **i** from (iv) is either designated or undesignated. Suppose first that **i** is

Theorem 9

Suppose:

i. The logical consequence relation for a many-valued logic k is defined in terms of the preservation of designated truth values.

ii. $A \vDash_{\mathrm{k}} B$

iii. If $w(X) = \mathbf{t}$, then $v_w(X) = 1$; and if $w(X) = \mathbf{f}$, then $v_w(X) = 0$.

iv. There is truth value **i** such that, if $w(X) = \mathbf{i}$, then $0 < v_w(X) < 1$.

v. $\mathrm{EU}(c, w) = \sum_{X \in \mathcal{F}} \mathrm{eu}\,(w(X), c(X)) = -\sum_{X \in \mathcal{F}} \mathfrak{d}\,(v_w(X), c(X))$

Then:

a. There is an undominated credence function c such that $c(A) > c(B)$.

For instance, suppose we define v_w as follows for strong Kleene logic:

$$v_w(A) = \begin{cases} 1 & \text{if } w(A) = \mathbf{t} \\ \frac{1}{2} & \text{if } w(A) = \mathbf{u} \\ 0 & \text{if } w(A) = \mathbf{f} \end{cases}$$

This seems natural when the propositions in question include vague properties and they are governed by strong Kleene logic. If we do this, we can no longer establish the following version of (BP15):

(BP15 $_{\mathrm{skl}}$) If $\vDash_{\mathrm{skl}}$ governs A, B, and $A \vDash_{\mathrm{skl}} B$, then $c(A) \leq c(B)$.

Suppose $A \vDash_{\mathrm{skl}} B$. This does not preclude a possible world w such that $w(A) = \mathbf{u}$, but $w(B) = \mathbf{f}$. But in that world $v_w(A) = \frac{1}{2}$ and $v_w(B) = 0$. Thus, $v_w(A) > v_w(B)$.

The upshot of this section is similar to the upshot of our earlier section on bridge principles for beliefs in the presence of non-classical logics: the fate of logic-rationality bridge principles is sensitive to the logic that governs the propositions in question, the interpretation of the truth values in that logic, and the credences we thereby identify as vindicated.

designated. Then there is a logical possibility w^*, where A takes **t** and B takes **i**. And in this case $v_{w^*}(B) < v_{w^*}(A)$, by (iv). Since v_{w^*} is a vindicated credence function, it is in the closed convex hull of the vindicated credence functions. Thus, it is undominated. Next, suppose that **i** is undesignated. Then there is a logical possibility $w^\dagger$, where A takes **i** and B takes **f**. And in this case $v_{w^\dagger}(B) < v_{w^\dagger}(A)$, by (iv). Since $v_{w^\dagger}$ is a vindicated credence function, it is in the closed convex hull of the vindicated credence functions. Thus, it is undominated. QED.

Conclusion

In this paper, we have explored a novel way to adjudicate between the vast variety of putative logic-rationality bridge principles that purport to govern our full beliefs, disbeliefs, and suspensions of judgment, as well as the bridge principles that purport to govern our credences. We have deployed epistemic utility theory to discover which bridge principles are justified by considerations of the epistemic value that accrues to our doxastic attitudes in virtue of their accuracy. Our conclusions are a mixed bag. With very weak and natural assumptions about the epistemic utility of categorical doxastic attitudes the classical single-premise case for full belief, we found compelling arguments for the principles that most of the literature agree upon: if A entails B, then you ought not to believe A and disbelieve B, you ought not to believe A and suspend on B, and thus you ought to see to it that, if you believe A, and you adopt any attitude towards B, that attitude should be belief, providing A is not a contradiction. However, the picture is more complicated when we move to the classical multi-premise case and the non-classical single- and multi-premise cases. In these cases, the ways in which we assign epistemic utility to doxastic attitudes becomes very relevant. For instance, we obtain an epistemic utility argument for the multi-premise version of the principle that we justified in the single-premise case only if we assume that the badness of believing incorrectly is much greater than the goodness of believing correctly. And, in the non-classical case, whether or not the corresponding version of this principle holds depends on our account of the ideal doxastic attitude towards propositions with non-classical truth values.

WHY ANTI-LUCK VIRTUE EPISTEMOLOGY HAS NO LUCK WITH CLOSURE

Maura PRIEST

ABSTRACT: In Part I, this paper argues that Duncan Pritchard's version of safety is incompatible with closure. In Part II I argue for an alternative theory that fares much better. Part I begins by reviewing past arguments concerning safety's problems with closure. After discussing both their inadequacies and Pritchard's response to them, I offer a modified criticism immune to previous shortcomings. I conclude Part I by explaining how Pritchard's own arguments make my critique possible. Part II argues that most modal theories of knowledge will run into problems similar to those found in Pritchard's Anti-Luck Virtue Epistemology. I hence offer my own theory grounded in risk assessment and explain why and how it does much better.

KEYWORDS: safety, closure, barns, risk. Edmund Gettier

Preliminary Remarks

"Anti-Luck Virtue Epistemology" is Duncan Pritchard's response to what he perceived as an inability of a pure anti-luck theory to accommodate the widespread intuition that knowledge is a product of the knower's cognitive abilities. Pritchard hence modifies his account by incorporating a virtue component; therein lies his move from an anti-luck epistemology to Anti-luck Virtue Epistemology (ALVE).[1] According to ALVE, knowledge consists of two related epistemic criteria to satisfy two compelling intuitions. The anti-luck criterion, the intuition that knowledge is incompatible with luck; the virtue criterion, that knowledge is a product of the knower's cognitive abilities.[2]

This paper argues that safety, Pritchard's *anti-luck* criterion, is incompatible with closure and then offers an alternative solution which does not run into the same problem. I divide the paper into two parts, a negative part and a positive one. Part one is my negative argument against Pritchard's anti-luck virtue epistemology (ALVE). It begins by reviewing past arguments concerning safety's

[1] Duncan Pritchard, "Anti-Luck Virtue Epistemology," *The Journal of Philosophy* 109 (2012): 247-248.

[2] Pritchard, "Anti-Luck," 247-248.

problems with closure. I then explain both the inadequacies of such arguments and Pritchard's response to them. Third, I offer a modified criticism immune to the previously mentioned shortcomings. I conclude my negative argument with an explanation of how Pritchard's arguments for ALVE push him into this predicament.

The positive part of my paper argues in favor of an alternative theory of knowledge that avoids the closure dilemma. I suggest the modal aspect of Pritchard's theory forces him to deny closure. I hence argue that a risk grounded account fairs much better than a modal one, at least in this respect. I will not offer an 'all things considered' argument in favor of my theory. Rather, I hope to get epistemologists interested in risk-centered theories by demonstrating their superiority in this small but important aspect of epistemology. In other words, a theory's ability to explain fake barn examples while also accommodating closure gives us reason to take it seriously.

Part I: No Luck with Closure

1. Kripke's Farm

> One version of Pritchard's safety principle is below.
>
> SAFETY: SP**) S's true belief is safe iff in most near-by possible worlds in which S continues to form her belief about the target proposition in the same way as in the actual world, and in all very close near-by possible worlds in which S continues to form her belief about the target proposition in the same way as in the actual world, the belief continues to be true.[3]

The gist of the above is that safe beliefs could not have easily been false, when 'easily' is understood modally. If S's belief that p is safe, there is no close world in which S believes p but p is false.[4]

As mentioned, this safety condition, allegedly, has problems with closure. Let us begin by noting that there are many variations of the closure principle. We can first look at the 'naïve' closure principle, which at first glance seems common sense but upon closer inspection seems implausible:

> *Naïve Closure Principle:* If S knows P, and if P necessitates Q, then S knows Q.

[3] Pritchard, "Safety Based Epistemology," *Journal of Philosophical Research* 34 (2009): 34.

[4] As seen in the above quotation, Pritchard makes nuanced distinctions between "close worlds," "very close worlds," and "near-by worlds." I think there are many instances in which such nuance is important. For the purposes of this paper, however, it is unnecessary. For simplicity, we will only distinguish between worlds that are close and those that are not. If a believer would falsely believe p in a 'close' world, his belief p is unsafe.

