12-2022

Content Externalism and Self-Knowledge

Donnie Gene Barnett Jr.
University of Tennessee, Knoxville, dbarne23@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Donnie Gene Barnett Jr. entitled "Content Externalism and Self-Knowledge." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Philosophy.

Eldon Coffman, Major Professor

We have read this dissertation and recommend its acceptance:

Jon Garthoff, Georgi Gardiner, Garriy Shteynberg

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

CONTENT EXTERNALISM AND SELF-KNOWLEDGE

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Donnie Gene Barnett, Jr.
December 2022

For JLT,
who, among many other things,
mercifully resisted ever asking me to *finish writing already!*

Acknowledgments

I would like to thank everyone who helped and supported me throughout my grad school years.

Thank you to the many faculty members at UT who have at one point or another gone out of their way to help me succeed. Thanks in particular to the members of my dissertation committee: EJ Coffman, Georgi Gardiner, Jon Garthoff, and Garriy Shteynberg. Special thanks are due also to Nora Berenstain, John Nolt, Clerk Shaw, and Mariam Thalos. I would not have made it very far without your encouragement and generosity.

Thanks to Joseph Dartez, Michael Ebling, Paige Greene, Linh Mac, and Jacob Smith for valuable and interesting discussions on drafts of my dissertation chapters. Thanks also to Alex Richardson, who tried his darndest to break me out of my shell. Those efforts are much appreciated.

Finally, thank you to Jennie Tan for her unwavering patience and support even when this project seemed like it would go on forever.

Abstract

There appears to be a tension between two widely held philosophical theses: content externalism and what is often called "privileged access". The first is the metaphysical thesis that the contents of many propositional attitude-types are at least partially determined by properties external to the thinking subject. The second is the epistemological thesis that we have *a priori* access to the contents of our own propositional attitudes. Those who hold that at least one of these theses must be false are called incompatibilists. My goal is to show that the incompatibilists are wrong, that content externalism and privileged access can both be true.

In Chapter 1, I briefly introduce content externalism and review the source of the alleged tension between the latter and privileged access. In Chapter 2, I address the so-called "discrimination argument" for incompatibilism. This argument appeals to the fact that, if content externalism is true, then we will not always be able to discriminate one thought-type from another. This generates problems for privileged access if we think that knowledge requires the ability to discriminate between relevant alternatives. I argue, however, that knowledge does not require such an ability. In Chapter 3, I address Jessica Brown's "illusion argument" for incompatibilism. While this argument might show that *singular* externalism is incompatible with privileged access, I argue that it does not generalize to other forms of content externalism. In Chapter 4 I evaluate Boghossian's "memory argument". First, I draw on existing criticism to show that the original 1989 version of the argument fails because it relies on false premises about memory. I then consider and reject the possibility, originally proposed by Sanford Goldberg, that the argument can be reconstructed without these false premises. Finally, in Chapter 5, I discuss and evaluate McKinsey's reductio. First, I argue that the

externalist cannot be expected to accept the closure principle on which McKinsey relies. Second, I argue that though the compatibilist may be committed to the apriority of certain environmental propositions, these propositions are modest enough that it is not obviously absurd to suppose that they might be *a priori*.

Table of Contents

Chapter 1: Introduction

1.1 Project Outline

There appears to be a tension between two widely held philosophical theses. The first is the metaphysical thesis that the contents of many propositional attitude-types are at least partially determined by properties external to the thinking subject. This is the thesis referred to herein as "content externalism" (or simply "externalism"). The second is the epistemological thesis that we have introspective access to the contents of our own propositional attitudes. This is the thesis commonly referred to as "privileged access". My goal is to show that this tension is illusory, that content externalism and privileged access are in fact compatible.

Content externalism is primarily motivated by the now famous Twin Earth thought experiments (Putnam 1975, Burge 2007) and by Burge's work on the relevance of linguistic community to concept acquisition (Burge 1979).[1] The Twin Earth argument for externalism usually goes as follows: Twin Earth is exactly like Earth except that Twin Earth's oceans, lakes, and rivers are not filled with H_2O, but with the qualitatively identical liquid XYZ. Oscar, a resident of Earth, believes that water is wet. Intuitively, however, Oscar's physically identical counterpart on Twin Earth, Toscar, doesn't believe anything about water. Toscar's propositional attitudes are instead about XYZ. When prompted, he will say things like "I think water is wet" just like Oscar does. Coming from Toscar, however, the utterance expresses a thought not about H_2O, but about XYZ. Thus Toscar's propositional attitude content is not *that water is wet*, but *that* (say) *twater is wet*. Since all other variables are held constant, we

[1] See also Majors & Sawyer 2005.

conclude that Toscar's propositional attitudes are different precisely because his environment is different.

The argument generalizes. We can run the thought experiment using other natural kinds and get the same result. And Burge (1979, p. 77-79) employs similar thought experiments in which counterfactual linguistic environments are used to show that attitudes involving social concepts, like *arthritis*, similarly depend on external factors. For example, suppose a member of an English-speaking linguistic community has a belief that he would express by uttering, "I have arthritis in my thigh". According to Burge, we would interpret this person as expressing a false belief about arthritis. (Arthritis is an aliment of the joints. One cannot have arthritis in one's thigh.) But now imagine a counterfactual scenario which differs *only* with respect to the semantic norms of the man's linguistic community. In the counterfactual scenario, his linguistic community uses "arthritis" to refer not only to arthritis, but to various other rheumatoid ailments as well – including the condition the man has in his thigh. In this case, the man's utterance would express a *true* belief, not about arthritis, but about a condition we might refer to as *tharthritis*. This shows that intentional content is partly determined by the semantic norms of the linguistic community to which one belongs and in the context of which one acquired one's concepts.

Because of examples like these, it is now widely accepted that content externalism is true.[2]

But if externalism is right, then we are faced with a puzzle. If what I am thinking depends on what my environment is like, then how can I know what I am thinking without first

[2] Brian Loar (1988, 2003) is a notable holdout.

knowing something about my environment? This question is an expression of what might be called *the incompatibilist intuition*. Broadly speaking, this intuition is usually developed in one of two ways.

First, it is pointed out that one consequence of externalism is that we will not always be able to discriminate one thought-type from another.[3] An analogy is helpful here. If externalism is true, then thoughts are like sunburns (to use Davidson's 1987 analogy). What makes a sunburn a *sun*burn is its causal history, particularly the fact that it was caused by the sun. But this is not a property intrinsic to the skin condition itself. It is conceivable that a skin condition with the same intrinsic properties be caused instead by some chemical agent. Call this a *chem*burn. Because chem- and sunburns have the same intrinsic properties, they would be indistinguishable. If, then, chemburns were frequent enough to constitute relevant alternatives to sunburns (imagine e.g. a situation in which a chemical spill has made chemburns a common occurrence), one would not be in a position to know of a given sunburn-esque rash whether it was a sunburn or a chemburn. To know which, one would have to know something about its causal history.[4]

The idea is that it follows from the truth of externalism that the epistemic situation with respect to thought contents is like that of the sunburn case just described. If externalism is true, then a particular thought is e.g. a *water*-thought (and not a *twater*-thought) partly in virtue of its causal history – for example, its being causally connected to contact with instances

[3] It is almost universally accepted that this is a consequence of externalism. For discussion see Owens (1992), Boghossian (1994), Falvey and Owens (1994), Brown (2004), and Schroeter (2007).

[4] To figure it out, for example, one would have to ask questions like "Have you been swimming in that lake near the site of the chemical spill?" or "Did you forget to wear sunscreen when you went to that pool party earlier?"

of water or with members of a linguistic community that possesses the concept *water*. But it is conceivable that a thought with the same intrinsic (i.e. non-intentionally described physical) properties have a *different* causal history. It might, like Toscar's "water" thoughts, be causally connected instead to contact with instances of XYZ. In that case, the thought in question would not be a *water*-thought, but a *twater*-thought. This is precisely what the Twin Earth thought experiment is designed to illustrate.

The incompatibilist points out that these two thought-types would be indistinguishable from one another. Notice, for example, that Toscar's (non-intentionally described) inner-life would be exactly the same as Oscar's. Consequently, the argument goes, given a situation in which the possibility that I am thinking some twin thought is a relevant alternative, I may not be in a position to know immediately what I am thinking. As in the sunburn case, it looks like I would have to learn something about my thought's causal history in order to know e.g. whether it is a *water*- or *twater*-thought. Since this is not something I could do by introspecting, I would not be in a position to know introspectively what I am thinking.

What has come to be known as the "discrimination argument" is an argumentative strategy that exploits the fact that there *do* appear to be situations in which some twin thought would constitute a relevant alternative.[5] In fact, there is good reason to believe that such situations are not merely hypothetical but occur regularly in the actual world (Ludlow 1995a). They will occur whenever there is an instance of so-called "slow-switching" (Burge 1988, Boghossian 1989). Slow-switching occurs when one is unknowingly integrated into a new

[5] See Boghossian (1989) for the original version of this argument. Brown (2004) provides a developed and in-depth discussion of how the logic of the argument is supposed to work.

linguistic community with the result that there is an undetectable shift in one's conceptual repertoire. This would happen if, for example, Oscar were transported to Twin Earth without his knowledge and resumed his life as if nothing had changed. Eventually, after frequent enough interaction with XYZ and with a linguistic community for whom "water" refers to XYZ, Oscar's own tokens of the word "water" would begin to express attitudes about XYZ. Thus, for those who are slow-switched, thoughts and expressions of the same syntactic[6] type will come to have different contents at different times. This state of affairs opens up the possibility for relevant alternatives.

The second way of developing the incompatibilist intuition is to draw out the implications of the conceptual relationship that externalists posit exists between content and environment. Michael McKinsey (1991, 2002, 2007) takes this route, arguing that if externalism is true, then *a priori* knowledge of one's thoughts entails *a priori* knowledge of one's environment. This, he argues, follows as long as we assume the following closure principle.

> *Closure of Apriority under Conceptual Implication (CA)*
> Necessarily, for any person x, and any propositions P and Q, if x can know a priori that P, and P conceptually implies Q, then x can know a priori that Q (McKinsey 2002, p. 207; 2007, p. 55).

According to the externalist, my thinking a *water*-thought conceptually implies E, where E is some proposition about my environment. Thus, given CA, it follows that if I can know *a priori* that I am thinking that e.g. water is wet, then I can know *a priori* that E. But, McKinsey argues, the notion that one could know E (*whatever* E ends up being) *a priori* is absurd. Therefore, he

[6] For ease of exposition, I assume for a moment that thoughts have syntactic structure. Nothing hinges on this.

reasons, externalism forces us to give up on the idea of privileged, *a priori* access to the contents of our thoughts.

In broad outline, Chapters 2-4 of my dissertation address the first way of developing the incompatibilist intuition: slow-switching arguments. Chapter 5 addresses McKinsey's reductio. In the next section, I provide a chapter-by-chapter outline of how my argument will go. But first, a few notes on methodology.

The term "privileged access" has been used to mean different things in different places. The particular conception of privileged access I am concerned to defend here is one that I will call *trans-world privilege* (or TWP).

(**TWP**) Necessarily, if S is thinking that p and her faculty of introspection is functioning properly, then S can know introspectively that she is thinking that p.[7]

There are two reasons for this. First, TWP (or something very much like it) is often the operative conception of privileged access in the literature with which I am interacting (e.g. McLaughlin & Tye 1998a & 1998b, McKinsey 2002 & 2007, Goldberg 2000 & 2003a, Morvarid 2013). Indeed, even when it is not explicitly stated, something like TWP is often presupposed. For example, it is clearly presupposed by the logic of the discrimination argument, which I discuss in Chapter 2. Second, TWP is modest enough that it would be genuinely surprising if it turned out to be false. I follow Sawyer (2002) in drawing a distinction between "is thinking" and "thinks". *Thinking* is an occurrent, conscious activity; "thinks", on the other hands,

[7] This is distinct from what we might call *actual world privilege* (which would just be TWP minus the necessity operator), which is closer to what e.g. Warfield (1992, 1997) defends as being compatible with externalism.

suggests something like "believes", a state that need not be either occurrent or conscious.[8] Thus, the domain of TWP extends only to one's *conscious, occurrent propositional attitudes*. It does not say, for example, that one necessarily has introspective access to the contents of one's beliefs or to any other of one's standing or dispositional states.

Finally, I do not presuppose any specific account of introspection. By "introspection", I simply mean that mechanism (whatever precisely it consists in) by which we come to know our own thoughts and that we cannot use to discover others' thoughts. I do assume throughout that the self-knowledge acquired via this mechanism is *a priori*.[9] But save for the chapter on McKinsey's reductio (where what is at stake is precisely the limits of *a priori* knowledge), this assumption doesn't materially affect anything. And even in that context, the assumption is concessive. (If it turns out that introspective self-knowledge is not *a priori*, then we cannot use CA to get the allegedly absurd conclusion that E can be known *a priori*.)

1.2 Chapter Summaries

In Chapter 2, I directly address the discrimination argument. The standard strategy for dealing with it is to emphasize the reliability of second-order judgments (that is, judgments about what one is thinking). It is pointed out that first- and second-order thoughts are individuated by the same environmental factors (Davidson 1987, Burge 1988, Heil 1988, Stalnaker 1991/1999, Falvey & Owens 1994, Forbes 1995, Gibbons 1996, Peacocke 1996).

[8] An occurrent state is one that is in some sense active. An occurrent state need not be conscious. For example, a belief (think e.g. of a racist belief) can be active in the sense that it is influencing one's behavior while at the same time remaining unconscious in the sense that one is not aware of having it. See Bartlett 2018.

[9] Thus, for example, I will say things like "TWP entails that S can have *a priori* knowledge that she is thinking that p".

Therefore, one cannot misidentify a first-order thought in the same way that one might misidentify a sunburn. But, Brown (2004) and others (McLaughlin & Tye 1998a, Vahid 2003) have argued that this is not to the point. For it simply assumes away the basic contention underlying the discrimination argument: the idea that knowledge requires the ability the discriminate between relevant alternatives. My primary goal in this chapter is to address this problem by showing that such an ability is *not* necessary for knowledge. To do this, I build on Goldberg's (2006) work showing that a discrimination requirement would conflict with other plausible epistemic principles.

Chapter 3 takes on Jessica Brown's (2004) "illusion argument". This argument focuses on the epistemic consequences of no-reference cases, such as when one thinks that *phlogiston is interesting*. Some externalists would say that since "phlogiston" does not refer, such a thought has no representational content and is therefore not really a *thought* at all (see e.g. Boghossian 1998).[10] Nonetheless, it will appear to the thinker to be a genuine thought. Thus, a thinker in these circumstances will be disposed to form many false second-order beliefs to the effect that she is thinking a thought. To Brown, this indicates that the belief-formation processes underlying the second-order judgments of those who are (or frequently find themselves) in no-reference situations are unreliable. This, she argues, will result in failures of self-knowledge that are inconsistent with privileged access. I grant that the argument is compelling if interpreted as an argument for the claim that *singular* externalism (i.e. the view

[10] It would still constitute a mental state. The idea is just that since it has no representational content, it is not a *thought* per se. One might, following Morvarid (2013), call it a "pseudo-thought".

that singular thought is object-dependent) is incompatible with privileged access; but I argue that the argument does not apply to content externalism more generally.[11]

In Chapter 4 I address what is known as the "memory argument". In one of his seminal papers on the subject, Tyler Burge claims that a person who learns of having been slow-switched may ask, "'Was I thinking yesterday about water or twater?' – and not know the answer" (Burge 1988, p. 659). This suggests that, given externalism, a subject may lack access to the contents of yesterday's propositional attitudes. But, if that is correct, then it looks like a subject may lack access to *today's* [read: conscious, occurrent] propositional attitudes as well. As Boghossian puts it: "It is not as if thoughts with widely individuated contents might be easily known but difficult to remember. The only explanation […] for why S will not know tomorrow what he is said to know today, is not that he has forgotten but that he never knew" (Boghossian 1989, p. 23). I reject this argument, making my case in two stages. First, I argue that Boghossian's original argument invariably relies on at least one false premise about memory. Second, I consider the possibility that these premises are incidental to the argument and that it can be reconstructed without them. This idea was originally proposed by Sanford Goldberg (1997, 2003a). He has shown that a subject may not know of a thought she is *currently entertaining* whether it is about water or twater (1997, 2003a, 2003b). Conceding the possibility, I consider and reject two reasons to think that one must be able to identify one's thought contents in this way in order to know what one is thinking.

[11] Of course, any version of the illusion argument presupposes that the illusion interpretation of no-reference cases is correct. That presupposition may well be wrong (see Gerken (2007) for good reasons to think that it is). But since I think that the illusion argument can be defeated *even if* the illusion interpretation is correct, I will not attempt to refute it in my dissertation.