At first glance this principle appears plausible. We think that if a subject knows some proposition, and that proposition necessitates a second proposition, then he or she will know that second proposition. We can even point to a few cases in which this Naïve Closure Principle seems to hold. Suppose, for instance, that Steve knows that Mary is a women. Steve also knows that being a women implies (necessarily) that one is not a bachelor. It would seem to obviously follow that Steve knows that Mary is not a bachelor.

In spite of some intuitive plausibility, there are various problems with naïve closure. The most obvious is this: one can know P while lacking knowledge, or even awareness of, P's entailments. S might know P and P might imply Q, but if S is unaware that P implies Q, then clearly S will not know Q. It is possible, however, to amend the closure principle to account for this issue. Let us call this modified version "Less Naïve Closure Principle"

> *Less Naïve Closure Principle* If S knows P, and if P necessitates Q, and If S knows that P necessitates Q, then S knows Q

The above closure principle is more plausible than naïve closure, but it still has its problems. The most obvious is this: S can know P, and also know that P implies Q, and also believe Q, but nonetheless believe Q *for the wrong reasons.* Philosophers are capable of dreaming up very strange scenarios. And they may easily dream up one in which, for instance, (1) Sam knows that Paul runs slowly, and, (2) also knows that Paul running slowly implies that Paul does not run quickly. In addition, in this odd world, Sam believes (3) that Paul does not run quickly. He believes this entailment, however*, not because* he has inferred (2) from (1), but rather because he trusts tea leaf readings. In this instance the Less Naïve Closure Principle holds, but because Paul believes for the wrong reasons, it appears he lacks knowledge. Tea leaf readings provide no justificatory grounds. We are once again left with an unsatisfactory closure principle.[5]

In the first chapter of his book, *Epistemic Angst,* Pritchard discusses the problems with the naive versions of the closure principle I just explained (although he does not use my terms). Pritchard argues that the true intuitive aspect of closure is that, "…such principles attempt to codify how one might legitimately extend one's knowledge via competent deduction from what one already knows."[6] Pritchard goes on to formulate this preferred version of the

[5] Pritchard himself has an extensive discussion of the different variations of the closure principle (including the types I am calling the "Naive Closure Principle" and the "Less Naive Closure Principle") in the first chapter of his book, *Epistemic Angst*

[6] Pritchard, *Epistemic Angst: Radical Skepticism and the Groundlessness of Our Believing* (Princeton: Princeton University Press, 2016), 13.

closure principle that accounts for the 'competent deduction' intuition as follows, "If S knows that p, and S competently deduces from p that q, thereby forming a belief that q on this basis while retaining her knowledge that p, then S knows that q."[7]

I agree with Pritchard that the intuitive version of closure is something very close to the above; throughout the rest of the paper when I reference 'closure,' I mean something just along these lines. Or, to use Pritchard's own words, "henceforth when we refer without qualification to the "closure principle" we will have this highly compelling articulation of the principle in mind."[8]

Now that we understand our terminology, let us go back to safety and the potential problems it runs into with closure. Here is one problematic scheme discussed in the literature:

> (1) An agent forms a belief about an object and a quality of that object
>
> (2) The agent forms a general belief about that object because it is entailed by (1)
>
> (3) (1) is safe but (2) is unsafe
>
> (4) The agent thereby knows a proposition but not the entailment

The most well-known example comes from Saul Kripke in a criticism of Robert Nozick's sensitivity condition. Kripke alters the traditional fake barn Gettier case along the following lines:

> RED BARN, GREEN BARN: Henry* is driving past a farm with one real green barn and many red fakes. His eyes fall upon the green barn and he believes "That is a green barn." From this he forms a belief in the entailment, "That is a barn."[9]

Although Kripke's case was aimed against sensitivity, as others noticed, it also appears applicable to safety. According to ALVE, Henry* knows 'that is a

[7] Pritchard, *Epistemic Angst,* 13. After offering this formulation, Pritchard gives credit in the following footnote, "This is essentially the formulation of the closure principle put forward by Williamson (2000a, 117) and Hawthorne (2005, 29). See also David & Warfield (2008)" (*Epistemic Angst*, 191). In addition to these authors cited by Pritchard, others places readers can find an extensive discussion of closure include Peter Bauman, "Epistemic Closure," in Sven Bernecker and Duncan Pritchard's, *The Routledge Companion to Epistemology* (New York: Routledge, 2011), 597-608 and Sven Bernecker, "Sensitivity, Safety, and Closure," *Acta Analytica* 27 (2012): 367-381.

[8] Pritchard, *Epistemic Angst*, 14.

[9] Kripke's example is a modification of one described by Alvin Goldman in "Discrimination and Perceptual Knowledge," *The Journal of Philosophy* (1976): 771-791. Goldman credits the case to Carl Ginet. See Saul Kripke "Nozick on Knowledge," in *Philosophical Troubles: Collected Papers* (New York: Oxford University Press, 2011), ch. 7.

green barn,' but not 'that is a barn.' He forms his green barn belief via his properly functioning visual abilities, thereby meeting the ability criterion, and because the belief is safe he meets the anti-luck criterion. Regarding the entailment barn belief, however, the ability criterion is meet but *not* the anti-luck criterion. In a close world Henry* falsely believes he is looking at a red barn and so falsely believes the entailment. Hence the belief in the proposition, 'that is a barn,' is unsafe and fails to qualify as knowledge. This is a troubling closure violation. How can you know that there is a green barn but not that there is a barn? Recognizing the unfortunate consequences, Pritchard construes a response to preserve both safety and closure. Contrary to first appearances, he argues, Henry*'s green barn belief is *unsafe*. Here is his reply to a 'Kripke barn' challenge posed by Jonathan Kvanvig.[10]

> The trouble with examples such as this is that it is far from plausible that the agent has knowledge of the antecedent proposition-in this case that this is a green barn-in the first place… it seems that the agent in this example does not have a safe belief in the target proposition, since in an environment where there is barn-deception going on there will be a wide class of nearby possible worlds where, for example, the agent is looking at a green barn facade and yet is nevertheless forming a belief that she is looking at a green barn (it could be, for instance, that this is one of the barn facades that the townsfolk haven't got around to painting red yet).[11]

This reply is puzzling; it is unclear why there *must* be a close world with green fakes. Surely this depends on the details of the example. And even if Kripke's or Kvanvig's particular construction doesn't apply, surely we might imagine a case in which no close world has green fakes. Consider this one:

> RED BARN, GREEN BARN 2: As Henry** drives through fake barn country, he looks at the one real (green) barn, believing, "That is a green barn." Unbeknownst to him, the surrounding barns are fakes, some red, some green. Neither is he aware that this particular fake barn country is managed by Seuss, a demon epistemologist. Bored with the usual Gettier problems, Seuss behaves as follows: If Henry**'s eyes veer toward a green fake, Seuss magically erects a real green barn in front; he does nothing when it comes to red fakes.

The point, of course, is just that of RED BARN, GREEN BARN. Henry**'s green barn belief is safe, his entailment belief unsafe: He knows that there is a green barn but not that there is a barn. The green barn belief is safe, for the demon guarantees that Henry can only view real green barns. However, since the

[10] Jonathan Kvanvig, *The Knowability Paradox* (New York: Oxford University Press, 2006), ch.4.

[11] Pritchard, *Epistemic Luck* (New York: Oxford University Press, 2005), 168.

demon does nothing with red fakes, there is a close world in which Henry** falsely believes a fake red barn is real, and from this he forms a false belief in the entailment. Notice that RED BARN, GREEN BARN 2 leaves little room for Pritchard's already suspect rejoinder. In response to the original RED BARN, GREEN BARN, he argued that there must be a close world with green fakes. It seemed a strange retort, because it seems we can stipulate otherwise. This is displayed in RED BARN, GREEN BARN 2. Thanks to Seuss, in no close world does Henry** have false green barn beliefs. Our demon ensures truth. In Section 4, we see that to insist otherwise, to argue against the stipulation of such demons, is bound to undermine Pritchard's own methodology.