Finally, in Chapter 5, I address McKinsey's reductio (outlined near the end of the previous section). First, I show that if externalism is true, then there are counterexamples to CA. This means that the externalist cannot be expected to accept it. After that, I evaluate the claim that *a priori* knowledge that E is absurd. Whether it is seems to me to depend on what exactly E is. I show that E will be a relatively modest proposition – modest enough that it is not implausible that it can be known *a priori*. Finally, I address the worry that the conjunction of externalism and privileged access absurdly implies that it is possible to know *a priori* that one is not a brain in a vat in an otherwise empty world. Again, my strategy here is to simply accept that one can know this *a priori*. Indeed, I think it is possible to deduce it from *a priori* premises about one's thought contents. This is sometimes called "McKinsey-style" reasoning (e.g. Pryor 2007). McLaughlin (2003) has argued that McKinsey-style reasoning is necessarily question-begging. I structure my own argument around McLaughlin's objections, using them as a vehicle for launching a positive case to think that McKinsey-style reasoning can be perfectly cogent.

Chapter 2: The Discrimination Argument

In this chapter we examine the discrimination argument for incompatibilism. The argument, in a nutshell, is this: Given externalism, it is possible to find oneself in a situation where one cannot know by introspection that one is thinking that p because one cannot discriminate between the actual situation and a relevant alternative situation in which one is thinking that q. If this is right, it follows that externalism is incompatible with privileged access, as articulated by TWP (see section 1.1 of the Introduction). Various approaches to solving this problem have been proposed. What Butler (1997) has dubbed "the standard strategy" for dealing with it is to deny that knowledge requires the ability to discriminate among relevant alternatives. What is required instead is a kind of *reliability*, and externalism, the argument goes, does nothing to threaten the reliability of self-ascriptive judgments.

But, Brown (2004) and others (McLaughlin & Tye 1998a, Vahid 2003) have argued that this does not address the problem. For the compatibilist cannot simply *deny* that the ability to discriminate is required for knowledge. What is needed is an independent reason for rejecting a discrimination requirement – and this, says Brown, has not been offered. My aim in this paper is to defend the standard strategy by accepting Brown's challenge – that is, by finding an independent reason for denying that knowledge requires the ability to discriminate between relevant alternatives. Building on Goldberg's (2006) work, I accomplish this by showing that the discrimination requirement conflicts with other plausible epistemic principles.

The chapter is structured as follows: In the next section, I outline the discrimination argument itself. Next, I turn to the standard strategy, explaining how compatibilists have

typically responded to the discrimination argument. I then take a closer look at Brown's objections and at the difficulties facing the compatibilist in light of these objections. Finally, I address these objections, showing that there is a way to reject the discrimination requirement and vindicate the standard strategy.

2.1 Slow-Switching

The discrimination argument stems from Boghossian's 1989 "slow-switching" thought experiment. Boghossian asks us to imagine that, unbeknownst to him, Oscar is repeatedly moved back and forth between Earth and Twin Earth.[12] The externalist intuition is that if he is left on Twin Earth to interact with XYZ long enough, his water-thoughts[13] will eventually come to be about XYZ. Suppose he is left long enough for this to occur. At this point, he is brought back to Earth and allowed to stay long enough for his water-thoughts to once again become H_2O thoughts. While still on Earth, Oscar thinks a thought that he would express by saying "water is wet". The question is this: Is Oscar in a position to know introspectively what he is thinking?

Boghossian argues that he is not. His reason is that the possibility that Oscar is thinking a *twater*-thought is now a relevant alternative. He draws on an epistemology developed by Alvin Goldman (1976) to make his point.[14] Goldman argues that in order for S's perceptual

[12] Suppose he and his twin, Toscar, take one another's place each time.
[13] By this I mean the thoughts that Oscar expresses using the term "water". In addition, I adopt the following conventions. I will use "water" to refer to water – that is to H_2O; I use ' "water" ' to refer to the linguistic sign that we use to refer to water (and that Twin Earthlings use to refer to XYZ); and I use "*water*" to refer to the concept of water. Thus, "*water*-thought" refers to a thought that involves the concept *water*. Similarly, "*twater*-thought" refers to a thoughts that involves the concept *twater*.
[14] See also Dretske 1970 and McGinn 1984.

belief that p to count as knowledge, S must be in a position to rule out all relevant alternative possibilities q. And this, Goldman argues, requires that S be able to *discriminate* between the situation in which p holds and that in which q holds. His Fake Barn Country is meant to demonstrate this. Imagine that Henry is driving through the countryside. The surrounding land is populated by what appear (from the road) to be barns. But unbeknownst to Henry, a good deal of them are really just façades – they're not actually barns at all. Looking at one that happens to be real, Henry forms the (true) belief that he is looking at a barn. Does he *know* that he is looking at a barn? Intuitively, the answer is "no". The reason we do not want to ascribe knowledge in this case, Goldman argues, is because Henry cannot discriminate between real and fake barns. Thus, he is not in a position to rule out the relevant alternative possibility that he is actually looking at a barn façade.

Normally, of course, one can know, just by looking, that one is driving by a barn. One does not need to rule out the possibility that what one is seeing is actually a very realistic fake. For unless one is driving through Fake Barn Country, that is not normally a *relevant* alternative. But, for Henry, it is. The fact that there are façades present in the countryside through which he is driving makes it relevant. For if he were not looking at a barn, *he could easily be looking at a façade*.[15]

According to Boghossian, this is precisely Oscar's situation. The fact that he is being slow-switched makes the possibility that he is now thinking a *twater*-thought a relevant

[15] Here I'm presupposing something like Nozick's (1981, p. 175) account of when an alternative is relevant. There is, however, disagreement about this. For other accounts see Goldman (1976, p. 775-777), Dretske (1981, p. 373-378), Cohen (1988), Luper (2006, p. 380), and Dierig (2010a & 2018). This disagreement is not important for present purposes. What is important is *that* the possibility that Henry is looking at a façade is a relevant alternative, not *why* it is relevant.

alternative. Because it is a relevant alternative, Oscar must be able to rule it out before he can know what he is thinking. But, since he is unable to discriminate between *water*- and *twater*-thoughts, he is not in a position to do that (at least, not introspectively). Thus, he is not in a position to know what he is thinking (at least, not introspectively). The upshot is that self-knowledge is subject to environmental contingencies. It is possible to find oneself in a situation where one cannot know what one is thinking without first learning something about one's environment – something that cannot be learned introspectively. Therefore, if externalism is true, then TWP is false.

As I've just outlined it, this argument relies on the claim that there is a necessary condition on knowledge that can be articulated by the following discriminability requirement.

(**DR**) Where q is a relevant alternative to p, S can know that p on the basis of w *only if* w allows S to discriminate between the actual situation in which p is true and the counterfactual situation in which q is true.[16]

Given DR, we can reconstruct the argument more precisely. From here on, when I refer to the "discrimination argument", I refer specifically to the argument that follows. Let "W" be the

[16] In much of the literature (e.g. Brown 2004, p. 37-41; Vahid 2003, p. 377-378; and especially Morvarid 2012, p. 28), the discriminability requirement is articulated (more or less) as follows:

(**DR'**) Where q is a relevant alternative to p, S knows that p *only if* S can discriminate between the actual situation in which p is true and the counterfactual situation in which q is true.

However, DR' is clearly false unless interpreted as saying the same as DR. Suppose, for example, that Henry and Bob are driving through the countryside together. The surrounding land is peppered with red and green barns. Let R be the proposition that they are now passing by a red barn and G the proposition that they are now passing by a green barn. Henry is color blind; Bob isn't. Henry correctly believes that R on the basis of Bob's testimony. If G were true, Bob would have said so, and Henry would have believed him. Intuitively, Henry *knows* that R despite the fact that, being colorblind, he cannot discriminate between red and green barns. This is because he bases his belief on Bob's testimony rather than on his own perceptions. This is a counterexample to DR' *unless* we interpret it as taking into account the *grounds* (w) of one's belief . Hence DR. DR makes clear the distinction between Henry's believing R on the basis of perception and his believing R on the basis of Bob's testimony.

proposition that Oscar is thinking that water is wet and "T" the proposition that he is thinking that twater is wet.[17]

(D1) **DR**

(D2) Oscar cannot introspectively discriminate between the actual situation in which W is true and the counterfactual situation in which T is true.

(D3) T is a relevant alternative to W.

(D4) Oscar can't know introspectively that W. (D1,D2,D3)

(D5) ~**TWP** (D4)

Many have responded to this argument by pushing back on the idea that knowledge requires the ability to discriminate *per se*. The primary advocates of this response are Falvey and Owens (1994). The idea here is that what is important is a kind of reliability. The ability to discriminate is important only when it is necessary for satisfying the relevant reliability requirement, as in the case of perception. But it is not necessary when it comes to self-knowledge. On this view, Boghossian's mistake is to assume that perception and introspection are relevantly analogous and hence that Goldman's lessons apply to the latter as well as to the former (cf. Wikforss 2008). This strategy involves denying DR.

In the next section, we take a closer look at this response – the so-called "standard strategy". Then, in section 2.3, we look at the problems facing it.

2.2 The Standard Strategy

In general, advocates of the standard strategy attempt to establish two things. First, they argue that (i) Goldman's thought experiments show only that knowledge requires a kind of

[17] Note that the following argument assumes that Oscar's faculty of introspection is functioning properly.

reliability, not the ability to discriminate *per se*. And, second, they try to show that (ii) externalism poses no special threat to the reliability of self-ascriptive judgments. Having established (i) and (ii), they conclude that slow-switching cases pose no special threat to privileged access.

Let's begin this section by looking at the arguments given in defense of (i). Is the ability to discriminate required for knowledge? Both Falvey & Owens (1994) and Gibbons (1996) argue persuasively that Goldman's barn case does not establish that it is.

Gibbons begins by pointing out that the language of "ruling out" is misleading. It suggests that "in order to know that p in a relevant alternative situation, you need to go through a certain process of reasoning by which you rule out the relevant alternative q" (1996, p. 296). This way of conceiving things, he thinks, obscures the fact that what's important in the fake barn case is the truth of certain counterfactuals.

Before we can say whether Henry knows he is seeing a barn, we need to know whether he would still believe he were seeing a barn in the counterfactual situation in which he is actually seeing a barn façade. For example, Henry's belief will *not* count as knowledge if the following counterfactual is true:

(**P**) If Henry had been looking at a barn façade, he would have falsely believed that it was a barn.

But now suppose that we change the case by stipulating that the barn façades peppered throughout Fake Barn Country are not very realistic and that, as a result, the following is true:

(**C**) If Henry had been looking at a barn façade, he would have correctly believed that it was a barn façade.

Given that (C) is true and (P) is false, Gibbons argues, we ought to say that Henry knows that he sees a barn. And he knows this without having to do anything that might be described as "ruling out" the possibility that he is looking at barn façade. It is necessary only that (C) be true and (P) be false – Henry need not reason his way to or believe anything about either counterfactual.

Falvey & Owens take a similar approach, arguing that Goldman's barn case demonstrates that something like the following principle[18] is true.

(**RA'**) If (i) q is a relevant alternative to p, and

(ii) S's justification for his belief that p is such that, if q were true, then S would still believe that p, then

S does not know that p.

Like Gibbons, this principle tells us that what's important is what S would believe in certain counterfactual scenarios – those in which a relevant alternative is true. According to (RA'), Henry fails to know that he sees a barn because in the relevant alternative situation in which he is seeing a façade, his grounds for belief (i.e. his perceptual experience *as of* a barn) would still lead him to believe that he was seeing a barn.

(RA') specifies a necessary condition for knowledge. It says that where q is a relevant alternative to p, S knows that p only if *it is false that if q were true, S would still believe that p*. Let's call this *the reliability requirement* for short. This requirement can be specified more precisely as follows.

[18] I follow Falvey & Owens in naming this principle " RA' ".

(**RR**) Where q is a relevant alternative to p, S knows that p on the basis of w *only if* ~[w is such that were q true, S would still believe that p].

According to Falvey & Owens, the idea that knowledge requires the ability to discriminate between relevant alternatives gets whatever plausibility it has from RR.[19] For satisfying the reliability requirement (i.e. the condition specified in the consequent of RR) will often require the ability to discriminate. For example, it is plausible that (C) is true and (P) is false only when Henry is able to *discriminate* between real and fake barns. And (P) will be true as long as Henry is *not* able to discriminate between real and fake barns. Indeed, Falvey & Owens claim, as long as we restrict ourselves to talking about *perceptual* knowledge, it will plausibly always be the case that where S is unable to discriminate between the actual situation in which p and the counterfactual situation in which q, S's belief that p will fail to satisfy the reliability requirement. Hence, when restricted to perceptual knowledge, (RR) *entails* that S will fail to know as long as S lacks the ability to discriminate. But it is not the ability to discriminate *per se* that is important. That is important only to the extent that it is necessary for satisfying the reliability requirement.

It may be worth noting that Goldman himself seems to think that the ability to discriminate is important only because it is necessary for reliability. On the first page of his 1976, he asks, "What kinds of causal processes or mechanisms must be responsible for a belief if that belief is to count as knowledge?" He answers: "They must be […] 'reliable'" (Goldman 1976, p. 771). He goes on to say that in order to be reliable, "a cognitive mechanism must

[19] Falvey & Owens do not explicitly endorse RR, but it can be derived from RA'. The only change I've made was to replace their use of the term "justification" with "w" (short for "warrant") to make clearer the fact that what is being referred to are the *grounds* of S's belief.

enable a person to *discriminate* or *differentiate* between incompatible states of affairs" (Ibid., emphasis is Goldman's). It is only after making this connection between reliability and the ability to discriminate that Goldman emphasizes the necessity of the latter.

Goldman is dealing explicitly with perceptual knowledge (the title of his 1976 paper is "Discrimination and Perceptual Knowledge"). It is therefore not unreasonable for him to neglect the differences between reliability and the ability to discriminate. For it is probably the case that in the realm of perceptual knowledge, these two things go hand in hand. But, as many compatibilists have been quick to emphasize, they come apart in the realm of self-knowledge. Though Oscar may not be able to discriminate between his *water*- and *twater*-thoughts, his second-order beliefs regarding those thoughts may still be reliable.

This brings us to the second strand of the standard strategy: the claim that externalism poses no special threat to the reliability of self-ascriptive judgments. Much of the discussion here begins with Burge's idea of *basic self-knowledge* (1988). Burge points out that for some occurrent mental states *m*, judging that one is in *m* makes it true that one is in *m*. These judgments are self-verifying in much the same way that Cartesian *cogito*-judgments (like "I am now thinking") are self-verifying. In particular, any judgment that r, where r can be expressed by a proposition of the form

(r) I am thinking (with this very thought) that p[20]

is self-verifying in this way. Let's call judgments like these "basic judgments". Consider, for example, judgments of the kind expressed by claims like "I am thinking that water is wet". The

[20] If I understand Burge correctly, the purpose of adding "with the very thought" is to indicate that in cases of basic self-knowledge, one thinks that p *self-consciously*. That is, one thinks the thought *and* thinks of it as one's own in the same mental act. See Sawyer 2002 for a helpful discussion.

idea is that such judgments *contain* the very thought one ascribes to oneself in making the judgment. By judging that I am thinking about water I *thereby* think about water. In these cases, there can be no gap between what one *thinks* one is thinking and what one *is* thinking.

So, Oscar's basic judgments will satisfy the reliability requirement despite his history of slow-switching. For suppose that Oscar makes the basic judgment that he is thinking that water is wet. In this case, it is false that if he were thinking that twater is wet, he would still believe that he was thinking that water is wet. For the closest possible world in which Oscar is thinking that twater is wet is one in which he instead makes the basic judgment that he is thinking that twater is wet. So the reliability requirement is satisfied despite the fact that Oscar is in no position to distinguish between his *water*- and *twater*-thoughts.[21]

This shows that it is possible for a subject S's belief that p to satisfy the reliability requirement despite S's inability to discriminate between the actual situation in which p and the counterfactual situation in which q. But this alone will only get the compatibilist so far. The reason that S's basic judgment that r is reliable even where S cannot discriminate between relevant alternatives is because basic judgments are self-verifying. But not all self-ascriptive judgments *are* self-verifying (see Boghossian 1989, Peacocke 1996, Burge 2003, Dierig 2014). In particular, any self-ascriptive judgment that is not perfectly coincident with the attitude ascribed will not be self-verifying. This includes, for example, judgments about what one was *just* thinking. It also includes judgments about standing states or about perceptual beliefs. For

[21] It follows that *if* basic judgments count as knowledge, then they constitute counterexamples to the claim that the ability to discriminate between relevant alternatives is required for knowledge. We consider this possibility in section 2.4, where I argue that this is the way out of the discrimination argument's challenge. Here, we are only trying to show that externalism poses no special threat to the reliability of self-ascriptive judgments. Whether basic judgments *do* count as knowledge is part of what's at stake here.

example, suppose I look outside and see a lot of tall pine trees. I then form the belief that pine trees are tall. The formation of this belief need not be conscious. I may not consciously engage it at all until e.g. someone asks me what I thought of the pine trees.[22] Only then, perhaps, will I make any second-order judgments about the content of this belief. What, then, of the epistemic status of judgments like these? Does externalism threaten the privileged status of *non*-basic self-ascriptive judgments?

The standard response is to point out that the contents of second-order thoughts are individuated by the same external factors that individuate first-order thoughts (Davidson 1987, Burge 1988 & 1996, Heil 1988, Stalnaker 1991/1999, Falvey & Owens 1994, Forbes 1995). As Burge puts it, "if background conditions are different enough so that I am thinking different thoughts, then they will be different enough so that the objects of […] self-ascription will also be different" (Burge 1996, p. 96; cf. Burge 1988, p. 659). And this is held to be true of *non*-basic judgments as well as basic judgments.

Apparently, then, the compatibilist can say something like the following: Any situation in which Oscar *thinks* he is thinking about water will be (barring some unusual cognitive malfunction) one in which he *is* thinking about water. For if he were thinking about motorcycles, weather vanes, or candy bars, he would (barring some unusual cognitive malfunction) not think he was thinking about water. And, if he were thinking some twin thought – one about twater, say – then his environment would be different enough to affect

[22] The idea is that at this point, the belief becomes occurrent and conscious and so falls within TWP's purview. But since the corresponding second-order judgment is not perfectly coincident with the formation of this belief, it does not seem to constitute a basic judgment. Essentially, what I am doing here is accounting for the possibility that not all second-order judgments regarding one's conscious, occurrent thoughts constitute basic judgments.

the contents of his second-order thoughts as well. So, if he were thinking about twater, he would not mistakenly think he was thinking about water. Hence, externalism poses no special reason to doubt the reliability of even non-basic self-ascriptive judgments.

This story makes sense in cases where no slow-switching is involved. A resident of counterfactual Twin Earth, for instance, will have beliefs about twater and will *believe* that those beliefs are about twater (assuming normal functioning). Certainly, she will not think that the beliefs in question are about *water*. She will not even have the concepts necessary to think this for the same reason that she is unable to think *first*-order thoughts about water – namely, because she has never been in a water environment and is not part of a linguistic community that possess the concept *water*.

But the situation gets murkier when slow-switching enters the equation. To see this, let's start by looking at Oscar's situation before he is slow-switched. Given his environmental circumstances, the thoughts he would express by saying things like "water is refreshing" are *water*-thoughts. And those same environmental circumstances ensure that the second-order thoughts he would express by saying things like "I believe that water is refreshing" are also *water*-thoughts. Oscar will not even possess the concept *twater*. Thus, if Oscar's non-basic self-ascriptive judgments fail to satisfy the reliability requirement, it is because he would still ascribe to himself a *water*-thought even in some relevant alternative scenario in which the relevant first-order thought is not about water but about, say, motorcycles. That is to say, before becoming a victim of slow-switching, Oscar's non-basic self-ascriptive judgments will fail to satisfy the reliability requirement only in the event of some unusual cognitive malfunction. The specter of twin thoughts poses no epistemic threat.

What about after he is slow-switched? Notice that in the story above, Oscar does not (yet) possess the concept *twater* and so cannot even think *twater*-thoughts. Hence, he cannot mistakenly ascribe to himself a *twater*-thought. (A parallel story can be told about Oscar's Twin Earth counterpart, Toscar.) After he is slow-switched, however, he does eventually acquire the concept *twater*. The question is: Does he *retain* the concept *water*, so that he is able to entertain both *water*- and *twater*-thoughts after the switch? If one concept replaces the other whenever he switches environments (so that he is no longer able to think about water after the switch to Twin Earth), then slow-switching will not pose any special epistemic problems. That is, there will be no epistemic difference between (a) always having been embedded in some environment E and (b) being embedded in E as a result of having been slow-switched. In either case, one's environment will determine both first- and second-order thought contents, ensuring that there will be no mismatch between contents self-ascribed and contents thought.

Some philosophers have taken this route (Ludlow 1998, Tye 1998). Indeed, standard strategists often appear to take it for granted. But most hold that one would in fact retain one's old concepts after having been slow-switched (Burge 1998, Boghossian 1989, Gibbons 1996, Goldberg 2005b). The implication of this is that Oscar would retain the ability to think about twater even after rejoining his old linguistic community back on Earth. I think this is correct, but will not defend it at length here.

What's important is that the possibility raises a problem for the prospects of the standard strategy as I've outlined it so far. This is simply because if Oscar is able to think both *water*- and *twater*-thoughts, then there is no obvious reason why it would be impossible for his

first- and second-order water-thoughts to have different content. Hence, no obvious reason why it would be impossible for him to mistakenly self-ascribe a *water*-thought.[23]

Some standard strategists, however, have added that second-order thought contents are externally individuated *because* (a) first-order thought contents are externally individuated and (b) second-order thoughts *inherit their content* from the first-order thoughts to which they refer. This is sometimes referred to as the *redeployment thesis*. Both Gibbons (1996) and Peacocke (1996) take this route, arguing that (besides self-interpretive cases where one forms a second-order belief by e.g. trying to psychoanalyze oneself) second-order self-ascriptive beliefs are caused by the first-order thoughts that they ascribe. It is because of this causal relationship, they argue, that a second-order attitude will have the same content as the corresponding first-order attitude.

Gibbons draws an analogy with intentional action to illustrate the point (1996, p. 291-292). Behaviors have the intentional properties that they do in virtue of having been caused by certain of your beliefs, desires, and intentions. Your walking around the kitchen, for example, counts as looking for water because it was caused in the appropriate way by your intention to be looking for water and your belief that some can be found in the kitchen. Your behavior *inherits* its intentional properties from the mental states that cause it. Similarly, says Gibbons, a second-order self-ascriptive belief will inherit its content from the first-order attitude that caused it. This ensures that there will be no mismatch between what one *thinks* one is thinking and what one *is* thinking.

[23] See Appendix A for an extended discussion of the problem.

Let us sum up what we have said so far. According to the discrimination argument, it is possible (given externalism) to find oneself in a situation where one cannot know by introspection that one is thinking that p because one cannot rule out the relevant alternative possibility that one is actually thinking that q. Oscar's case is meant to illustrate this. Because Oscar cannot discriminate between *water*- and *twater*-thoughts, the argument goes, he is in no position to rule out the relevant alternative possibility that he is actually thinking about twater. Thus, Oscar does not know that he is thinking about water. This shows that TWP is false.

Standard strategists respond that talk of "ruling out" relevant alternatives is misleading. What's important, they argue, is that one's belief satisfy the reliability requirement. The ability to discriminate, on the other hand, is important only when it is necessary for satisfying that requirement. It is not necessary for knowledge *per se*. Moreover, this is perfectly consistent with Goldman's relevant alternatives epistemology, which advocates of the discrimination argument rely on to make their case. Admitting that Henry doesn't know he is looking at a barn (as opposed to a façade) doesn't commit us to endorsing DR. The reliability requirement (as articulated by RR) can explain our intuitions in the fake barn case on its own.

Further, though Oscar is unable to discriminate between relevant alternatives, his second-order self-ascriptive beliefs *will* satisfy the reliability requirement. Or, at least, externalism gives us no special reason to suppose that they *won't*. That's because, as Gibbons and Peacocke argue, his second-order beliefs will inherit their content from the first-order beliefs to which they refer. This ensures that there will be no mismatch between what Oscar is thinking and what he *thinks* he is thinking.

Thus, if the standard strategist is right, the incompatibilist challenge here is disarmed: Goldman's relevant alternatives epistemology shows only that the reliability requirement is true, and externalism gives us no special reason to suppose that self-ascriptions will fail to satisfy that requirement.

2.3 A Dialectical Problem

Even if the standard strategist is right to think that externalism gives us no special reason to doubt that second-order judgments will satisfy the reliability requirement, we still might worry that this by itself does not solve the problem. For the proponent of the discrimination argument is making two claims: that (i) the ability to discriminate between relevant alternatives is necessary for knowledge (DR) and that (ii) it is possible, given externalism, to find oneself in a situation where one is unable to discriminate between the actual situation in which one is thinking that p and a relevant alternative situation in which one is thinking that q. So, to avoid the conclusion that TWP is false, it looks as though the compatibilist will have to show that either (i) or (ii) is false. But, as critics of the standard strategy have pointed out more than once, it isn't clear how emphasizing the *reliability* of second-order judgments is supposed to accomplish this (McLaughlin & Tye 1998a, Vahid 2003, Brown 2004).

There are two problems here. The first, emphasized by Brown (2004, p. 59-64), is that Goldman's barn case does not by itself favor one epistemic principle – RR or DR – over the other. We might say that Henry's belief that he is seeing a barn doesn't count as knowledge because it is unreliable – if he were looking at a façade, he would still believe he was looking at

a real barn. This is the interpretation favored by the standard strategist. But, we might with equal plausibility say instead that Henry's belief doesn't count as knowledge because he can't *discriminate* between real and fake barns. This is the interpretation favored by proponents of the discrimination argument. Unfortunately, the barn case by itself doesn't give us any grounds for adjudicating between these two approaches. Either principle explains our intuitions in the barn case just as well as the other.

What is needed, then, is an independent reason for favoring one approach over the other. We can't just *say* that the more basic epistemic principle is RR. But this, Brown suggests, is precisely what standard strategists like Gibbons and Falvey & Owens appear to do.

As far as explaining our intuitions in the barn case goes, however, it appears that the externalist is well within her dialectical rights to insist on RR. For as Dierig (2010b) has pointed out, RR is *weaker* than DR on the assumption that externalism is true, where "weaker" is understood as follows:

(**Weak**) H_1 is **weaker** than H_2 *iff* (i) H_2 entails H_1 but (ii) ~[H_1 entails H_2][24]

Slow-switching cases like Oscar's demonstrate that one's belief might satisfy the consequent of RR without thereby satisfying the consequent of DR.[25] It follows that RR does not entail DR. The converse, however, appears not to be true. *Prima facie*, DR does entail RR. It's hard to imagine how the consequent of DR might be true but the consequent of RR false.

For example, let's return for a moment to Fake Barn Country. Say I believe correctly that *that* is a barn and that: (a) I believe this on the basis of perception and (b) perception

[24] Dierig, writing in German, uses the verb *impliziert* where I would use *entails*. It seems clear from his discussion, however, that he has something like **Weak** in mind.
[25] We will consider another such case below.

allows me to discriminate between real barns and mere façades (perhaps the façades are not very convincing). Given (a) and (b), my belief satisfies the consequent of DR. Does it follow that the consequent of RR is satisfied too? To show otherwise, one would need to come up with a case in which both (a) and (b) are true but where my belief is such that: (c) I would still believe that *that* is barn even if it were a façade. I am not optimistic that this can be done.

Certainly, it is easy to imagine a case in which I am *able* to distinguish real from fake barns on the basis of perception but in which my belief is such that (c) is true. But this could happen only if I either fail to base my belief on perception (in which case (a) is not true) *or* I pay so little attention to what I'm seeing that I do not actually engage my ability to discriminate (in which case (b) is not true). In neither case do we have a counterexample to the claim that the consequent of DR entails the consequent of RR. And I think we would get the same result regardless of what we plug in for w (whether perception, testimony, introspection, or something else).

So, given the plausible assumption that all else being equal one should endorse the less restrictive principle, we should not expect the externalist to accept DR.

But, of course, showing that the externalist cannot be expected to endorse DR *on the basis of the fake barn case* isn't to show that DR is false or even that there isn't some other reason one might have to endorse it. This brings us to the second problem. Slow-switching cases are not the only ones in which RR and DR render different verdicts (cf. McLaughlin & Tye 1998a, p. 356, esp. footnote 15). If they were, then we would face a stalemate (we obviously can't appeal to slow-switching cases in order to solve the dispute since the verdict in those cases is precisely what's at issue), and the tie would go to the less restrictive (or weaker)

principle – that is, RR. But we can construct non-slow switching cases in which the consequent of RR is satisfied, the consequent of DR isn't, and where the subject fails to know. For example, consider the following case.

VIALS

Jen is one of a number of test subjects who have agreed to receive via injection either of two solutions: Solution A or Solution B. Solution A renders one unconscious immediately upon injection. Solution B has no effects whatsoever. The subjects do not know that there are two solutions or anything about what effects, if any, they might experience. Who receives which solution is determined randomly. Jen is to receive Solution B. Just before the injection, she is shown the vial from which the solution she is to receive was drawn. Shortly after the injection, she is taken into another room where there is a table on which sits two completely indistinguishable vials of solution. One is a vial of Solution A, the other a vial of Solution B. She is told which is which. She is also told that one (but not *which* one) of the vials is the same one she was shown earlier. She is then asked whether she believes she was given Solution A or Solution B. Considering what she has been told, Jen, *thinking* that she recognizes the vial from earlier, forms the (true) belief that she was given Solution B. Had she received Solution A, she would still be unconscious and would have no beliefs whatsoever about which solution she had received.

Intuitively, Jen does not know that she received Solution B. But we cannot appeal to RR to explain why. For it is false that had Jen received Solution A, she would still believe that she had received Solution B. That's simply because had she received Solution A, she would be unconscious and would have *no* occurrent beliefs whatsoever or any dispositional beliefs about

which solution she had received.[26] Thus, *as far as RR is concerned*, Jen's belief might count as knowledge.

On the other hand, DR *does* seem to provide a satisfying explanation for why Jen's belief doesn't count as knowledge. The possibility that Jen had received Solution A is clearly a relevant alternative. So, according to DR, Jen knows that she received Solution B only if her grounds for belief allow her to discriminate the actual situation from that in which she received Solution A. In this case, Jen's belief is grounded in perception, which she is convinced allows her to discriminate between the two vials well enough to pick out the one she saw earlier. But, clearly, she cannot discriminate between the actual situation and the relevant alternative situation on these grounds alone. That is simply because the two vials are *ex hypothesi* indistinguishable (despite what Jen may think).

In light of this case, the standard strategist has a couple of options. She could try to amend RR so that it returns the verdict we want in **Vials**. Alternatively, she could try to think up another necessary condition on knowledge that would explain why Jen's belief doesn't count as knowledge but that *doesn't* rule out self-knowledge in slow-switching cases. That is, she could try to show that DR is *not* the best explanation for why Jen fails to know in **Vials**.

Neither of these options are ideal, however. The first option is at best a temporary stop gap. For it doesn't foreclose the possibility that the incompatibilist could simply cook up another thought experiment – a **Vials 2.0** – that takes the proposed amendments to RR into account. Then the standard strategist would be back where she started, looking for yet another

[26] Remember, Jen did not even know that there *were* two solutions until she was shown the two vials. If she had received Solution A, she would not have learned this fact.

ad hoc amendment to RR that would evade the problems posed by **Vials 2.0**. In short, the standard strategist would risk getting trapped playing Whac-A-Mole with thought experiments.

The second option is not much better. Again, it does nothing to foreclose the possibility that the incompatibilist could simply come up with another thought experiment, one in which DR *is* the best explanation for why the subject in question fails to know some proposition. But even ignoring this possibility, we would end up stuck in a new debate over the relative merits of DR and the newly proposed necessary condition on knowledge.

Clearly, it would be best if we could simply show that DR is *false*. Arguing that DR fails to be the best explanation for why a subject fails to know in this or that case provides at best only a temporary solution to the problem posed by the incompatibilist.

2.4 Self-Knowledge and Infallibility

Goldberg has proposed a way out of this dilemma (Goldberg 2006; cf. Goldberg 2005a). He argues that there is an eminently plausible epistemic principle that is inconsistent with DR. We'll call this **Goldberg's Principle**.

(**GP**) If (G) S's belief that p is based on grounds that guarantee its truth, and

(A) that G is reflectively accessible to S, then

S knows that p.[27]

GP specifies a sufficient condition for knowledge. As long as one's belief satisfies conditions (G) and (A), that belief counts as knowledge. Goldberg points out that basic judgments are (or

[27] Goldberg himself never precisely articulates his principle, leaving some interpretive work for his reader. GP is based on the clearest articulation of the relevant principle that I could find in Goldberg's work (it can be found on page 310 of his 2006). We will consider other versions of GP below, some of which may diverge somewhat from what Goldberg has in mind.

produce) beliefs that meet these criteria. Since basic judgments are self-verifying, condition (G) is met. Further, that fact that basic judgments are self-verifying is reflectively accessible – one need only imagine a subject that is "familiar with Burgean reasoning" to see this (Goldberg 2006, p. 310). So, given GP, basic judgments count as knowledge. But, importantly, they do *not* satisfy the requirement specified by DR. Hence, Goldberg reasons, DR is false. Call this the Argument from Basic Self-Knowledge (**AB**).

Let's reconstruct the argument more precisely. Again, let "W" be the proposition that Oscar is thinking that water is wet and "T" the proposition that he is thinking that twater is wet. Now suppose that Oscar (after having been reintegrated into his old English-speaking community) makes the basic judgment that W.

(AB1) **GP**

(AB2) When Oscar makes his basic judgment that W, he forms a belief (namely, that W) that is based on grounds that guarantee its truth, a fact that is reflectively accessible to Oscar.

(AB3) So, Oscar knows that W. (AB1,AB2)

(AB4) But, Oscar's grounds for believing that W do not allow him to discriminate between the actual situation in which W and the counterfactual situation in which T.

(AB5) T is a relevant alternative to W.

(AB6) So, ~**DR**. (AB3,AB4,AB5)

Unfortunately, this argument is unsound as it stands. To see this, we should first note that both of the conjuncts of GP's antecedent, G and A, are ambiguous. First, there are at least two ways to interpret G. The first way is to understand it as being equivalent to G_1.