2. Methods Rejoinder

One alternative line of response open to Pritchard involves an appeal to method relativization. Most versions of safety are defined in terms of belief acquisition method: to determine whether a belief is safe, one must consider the method that the agent used to acquire it. Remember that Pritchard has defined his safety condition as follows:

> SAFETY: SP**) S's true belief is safe iff in most near-by possible worlds in which S continues to form her belief about the target proposition in the same way as in the actual world, and in all very close near-by possible worlds in which S continues to form her belief about the target proposition in the same way as in the actual world, the belief continues to be true.[12]

Notice that Pritchard stipulates that S's belief is safe if there are no very close worlds in which S forms a false belief in the target proposition, '*in the same way as in the actual world.*' We can think of 'the same way' as referring to a method of belief formation. In order for S's belief that p to be safe, it is *not* necessary that whenever S believes p in very close nearby worlds, she believes truly. Rather, S cannot falsely believe p in a very close world *via the same method.* I can know that my brother is home when I see him sitting on the couch. I can know this even though there is a close world in which I would falsely believe as much. For although I hold a false belief in a close world, it would be acquired via a method that is distinct from the one used in the real world. I might, for example, falsely believe my brother is home via my mother's lying testimony. However, the method by which I would acquire this false belief (testimony) is distinct from the method via which I acquire my true belief in the actual world (vision).[13] Some

[12] Pritchard, "Safety-Based Epistemology," 34.

[13] This example is based off of a similar example provided by Robert Nozick, *Philosophical*

might argue that Pritchard's theory can be saved via method relativization. We just need to suppose that Henry acquires his barn belief via the following method (M1):

> M1: Henry deduces the belief, "I am looking at a barn," from his other belief, "I am looking at a *green* barn."

Given M1, Henry can indeed know he is looking at a barn. Because we stipulated that there are no close worlds with green fakes, there are no close worlds where Henry, *via M1*, falsely believes he is looking at a barn. If Henry indeed has knowledge, there is no longer a problem with closure. Consider: Henry has a safe, true, green barn belief acquired via his own abilities. If Pritchard's ALVE is right, then it follows that Henry also has a safe belief in the entailment which amounts to knowledge. The belief qualifies not only as safe but also as what Pritchard would call a 'safe cognitive success.' In order for the belief to be a 'safe cognitive success,' we must be able to credit the success (true belief) to the agent's cognitive abilities. And in this case we can. First, Henry used his visual abilities to form his belief about the green barn and also his competent deductive abilities to form his true belief in the entailment, "I am looking at a barn." This entailment belief is also safe, because there is no close world in which, *via M1* (deduction from 'green barn belief), Henry holds a false belief in the target proposition ('I am looking at a barn'). Hence Henry *does know* that he is looking at a barn, and there is no longer a problem with closure.

Before evaluating the counter reply at hand, we should note that method relativization has faced criticism that it falls victim to a generality problem. Generality worries amount to the following: If we define a method too broadly, method relativization will not work as desired. (A broadly defined method might not be much better than no method at all). However, if we define a method too narrowly, satisfaction of the safety condition becomes trivial.

Suppose that I form a true belief about the number of grains of sand on a beach. My method, *prima facie*, is 'guessing.' But we might define the method more narrowly as: 'guessing while walking a Dalmatian on a cold winter day in December approximately 7 minutes after 5pm.' This method characterization makes my true sand belief trivially safe. It is safe because I will never have the relevant false belief via that uniquely defined method. Yet it is trivially safe because the method is so obscure that there is no close world in which I would ever use it again.

A similar triviality objection is applicable to Henry if we define his method

Explanations (Cambridge: Harvard University Press, 1981).

of belief formation as "deriving the belief that one is looking at a barn from the belief that one is looking at a green barn." The belief is safe, but only trivially so. Henry will never form a false belief via the method in question, because by stipulation, there is no close world with green barns and so no close world in which he could possibly use the same method. Even more, if we get into the habit of relativizing methods this narrowly, it is bound to set safety up for many accusations of triviality in similar cases.

Generality worries, although problematic, are not the biggest problem with the method relativization response. In the next section, we will discuss Pritchard's move to ALVE, which is at least in part a response to safety and triviality worries, and might have some potential to help Pritchard out with the generality problem. The biggest problem for Pritchard is not method generality, but that method relativization will get Pritchard out of the closure dilemma only by leaving him with a completely different problem. Because of method relativization, we are supposed to admit that Henry *does* know that he is looking at a barn. Yet, intuitively, Henry does *not* know this. Recall that Pritchard's initial response to the Kripke barn challenge was to argue that Henry lacked knowledge both about the green barn and the barn itself. Pritchard argued as much because he wanted to show that ALVE aligns with the widespread intuition that Henry lacks knowledge about the barn. The safety condition is supposed to be a preferred epistemic criterion specifically because it gives us the intuitive result in 'Fake barn Gettier cases' (i.e. the result that Henry does not know that he is looking at a barn).

To conclude this section, method relativization cannot save Pritchard from the dilemma at hand. It cannot do this for at least two reasons. First, we must describe Henry's method of belief formation in an unnaturally narrow way if it is to be of any help with the closure dilemma. This excessive narrowness risks subjecting safety to further generality accusations. And even more importantly, method relativization will only get Pritchard out of the dilemma at the cost of giving us the wrong result in fake barn Gettier cases. If one admits that Henry knows he is looking at a barn, one simply bites the bullet on the Gettier case. A purported advantage of ALVE, however, was that it *does not* bite the bullet in this way.

3. Safety & Ability

When Pritchard argued for his change from a mere anti-luck theory of knowledge to an anti-luck *virtue* epistemology, he motivated the switch with the following case:

> TEMP: Temp forms his beliefs about the temperature in the room by consulting a

> thermometer. His beliefs, so formed, are highly reliable, in that any belief he forms on this basis will always be correct. Moreover, he has no reason for thinking that there is anything amiss with his thermometer. But the thermometer is in fact broken, and is fluctuating randomly within a given range. Unbeknownst to Temp, there is an agent hidden in the room who is in control of the thermostat whose job it is to ensure that every time Temp consults the thermometer the "reading" on the thermometer corresponds to the temperature in the room.[14]

Pritchard claims that Temp lacks knowledge, even though his belief is safe. The missing ingredient is ability, "[W]hat is wrong with Temp's beliefs is that… their correctness has nothing to do with Temp's abilities and everything to do with some feature external to his cognitive agency."[15] Because what explains Temp's success is not ability but the hidden agent, we are disinclined to attribute knowledge. Safety is too weak on its own and must be supplemented with an ability criterion.[16]

Notice that for the Temp example to work, the hidden helper must *guarantee* Temp's safe beliefs. In Pritchard's words, "[W]hatever one wishes to say about what is epistemically deficient in Temp's beliefs, it does not seem that his beliefs fail to satisfy the anti-luck intuition. After all, his beliefs are *guaranteed* to be true…"[17] (my emphasis). Let us review the structure of the Temp case:

Temp has true temperature beliefs

The beliefs are safe because a hidden agent ensures their truth

Intuitively Temp lacks knowledge

Hence safe belief is insufficient for knowledge

Hence we must amend our theory of knowledge (with an ability condition)

The Temp case is critical for Pritchard's move from an anti-luck theory of knowledge to a theory that incorporates a virtue component. It also guarantees problems in RED BARN, GREEN BARN 2. The Temp case needs a hidden helper to ensure safe beliefs. Again, in Pritchard's words, "…Temp's belief *satisfies the safety principle*. This is ensured by the fact that the manner in which Temp is forming his beliefs, such that success is guaranteed, means that *it can hardly be the*

[14] Pritchard, "Anti-Luck," 260.

[15] Pritchard, "Anti-Luck," 260.

[16] "Anti-Luck Virtue Epistemology (ALVE): S knows that p if and only if S 's safe true belief that p is the product of her relevant cognitive abilities (such that her safe cognitive success is to a significant degree creditable to her cognitive agency)" (Pritchard, "Anti-Luck," 260).

[17] Pritchard, "Anti-Luck," 261.

case that he could easily have formed a false belief" [18](my emphasis). If Pritchard can stipulate a helper that ensures Temp's safe temperature beliefs, others can stipulate demons who ensures that fellows named Henry have safe green barn beliefs. Pritchard cannot, then, dismiss our objection on the grounds that "…there will be a wide class of nearby possible worlds where… (Henry**) is looking at a green barn facade and yet…forming a belief that (he) is looking at a green barn."[19] Thanks to Seuss, in RED BARN, GREEEN BARN 2, there is no such close world. Henry**'s belief is safe according to Pritchard's own standards.