(G_1) The ground of S's belief that p *entails* that S's belief is true.

But this interpretation is problematic. As Gibbons (2001) makes clear, reflectively accessible infallibility is *not* sufficient for knowledge. To see this, suppose that I am a mathematician who

believes some theorem, a theorem which happens to be a necessary truth. Suppose that I believe it, however, not because I have proven it (though it *is* provable), but because I *want* it to be true for some reason or other. Intuitively, I do not *know* that the theorem is true.

Notice, however, that if we interpret G as being equivalent to G_1, then GP entails that my belief *does* count as knowledge.[28] First, the ground of my belief (my desire for the theorem to be true) does in fact entail that my belief is true. Since this theorem, being a necessary truth, is true in all possible worlds, it trivially follows that it is true in all possible worlds in which I desire it to be true. In other words, there is no possible world in which I base my belief on a desire that the theorem be true and end up with a false belief. Second, since we're supposing the theorem to be provable, the fact that it is guaranteed to be true is reflectively accessible to me. So, GP_1 is false.

The second way to understand G is as being equivalent to G_2.

(G_2) S's belief that p is true *in virtue of* its grounds.

In the previous example, my belief is guaranteed to be true, but *not* in virtue of the desire on which it was based. So it is not a counterexample to GP_2. However, Gibbons has come up with another case that might be. It goes as follows.[29]

> Suppose an insecure student, Harry, has read Descartes and knows that thinking is the most general propositional attitude. […] Unfortunately, Harry also believes that in order to understand a proposition, you have to grasp it through the natural light of reason. Harry tries to

[28] From here on, I'll adopt the following convention: Where G is interpreted as G_1, I'll refer to the corresponding version of GP as " GP_1 ". Where G is interpreted as G_2 (see below), I'll refer to the corresponding version of GP as " GP_2 ".

[29] Dierig (2014, p. 221-224) also discusses the following case as a potential counterexample to GP. Notice that it constitutes a problem for GP_1 as well. Thus, we may not need to appeal to necessary truths in order to refute it.

> grasp a proposition through the natural light of reason. But whatever else goes on, he is never aware of any light, natural or otherwise. Harry begins to suspect that he is incapable of thinking. […] Harry believes that he is *not* thinking about Descartes even though he is thinking about Descartes. He does not believe any proposition of the sort relevant to Burge's basic self-knowledge. In fact, he believes, dispositionally or otherwise, the negations of a great number of these propositions. Ashamed at his imagined disability, Harry begins daydreaming about how nice it would be to think. Sometimes during these reveries Harry believes for a moment that he really is thinking. [He believes this] in spite of what he takes to be good evidence against this claim about his own mind. He accepts it, not because the belief seems so obvious to him that it outweighs all evidence against it, but simply because he wants it to be true.
>
> (Gibbons 2001, p. 22)

Gibbons argues that "no belief formed in this way, regardless of truth-value, counts as knowledge" (Gibbons 2001, p. 24). Suppose that he is right about this, that Harry's belief fails to constitute knowledge. Does GP_2 entail that it does?

Notice that Harry's belief is grounded in a *desire* to be thinking. Desiring is a propositional attitude. So, if "thinking is the most general propositional attitude", as is stipulated in the example, then S's desiring that p guarantees the truth of p if p is the proposition that S is thinking. *If one desires, then one thinks*. So, G_2 is satisfied: Harry's belief is made true by the desire that grounds it.

The status of A, however, is less clear. A is satisfied only if it is reflectively accessible to Harry that his belief is true in virtue of its ground.[30] But this requires that it be reflectively

[30] Remember, at this point in the dialectic we're assuming that G is equivalent to G_2. Thus, A expresses the condition that G_2 be reflectively accessible to S.

accessible to him that his belief is grounded in a desire to be thinking. The problem is that it is not obvious that Harry is in a position to justifiedly believe even that he *has* a desire.

It comes down to what precisely is meant by saying "p is reflectively accessible to S". I assume it means (at least) that S is in a position to arrive via reflection at a doxastically justified belief that p. But this too is ambiguous. For example, it is probably the case that Harry's reflections (as described in the example) are sufficient to generate *prima facie* propositional justification for him to believe that he desires to be thinking. Does this mean that he is in a position to arrive via reflection at a doxastically justified belief to that effect? Yes and no. The answer is "no" if we hold constant Harry's other commitments (about "the natural light of reason" and so on). In other words, if we take Harry *as he is*, then we should say that he is *not* in a position to justifiedly believe that he desires to be thinking. Such a belief would be irrational given Harry's other beliefs, beliefs which Harry explicitly realizes commit him to *denying* that he is thinking (and *a fortiori* desiring). But, of course, Harry *is* in a position to realize the falsity of his beliefs about "the natural light of reason" and to revise them accordingly. He might then come to believe, justifiedly, that he has thoughts after all. Given this, we might want to say that the answer to our question is "yes": Harry *is* in a position to justifiedly believe that he desires to be thinking.

Given the above, it is not at all clear that Harry's case constitutes a counterexample to GP_2. For the proponent of GP_2 might simply stipulate that A is to be understood in such a way as to imply that we must take Harry as he is. But, as he is, Harry is not in a position to justifiedly belief that he desires to be thinking. Hence, he is not in a position to justifiedly believe that the

relevant instantiation of G_2 holds. Hence, condition A is not satisfied. So, Harry's case is not a counterexample to GP_2.

But there is another problem. If we adopt G_2 as the correct interpretation of G, then it looks like AB2 is false (or, at least, the inference from AB1&2 to AB3 is invalid). Basic judgments are *not* guaranteed to be true in virtue of their *grounds*. They are guaranteed to be true in virtue of their self-verifying natures. They are like beliefs in necessary truths in that respect. They are guaranteed to be true regardless of how they are grounded. Consequently, it does not follow from GP_2 that basic judgments count as knowledge.[31]

There is, however, a way to revise GP so that it avoids the various problems just discussed. Consider GP_3 below.

(**GP_3**) If (G_1) The ground of S's belief that p *entails* that S's belief is true, and

(A') S justifiedly believes that G_1, then

S knows that p.

Changing A to A' allows us to avoid the problems that plague GP_1.[32] Unlike GP_1, GP_3 does *not* entail that reflectively accessible infallibility is sufficient for knowledge. But, unlike GP_2, GP_3 allows us to infer that at least *some* basic judgments count as knowledge. In particular, it will follow that the basic judgments of those who *realize* that such judgments are infallible will constitute knowledge.

I think that GP_3 is a step in the right direction, but we are not out of the woods yet. For a moment's reflection reveals that G_1, our attempt at disambiguating G, is *itself* ambiguous.

[31] Notice that is *does* follow from GP_1 that basic judgments count as knowledge. But, as we saw earlier, GP_1 is false.

[32] It also avoids vague phrases like "reflectively accessible" and prevents us having to parse various interpretations of what it might mean to say of a subject that she is "in a position" to justifiedly believe some proposition.

This is because "ground" is ambiguous. Consider that S might base her belief that p on another belief – the belief that q, say. In this case, S would cite q as her reason for believing that p. Given this, should we say that the ground of S's belief that p is her *belief* that q (that is, a token mental state) or the proposition q itself, which is S's reason for believing that p? Either way, we run into a problem.

Consider the latter interpretation first. Suppose we allow that the ground of S's belief that p might consist in the propositional contents of other belief states.[33] If we do this, then GP_3 becomes straightforwardly vulnerable to Gettier cases. Consider the classic example below.

> I see two men enter my office whom I know to be Mr. Nogot and Mr. Havit. I have just seen Mr. Nogot depart from a Ford, and he tells me that he has just purchased the car. […] Mr. Nogot is a friend of mine whom I know to be honest and reliable. […] However, imagine that, contrary to my evidence, Mr. Nogot has deceived me and that he does not own a Ford. Moreover, imagine that Mr. Havit, the only other man I see in my room, does own a Ford, though I have no evidence that he (or I) owns a Ford.
>
> (Lehrer 1965, p. 169-170)

I believe that: (N) Nogot, who is in the office, owns a Ford. On the basis of this belief, I form the further belief that: (H) Someone in the office owns a Ford. N entails H. Thus, if N is the ground of my belief that H, then the ground of my belief that H entails that it is true. G_1 is satisfied. Now suppose I justifiedly believe (thanks to a bit of reflection) that: N, the ground of my belief

[33] Obviously, if we allow this, we'd have to add a third conjunct to GP_3's antecedent, one requiring that S's believe that p be true.

that H, entails H and *a fortiori* that my belief that H is true. Then, A' is satisfied. It follows by GP that I know H. But, intuitively, I don't.

To get around this problem, we'll have to restrict the extension of "ground" so that it does not include the propositional contents of other belief states. Only other token mental states can constitute grounds in our sense. But even this does not solve the problem. To illustrate why, let's return for a moment to Fake Barn Country.

Henry sees what looks to him just like a barn. On the basis of that perception, he believes that he sees a barn. Intuitively, his belief is justified. Suppose also that it happens to be correct. If so, then the token mental state that grounds his belief is a perceptual representation of a barn. Now, Henry's experiencing a perceptual representation of a barn (a factive notion) entails that Henry is in fact seeing a barn. Thus, the ground of Henry's belief entails that it is true. Condition G_1 is satisfied. What about condition A'? Plausibly, since Henry is justified in believing that he is seeing a barn, he is justified in believing that he is experiencing a perceptual representation of a barn.[34] Suppose, then, that Henry does justifiedly believe that he is experiencing a perceptual representation of a barn and is aware that this mental state constitutes the ground of his belief. Now suppose he realizes that his experiencing a perceptual representation of a barn entails that he sees a barn and *a fortiori* that his belief is true. Given this, it follows from GP_3 that Henry knows that he sees a barn. But, since he is in Fake Barn Country, he *doesn't* know. So GP_3 is false.

[34] Clearly, this assumes that justification is closed under entailment. Thus, a defender of GP_3 could try to show that justification is *not* closed under entailment. But, I think, GP_3 is not worth assuming such a burden.

We *could* perhaps try to save GP_3 by amending A' as follows: S justifiedly believes, *on the basis of reflection*, that G_1. Then, it would not follow from GP_3 that Henry knows that he sees a barn since Henry's justification for believing that G_1 holds is empirical. It derives from the empirically justified belief that he sees a barn rather than from some purely reflective process. But this amendment, in conjunction with the stipulation that "ground" refers only to token mental states (and not to propositions), would leave us with quite an odd epistemic principle. Alternatively, a defender of GP_3 could stipulate that the mental states that constitute the grounds in question must be non-intentionally described. Accordingly, the relevant mental state in Henry's case would be characterized as a perceptual representation *as of* a barn. Obviously, one's experiencing a perceptual representation as of a barn does not entail that one is actually seeing a barn. So, again, Henry's case would not be a counterexample to GP_3.

But this too seems intolerably ad hoc. One could fairly complain of the resulting principle that it is tailor-made to get the standard strategist precisely what she needs to refute the discrimination argument. Outside of that very specific goal, it seems entirely unmotivated. Indeed, one might worry, it comes as close as one can to stipulating that (at least some) basic judgments count as knowledge without explicitly doing so. To the extent that that is true, (our heavily amended version of) GP_3 may even be question-begging.

There is, however, one more way to interpret GP_3. Let's leave A' alone, and interpret G_1 along the following lines.

(G_1') There is no possible world in which S forms her belief in the way that she actually does and ends up with a false belief.

The resulting version of GP_3 (let's call it GP_3') takes care of the Gettier and fake barn cases just discussed. My belief that H is false in, for example, the nearby possible world in which Havit does not come into the office at just that moment. And Henry's belief that he sees a barn is false in all nearby possible worlds in which he is actually seeing a barn façade.

It also takes care of the objection that GP_3 is unmotivated at best and question-begging at worst. For GP_3' can be derived from a more general, less arbitrary principle. Notice that G_1' is a stronger version of the familiar epistemic property *safety*, understood as given below (cf. Sosa 1999, Pritchard 2007).

(**SF**) S's belief that p is **safe** *iff* there is no nearby possible world in which S forms her belief in the way that she actually does and ends up with a false belief.

Thus, GP_3' follows from the more general principle GP_S.

(**GP_S**) If (G_S) S's belief that p is safe, and

(A') S justifiedly believes that G_S, then

S knows that p.

Now, GP_S is highly plausible. Further, it is not, as perhaps GP_3 is, an odd epistemic principle clearly motivated by, and specifically engineered to suit the purposes of, some dialectical strategy. On the contrary, it appeals to a common epistemic property. It says simply that safe belief counts as knowledge in at least those cases where A' holds.

Notice, however, that if GP_3' gets whatever plausibility it has from GP_S, then any counterexample to GP_S severely weakens the case for GP_3'. Given this, GP_S is the principle we're really interested in defending. If GP_S turned out to be false, it would be that much easier for the proponent of DR to throw GP_3' out with it.

In his 2007, David Manley points out several problems with supposing that safety, as traditionally conceived, is sufficient for knowledge. One might worry about GP_S in light of these concerns. One of these problems stems from the fact that, if we suppose that perceptual demonstrative beliefs are object-dependent, then (on some standard definitions of safety) we can easily come up with cases in which some perceptual demonstrative belief is safe but fails to count as knowledge. To use one of Manley's examples:

> Suppose I have a true demonstrative thought, one that I might express by saying, 'That is a lark'. If [many lark-imitating] imposters are nearby, I am still in danger of messing up. But the problem is not that I could easily have falsely believed that very proposition. (Arguably it is a necessary truth.) Had an imposter been singing, I would have believed something else – a proposition with different truth conditions – though I would have expressed my belief the same way.
>
> (Manley 2007, p. 403-404)

There is no nearby possible world in which I falsely believe the very proposition that I actually believe – call it "P". Thus, my belief that P appears to be safe. Regardless, Manley argues, it doesn't count as knowledge since I could easily have been wrong. Though I couldn't easily have falsely believed that P, I *could* easily have ended up with a false belief since I could easily have ended up believing *of a lark-imitator* that it was a lark.

To see the problem this poses for GP_S, we need only add that I, the subject of the thought experiment, am a philosopher who knows about object-dependent thought. Then, realizing that my belief that P is object-dependent, I might form the justified belief that my belief that P is safe. In that case, condition A' would be satisfied. Then, it would appear to follow from GP_S that I *know* that P. But since I don't, it looks like GP_S is false.

SF, however, is formulated so as to accommodate cases like this. According to SF, a belief B counts as safe only if there is no nearby possible world in which the very belief-formation process that yielded B yields a false belief. Since, in the lark case, I could easily have ended up with a false belief (even if I couldn't have easily ended up with a false belief *that P*), my belief is not safe. Thus, it doesn't follow from GP_S that I know that P. Manley's case is therefore not a counterexample.[35]

Given GP_S, we can amend the argument as follows. As before, suppose that Oscar makes the basic judgment that W. This time, however, suppose that Oscar employs Burge-style reasoning to arrive at the conclusion that basic judgments like the one he is now entertaining are safe. Then:

(AB1*) **GP_S**

(AB2*) When Oscar makes his basic judgment that W, he forms a belief (namely, that W) that is safe, a fact that he justifiedly believes to be the case.

(AB3) So, Oscar knows that W. (AB1*,AB2*)

(AB4) But, Oscar's grounds for believing that W do not allow him to discriminate between the actual situation in which W and the counterfactual situation in which T.

(AB5) T is a relevant alternative to W.

---

[35] There is another version of the case that Manley considers. Suppose, not that there are any lark-imitating imposters nearby, but that I often hallucinate lark calls. I hear a real lark call on the basis of which I form a demonstrative belief that I would express by uttering, "*That* is a lark". Given my condition, I could easily have hallucinated the lark call, in which case my belief would have had no content. But if a belief has no content, then it can't be *false*. So, according to SF, my belief is safe (or, to be more precise: the fact that I might have hallucinated does not by itself render my belief unsafe). Plausibly, however, my belief doesn't constitute knowledge. So, again amending the case so that A' is satisfied, we have a counterexample to GP_S. To get around this problem, we need only amend SF by replacing "false" with "untrue". I have no objection to this. If we do, we end up with a definition of safety that is equivalent (or at least very close) to the one Manley ends up endorsing, what he calls "revised safety".

(**RS**) S could not easily have had a failed counterpart thought.

A failed thought is one that "has no content, or has as its content a false, gappy, or paradoxical proposition" (Manley 2007, p. 406). Manley, however, thinks that RS is necessary for knowledge. I do not think he is right about this.

(AB6) So, ~**DR**. (AB3,AB4,AB5)

I think that this argument is sound. It comes down to our revised version of Goldberg's Principle – GP_S. (It certainly seems possible to come up with a story that makes AB2* true. And advocates of the discrimination argument aren't in a position to deny either AB4 or AB5.) And I think that GP_S is highly plausible *regardless* of one's prior commitments with respect to the compatibilist/incompatibilist debate. Importantly, one can accept it while at the same time *denying* that we have self-knowledge in the vast majority of cases. For, as far as GP_S is concerned, a second-order judgment might constitute knowledge only when accompanied by a justified belief to the effect that it is safe. It is silent in all other cases.

Regarding AB, one might respond that a proponent of DR could just as easily reason in the other direction, arguing that GP_S is false since it is inconsistent with DR. Logically speaking, of course, one can always protect a favorite principle by obstinately denying anything that might conflict with it, so long as the principle in question is consistent. But, surely, the compatibilist is within her dialectical rights to draw the line somewhere. GP_S seems to me a good place to draw it.

2.5 An Objection Considered: The Case of Temp

Duncan Pritchard's TEMP example appears to straightforwardly undermine GP_S. It goes as follows:

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.
>
> (Pritchard 2012, p. 260)

Temp's beliefs about the temperature in the room are safe. Because the hidden agent will guarantee that Temp is always given the correct reading, there is no nearby world in which the thermometer leads him astray. Further, it is reasonable to think that since "he has no reason for thinking that there is anything amiss", Temp is justified in believing that consulting the thermometer is a safe way to form beliefs about the temperature in the room. Assuming that Temp does justifiedly believe this, GP_S commits us to saying that Temp *knows* (or is in a position to know) the temperature in the room. But, according to Pritchard, Temp does not know.

I confess that if Temp really does not know the temperature in the room, then GP_S is in trouble. However, it seems to me that Temp does know. While I do not have the space to offer a full analysis of the case here, I will briefly outline two reasons in my defense. First, other than our intuitions to the contrary, there is no obvious reason why we should deny that Temp knows. Second, I think the intuition that Temp doesn't know can be explained away. Let's take these in turn.

With respect to the first point, several explanations for why Temp might fail to know have been offered. None, I think, have been successful. Pritchard's own explanation is that while Temp's beliefs about the temperature in the room are correct, their correctness is not to "a significant [enough] degree creditable to [his] cognitive agency" for them to constitute

knowledge (Pritchard 2012, p. 273). But, as Christoph Kelp (2013) has suggested, if belief based on testimony will under normal conditions satisfy Pritchard's creditability standard, then so should Temp's beliefs about the temperature in the room. As Pritchard himself admits (2010, p. 41; 2012, p. 270), the success of a testimonial belief is sufficiently creditable to the cognitive abilities of the agent who holds it so long as, for instance, she wouldn't have consulted someone manifestly unreliable (e.g. a child), would be sensitive to obviously false testimony, and so on. But, Kelp notices, this suggests that Temp too will satisfy the creditability standard – that is, so long as we can suppose that he wouldn't consult a manifestly unreliable thermometer (e.g. one that is clearly broken or labeled as out of order), would be sensitive to obviously false readings, and so on. Thus, if Temp really does not know, it must be for some other reason.[36]

Kelp's preferred explanation is that Temp's beliefs are not *apt*. Roughly, an agent's belief is apt when its truth manifests the agent's cognitive competence.[37] The idea here is that while *ex hypothesi* Temp forms his beliefs competently, their truth does not manifest this competence. Temp is cognitively successful thanks to the interventions of the hidden agent, not to his own competence.

This explanation has its own problems. To see this, consider the following case, which is structurally similar to TEMP:

JELLY BEANS

[36] See Hudson (2014) for a more comprehensive critique of Pritchard's explanation.
[37] The idea that beliefs should be assessed for aptness is Ernest Sosa's innovation. See his 2007 for further discussion.

> At the county fair, a jar full of jelly beans is on display. As people walk by, they are invited to guess how many jelly beans are in the jar. Once they have made their guess, they can check to see if they are correct by lifting a small flap beneath which is an LED display with the correct number on it. The number displayed is the output of a sensor that is meant to detect the exact number of jelly beans in the jar. But, unbeknownst to the passersby, the sensor is broken, causing the number on the LED display to fluctuate randomly within a given range. Luckily, however, the number always remains stable for at least 30 seconds. To remedy the situation, a fair employee is assigned to ensure that every time someone consults it the LED display shows the correct number. Whenever someone is seen approaching (and before any guessing takes place), the employee adjusts the number of jelly beans in the jar to match the number shown on the LED display. And, to ensure that the number doesn't change in the middle of a game, each fairgoer is allowed only 30 seconds to make a guess.

My intuition is that whenever a fairgoer checks the LED display, they come to *know* how many jelly beans are in the jar at that time.[38] After all, they clearly come to know whether or not they guessed correctly and, if not, by how much they were off. Thus, either JELLY BEANS is a counterexample to the idea that knowledge requires apt belief, or the fairgoers' respective beliefs about the number of jelly beans in the jar are apt. If the former, then Kelp's explanation straightforwardly fails. But, if the latter is true, then, given the structural similarities between

[38] Of course, this knowledge can be defeated. This might happen if I consult with someone who had played the guessing game at a time when there was a different number of jellybeans in the jar. If in the course of conversation a discrepancy is discovered, I would have reason to believe that I'd made some kind of mistake.

the two cases, Temp's beliefs about the temperature in the room must also be apt.[39,40] So, Kelp's appeal to aptness fails either way.[41]

What, then, of the intuition that Temp doesn't know? I suspect that when we read through the case, we picture to ourselves something like the following sequence of events: Temp begins to wonder about the temperature in the room at some time t_1; to satisfy his curiosity, he begins to head over to the thermometer at t_2; the hidden agent notices this and immediately adjusts the temperature in the room to match the reading on the thermometer; at

[39] According to Sosa, "[w]hat is required for aptness is that the performance succeed through the exercise of a competence in a situation appropriately normal for that exercise" (Sosa 2007, p. 84). Kelp's explanation for why Temp's beliefs are not apt is that, because the thermometer is broken, his situation is not appropriately normal. I think JELLY BEANS helps to show that things are not that simple. Though the sensor is broken, we still want to ascribe knowledge to the fairgoers. Why? What I think is going on is that the fair employee is effectively *restoring* a degree of normalcy to the situation. She is changing the situation so that it no longer matters so much that the sensor isn't working normally. I think we can say the same thing about the hidden agent in TEMP. The hidden agent complicates the simple "you can't rely on a broken thermometer" argument. For given the hidden agent's systematic efforts to ensure that the thermometer never gives the wrong reading, it is probably misleading (though not technically false) to continue to describe the thermometer as "broken". In any case, determining whether or not Sosa's normality condition is satisfied in a given situation is probably less straightforward than Kelp acknowledges.

[40] As I suggest below, there is one difference between the two cases. In TEMP, the temperature in the room is liable to change between the time Temp forms his intention to check the thermometer and the time he actually checks it. In JELLY BEAN, however, the case is set up so that there is no risk that the number of jellybeans will change between the time the fairgoers make their guesses and form the intention to check the answer and the time they actually look at the LED display. Thus, while Temp is liable to form false beliefs about what the temperature was *just now*, the fairgoers are not at risk of forming false beliefs about how many jelly beans were in the jar just now (as they were guessing). It is not immediately obvious to me why this should make any difference with respect to the aptness of our target beliefs – that is, beliefs about what the temperature *is* and about how many jelly beans are in the jar, respectively. In both cases, we have a hidden agent changing the way the world is to match our protagonists' beliefs.

[41] I know of one other proposed explanation for why Temp doesn't know. Robert Hudson (2014) argues that because Temp thinks the thermometer is working normally, he "misdescribe[s] the causal processes that generate [his] beliefs" (Ibid., p. 811). This is a problem, he thinks, because it "could lead to a situation where he doesn't think he should be believing p [where p is a proposition about the temperature in the room], given how he thinks such a belief is caused in his mind" (Ibid., p. 814; I have changed the pronouns for the sake of continuity). I am not sure how exactly to interpret this. It sounds like Hudson thinks that S's belief B constitutes knowledge only if S's understanding of the causal process that generates B doesn't leave S open to defeaters. But, as stated, this principle is probably vulnerable to the old Tom Grabit case (Lehrer & Paxson 1969). Perhaps Hudson is interested only in *mis*understandings. Still, the principle is spurious. Suppose I falsely believe that all functioning thermometers must have mercury in them. This misunderstanding may expose me to defeaters (if I discover that my thermometer does not contain mercury, it may no longer be rational for me to trust it), but surely thermometers can be a source of knowledge for me all the same. Perhaps there is a more promising interpretation of Hudson's principle, but that is another paper.

t_3, Temp arrives at the thermometer and forms his belief about the temperature in the room. Put this way, it is natural to suppose that at t_3 Temp forms a belief about what the temperature in the room *was at* t_1. But if that's right, then his belief is (or could easily have been) false. It is true that if, upon checking the thermometer, he forms *only* beliefs about what the temperature in the room is at t_3, then his beliefs will be safe. But it is unnatural to read the case this way. No one checks a thermometer and forms only beliefs about what the temperature is at that very moment. On the contrary, we consult thermometers to learn something about what the temperature was *just now*, when we first started wondering about it. We take for granted that the temperature will not change from the time we form our intentions to check the thermometer to the time we actually check it. But this is precisely what happens in TEMP.

I think that this explains our initial reactions to the case. We latch on to the fact that Temp is systematically misled with respect to his beliefs about what the temperature was just now, when he first started wondering about it. And this affects our judgments about whether he can know what the temperature *is* because, at the level of intuition, we are failing to

distinguish between these two types of belief.[42] That is predictable. Outside the world of odd thought experiments, there is no reason ever to make such a distinction.[43]

For these reasons, I do not think that TEMP works as a counterexample to GP_S. First, as I have argued, there is no good reason to suppose that Temp doesn't know. And, second, our intuitions seem to suggest otherwise only because we imagine that he is forming *unsafe* beliefs. With this result, the conclusion of section 2.4 stands.[44]

---

[42] This way of putting things may suggest another reason to suppose that Temp doesn't know. I have suggested that checking the thermometer causes Temp to form many false beliefs about the temperature in the room. That's because it leads him to form false beliefs about what the temperature was at some time in the recent past. But if that's right, then it looks like forming beliefs about the temperature in the room by consulting the thermometer is a globally unreliable belief-formation process type. So, one might think, Temp's beliefs about what the temperature in the room is *right now*, however safe they may be, do not constitute knowledge because they are the products of tokens of an unreliable process type. Notice, however, that going this route would commit us to saying that even perfectly functional thermometers in normal situations can fail to be sources of knowledge about the temperature in a room. For example, imagine a sun room the temperature inside of which fluctuates frequently due to, say, changes in cloud coverage. Or perhaps a room whose temperature fluctuates simply as a result of the AC unit kicking on and off. In other words, imagine a room whose temperature fluctuates not because a hidden agent is messing with the thermostat, but for more mundane reasons. Now suppose that there is a thermometer in this room and that it is functioning perfectly. In this case, as in TEMP, we have a thermometer whose readings are always correct. The only real difference is that now the thermometer is reacting to the temperature rather than the other way around. The important thing to notice is that checking the thermometer will be no more or less reliable a belief-formation process type in the one case than in the other. Should we say, then, that checking the thermometer in these more mundane circumstances is a globally unreliable way to form beliefs and that, as a result, one cannot come to know the temperature in the room by doing so? This strikes me as implausible. *Even if*, as a result of temperature fluctuations caused by e.g. an aggressive AC unit, I am induced to form false beliefs about what the temperature was, say, ten minutes ago, it still seems to me that I can come to know what the temperature *is* by checking the thermometer. The mistake here probably lies in the assumption that beliefs about what the temperature *was* and beliefs about what it *is* are products of tokens of the same process type. But that is a complicated issue best left for another occasion.

[43] This also explains why JELLY BEANS elicits a different intuition. The reader is not tempted to suppose that the fairgoers are forming false beliefs about how many jelly beans were in the jar *before* the employee intervened. We might also note that if we stipulate that the hidden agent in TEMP ensures that the temperature in the room *always* matches the "reading" on the thermometer (whether or not Temp is looking at it), then the intuition that Temp doesn't know is much weaker. I think that's because, if we make this stipulation explicit in our description of the case, we're less inclined to imagine the hidden agent changing the temperature *as* Temp is headed to the thermometer – that is, less inclined to imagine a temperature change happening between the time Temp forms his intention to check the thermometer and the time he actually checks it. But then we are less inclined to imagine that Temp is forming false beliefs about what the temperature was *just now*. This, I think, is further evidence that when we have the intuition that Temp doesn't know, it's because we're led to imagine that Temp is forming false, unsafe beliefs about the temperature in the room.

[44] I consider one final objection to GP_S in Appendix B.

Chapter 3: The Illusion Argument

Let's turn now to Jessica Brown's (2004) "illusion argument". The formula behind the illusion argument is the same as that behind the discrimination argument: (1) Identify some necessary condition on knowledge and then (2) come up with a thought experiment in which a thinker's introspective beliefs about what she is thinking fail to satisfy that necessary condition even though her cognitive faculties are functioning properly. Externalism is implicated because the thought experiment is supposed to be an illustration of the kind thing that would be possible if externalism were true. Thus, if these two steps are carried out successfully, the externalist is forced to admit that a thinker might fail to know what she is thinking even when nothing goes wrong cognitively. In other words, she must admit that TWP is false. The upshot is that she must give up either TWP or externalism.

Rather than relying on the controversial idea that knowledge requires some kind of discriminative ability, however, the illusion argument takes as its starting point the influential idea that in order to constitute knowledge a belief must be *globally reliable*. That is, it must result from a process token whose relevant process type is generally truth-conducive. Thus, the illusion argument attempts to show that, given externalism, it is possible to find oneself in a situation where one cannot know introspectively that one is thinking that p, not because one lacks the relevant discriminative ability, but because the relevant belief is not globally reliable.

This sets the stage for step (2) of the argument: the thought experiment. Here the argument presupposes what Brown (2004) refers to as the "illusion version" of externalism. The illusion version holds that in no-reference cases, such as when one thinks that *phlogiston is*

interesting, one suffers an illusion of thought. The idea is that since "phlogiston" does not refer, such a thought has no representational content and is therefore not really a *thought* at all (see e.g. Boghossian 1998). Nonetheless, it will appear to the thinker to be a genuine thought. Thus, a thinker in these circumstances will be disposed to form many false second-order beliefs to the effect that she is thinking a thought, rendering the responsible belief-formation processes unreliable. Given this, Brown argues, it is possible that a subject might genuinely be thinking that p, correctly introspect that she is thinking that p, but fail to know that she is thinking that p because the relevant belief-formation process has been rendered unreliable. The thought experiment she offers (the "wasp" example described below) is meant to illustrate this possibility.

In the previous chapter, I argued that the discrimination argument fails at step (1). DR, I tried to show, is false. The goal of this chapter is to show that the illusion argument fails at step (2). I will not take issue here with the assumption that the illusion version of externalism is correct. My concern lies instead with the scope of the illusion argument. Strictly speaking, Brown's thought experiment shows only that *singular* externalism (i.e. the view that singular thought is object-dependent) is incompatible with TWP. She contends, however, that the argument can be generalized to apply to natural-kind externalism as well (Brown 2004, p. 115 & 134-135). The aim of the present chapter is to show that this is mistaken. That is, I will try to show that there is no thought experiment analogous to the one Brown offers that can be used to implicate natural-kind externalism, despite her assumption to the contrary.

I begin in the next section with a more detailed outline of the illusion argument.

3.1 Global Reliability and Illusions of Thought

As I have already mentioned, the illusion argument makes use of the idea that *global reliability* is necessary for knowledge. This is not to be confused with the kind of *local* reliability exemplified by RR (see Ch. 2).[45] Local reliability has to do with the reliability of a process with respect to a *particular* belief token. It is usually understood in modal terms. Thus, to ask if a belief is locally reliable is to ask whether, given the process by which it was formed, it could easily have been false. In addition to RR, both safety and sensitivity are versions of local reliability. Global reliability, on the other hand, has to do with the *general* truth-conduciveness of a belief-formation process type. If we want to know whether a belief or the process responsible for it are globally reliable, we ask something like, "what proportion of the beliefs produced by this process are true?"

It's important to note that these two conceptions of reliability are not mutually exclusive. Thus, even if we think that RR is correct, we should not assume that it expresses the *only* reliability condition on knowledge. It might be that a belief must be both locally *and* globally reliable in order to constitute knowledge. In fact, some philosophers have argued that this is the case (e.g. Goldman 1986, Becker 2008). For example, Becker (2008) endorses both forms of reliability as necessary for knowledge both as a way of eliminating epistemic luck and as a way of solving the generality problem. And Brown (2004, p. 124-129), following McGinn (1984) and Goldman (1986), points to cases where a subject's lack of knowledge cannot be explained by a local reliability condition alone but *can* be explained by a global reliability condition (we'll discuss one such case in section 3.2.2 below).

[45] The distinction between local and global reliability was introduced in McGinn (1984).

The upshot is that even if the compatibilist is right to insist that certain kinds of second-order judgment are self-verifying or are otherwise highly reliable in the *local* sense, this does not solve the problem posed by the illusion argument. That is simply because a belief's being locally reliable is not sufficient for its being globally reliable.

Let's now flesh out the relevant concepts more clearly. For the rest of this chapter, I will understand global reliability as follows.

> *Global Reliability*
> A token belief B is **globally reliable** *iff* the process leading to B is a process token whose relevant process type is truth-conducive.[46]

Then, to say that global reliability is necessary for knowledge is simply to endorse the following conditional.

(**GR**) A token belief B constitutes knowledge *only if* B is globally reliable.