At this point some are probably thinking, 'Wait a minute, doesn't TEMP *vindicate* Pritchard'? Seuss, just like Temp, ensures safe belief. The demon's help, some might argue, disqualifies Henry** from meeting the ability criterion and hence the complete demands of knowledge. And if Henry** thereby lacks knowledge of the green barn, it is irrelevant that he also lacks knowledge of the entailment. Here are the problems with that claim. In the original fake barn case, Pritchard himself argues that Henry's safe cognitive success is indeed explained by his own abilities. In Pritchard's words, "…given that (Henry) does undertake, using his cognitive abilities, a genuine perception of the barn, it seems that his cognitive success is explained by his cognitive abilities…"[20] (original emphasis). Notice that just like Henry from the original fake barn case, Henry** "undertakes, using his cognitive abilities, a genuine perception of the barn." There is, then, no grounds to say that Henry does not meet the standards of ALVE which demand a 'safe cognitive success.' Henry's belief, after all, is safe. And, as Pritchard himself admits, Henry uses his cognitive abilities in acquiring this safe belief. Moreover, Henry and Henry** exercise the very same abilities. We must then conclude that Henry**'s barn belief qualifies as a safe cognitive success (just like Henry) thereby meeting the standards of Pritchard's very own ALVE.

Given Pritchard's decision to prescribe to a *weak* virtue theory, it is especially difficult for him to reply to the above criticism. Describing ALVE, he argues, "…the ability condition in play here is that proposed by a weak virtue epistemology rather than a strong virtue epistemology…the agent's *safe* cognitive success should be to a significant degree creditable to her cognitive agency"[21] (original emphasis). Pritchard rejects that safe belief must be *because of* or *primarily* creditable to the believer's cognitive ability. What matters is that the agent's abilities are involved to 'a significant degree.' Henry** passes the bar in this

[18] Pritchard, "Anti-Luck," 259-260.

[19] Pritchard, "Anti-Luck," 260.

[20] Pritchard, "Anti-Luck," 272-273.

[21] Pritchard, "Anti-Luck," 274.

regard; his reliable vision is indispensably involved in acquiring his safe belief. We might even stipulate that Henry** has unusually excellent visual abilities, thus highlighting his contribution even further. If Henry of the original fake barn case utilizes his abilities in a way that satisfies Pritchard's self-described standards (i.e. so that 'his cognitive success is explained by his cognitive abilities'), then so must Henry**.

What if Pritchard contends that the demon's help diminishes Henry**'s import to an insignificant level? I find this unconvincing, given that (1) 'significant' involvement of the agent's cognitive abilities is enough to secure the virtue component, and (2) as mentioned, we can stipulate that Henry** had especially exquisite vision. Exquisite vision, it seems, should qualify as 'significant' involvement in the acquisition of a safe belief. Notwithstanding, although I am not convinced of Pritchard's imagined retort, there is indeed a way to counter it. Let us consider one last barn example.

> RED BARN, GREEN BARN 3 A demon named Henry*** is driving through fake barn country, and like the other Henrys, believes he is in regular barn country. Henry*** views the one real (green) barn surrounded by red fakes, forming the belief, "That is a green barn." Connecting the logical dots, he then believes the entailment, "That is a barn." Henry*** is especially confident in his green barn belief, because many years ago he cast a special spell. (Demons can cast one personally enhancing spell as soon as they are of age to exercise demonic powers.) The spell was as follows: *Every belief I form about green objects will qualify as safe, according to Duncan Pritchard's safety condition.*

We can see that Henry***'s green barn belief is safe; the spell guarantees as much. Casting the spell himself, moreover, allows him to fulfill the ability criterion. (Let us imagine that spell casting requires advanced analytic ability. This is why demons cannot cast spells upon birth, but must wait until their cognitive capacities are more fully developed.) Henry*** meets all the requirements of ALVE and so his green barn belief qualifies as knowledge. His entailment belief, however, is unsafe and thereby not knowledge. The following criticism stands: According to ALVE, Henry*** knows there is a green barn but not that there is a barn. This, of course, is a glaring closure violation. And there are likely to be other cases with the same structure.

4. Hard Choices

While it may be possible to argue something stronger, clearly we must abandon one of the following:

(1) Safety

(2) Closure

(3) The stipulation of creatures that guarantee safety.

We see that (3) from above allows us to modify Kripke's original criticism against sensitivity as to make it equally forceful against safety. Pritchard cannot have it all. Given the foundational import to his theory, rejecting safety itself is unlikely. Abandoning closure is an option, one that others have taken before. But closure's compelling intuitive force makes this less than ideal. Pritchard, moreover, argues that an *advantage* of safety is its' compatibility with closure.[22] Rejecting (3), then, may seem the most palatable. But this too comes with undesirable consequences. If Pritchard loses (3), he loses the Temp case. If he loses the Temp case, he loses what grounds his argument for an ability criterion. There are also problems apart from ALVE. Without justification, a prohibition on safety ensuring creatures seems arbitrary. On the other hand, arguments against such magic may cover too much. If, for example, demons cannot ensure safe belief, what else? Epistemologists who limit demonic powers run the risk of biting the hand that feeds them. The next thing we know epistemic villains are unable to deceive the senses. ALVE then runs into trouble not due to any specific flaw, but because by first putting demons out of business, and thereby the skeptic, epistemologists may unintentionally do the same to themselves.

PART II: A Non-Modal Solution

This second-half of the paper is devoted to showing how a non-modal theory of knowledge avoids the problems with closure that troubles ALVE. Indeed, I think that most modal theories will run into the same problems as ALVE. The problem is it seems ever possible to design an entitlement counterexample that simply does not mesh with modal accounts of knowledge. If I can know a proposition about an object O and a quality of that object Q, then it seems intuitively plausible (in the usual cases) that I can derive a belief in the entailment that is propositional knowledge about O itself. However, as long as the theory is a modal one, the skeptic is there awaiting with a counterexample to deny the epistemic agent his entailment belief. The scope of possible worlds, even close ones, is quite expansive, especially when we include demon worlds. Hence there will be cases in which an agent cannot fulfill the modal criterion for the entailment belief (even though she can fulfill the modal criteria for the entailing belief).

[22] Pritchard, *Epistemic Luck*, 94. See also, Pritchard "Sensitivity, Safety, and Anti-luck Epistemology," in John Greco's *The Oxford Handbook of Skepticism* (Oxford: Oxford University Press, 2008), 447.

My solution to the problem of skeptics and closure, explained in what follows, is to bypass the modal criterion all together. Without a modal requirement, the skeptic's challenges will prove irrelevant. Closure comes out alive (that is, it comes out alive in the most important world, the actual one). In my account a *risk* criterion will in some sense replace the modal criterion. Section 2.1 will detail my risk criterion. Section 2.2 will explain how the risk criterion addresses fake barn Gettier cases. Section 2.3 addresses potential confusions. Lastly, Section 2.4 explains specific advantages of a risk grounded theory compared to modal ones.

Before moving on, I want to clarify a potential confusion. I am about to argue for a 'risk-centered' approach I call, 'risk sensitive credit. I should note that Pritchard's most recent work now describes his theory as 'anti-risk' rather than 'anti-luck.' The risk that Pritchard's theory is now committed to, however, is not the type of risk that is relevant for my own 'risk-sensitive credit.' I agree with Pritchard that "risk assessments seem to be essentially forwards-looking…Luck assessments, in contrast, seem to be essentially backwards-looking…" (Pritchard: forthcoming). It is true that whether we are understanding risk in my probabilistic sense or in Pritchard's modal one, risk assessments are forward looking. According to my own theory, for instance, an agent looks forward toward the probability that her future belief will or will not hold true. This agreement on the forward looking characteristic of risk, however, is where the similarities between my own view of risk and that of Pritchard's grinds to a halt.