Notice that I am understanding global reliability as a property that token *beliefs* may possess.[47] It is a property that a belief has when the process type responsible for it produces (or tends to produce) a sufficiently high proportion of true beliefs. And a belief counts as knowledge, according to GR, only when it possesses this property.

I acknowledge that this definition of global reliability is under-described. It would be almost impossible to apply it in particular cases without further specification. For one thing, it is silent on whether what matters is the proportion of true beliefs *actually* produced or the proportion that *would be* produced in a suitable range of instances (see e.g. Alston 1995). Nor,

[46] Note that our definition does not account for belief-dependent processes. It need not do so for present purposes.
[47] Though, for stylistic purposes, I will sometimes describe process types as being "unreliable". When I do so, I mean that they fail to be truth-conducive.

relatedly, does it specify whether truth-conduciveness is determined by a process's performance in the actual world or in some domain of possible worlds (see e.g. Goldman 1986, Goldman 2008, Henderson & Horgan 2010). It also ignores the fact that a given process type will perform differently at different times (see Frise 2018, Tolly 2019) and in different circumstances. And, finally, it leaves open how we ought to determine what the "relevant" process type is.[48]

In short, I have intentionally offered a neutral definition of global reliability. I have done so because, for the most part, neither Brown's illusion argument nor my objection to it presuppose any very specific account of global reliability. And any important assumptions that *are* made can be addressed along the way.

Brown argues that given externalism and GR, it can be shown that TWP is false. The following thought experiment forms the core of her argument. Suppose that Sally hears a wasp, *w*, buzzing nearby and thinks to herself the following perceptual demonstrative thought: *That wasp (*w*) is near*. Let's call this token thought "T_W". Via introspection she then forms the second-order belief that she is thinking that *that wasp (*w*) is near*. Call this token belief "B_W". Suppose further that, because there is a wasp nest close to her office, Sally has become very sensitive to wasp sounds. Consequently, she's started "hearing" them even when there are none. In other words, she frequently suffers illusions of hearing a wasp. When this happens, she thinks an empty analogue of T_W. Unlike T_W, the analogue does not refer. Now, according to the illusion view, a perceptual demonstrative thought without a referent has no content and is

[48] This, of course, is the well-known generality problem (see Feldman 1985, Conee & Feldman 1998). Any attempt to solve it is well beyond the scope of this chapter. But, as we'll see, it will end up being a problem not only for Brown's original version of the illusion argument but for any attempt to generalize it as well.

therefore not really a *thought* at all. So, in these cases, Sally suffers an *illusion* of thought. She believes she is entertaining a genuine thought about a wasp when in fact she is not. Due to the frequency of these illusions, Sally forms many false second-order beliefs to effect that she is thinking a thought about a wasp[49] – enough, we can suppose, that the responsible belief-formation process type fails to be truth-conducive. Let's call this process type "P".

Now, according to TWP, B_W should constitute knowledge. Sally's introspective faculties are functioning properly, and B_W is grounded in the right way via introspection. However, Brown contends that it's reasonable to suppose that the process token that leads to B_W is an instance of P (2004, p. 129-133). After all, B_W is a second-order belief that self-ascribes a wasp-thought. But, *ex hypothesi* P is not truth-conducive. It would follow that B_W is not globally reliable. Then, given GR, B_W wouldn't constitute knowledge. So TWP would be false. Here is the whole argument put together[50]:

(I1) **GR**

(I2) The process leading to Sally's belief B_W is a process token whose relevant process type is P.[51]

(I3) P is not truth-conducive.

(I4) B_W is not globally reliable. (I2,I3)

(I5) Sally's belief B_W does not constitute knowledge. (I1,I4)

(I6) ~**TWP** (I5)

[49] Brown is not clear about what exactly these false second-order beliefs are. Morvarid (2013) offers a helpful discussion of the options. (Does e.g. Sally falsely believe that *there is an x such that x is a wasp and she is thinking that (x is near)*? That her utterance "that wasp is near" expresses a genuine thought? Or perhaps that *that* [indicating a mental state via "mental pointing"] is a genuine thought?) See especially section 3 of his paper (p. 309-313). See also section 3.2.4 below.

[50] The following argument assumes that Sally's faculty of introspection is functioning properly.

[51] As Brown anticipated, I2 has borne the brunt of the criticism (Dierig 2010b, Morvarid 2013). I discuss this issue in some detail in section 3.2.4 below.

It's important to be clear about what exactly this argument accomplishes, assuming that it is sound. Sally's thought T_W is an instance of *singular* thought. Thus, Brown's thought experiment technically only shows that the illusion version of *singular* externalism is incompatible with TWP. Nothing at all has been said that might implicate natural-kind externalism. Nonetheless, it's perhaps natural to suppose that the argument can be generalized. Brown herself seems to take for granted that it can (2004, p. 115 & 134-135). All we would need to do, it seems, is put Sally in an environment in which she frequently suffers illusions, not of hearing wasps, but e.g. of seeing *water*. We might imagine that she is on Dry Earth.[52] In that case, as in the original wasp example, Sally will suffer illusions of thought. She will believe that she is entertaining genuine thoughts about a watery substance when in fact she is not. So, it looks like we have the same problem as before – only this time, *natural-kind* externalism is on the chopping block.

In the next section, I argue that the situation is more complicated than that. As I will show, simply putting Sally on Dry Earth (or some other illusion-inducing environment) will not do the trick. In short, the reason is that we need a *non-empty* analogue of B_W in our new thought experiment. That is, we need a second-order belief that *should* count as knowledge (according to TWP), but that doesn't because it isn't globally reliable. I suggest that there is no way to build the thought experiment to satisfy those conditions. In other words, there is no natural-kind analogue to the wasp example.

[52] Here is Boghossian's description of Dry Earth: It is "a planet just like ours in which, although it very much seems to its inhabitants that there is a clear, tasteless and colorless liquid flowing in their rivers and taps and to which they confidently take themselves to be applying the word 'water', these appearances are systematically false and constitute a sort of pervasive collective mirage" (Boghossian 1998, p. 206).

3.2 Does the Argument Generalize?

Let's assume that the illusion argument, as given above, is sound. Then, we must give up either TWP or the illusion version of singular externalism.[53] The question that concerns us in this section is whether the same argumentative strategy can be used against natural-kind externalism as well. I argue that it cannot. To make my case, I will run through three different versions of the Dry Earth scenario. None, I argue, can be used to construct an analogue to the wasp example.

Recall that in the original wasp example, Sally introspectively forms a *true* second-order belief about an externally individuated first-order thought that she is entertaining. That belief we called "B_W". It is this second-order belief that TWP tells us should constitute knowledge but that allegedly doesn't because it fails to be globally reliable. So, any natural-kind analogue to the wasp example will need to feature an analogue to B_W – a *true* second-order belief about an externally individuated first-order thought. Only this time, the relevant belief must involve some natural-kind concept. And, it must occur in the context of an illusion-inducing environment, one that renders the belief-formation process type responsible for it unreliable. As I've mentioned, we'll use Dry Earth as our test case. But note that there is nothing special about Dry Earth – any illusion-inducing environment would work for the purposes of this section.

[53] Its early proponents (Evans 1982, McDowell 1986) argue that singular externalism is wedded to the illusion interpretation of no-reference cases. If so, then to give up on the illusion version of singular externalism is to give up on singular externalism altogether. Not everyone, however, agrees that they are so wedded (see e.g. Aasen 2017). If they are not, then one can coherently accept the illusion argument while maintaining both TWP and some version of singular externalism.

3.2.1 First Attempt

For our first attempt at an analogue case, let's place a new protagonist, Wally, on Dry Earth. Let's suppose that Wally is a resident of Dry Earth and has never been anywhere else. In particular, he's never been to an environment that contains water. Now suppose Wally sincerely utters the sentence, "Water is refreshing". Since he is on Dry Earth, of course, his use of the term "water" fails to refer. He is in a no-reference situation. Thus, on the illusion view, the mental state corresponding to this utterance has no determinate content. So, it fails to constitute a thought. We might, for ease of exposition, call it a *pseudo-thought*.[54] Let's call this particular pseudo-thought "T_{pseudo}". Being on Dry Earth, Wally will entertain many pseudo-thoughts like T_{pseudo}. He will do so any time he attempts to think about the clear, watery substance that he falsely believes exists in his environment. In these cases, he will, like Sally, suffer illusions of thought. He will falsely believe that he is thinking a genuine thought about a watery substance when in fact he is not. Let's call the process type responsible for producing these false second-order beliefs "P*".

Let's grant that, due to the frequency of the false second-order beliefs that it produces, P* is not truth-conducive. Then, any belief produced by any process token whose relevant process type is P* will, given GR, fail to constitute knowledge. The question now is: Are there any beliefs that fit this description *and* that TWP tells us should constitute knowledge? The answer must be "yes" if we are to have a successful analogue to the wasp example. For only then do we have an analogue to B_W.

[54] I'm borrowing this term from Morvarid (2013). More on pseudo-thoughts below (in section 3.2.4).

No such analogue seems possible, however. The reason, in short, is that B_W has as its object a *genuine* first-order thought about a wasp (viz. T_W). But, on Dry Earth, there are no genuine first-order thoughts about water. There are only pseudo-thoughts like T_{pseudo}. There are a couple of ways to spell out the problem this poses. First, TWP implies that *if* Wally is thinking a first-order water-thought, then he can know this introspectively. But, in Wally's case, the antecedent is never satisfied. In other words, there *are* no first-order water-thoughts such that Wally should be able to know that he is thinking them. This means that there are no corresponding second-order beliefs that, given TWP, ought to constitute knowledge. Second, suppose Wally does attempt to express a genuine second-order belief by sincerely uttering the sentence, "I am thinking that water is refreshing". This, if it did express a genuine belief, would be the obvious candidate for our B_W analogue. But, according to the illusion view, it too must be a pseudo-thought. If "water" fails to refer when Wally utters, "Water is refreshing", it must fail to refer here as well. So, the best candidate for our B_W analogue is *not even a belief*. Our first attempt, then, is not going to work.

3.2.2 Second Attempt

Let's try again. The previous paragraph suggests that in order to have an analogue to B_W, we need an analogue to T_W as well. Just as Sally sometimes has genuine thoughts about particular wasps, we need to somehow amend the Dry Earth scenario so that Wally sometimes has genuine thoughts about water. One way to do this is to imagine that Wally is slow-switched between Earth and Dry Earth. Then, when he is on Earth, his utterances of "Water is refreshing" will express genuine first-order thoughts about water. When he is on Dry Earth,

however, they will express pseudo-thoughts. This move gets us closer to the structure of the original wasp example. Wally's being on Earth parallels cases where Sally really *does* hear a wasp, and his being on Dry Earth parallels cases where Sally suffers illusions of hearing a wasp. More important, though, is that it allows Wally to have genuine second-order beliefs to the effect that he is thinking about water.

Here's how a second attempt might go: Suppose that Wally, now a victim of slow-switching, is currently on Earth and has been there long enough so that his conceptual repertoire has had time to adjust. In particular, he now possesses the concept *water*. Now, suppose he thinks to himself the following first-order thought: *Water is refreshing*. Let's call this token thought "T_Y". Via introspection he then forms the second-order belief that he is thinking that *water is refreshing*. Call this token belief "B_Y". Like B_W, B_Y should, given TWP, constitute knowledge. After all, we can suppose that Wally's introspective faculties are functioning properly and that B_Y is grounded in the right way. Unfortunately, however, Wally spends most of his time on Dry Earth thinking pseudo-thoughts like T_{pseudo}. That means that when Wally believes that he is thinking a genuine thought about a watery substance, he usually believes falsely. Because such second-order beliefs are usually false, the process type responsible for them, P*, fails to be truth-conducive. But it's reasonable to suppose that the process token that leads to B_Y is an instance of P* (since B_Y is also a second-order belief that self-ascribes a thought about a watery-substance). Thus, B_Y is not globally reliable. It therefore fails to constitute knowledge, contrary to what TWP might suggest.

At first glance, this argument appears to be no less plausible than Brown's original. It is certainly very similar structurally. However, I argue that the introduction of slow-switching

causes problems absent from the original version of the argument. In particular, Wally's epistemic circumstances significantly change every time he is slow-switched. While he is on Earth, P* is highly reliable. When he believes that he is thinking a genuine thought about a watery substance, he will (barring some unusual cognitive malfunction) be correct. It is only when he is on Dry Earth that P* becomes highly *un*reliable. This suggests that when assessing P* for truth-conduciveness, we must account for the *circumstances* in which it is being used. But, if we take circumstances into account, we do not get the conclusion that B_Y is globally unreliable. That's because B_Y is formed while Wally is on Earth. And in these circumstances P* is highly truth-conducive. It would follow that, as far as GR is concerned, B_Y very well may constitute knowledge.

Brown in fact appears committed to this conclusion. Consider the following variation on the fake barn case that she offers as part of her argument for GR (2004, 126-129). Suppose that Laura is passing through Fake Barn Country. Looking at a real barn, she forms the true belief that *that* (b) is a barn.[55] Now, if Kripke is correct, then there is no possible world in which b exists but is not a barn (Kripke 1980). In those possible worlds where b doesn't exist, then of course Laura will have no beliefs about b. But in all possible worlds where b does exist, the proposition *b is a barn* is true. So, there is no possible world in which Laura falsely believes that *b is a barn*. In short, Laura's belief is by almost any metric locally reliable. This means, Brown

[55] Recall that in the version of the case we considered in Ch. 2, Henry's belief was something like: (∃x)(x is a barn ^ I [Henry] am looking at x). The content of Laura's belief, by contrast, has the form of a singular proposition. That's why the fact that Henry's belief fails to constitute knowledge can be explained by a local reliability condition (like RR) while the fact that Laura's belief fails to constitute knowledge cannot.

argues, that if her belief fails to constitute knowledge, the best explanation for this is that it is not *globally* reliable.[56]

Let's call the process type by which Laura forms perceptual demonstrative beliefs of the form *x is a barn* "P_B". Then, to say that Laura's belief is not globally reliable amounts to saying that P_B fails to be truth-conducive. Obviously, it fails to be truth-conducive in Laura's case thanks to the presence of deceptive barn façades. So, the explanation for Laura's lack of knowledge is simply that her being in Fake Barn Country has rendered P_B unreliable.

But what happens when Laura leaves Fake Barn Country and returns to more epistemically friendly circumstances? Naturally, we want to say that her ability to *knowingly* identify barns is at some point restored. One trip through Fake Barn Country cannot be supposed to permanently destroy this ability. The reason we want to say this is clear enough. Although P_B fails to be truth-conducive in the context of Fake Barn Country, we suppose that outside of that context it is highly truth-conducive. So, we want to say that, as far as global reliability is concerned, the beliefs P_B produces outside of Fake Barn Country should constitute knowledge.

The point I want to make is that Wally's case is like Laura's in this respect. His being slow-switched from Dry Earth to Earth is analogous to Laura's leaving Fake Barn Country. He is

[56] Note the "almost" in the previous sentence. I hedge because Laura's belief might in fact be unsafe, depending on how we understand safety. This will be the case if we adopt something like Manley's "revised safety" (Manley 2007) or even my own SF (see Ch. 2). However, I do not believe that safety is necessary for knowledge (for alleged counterexamples, see Neta and Rohrbaugh 2004, Comesaña 2005, Kelp 2009, Bogardus 2014). So I'm inclined to agree with Brown that appealing to global reliability is likely the best way to explain why Laura doesn't know. Incidentally, even if we do want to say that Laura doesn't know because her belief is unsafe, this would not be fatal to Brown's strategy for defending GR. For, as she points out (2004, p. 124-126), there are other examples in which a subject's lack of knowledge cannot be explained by a local reliability condition but *can* be explained by GR (e.g. McGinn 1984, last full paragraph of p. 534).

escaping a deceptive environment and entering a more epistemically friendly one. So, if for this reason we want to say that Laura's ability to knowingly identify barns is restored when she leaves Fake Barn Country, we should say the same thing about Wally's ability to knowingly self-ascribe water-thoughts when he is slow-switched to Earth. In particular, we should say that, as far as global reliability is concerned, the beliefs P* produces on Earth (including B_Y) should constitute knowledge.

3.2.3 An Objection Considered: The Case of New Laura

One might object that there is an important difference between Laura's case and Wally's. Though it's true that P* is highly truth-conducive on Earth, Wally spends most of his time on Dry Earth – that is, in circumstances in which P* is *not at all* truth-conducive. Laura, on the other hand, is only "passing through" Fake Barn Country. The implication is that most of *her* time is spent in more epistemically favorable circumstances. So, one might reason, a more apt comparison would be to a case in which Laura spends most of her time in Fake Barn Country. We might, for instance, suppose that she lives there and that, though she must sometimes go out in order to buy groceries or to run other errands, she leaves very infrequently and for only brief amounts of time. But then it is not at all clear that Laura regains her ability to knowingly identify barns in those rare instances when she leaves Fake Barn Country.[57] For however truth-conducive P_B (the process type responsible for Laura's perceptual demonstrative barn beliefs) might be in those particular circumstances, it is no longer truth-

[57] For now, we'll just say that one "leaves" Fake Barn Country when, roughly, one reaches a point beyond which there are no more deceptive barn façades. This issue will be discussed in more detail below.

conducive *overall*. And that seems to make a difference. But, one might conclude, if it makes a difference here, then it makes a difference in Wally's case as well. So, although P* is truth-conducive on Earth, we should still deny that Wally's belief B_Y constitutes knowledge. And the reason is the same as before: because P* fails to be truth-conducive overall.

Let's call the Laura of this new version of the case "New Laura" (reserving the name "Laura" for the subject of the original version of the thought experiment). I find it plausible that New Laura can at no point knowingly identify barns. But even granting this, I do not think the objection succeeds. One reason is that the objection appears to assume a principle we might call "Context Insensitivity".

> *Context Insensitivity (CI)*
> A given belief-formation process type is truth-conducive in the relevant sense *only if* it has produced more true beliefs than not over the entire range of its actual use.

From CI and *Global Reliability*, it follows that

> A token belief B is **globally reliable** *only if* the process leading to B is a process token whose relevant process type has produced more true beliefs than not over the entire range of its actual use.

If this is true, we get the conclusion that neither Wally's belief B_Y nor any of New Laura's perceptual demonstrative barn beliefs are globally reliable.[58] If they are not, then they violate GR and therefore fail to constitute knowledge.

[58] Interestingly, unless we suppose that *all* of the barns in Fake Barn Country are fakes, it's *possible* (though statistically unlikely) that P_B satisfies *Context Insensitivity* – that is, that it has produced more true beliefs than not overall. For it's possible that by sheer chance New Laura never, or only very rarely, attempts to identify what turn out to be barn façades. In that case, P_B will not produce very many false beliefs – perhaps none. But it's my sense that this would be epistemically irrelevant. If our initial reaction to New Laura's case was that she doesn't know, I don't think this kind of statistical anomaly would give us much reason to change our minds. This is perhaps more evidence that if New Laura's perceptual demonstrative barn beliefs do not constitute knowledge, it is not because the process responsible for them violates a principle like *Context Insensitivity*.

However, I believe that *Context Insensitivity* is false. In conjunction with *Global Reliability*, it implies that a token belief should be considered globally reliable only if the process type responsible for it has had a good track record up to the point at which the belief was formed. But I doubt that this is the right way to think about reliability.[59] Consider an analogy.[60] Suppose a manufacturing company purchases a new thermometer, one that has just come off the assembly line. They plan to put it into one of their industrial strength freezers in order to monitor its temperature. However, the thermometer was not designed to operate in temperatures that low. Failing to realize this, the company installs the thermometer anyway. After a few weeks of consistently false readings, the employees finally notice that something is amiss. They uninstall it, and one of the employees takes it home to use in his garage. The thermometer now functions perfectly well.

Now, up to this point the thermometer has produced far more false readings than correct ones. It does not have a good track record. Nevertheless, we want to say that the thermometer is reliable. It is certainly perfectly reasonable for the employee to rely upon it to tell him the temperature in his garage. Indeed, it's clear that he can come to *know* the temperature in the garage by consulting it. If so, then reliability is not a matter of track record in the way that CI suggests.

---

[59] Frise (2018) considers and rejects a very similar conception of reliability, what he calls "Historical Reliability" (Frise 2018, p. 928). Historical Reliability differs from the conception of reliability considered here only in that it takes a good track record to be both necessary *and* sufficient for reliability.

[60] The following analogy is inspired by Alston's thermometer-on-the-sun example. He notes that "[i]f I claim that my thermometer is reliable, it is no refutation to point out that it would not give an accurate reading on the sun" (Alston 1995, p. 10). *His* point, however, is that "[w]hen I make a judgment of reliability – whether for an instrument, a documentary source, a psychological mechanism, or whatever – I have in mind, at least implicitly, a range of situations with respect to which the claim is being made. What happens outside that range *is simply irrelevant to the claim*" (Ibid., emphasis added). I am not here endorsing this latter claim (nor am I denying it). My point is simply that *Context Insensitivity* is false.

Not only does CI presuppose a counterintuitive conception of reliability, but it also appears to give the wrong verdict in a number of cases. Suppose, for example, that in New Laura's version of Fake Barn Country there are *only* barn façades – no real barns at all. Suppose further that, on those rare occasions when she is outside of Fake Barn Country, she never attempts to identify any of the real barns she happens to encounter. Perhaps she simply never notices them. From these two suppositions it follows that New Laura has *never* successfully identified a barn, that P_B's outputs have to date all been false. But now suppose that, after living there her whole life (let's say, 30 years), New Laura moves far away from Fake Barn Country (at time t). For good measure, we might add that Fake Barn Country is then razed to the ground. No more deceptive barn façades. Given this, what should we say about New Laura's ability to knowingly identify barns? Surely, she will *at some point* regain (acquire?) this ability. This much an advocate of CI could concede. But, assuming that (i) New Laura had attempted to identify at least one barn before t (that is, the number of P_B's lifetime outputs by t is > 0) and that (ii) the rate at which she attempts to identify barns is no greater after t than it is before t, it follows from CI (in conjunction with *Global Reliability* and GR) that New Laura won't be able to knowingly identify barns for roughly *30 years* after t! For, given assumptions (i) and (ii), that's how long we can expect it to take for P_B to be such that it has produced more true beliefs than not over the entire range of its actual use. Of course, we should expect *some* period of transition. It is probably not the case that New Laura acquires the ability to knowingly identify barns *immediately* upon Fake Barn Country's razing. But it is implausible that she must spend 30 years balancing P_B's ledger before this can happen.

At this point, one might respond as follows: Perhaps *Context Insensitivity* is false. But that is irrelevant. For the *real* reason that New Laura cannot knowingly identify barns has been obscured by our description of the case. Up until now, we have been talking about "leaving" Fake Barn Country as if it were unproblematic, as if "Fake Barn Country" named a subdivision or some other place with clear physical boundaries. But Fake Barn Country is not really a *place* at all. It is better thought of as a set of epistemic circumstances. And what does it take to "leave" these circumstances? It is clear that simply reaching the nearest physical location beyond which there are no more deceptive barn façades is not enough. For it is not plausible that this *by itself* changes one's epistemic situation in any meaningful way. (For what if one immediately turns back? Can one *really* be said to have left Fake Barn Country in that case?) Admittedly, what is required beyond this is unclear. Imagining the smallest possible circle that contains all barn façades, perhaps it is simply a matter of spending enough time outside of this circle. Perhaps in addition one must encounter a certain number of real barns before one can be said to have "left" Fake Barn Country. Whatever the criteria, it's clear that New Laura never *really* leaves. This is why she cannot knowingly identify barns even when she is supposedly "outside" of Fake Barn Country.

One might go on to argue that we can say the same thing about Wally. Recall that we have already stipulated that Wally spends most of his time on Dry Earth. We might add to this that his trips to Earth are brief and infrequent. Then, we might doubt that these trips materially affect his epistemic circumstances. We might think that just as New Laura never *really* leaves Fake Barn Country, Wally never *really* leaves Dry Earth. And since P* is not truth-conducive on

Dry Earth, it would follow that B_Y does not constitute knowledge (despite the fact that it is formed while Wally is physically located on Earth).

I think this line of reasoning is half right. I find it plausible that if New Laura is unable to knowingly identify barns, it is because she never really leaves Fake Barn Country. But the idea that B_Y fails to constitute knowledge because Wally never really leaves Dry Earth cannot be correct. Remember, Wally is slow-switched to Earth. And slow-switching is *slow*. It takes time for shifts in one's conceptual repertoire to occur. So, if Wally is not left on Earth long enough, then the shift will not have time to occur and he will not think any *water*-thoughts. In particular, he will not entertain thoughts like T_Y or hold beliefs like B_Y. If, on the other hand, he *is* left on Earth long enough for the shift to occur, then it becomes implausible to say that he has not "really" left Dry Earth. In short, if our second attempt at a natural-kind version of the illusion argument is to work, we must suppose that Wally is left on Earth for an extended period of time – enough time for a conceptual shift to occur. But then we cannot say he never "really" leaves Dry Earth.

For these reasons, I believe that the point I made at end of the previous section (3.2.2) still stands. If we suppose that B_Y is formed after Wally is slow-switched to Earth, then, as far as GR is concerned, B_Y should constitute knowledge. That is simply because the process-type responsible for B_Y, P*, is highly truth-conducive in those circumstances. And, as I have just tried to show, the analogy to New Laura's case doesn't give us much reason to doubt this. For the best explanation for New Laura's lack of knowledge (viz. that she never *really* leaves Fake Barn Country) clearly does not apply to Wally and his belief B_Y. Thus, it looks like our second attempt at a natural-kind version of the illusion argument is not going to work.

3.2.4 Third Attempt

The third attempt requires a bit of background. Let's return for a moment to the original version of the illusion argument outlined in section 3.1. As Brown anticipated, premise I2 of the argument has borne the brunt of the criticism (Dierig 2010b, Morvarid 2013). After all, we do not yet have a solution to the generality problem. That means that we have no noncontroversial way to determine relevant process types independent of our intuitions about which beliefs do and do not count as knowledge. But then there is no non-tendentious way to defend I2. For compatibilists will have the intuition that B_W constitutes knowledge, incompatibilists will have the opposite intuition. Naturally, then, they will type the relevant belief-formation process differently.

An incompatibilist might respond that the content of the false-second order beliefs produced by P is similar enough to the content of B_W that it is *prima facie* reasonable to suppose that, however the generality problem resolves itself, P will turn out to be the process type relevant to the global reliability of B_W. But what exactly *is* the content of the false second-order beliefs that P is supposedly producing? Notice that when Sally suffers an illusion of thought, she will *not*, for some particular wasp *x*, falsely believe that she is thinking that *that wasp (x) is near*. She will not, in other words, believe a false analogue of B_W. For in an illusion situation, there is no particular wasp to which she might refer. This means that whatever it is that Sally falsely believes, it will differ from B_W in both form and content. Brown, for instance, says that what Sally falsely believes is "that she is thinking about a wasp to the effect that it is near" (2004, p. 115, 118). But, as Morvarid has pointed out, if this amounts to the belief that (∃x)(x is a wasp ^ I [Sally] am thinking that *x is near*), then we have a belief that is not just about

Sally's thought contents but about external affairs as well (Morvarid 2013, p. 309-310).[61] Specifically, we have a belief to the effect that there exists a wasp that Sally is related to in a particular way. That is quite different from the content of B_W.

Morvarid (2013) has persuasively argued that the illusion argument can be reconstructed to get around this problem.[62] Following Alston (1995), he argues that the process type relevant to the global reliability of a belief B is determined by the function instantiated by the narrow (i.e. non-intentionally described) psychological mechanism that produced B. What, then, is the process type relevant to B_W's reliability? Let "M" refer to the narrow psychological mechanism responsible for mapping T_W onto B_W – the one that takes T_W as input and outputs B_W. Then, the process type relevant to B_W's reliability is determined by the function that is, as a matter of fact, instantiated by M. Let's suppose that the function instantiated in this case is one that maps *T_W-type* states onto *B_W-type* states[63], where a T_W-type (B_W-type) state is just one that has the same narrow causal-functional profile as T_W (B_W).[64] We can call this function *f*(T_W-type, B_W-type). Then, the process type relevant to B_W's reliability is

[61] We might think that what Brown is saying is that Sally falsely believes that *I [Sally] am thinking that (∃x)(x is a wasp ^ x is near)*. But this belief would be true. And, in any case, Brown explicitly admits that the relevant belief "is not of the cogito form (I believe that I think that *p*)" (Brown 2004, p. 131).

[62] Disclaimer: In what follows, I have adapted Morvarid's argument to better integrate it into the structure of the dialectic as I have set it up. I do not believe that anything crucial has been lost in translation, but it is worth noting that changes have been made. The biggest change is that I have dispensed with the assumption that when Sally suffers an illusion of hearing a wasp she is "intrinsically the same" as when she really does hear a wasp (Morvarid 2013, p. 313). I do not think his argument requires this assumption. Mostly, though, I have simply used different terminology to explain (what I take to be) the same ideas.

[63] How we characterize the function instantiated by M is not very important (we could just as easily have characterized the relevant function as one that maps T_W-type *and other similar mental state types* onto B_W-type states). What will be important going forward is that a difference in mechanism means (or at least suggests) a difference in function.

[64] So, T_W counts as an instance of a T_W-type state in the same way that a particular pain counts as an instance of the type *pain* (that is, in virtue of having the same narrow causal-functional profile characteristic of that type).

just the one that takes T_W-type states as input and outputs B_W-type states. Let's call this process type " P_f ".

Now, suppose that shortly after forming belief B_W, Sally suffers an illusion of hearing a wasp. Earlier I said that when this happens, she will think an empty analogue of T_W. Let's refer to the one she entertains in this case as " T_W* ". Morvarid notices that when Sally suffers an illusion of hearing a wasp she will *also* think an empty analogue of B_W. Let's refer to the one she entertains in this case as " B_W* ". Importantly, although both T_W* and B_W* lack any representational content, they have the same *narrow* causal-functional profiles as their genuine counterparts. Thus, T_W* (B_W*) is an instance of a T_W-type (B_W-type) state. This suggests that the same narrow psychological mechanism responsible for mapping T_W onto B_W, M, is also responsible for mapping T_W* onto B_W*. So, since the function instantiated by M is *f*(T_W-type, B_W-type), it follows that the process type responsible for B_W* is P_f.

Thus, Morvarid continues, given the frequency with which Sally suffers illusions of hearing a wasp, P_f will produce many states like B_W*. Referring to states like B_W* as *pseudo-thoughts*, which he characterizes as mental states that lack any representational content but have the same narrow causal-functional profiles as their genuine counterparts, Morvarid then argues that P_f is not truth-conducive because it produces too many pseudo-thoughts. He therefore assumes that

> *Strong Reliability (SR)*
> A given belief-formation process type is truth-conducive in the relevant sense *only if* it is such that were it exercised sufficiently many times, the ratio of true beliefs it would produce to failed beliefs would surpass a certain threshold

where a failed belief is just one that is either false or a pseudo-thought.[65] Supposing that Sally suffers illusions frequently enough so that P_f fails to satisfy the consequent of SR, we end up with the conclusion that B_W is not globally reliable and therefore fails to constitute knowledge.

In short, Morvarid argues that as long as we adopt SR and replace "P" with "P_f" in premises I2 and I3, we get a version of the illusion argument that avoids the problems that plague Brown's original. What's important for our purposes is that Morvarid's reconstruction also opens up the possibility of a third way to construct an analogue to the wasp example.

For suppose that Wally, after having spent enough time on Earth to acquire the concept *water*, is now back on Dry Earth. And suppose that he has been back long enough to have reintegrated into his old linguistic community – one where the utterance "water" fails to refer. Now, as I discussed in Chapter 2 (section 2.2), most externalists deny that being slow-switched causes one to lose previously acquired concepts. Let's suppose that's correct. Then, Wally retains the ability to think about water despite having been returned to Dry Earth. Arguably, given his new linguistic environment, the *presumption* is that when he utters sentences like "water is refreshing" we should interpret him as having thereby expressed a pseudo-thought. But, following Sanford Goldberg (2005b), we might argue that this presumption is defeated in cases where Wally specifically *intends* to be referring to a substance that happens to be water (whether he knows it or not).[66] In these cases, it is plausible that we should interpret him instead as having expressed a genuine thought about water. Thus, when paired with certain intentions, Wally's utterances of "water is refreshing" will express genuine first-order thoughts.

[65] I will not go into his argument here, but Morvarid defends SR on pages 318-320 of his 2013.
[66] See (esp. the first two pages of) Appendix A for a more detailed explanation of the assumptions being made here.

Otherwise, however, they will express pseudo-thoughts. Given this, a third attempt at an analogue case might go as follows:

Suppose that while still on Earth, Wally had a swim (in water) to cool down after a strenuous hike. Sometime after having been slow-switched back to Dry Earth, he begins reminiscing about that swim and how pleasant it was. Prompted by this memory, he sincerely utters the sentence "water is refreshing". Given that Wally clearly intends to be saying something about the kind of stuff in which he remembers swimming, it is plausible to interpret him as having expressed a genuine first-order thought to the effect that *water is refreshing*. Let's call this token thought "T_Z". Now suppose that via introspection he then forms the second-order belief that he is thinking that *water is refreshing*. Call this token belief "B_Z". What process type is relevant to B_Z's reliability? Let " M* " refer to the narrow psychological mechanism responsible for mapping T_Z onto B_Z. Then, the process type relevant to B_Z's reliability is determined by the function that is instantiated by M*. Without attempting to specify it, let's call the function instantiated in this case " *f** ". Then, the process type relevant to the global reliability of B_Z is P_{f*}.

Now, suppose that shortly after forming belief B_Z, Wally sincerely utters another token of the sentence "water is refreshing". This time, however, his utterance is not prompted by any specific memory. He has no particular bits of watery stuff in mind that would allow us to interpret him as intending to refer to this or that specific kind. Thus, there is nothing to defeat the presumption that we should interpret Wally's utterance in accordance with the semantic norms of his current linguistic community. And since "water" fails to refer according to those semantic norms, the mental state corresponding to Wally's utterance has no determinate

content (given the illusion view). So, it fails to constitute a thought. Wally has therefore expressed a pseudo-thought, one we will call " T_Z* ". This will lead him to entertain another pseudo-thought, one he would express by uttering "I am thinking that water is refreshing" – call it " B_Z* ".