Pritchard makes clear that the type of risk which he is concerned with is a modal account. My account of risk is explicitly *not* a modal one. I am understanding risk, rather, as it is being used by cognitive scientists, i.e., as a Bayesian type of probabilistic risk. This is also the type of risk, for instance, that David Henderson and Terry Horgan defend in their own risk-centered theory. This type of risk, rather than concern itself with modal possibilities, is concerned with 'chance' understood in terms of Bayesian probabilities. If believing p is epistemically risky (in my sense), it follows that there is a high probability that 'p' is not true. If belief in p is not risky, then there is a high probability that S's belief in p is accurate. When a theory of knowledge is founded in this type of probabilistic risk-sensitivity, we will see that Henry is able to completely evade the closure related problems that pop-up with fake-barn Gettier cases.[23]

[23] Pritchard, "Epistemic Risk," *The Journal of Philosophy* 113 (2016): 550-571, admits that modifying his account of knowledge from an "anti-luck" account to an "anti-risk" account may seem like a minor change, stating that, "the differences between the two views will not be radical" He nonetheless defends the switch by arguing that an anti-risk account has at least two

2.1. Risk Sensitive Credit

For any proposition p, an agent might believe p, believe not-p, or withhold belief. Believing, however, can be epistemically risky, and at times the risk of false belief is not be worth the potential reward. Along these lines, I will argue that in order for a belief to qualify as knowledge, it must *be risk sensitive*. S's belief is what I call risk sensitive only if the likelihood of false belief is low enough that belief (as opposed to disbelief or withholding) is the best epistemic option. Ernest Sosa, for instance, has discussed ideas along these lines, arguing that "[One's] meta-competence governs whether or not one should form a belief at all on the question at issue, or should rather withhold belief altogether."[24] Elsewhere Sosa argues that "A performance can thus easily fail to be 'meta- apt,' because the agent handles risk poorly, either by taking too much or by taking too little. The agent may fail to perceive the risk, when he should be more perceptive; or he may respond to the perceived risk with either foolhardiness or cowardice…"[25]

What I call risk sensitivity is similar to Sosa's meta-aptness. S's belief lacks risk sensitivity if she takes too much risk or too little. What exactly I mean by 'risk assessment' and 'risk sensitivity' is obviously important. I am *not* thinking about risk assessment as a highly reflective cognitive process. Under my account, risk assessment need not include reflection, higher order beliefs, or even the possibility of either. (If 'assessment' sounds too reflective, you may prefer to think of risk 'accommodation')[26] To explain further, we can helpfully turn to the work of David Henderson and Terry Horgan:

> We ourselves find very plausible the idea that competent risk assessment, as an aspect of the process of forming a belief, is required in order for that belief to

significant advantages. First, he argues that between risk and luck, it is the former that has a better claim to the status of what we might call the more 'fundamental' concept. He argues that, "…we naturally explain a concern to eliminate (luck) in terms of a concern to eliminate (risk) rather than vice versa…" Lastly, although Pritchard himself if not working with my own probabilistic conception of risk, he acknowledges that "most contemporary treatments of risk" utilize "a probabilistic conception."

[24] Ernest Sosa, "Knowing Full Well: the Normativity of Beliefs as Performance," *Philosophical Studies* 142 (2009): 14.

[25] Sosa, "Knowing Full Well," 12.

[26] In Sosa's own words, "We can now see that knowing something full well requires that one have animal and reflective knowledge of it, but also that one know it with full aptness. It requires, that is to say, that the correctness of one's first- order belief manifest not only the animal, first-order competences that reliably enough yield the correctness of the beliefs that they produce. One's first-order belief falls short if it is not appropriately *guided* by one's relevant meta-competence" Sosa, "Knowing Full Well," 16.

> constitute fully human knowledge. But we doubt whether such competence needs to take the form of a higher-order belief; and we also doubt whether a first-order belief can qualify as any kind of knowledge if it is formed in a way that *utterly lack*s the aspect of competent risk assessment (original emphasis).[27]

The risk sensitivity I advocate aligns with Henderson and Horgan on both counts: S cannot know p unless S (or S's abilities or S's cognitive system) assessed (or accommodated) p's risk, but this can take place without higher order belief. Moreover, risk assessment is necessary for knowledge of any kind. H&H further suggest that, "[We] might have a trained capacity that manages to accommodate [risk] without articulation, automatically and quickly…"[28] I agree, but would add that we might also have innate cognitive capacities that *evolved* to accommodate risk. I suspect that H&H were thinking of 'trained' loosely, and this was what they meant. In any case, visual studies confirm that automated cognitive processes can classify sensory data according to a risk sensitive framework. Consider the following commentary on a recent study,

> …Bayesian concepts are transforming perception research by providing a rigorous mathematical framework for representing the physical and statistical properties of the environment… describing the tasks that perceptual systems are trying to perform, and deriving appropriate computational theories of how to perform those tasks, given the properties of the environment and the costs and benefits associated with different perceptual decisions.[29]

The above suggests that perception works within a cost benefit framework that balances the benefits of perceptual belief versus the risks. Further studies provide evidence that we update statistical frameworks according to perceived environment. In short, there is much more to perception than sensory data. To ensure accuracy, our perceptual system first receives sensory information, and then second and separately, accommodates this data in accordance with the environment and other circumstantial contingencies. Environmental awareness, combined with sensory input, leads to risk assessment. This again is supported with research in cognitive science:

> [T]he objects that are likely to occur in a scene can be predicted probabilistically from natural scene categories that are encoded in human brain activity. This

[27] David Henderson and Terry Horgan, "Risk Sensitive Animal Knowledge," *Philosophical Studies* 166 (2013): 601. This quote was aimed at Sosa. Sosa since responded to the criticism and argues that his own account does not demand as much reflection as H&H may have assumed.

[28] Henderson and Horgan, "Risk Sensitive," 603.

[29] Wilson Geisler and Daniel Kersten, "Illusions, Perception and Bayes," *Nature Neuroscience* 5 (2002): 508.

> suggests that humans might use a probabilistic strategy to help infer the likely objects in a scene from fragmentary information available at any point in time.[30]

Our perceptual system matches visual sensations to familiar objects given other information about the environment and contextual circumstance. Suppose you experience a visual stimulus of a small furry animal. If you believe you are in the forest, this stimuli might indicate a squirrel. Contrastingly, if you were at home, your unconscious cognitive processes might suggest that the animal is a cat. In order to acquire perceptual knowledge, your sensory data must first accurately reflect the perceptual object. In other words, your vision is not blurry, you are an appropriate distance from the object, and you are not under the influence of hallucinogens. If this holds, you have data to make a probability assessment in accordance with the environment and other relevant conditions. Back to our visual studies:

> [A]n ideal observer convolves the posterior distribution with a utility function (or loss function), which specifies the costs and benefits associated with the different possible errors in the perceptual decision. The result of this operation is the expected utility (or Bayes' risk) associated with each possible interpretation of the stimulus. Finally, the ideal observer picks the interpretation that has the maximum expected utility.[31]

The above quote nicely explains how sensory input prompts the following evaluation: What are the chances that this stimulus comes from object O given environment E and circumstances C? The answer determines whether it is best to believe p, withhold belief, or believe not-p. Assume that a true belief is an epistemic benefit and a false belief a cost. Ideal agents, we might say, believe p only if belief has the highest expected epistemic value. I do not think, however, that in order to acquire perceptual knowledge one needs to be an 'ideal observer.' Indeed, in order to acquire any type of knowledge one need not be epistemically 'ideal' in any sense at all. Yet I do want to argue that knowledge demands a type of 'creditworthiness.' Hence the account I am arguing for falls under the umbrella of 'credit theories' of knowledge.

While there is all sorts of disagreements between credit theorists, most agree that an agent acquires knowledge when she forms her belief through a process which is 'epistemically creditworthy.' Credit theorists further argue that their accounts are especially well fit to explain the value of knowledge. For while both true belief and knowledge are in some sense epistemically desirable, the

[30] Dustin Stansbury, Thomas Naselaris, and Jack Gallant "Natural scene statistics account for the representation of scene categories in human visual cortex," *Neuron* 79 (2013): 1031.

[31] Geisler and Kersten, "Illusions, Perception," 508.

latter is preferable for it is an 'achievement' or an act of 'creditworthiness.' It is these creditworthy beliefs alone that count as 'knowledge.' The crux of the issue, of course, is just what belief forming mechanisms count as creditworthy. I am arguing that beliefs formed through a process that assesses epistemic risk are those special sort of beliefs that we might deem creditworthy. The creditworthy agent believes p only if believing presents minimal epistemic risk. We can call this Risk Sensitive Credit (RSC). More formally,

> RSC: An agent's belief p is risk sensitive and hence creditworthy if (1) her own abilities assess belief risk, and (2) she correctly believes p because (1) indicates a reasonably low chance of p's falsity.

Some might object to the vagueness of 'reasonably low.' It is used for two reasons. First, it seems a fruitless effort to determine whether the risk of falsehood must be below 15, 10, or 5 percent. Second, philosophers who disagree about justificatory *degree* might still agree on justificatory *kind*. But if we agree that risk sensitive belief is belief in accordance with reasonable risk assessment. What is risk assessment? Briefly, it is a means of analyzing and interpreting relevant data within an environment and set of conditions. Assessment goes about as follows: an agent's cognitive system, consciously or unconsciously, assesses the chances of p given what I call her *total information*. Total information consists of certain epistemic data D and epistemically relevant conditions C. That is, P(P/D&C). Risk assessment can go awry in at least three ways:

> Risk Assessment Errors
>
> (1) Inaccurate data
>
> (2) Inaccuracy regarding the conditions
>
> (3) Misinterpreting the meaning of the data given the conditions

Imagine a risk management company, SECURE, that is hired to assess the safety of a mansion hosting a prestigious fundraiser. SECURE might blunder through inaccurate data gathering, inaccurate conditional assessment, or misinterpretation of the data given the conditions. Examples of the first could include miscounting the fire alarms or misreading the thermostat. Either error would skew total assessment. But maybe there is no data inaccuracy. Problems ensure, however, because there is failure to consider a tornado warning. (A failure of conditional assessment).

A third possibility is that SECURE makes no error in data collection nor conditional assessment, yet still goes wrong in interpretation. They might judge that 7 fire alarms is appropriate when 15 are needed. To do their job, SECURE must collect good data, carefully apprise conditions, and then use both of the aforementioned to arrive at an all things considered risk assessment. Note that a safe event is not enough to fend off criticism. SECURE'S customers can demand a refund upon discovering the event unknowingly presented a high safety risk, even if no risk actualized. Each of us, when making an epistemic risk assessment, functions in a manner similar to SECURE. In other words, we attempt to make an accurate risk assessment (that is, an assessment of the chance of p's truth) given all relevant information. Things can go wrong when we either misinterpret the meaning of information or receive misinformation from the start. In the next section, I will explain the sad story of a misfired risk assessment by an innocent fellow named Henry deep within the land of barn facades.

2.2 Resolving The Fake Barn Dilemma

With the risk sensitive framework just described, we can now explain what goes wrong when Henry views the one real red barn. Although Henry has a true belief, he does not have a risk sensitive belief. Because risk sensitivity is required for knowledge, Henry's true belief fails to qualify.

Let us describe the process that Henry engages in in more detail. We can then see exactly where things go wrong. Henry, through his visual experience of the barn, receives data in need of epistemic analysis. Shortly after receiving this data his perceptual system gauges epistemic risk. Henry, however, assumes he is in a traditional barn environment; this skews assessment. We should think of epistemic evaluation in terms of 'total risk assessment.' In other words, creditworthy epistemic endeavors demand the proper processing of all relevant epistemic information. An agent might receive various information from many sources and over a long time period. Some of this information might be consciously accessible, while other information is not. An agent deserves credit (and so acquires knowledge) when she first accurately processes this data, second comes to the (correct) conclusion that not-p is improbable and therefore truly believes p. With this in mind we can recognize what goes wrong in fake barn country: Henry misinterprets a critical portion of epistemic information when he misjudges his environment 'traditional barn country.' This misinterpretation is critical to his misfired risk assessment and the ensuing failure to obtain knowledge.

Prima facie, we might be tempted to think that Henry's belief forming mechanism is nothing more than visual perception, and this would lead us to conclude he forms his belief via epistemically acceptable means. But things are not so simple. For instance, in challenging Fred Dretske's argument against closure, Pritchard himself has pointed out that beliefs ostensibly formed, 'just by looking,' are in reality much more complex. Suppose, for instance, that Zula looks at a zebra and forms the true belief that what she sees is a zebra. It may be tempting to say she forms her belief, 'just by looking.' But as Pritchard explains, this isn't quite right.

> I think that while there is a sense in which it is obviously true that Zula gains her knowledge just by looking…perceptual knowledge can…involve a wide range of specialist expertise and background knowledge…such expertise and background knowledge would surely have ramifications for the total evidence that you possess in support of your belief… to know a proposition just by looking need not entail that the only evidence you possess for your belief is the evidence you gained from the bare visual scene before you.[32]

Like Zula, Henry's 'evidence' (what I prefer to call 'information') consists in much more than just the bare visual scene before him. Background knowledge plays an important role; only from past experience does Henry know his percept has the appearance of an object called a 'barn,' and that open grassy areas are the types of places where barns are commonly found. Yet unfortunately for Henry, some of his background information misleads. If we assume Henry an ordinary fellow, he hasn't any reason to think that objects that appear like barns are actually barn facades. As far as he knows, it would be pointless to have a town full of barn facades, he has never heard of such things, and he would be prone to suspect (quite reasonably) that those who believe in fake barn country are conspiratorial loons. While these are all reasonable assumptions on his part, they have distorting consequences on his epistemic evaluation.

Total risk assessment is derived from various sources of epistemic information which are first individually interpreted and then collectively assessed. Going too far off the mark when interpreting information will corrupt the collective assessment. This is what happens with Henry. He misinterpreted his environment and unfortunately for him, this misinterpretation played a key role in his total risk assessment. Epistemic creditworthiness does not allow for these types of mistakes. In line with previous credit theorists emphasis on 'credit for *success*,' an understandable epistemic mistake is still a mistake. The idea is similar to the common externalist/reliabilist notion that justification goes beyond that which is internal to the believer. Even if an agent has good reason to think her method is reliable, she cannot be justified if it is unreliable. Similarly, even if we can understand why Henry made the risk assessment that he did (that the object

[32] Pritchard, "Relevant Alternatives, Perceptual Knowledge and Discrimination," *Nous* 44 (2010): 256-257.

before him was likely a barn), it was inaccurate (the object before him was *not* likely to be a barn) and therefore not creditworthy.

2.3 Clarifications

Let me make clear that RSC is not a variant of the so called 'no false lemmas' theory. As some may recall, shortly after Gettier introduced his Problem, a view often referred to as the 'no false lemmas' approach (NFL) suggested a simple solution.[33] According to NFL, Gettier's examples of troublesome beliefs are, in actuality, illegitimate (or unjustified) because they rely on false premises: Smith's true belief that 'the man who will get the job has 10 coins in his pocket' is acquired by reasoning through the false premise that 'Jones will get the job.' Similarly, Smith's true belief that "Either Jones owns a ford or Brown is in Barcelona,' is acquired only via reasoning through the false premise that 'Brown is in Barcelona.' NFL proponents argued that a necessary condition of knowledge was that the 'belief' in 'justified true belief' could not be acquired by reasoning through false premises. With this requirement, we see that the heroes of Gettier's puzzles rely on false premises and this therefore prevents them from acquiring knowledge.

Many problems with NFL soon came to light. First, with some imaginative effort, it is possible to come up with examples similar to those in Gettier's original paper that *do not* rely on false premises.[34] And second, a new breed of Gettier cases, those of the fake barn variety, were introduced onto the epistemological stage.[35] It seemed to many that simple visual beliefs (like the barn façade belief) do not rely on any premises at all, and hence even more so do not rely on false premises.

Because I emphasize the role false information plays in skewing risk assessment, some might confuse RSC with NFL. I want to be clear that RSC is entirely distinct from, and bears very little relation to any variant of the no false premise approach and *does not* suggest that NFL is necessary for knowledge. Let us return to Henry. I argued that his true barn belief, which might appear to arise

[33] See D.M. Armstrong, *Belief Truth and Knowledge* (Cambridge: Cambridge University Press, 1974).

[34] See Keith Lehrer, "Knowledge, Truth and Evidence," *Analysis* 25 (1965):170, and Richard Feldman, "An Alleged Defect in Gettier Counter-Examples," *Australasian Journal of Philosophy* 52 (1974): 68-69. For a challenge to Lehrer and Feldman, See Michael Levine, "Gettier Cases without False Lemmas?" *Erkenntnis* 64 (2006): 381-392.

[35] See Alvin Goldman, "Discrimination and Perceptual Knowledge," *The Journal of Philosophy* (1976): 771-791.

spontaneously, is actually dependent on a vast array of background information, much of which is really misinformation. Such misinformation plays a critical role in tipping Henry's risk assessment scales in the wrong way. However, we should not understand Henry's risk assessment failure in terms of false premises. First off, this would make knowledge requirements unreasonably strict. After all, much everyday knowledge is partly based on false or misleading background information.

Not only is false background information compatible with knowledge, it is unclear that background information necessarily consists of beliefs (beliefs to potentially serve the role of a false lemma). Our cognitive system can register information that never makes its way into the realm of explicit beliefs, and might not even rise to the level of implicit belief. But background information contributes to assessment of epistemic risk nonetheless. It is this *failure to accurately assess epistemic risk* which accounts for Henry's failure to obtain knowledge. Of course, there are many cases in which misleading background information (which may or may not consist of false beliefs) is not enough to prevent a reasonable assessment of epistemic risk. In such cases, one might have knowledge partly based on inaccurate information. However, in other instances, (like with Henry) inaccurate information *does* interfere with a reasonable risk assessment, and thus does prevent one from attaining knowledge.

Let us return to our analogy of the risk assessment company. Imagine that SECURE concludes that there is minimal safety risk at the mansion, but only because the company is unaware of the man-eating grizzly bears who reside in the courtyard. Clearly any valuation made without awareness of this environmental feature will interfere with a successful assessment. Similarly, Henry's ignorance of fake barn country prevents him from accurately assessing the riskiness of his situation.

2.4 Risk Sensitivity & Fake Barns

We can now see the benefits of a risk grounded theory as opposed to a modal one, at least in respect to fake barn Gettier cases. If one is committed to a modal theory and also wants to preserve intuitions in fake barn examples, Henry's lack of knowledge must be explained in terms of false beliefs in close worlds. Indeed, we all admit that Henry has a true barn belief formed through his reliable vision. The challenge is to explain *why* this seemingly true justified belief does not qualify as knowledge. Modal theories must turn to close worlds for an explanation. Pritchard in particular argues that knowledge demands 'safety'. A belief is unsafe if false in a nearby world. Henry, in turn, lacks knowledge because in a close world he falsely believes that a fake barn is real. But this is exactly where Pritchard leaves the door open to demons and skeptics. These challenging critics will quickly create a world in which an entailing belief is safe but the entailed belief is unsafe. And this is when Pritchard (and some other modal theorists) must come face to face with closure denial.

My risk sensitive theory avoids the problems just explained above, because it never gives the skeptic cause to dream about strange and troublesome close worlds. Rather, a risk sensitive theory explains that Henry lacks knowledge without any appeal to modal conditions. All of Henry's epistemic failures can be explained by reference to Henry's epistemic practices in the actual world. In order to acquire knowledge in the actual world, an agent must gather information and make a reasonably accurate epistemic risk assessment. Henry gathers information and makes a risk assessment. Unfortunately for Henry, it is *not* a reasonably accurate assessment.

Risk sensitivity demands reasonable accuracy regarding data, environment, and other relevant conditions. Mistakes about any of these can result in an assessment that either (1) misrepresents epistemic risk, or (2) makes an accurate assessment but only by luck. Both (1) and (2) are incompatible with creditworthiness and thereby knowledge. In the former case inaccuracy is the problem; in the latter accuracy is powerless because it does not derive from the agent's abilities. Henry's problem is with (1). His mistaken environmental assumption that he is in traditional barn country give rise to an inaccurate assessment and he gravely misrepresents epistemic risk. Because accurate risk assessment is required for knowledge, Henry lacks knowledge both about the barn and its color.

Conclusion

At first glance, Kripke's 'green barn challenge' to Nozickian sensitivity appears applicable to safety. Pritchard replied to the challenge, arguing that safety and closure get along just fine. This paper argued that his response works only for certain constructions of the green barn challenge; we have seen that an alternative version leaves no room for Pritchard's counterargument. Thanks to the help of demons, a subject can have a safe green barn belief while her belief in the entailment remains unsafe. Pritchard opened the door to this possibility in his argument for ALVE, which stipulated a safety ensuring hidden helper. Either safety and closure are incompatible, or Pritchard's argument for ALVE falls flat. If it was up to me, I would choose the former. I argued that we can keep both safety and closure if we replace ALVE with an alternative epistemic theory that is grounded in risk assessment rather than modality. My alternative theory, “Risk Sensitive Credit,’ preserves safety and closure while also explaining Henry’s lack of knowledge in fake barn Gettier cases.

AN ARGUMENT FOR THE SAFETY CONDITION ON KNOWLEDGE

Michael J. SHAFFER

ABSTRACT: this paper introduces a new argument for the safety condition on knowledge. It is based on the contention that the rejection of safety entails the rejection of the factivity condition on knowledge. But, since we should maintain factivity, we should endorse safery.

KEYWORDS: safety, knowledge, factivity

The safety condition on knowledge is a necessary condition for knowing that, recently, has been most systematically defended by Williamson, Sosa and Pritchard.[1] But it came into prominence in virtue of Nozik's analysis of knowledge, which was itself a reaction to earlier reliabilist accounts of knowledge and justification.[2] So, the safety condition is supposed to reflect the basic idea of the sort of reliability associated with bona fide knowledge that distinguishes it from mere belief and lucky true belief. The safety condition can be understood simply and informally as follows:

> If *A* knows that *p*, then *A* could not easily have falsely believed that *p*.

This relatively non-technical gloss on safety and it can be made more precise as follows:

> (Safety) $(w_i \vDash K_A p) \rightarrow \neg\, [<w_i> \vDash (B_A p \,\&\, \neg p)]$.

Here '$<w_i>$' is the set of world sufficiently close to w_i, '$K_A p$' represents *A*'s knowing that *p*, and '$B_A p$' represents *A*'s believing that *p*. So understood, the safety condition is the claim that if *A* knows that *p* at w_i, then *A* does not believe that *p* when *p* is false in worlds sufficiently similar to w_i. This regimentation captures the

[1] See Timothy Williamson, *Knowledge and its Limits* (Oxford: Oxford University Press, 2001), Ernest Sosa, "How to Defeat Opposition to Moore," *Philosophical Perspectives* 13 (1999): 141-54, Duncan Pritchard, "Anti-Luck Epistemology," *Synthese* 158 (2007): 277-98, "Knowledge, Luck, and Lotteries," in *New Waves in Epistemology*, eds. Vincent Hendricks and Duncan Pritchard (London: Palgrave Macmillan, 2008), 28-51, "Safety-Based Epistemology: Whither Now?" *Journal of Philosophical Research* 34 (2009) 33-45, and *Knowledge* (London: Palgrave Macmillan, 2009).

[2] See Robert Nozick, *Philosophical Explanations* (Cambridge: Harvard University Press, 1981).

© LOGOS & EPISTEME, VIII, 4 (2017): 517-520

core idea of the safety condition well.

One main issue involved in the debate about safety is determining what worlds count as close worlds and there is considerable controversy both about how to parse closeness and whether particular accounts of the factors involved in judging closeness are intuitively supported. For the purposes of this paper this does not, however, matter. Whatever turn out to be the correct factors involved in judgments of closeness it should be clear that any such account of closeness must be reflexive, that is to say $w_i \in <w_i>$. This is because, whatever the details involved, closeness is a similarity relation and every world is *maximally* similar to itself.

In any case, according to those who defend this condition on knowledge, safety is supposed to have independent merit as an intuitively plausible condition on knowledge. But, it would be advantageous to have a substantial argument in favor of this condition rather than having to depend on such weak and merely intuitive support for the principle and/or in light of conflicting and accounts of the closeness relation. The purpose of this paper is to provide such an argument and it is based on Kripke's recognition that safety and factivity are intimately related. Kripke made the relevant observation that is crucial to this argument in a 1986 talk in reference to Nozik's account of knowledge. In short, the argument presented here in support of safety involves the Kripke-inspired recognition that denying safety entails denying the factivity (or veridicality) condition of knowledge. It proceeds then by showing that since we should not deny factivity, we should endorse safety. Let us then look at Nozik's analysis of knowledge.

Nozik introduced the following account of knowledge as a particular form of epistemological reliabilism. *A* knowns that *p*, if and only if,

(1) *p* is true.

(2) *A* believes that *p*.

(3) If *p* weren't true, *A* wouldn't believe that *p*.

(4) If *p* were true, *A* would believe that *p*.[3]

(3) is, of course, Nozik's version of the safety condition. But, Kripke has pointed out that (2) and (3) jointly entail (1), in addition to pointing out a variety of other problems plaguing Nozik's analysis.[4] This point about the relationship between (1), (2) and (3) is particularly interesting because Kripke's observation can be

[3] Nozick, *Philosophical Explanations.*

[4] Saul Kripke, "Nozick on Knowledge," in *Saul Kripke: Collected Papers vol. 1* (Oxford: Oxford University Press, 2011), 162-224.

leveraged into a substantive argument for the safety condition on knowledge. This can be accomplished chiefly by considering what the denial of safety involves.

So what does denying safety entail? Denying safety entails this:

(Unsafe Knowledge) $(w_i \vDash K_A p)$ & $[<w_i> \vDash (B_A p \,\&\, \neg p)]$.

Knowing *p* at a given world is compatible with falsely believing *p* in worlds close to that given world. What then is the problem with respect to factivity? In order to see the problem we must have a clearer understanding of factivity in hand. The factivity condition on knowledge can be simply and informally understood as follows:

If *A* knows that *p*, then *p* is true.

As it is typically understood in epistemic logic, the factivity condition can then be parsed quasi-formally as follows:

(Factivity) $(w_i \vDash K_A p) \rightarrow [(w_i \vDash p)$ & $(w_j \vDash p$, for all w_j that are accessible from $w_i)]$.

To see the important implications of factivity consider the following basic model theory for standard epistemic logic. Let W be a set of worlds such that each $w_i \in W$, and R be the relation of epistemic possibility relating worlds. $<W, R>$ is then a frame in the usual sense and propositions will be subsets of W such that *p* is true in w_i if and only if $w_i \in p$. Let $R(w_i)$ be defined as follows: $R(w_i) = \{x \in W: R\, w_i\, x\}$. *p* is known at w_i then if and only if *p* follows from $R(w_i)$. In other words *p* is known at w_i if and only if *p* is true in all worlds that are epistemically accessible from, or are epistemic alternatives to, w_i. A world w_i is an *epistemic alternative* to world w_j for *A* just in case the accessibility relation holds between w_i and w_j. A bit more formally, factivity is the following condition on knowledge:

(Factivity) $(w_i \vDash K_A p) \rightarrow R(w_i) \subseteq p$.

Factivity holds in all frames in which the accessibility relation is reflexive, that is to say that factivity is an axiom of epistemic logic just in case w_i is accessible from itself. This is the case for all systems of epistemic logic at least as strong as the system KTD.

The issue then is that it should be clear that if one simultaneously accepts factivity and unsafe knowledge then one is committed to contradiction. This will be the case if there is at least one world where *p* is false that is close to a given world where *p* is known that is also an epistemic alternative to that world, and there is *always* at least one such world.[5] Consider a given proposition *p* known at

[5] There will actually be many such worlds.

w1 and the definition of unsafe knowledge. Since the notion of closeness involved in the safety condition is reflexive, if *p* is known at w1, then it can be the case that *p* is false at w1. Why? This is simply because unsafe knowledge permits an agent to have knowledge of a proposition in a given world w1 even when the agent falsely believes the proposition in worlds that are close to w1. But, since closeness is reflexive, w1 is itself one of those close worlds. So, unsafe knowledge permits an agent to know in w1 even when the agent falsely believes the proposition in question in w1. However, by factivity and the reflexivity of the epistemically access relation, if *p* is known at w1 it also follows that *p* is true at w1, since w1 is a member of the set of worlds that are epistemically accessible from w1. So, jointly endorsing unsafe knowledge and factivity leads to contradictions and one must go. But, since factivity is such a deeply entrenched and orthodox condition on knowledge and its denial invites all sorts of Morrean-like worries about false knowledge claims of the form "I know that *p*, but ¬*p*", we should simply treat Kripke's observation about Nozik's conditions (1), (2) and (3) as a reductio of the denial of safety and thereby as a substantive argument in favor of safety. In other words, since such Moorean "knowledge" claims clearly involve contradictions and are infellicitous we should maintain factivity and reject the denial of safety. What Kripke;s recognition allows us to see then is that arguments that support factivity are, ipso facto, arguments that support safety.

NOTES ON THE CONTRIBUTORS

Peter Baumann is Professor of Philosophy at Swarthmore College. His main areas of specialization are epistemology, philosophy of mind and questions concerning practical reasoning. He has published numerous articles in these areas and also a few monographs. Contact: pbauman1@swarthmore.edu.

David Coss is an Instructor of Philosophy at the University of Indiana Kokomo. His primary area of research is epistemology, where he currently focuses on issues surrounding contextualism and interest-relative invariantism (IRI). Both of his papers "The Pitfalls of Interest-Relative Invariantism" and "Interest-Relative Invariantism and Indifference Problems" are either published or forthcoming in *Acta Analytica*. He is particularly interested in the proper conceptualization of practical interests and epistemic contexts within the framework of contextualism and IRI. He has additional research interests in social and feminist epistemology (esp. epistemic injustice), as well as early modern philosophy and ethics. Contact: davidcossphilosopher@gmail.com.

Patrick Grim is Distinguished Teaching Professor Emeritus in Philosophy at Stony Brook and Philosopher in Residence with the Center for Study of Complex Systems at the University of Michigan. **Nicholas Rescher** is Distinguished University Professor of Philosophy at the University of Pittsburgh. Both have published extensively and across several disciplines, though Rescher's output puts just about anyone else to shame. Grim and Rescher's collaborative work has appeared most prominently in *Beyond Sets: A Venture in Collection-Theoretic Revisionism and Reflexivity: From Paradox to Consciousness*, both of which bear on the topic of the paper offered here. Contact: patrick.grim@stonybrook.edu; rescher@pitt.edu.

Richard Pettigrew is a Professor of Philosophy at the University of Bristol, UK. His research interests include: rational choice theory, formal epistemology, and the philosophy of mathematics. He has published numerous articles in these areas in journals such as *Philosophical Review*, *Philosophy and Phenomenological Research*, *Noûs*, *Philosophy of Science*, *British Journal for the Philosophy of Science*, *Review of Symbolic Logic*, *Synthese*, *Philosophia Mathematica*, and *Episteme*. His first book, *Accuracy and the Laws of Credence*, was published by

Oxford University Press in 2016. His second book, *Choosing for Changing Selves*, is under contract with Oxford University Press, and due to appear in 2019 or 2020. Contact: Richard.Pettigrew@bristol.ac.uk.

Maura Priest is a professional philosopher and bioethicist doing her best to live a good life. Currently she is an employed as an assistant professor of philosophy at Radford University. Her education includes a PhD from the University of California, Irvine, (philosophy) and a Master of Science degree from Columbia University (bioethics). Her research is normative and focuses on how we ought to behave, think, and feel, both as individuals and collectives. Her published papers have a wide scope, and include work in epistemology, ethics (including applied), political theory, and collective intentionality. Her current projects include work on the Gettier problem, epistemic altruism, hurt feelings, psychiatric and recreational drug use, and public health ethics surrounding obesity and medical autonomy. Contact: mp3588@columbia.edu.

Michael J. Shaffer is professor of philosophy at St. Cloud State University. His main areas of research interest are in epistemology, logic and the philosophy of science, and he has published many articles 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: mjshaffer@stcloudstate.edu.

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.

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.

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 take no more than six months. Authors not receiving any answer during the mentioned period are kindly asked to get in contact with the editors. Processing of papers in languages other than English may take longer.

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 şi Sociale „Gh. Zane" Academia Română, Filiala Iaşi, Str. Teodor Codrescu, Nr. 2, 700481, Iaşi, 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.