Now, although both T_Z* and B_Z* lack any representational content, they have the same narrow causal-functional profiles as their genuine counterparts. This suggests that the same narrow psychological mechanism responsible for mapping T_Z onto B_Z, M*, is also responsible for mapping T_Z* onto B_Z*. So, since the function instantiated by M* is *f**, it follows that the process type responsible for B_Z* is P_{f*}.

The third attempt concludes by having us suppose that Wally tends not to tie his utterances of "water" to any specific intentions (i.e. that T_Z is an anomaly). The result would be that if P_{f*} were exercised sufficiently many times, it would produce more beliefs like B_Z* than like B_Z (so that its ratio of true to failed beliefs would not surpass the relevant threshold). But then it would not satisfy the consequent of SR and would therefore fail to be truth-conducive. So, given that P_{f*} is the process-type relevant to the global reliability of B_Z, B_Z would not be globally reliable and would therefore fail to constitute knowledge.

I believe that there are two big problems with the line of reasoning sketched in the previous four paragraphs. The first is simply that T_Z* does *not* have the same narrow causal-functional profile as T_Z. T_Z interacts with a set of memories, intentions, and (arguably) beliefs in a way that mental states like T_Z* do not. Indeed, it is for precisely this reason that T_Z counts as a genuine belief about water while T_Z* doesn't. But if T_Z* has a different causal-functional profile than T_Z, then it is not obvious that M* is also responsible for mapping T_Z* onto B_Z*. A

different mechanism may be at work in that case. But a different mechanism would instantiate a function other than *f**. Then the process type responsible for B_Z* would not be P_{f*} and so not the one relevant to the global reliability of B_Z. So the fact that P_{f*} fails to be truth-conducive would simply be irrelevant.

Perhaps there is a way to get around this problem. For example, we might try to revise the case to get a version of T_Z* that *does* interact with memories and intentions in the way that T_Z does.[67] But even if that were successful, there is another problem. In order to get the third attempt off the ground we have had to assume a particular solution to the generality problem – namely, Alston's (1995) psychological approach. But it is not obvious that Alston's solution is correct (see e.g. Conee & Feldman 1998, p. 11-13). The jury is still out on the psychological approach. So the problem here is one that plagues the illusion argument more generally: the fate of the argument invariably depends on how the generality problem is resolved.

3.3 Conclusion

In this chapter, I have run through three ways that one might attempt to generalize Brown's illusion argument to apply to natural-kind externalism. None, I argued, succeed.

[67] Without going too far into the matter, I doubt such a strategy would work. The new version of T_Z* would still have to be a pseudo-thought. So, for example, if we supposed that it were prompted by a memory similar to the one that prompts T_Z, it couldn't be a memory of water. It would have to be a memory of an illusion of water, a mirage. But then (almost?) all T_Z*-type states would be ones prompted by memories of events that took place on Dry Earth. But, if so, I suspect that T_Z*-type states would interact with different sets of beliefs than T_Z-type states. To see this, consider how Wally's doxastic life would change were he to learn about his having been slow-switched. Plausibly, the network of beliefs built around a state like T_Z* would need to be revised in a way that the network of beliefs build around T_Z wouldn't. This suggests that the networks were different in the first place. But then, once again, we end up with the conclusion that T_Z- and T_Z*-type states have different narrow causal-functional profiles. Of course, the matter is complicated. I'm sure that an advocate of the illusion argument could come up with some response. But doing so would not be easy. It would take us into areas of inquiry quite different than the ones we started with. One starts to wonder if the illusion argument could possibly be worth all that trouble!

Simply placing Wally in an illusion-inducing environment like Dry Earth does not work (first attempt). For then we do not get an analogue of Sally's belief B_W – that is, a belief that should, given TWP, constitute knowledge but that doesn't because it fails to be globally reliable. And while introducing slow-switching gets us closer to the structure of the original wasp example (second and third attempts), it also generates problems absent from the latter. In particular, it becomes doubtful that the false (or failed) second-order beliefs induced on Dry Earth are relevant to the global reliability of the target belief (viz. B_Y in the second attempt, B_Z in the third attempt). I conclude that it is unlikely that the illusion argument can be generalized in this way. In any case, doing so is not as straightforward as Brown assumes.

Chapter 4: The Memory Argument

In his 1988 paper on externalism and self-knowledge, Tyler Burge admits that a person who learns of having been slow-switched may ask, "'Was I thinking yesterday about water or twater?' – and not know the answer" (Burge 1988, p. 659). This remark has set off a lively debate around what has come to be known as the "Memory Argument". For it suggests that, given externalism, a subject may lack access to the contents of yesterday's propositional attitudes. But, if that is correct, then it looks like a subject may lack access to *today's* [read: conscious, occurrent] propositional attitudes as well. As Boghossian famously puts it: "It is not as if thoughts with widely individuated contents might be easily known but difficult to remember. The only explanation […] for why S will not know tomorrow what he is said to know today, is not that he has forgotten but that he never knew" (Boghossian 1989, p. 23).

The primary aim of this chapter is to evaluate Boghossian's argument. I will argue that it does not succeed. I make my case in two stages. First, I draw on existing criticism to show that the original 1989 version of Boghossian's argument fails. As we will see, it fails because it relies on false premises about memory. Second, I consider the possibility that these premises are incidental to the argument and that it can be reconstructed without them. This idea was originally proposed by Sanford Goldberg (1997, 2003a).[68] He has shown that a subject may not know of a thought she is *currently entertaining* whether it is about water or twater (1997, 2003a, 2003b). Conceding the possibility, I consider and reject two reasons to think that one

[68] Goldberg begins to have doubts about this by the time he writes his 2003a. By his 2003b, he has abandoned it. However, the idea that externalism means that one may be unable to identify which concepts figure in one's thoughts remains throughout.

must know which concept figures in one's thought in order to know what one is thinking. In the final section of this chapter, I argue that subjects typically do know which concepts figure in their own conscious, occurrent thoughts. However, I concede that this knowledge is empirically defeasible. This fact has important implications for McKinsey's reductio, which I discuss in the next chapter.

4.1 The 1989 Original

Suppose that Oscar is a victim of slow-switching. At time t_1, during a stint on Earth, he introspectively forms a true second-order belief to the effect that he is thinking that water is refreshing. According to Boghossian, it is "quite clear" that Oscar will not know at t_2 what he was thinking at t_1 (Boghossian 1989, p. 23). But, Boghossian argues, since we can assume that Oscar hasn't *forgotten* what he was thinking at t_1, we should conclude that Oscar won't know simply because he never knew. A commonly cited reconstruction of this argument goes as follows, where W is the proposition that Oscar is [at t_1] thinking that water is refreshing.

(M1) If a subject S forgets nothing, then what S knows at t_1, S knows at t_2.

(M2) Oscar forgot nothing.

(M3) Oscar does not know that W at t_2.

(M4) So, Oscar does not know that W at t_1.[69]

If this argument is sound, then TWP is false. For M4 suggests that one may not be in a position to introspectively know the contents of a first-order propositional attitude *even as one consciously entertains it*.

[69] See Ludlow (1995b).

Whether this argument is successful is somewhat difficult to assess because a lot hinges on how we fill in the details. In one version of the case, Oscar is not informed of his having been slow-switched; in another version, he is so informed. Suppose he is not informed and at t_2 thinks to himself a thought he would express by uttering, "I remember thinking [at t_1] that water is refreshing". What is the content of this thought? That might depend on where Oscar happens to be at t_2. If he happens to be on Earth, then he is probably thinking a *water*-thought. But what if he is on Twin Earth at t_2? Then we have to consider what the externalist should say about the contents of *memories*. Do they shift as one is slow-switched, or does memory preserve the content of the original thought?

What we ought to say about the argument depends on how we answer these questions. In particular, it depends on (a) where Oscar is when he recalls (or attempts to recall) the thought he entertained at t_1, (b) the assumptions we make regarding what the externalist ought to say about memory contents, and (c) whether or not Oscar is told about his history of slow-switching. In the remainder of this section, I draw on existing criticism to make the case that the argument faces serious problems regardless of how we fill in the details – that at least one of M1-M3 ends up being false. This sets the stage for section 4.2 where I begin to discuss the possibility of a memory-free version of the argument. See Table 1 for a visual summary of the various versions of the 1989 Memory Argument along with brief descriptions of where each goes wrong.

Version 1. Let's first consider the prospects for the argument if we suppose that Oscar is told nothing of his slow-switching history. At time t_2, Oscar, seeming to recall his thinking at t_1, forms a belief he would express by uttering, "I was thinking that water is refreshing". Does this

Table 1: Four Versions of the Memory Argument. This table provides a visual summary of the various versions of the Memory Argument along with brief descriptions of the problems facing each one. For example, on the version of the argument where we assume that memory contents shift as one is slow-switched, Oscar is not told of his history of slow-switching, and he happens to be on Twin Earth at t_2, the primary issue is with premise M2.

<table>
<tr><td></td><td colspan="2">Not Told (Version 1)</td><td>Told (Version 2)</td></tr>
<tr><td rowspan="2">If contents shift:</td><td>(A) On Earth</td><td>(B) On Twin Earth</td><td rowspan="3">• M1 is false (Brueckner 1997, Burge 1998, Kobes 2003).</td></tr>
<tr><td>• M1 is false (Brueckner 1997).</td><td>• M2 is false (Gibbons 1996, Brueckner 1997).</td></tr>
<tr><td>If contents are preserved:</td><td colspan="2">• M3 is false (Burge 1998).</td></tr>
</table>

belief constitute knowledge? How we respond here depends on what we think the externalist ought to say about memory. We have two options. First, we could say that memory functions to *preserve* the content of Oscar's belief so that the belief Oscar expresses at t_2 is the belief that

(W) I was thinking that water is refreshing.

Second, we might hold that the contents of one's memories shift as one is integrated into a new linguistic environment. Then, what Oscar believes at t_2 will depend. If he is on Earth at t_2, he will believe that W. But if he is on Twin Earth at t_2, then the belief he expresses is the belief that

(W*) I was thinking that twater is refreshing.

If we take the first option, Oscar's belief at t_2 is true. If we take the second option, it *might* be false.

Burge (1998) defends the first option, invoking what he calls *preservative memory* to make the case that Oscar's memory-based second-order belief will be correct even if he is in a different linguistic environment at t_2 than he was in at t_1. The main idea is that a memory-based second-order judgment (that is, a judgment about what one *was* or *had been* thinking) inherits the content of the first-order thought it purports to recall in virtue of its causal connection to the latter (see esp. Burge 1998, p. 357-360).[70] If Burge is right about this, then it is not obvious that M3 is true – that Oscar *doesn't* know that W at t_2. The burden of proof would seem to lie with those who would maintain that a well-functioning, reliable memory

[70] Burge's account of preservative memory recalls the redeployment thesis endorsed by Gibbons (1996) and Peacocke (1996). See section 2.2 above. Note that aberrant causal chains will cause preservative memory to fail. We assume, therefore, that Oscar's memory is functioning properly and that his memory-based judgment at t_2 is connected in the right way to the thought he entertained at t_1. Otherwise, M2 is false.

whose accuracy one has no reason to doubt might nonetheless fail to be source of knowledge about one's past thoughts.[71]

But what if Burge is wrong? Some have argued that if externalism is true, then the contents of one's memories will shift as one is slow-switched (Ludlow 1995b & 1998, Tye 1998). This means that what Oscar believes at t_2 depends on where he is, whether on Earth or Twin Earth. Suppose he is on Twin Earth at t_2. In that case, Oscar falsely believes that W* at t_2. Obviously, then, Oscar does not know (since he doesn't even *believe*) that W at t_2. So, M3 is true on this version of the argument. But, as Gibbons (1996) and Brueckner (1997) point out, it now looks like M2 is false. It looks like the shift in content has caused Oscar to *forget* what he was thinking at t_1.

What if we suppose that Oscar is on Earth at t_2? In that case, he will correctly believe that W at t_2. Plausibly, however, he still doesn't *know* that W at t_2. Let's grant that he doesn't. The success of this iteration of the argument then comes down to premises M1 and M2. One could make the case that M2 is false – that even though Oscar's belief at t_2 correctly represents what he thought at t_1, he nonetheless doesn't genuinely *remember* what he was thinking at t_1. This is plausible if we suppose that Oscar has been slow-switched to Twin Earth and back in the time between t_1 and t_2. For given the corresponding shifts in the contents of Oscar's memories, we may want to say that the causal connection between Oscar's second-order judgment at t_2

[71] Or, rather, fail to *preserve* or *maintain* self-knowledge originally acquired by other means (cf. Dummett 1993, p. 420-421; Audi 1997, p. 410).

and his first-order thought at t_1 is too aberrant for the former to constitute a genuine memory of the latter (Brueckner 1997, n. 24).[72]

But even if we set this concern to the side and grant that M2 is true, there is good reason to think that M1 is false. Notice that M1 is false in general. For a subject might not know at t_2 what she knew at t_1, not because she has forgotten anything, but simply because her knowledge has been *defeated*. Do we have reason to believe that Oscar's knowledge has been defeated? It seems that we do. For the fact that Oscar is a victim of slow-switching seems in this case to constitute what Bernecker calls a "factual defeater" (Bernecker 2009, p. 75). A factual defeater, according to Bernecker, is one that defeats knowledge (or justification) just in virtue of being true.[73] A subject's being in Fake Barn Country, for instance, may constitute a factual defeater of her true belief that a barn stands before her. For given the circumstances her belief's being true is just a matter of luck. She could easily have been looking at a barn façade instead, in which case she would have believed falsely. Therefore, she doesn't know that a barn stands before her. Brueckner (1997, p. 9-10) argues that we can say something similar in Oscar's case. For given his history of slow-switching, it is just a matter of luck that his second-order belief at t_2 is true. He could easily have been on Twin Earth at t_2, in which case he

[72] Note that this won't be a problem if we assume that Oscar is not slow-switched at any point between t_1 and t_2. But, this assumption won't help with the concerns about M1 we're about to consider.

[73] This is in contrast to what he calls a "doxastic defeater", which "is a proposition that one *believes* to be true and that indicates that one's belief that p is either false or unreliably formed or sustained" (Bernecker 2009, p. 75; emphasis added). This kind of defeater is relevant to the next version of the Memory Argument that we will consider, the version where Oscar is told about his history of slow-switching.

would have believed falsely.[74] Thus, we can say that Oscar *did* know that W at t_1 but that, due to the presence of factual defeaters, his memory has failed to preserve that knowledge to t_2.[75]

Version 2. Let's now consider the version of the argument where Oscar *is* told of his history of slow-switching. Suppose that at t_2 Oscar is told (and, we can suppose, comes to *know*) that he has been slow-switched to and from Twin Earth at random intervals since before t_1 and that the contents of his "water"-thoughts[76] shift every time this happens. He is not told where he is now (t_2) or where he was at any specific point before t_2. Now, suppose Oscar is prompted to recall the *event* of his thinking about water at t_1 and is then asked, "Were you then [at t_1] thinking a *water*-thought or a *twater*-thought?"

In this version of the case Oscar is explicitly asked to identify the concept that figured in his earlier thought. Clearly, Oscar cannot knowledgably do so. He cannot do so because he does not know which concept it was. For ease of exposition, let's just say that at t_2 Oscar lacks *concept (or C-) knowledge* of his earlier thought. But what does this mean? Well, it *suggests* that M3 is true, that Oscar doesn't know that W at t_2. For one reasonably thinks: *If at t_2 Oscar*

[74] Remember, the version of the argument currently under consideration assumes that memory contents shift as one is slow-switched. So, had Oscar been on Twin Earth at t_2, he would have falsely believed that W* at t_2. Notice also that had Oscar been on Twin Earth at t_1 and had at that time been thinking instead about twater (a relevant alternative), his belief that W at t_2 would be false. Oscar's belief at t_2 therefore violates relevant alternative conditions like RR (see section 2.2 of this work). Nagasawa (2002) rejects the Memory Argument on similar grounds. However, in that paper he adopts a conception of memory according to which remembering something entails knowing it. This leads him to identify M2 rather than M1 as the faulty premise.

[75] I suspect that appealing to "factual defeaters" to explain a subject's lack of knowledge amounts to little more than a diagnosis of exclusion – something we can point to when we are unable or unwilling to explain it by appealing to the violation of a more specific epistemic principle (safety, global reliability, some relevant alternatives condition). But a diagnosis of exclusion will do for present purposes. For whatever more specific principle we decide best explains why Oscar doesn't know that W at t_2, it is clear that *some* other condition must be satisfied (other than simply not forgetting) in order for memory to preserve knowledge from one time to another. This means that M1 is false as is. But if we add a second conjunct to M1's antecedent in order to account for this other condition, then the corresponding version of M2 will be false. Again, see Brueckner 1997, p. 9-10.

[76] That is, thoughts he expresses or would express via use of the word "water".

doesn't know that at t_1 he was thinking a water-*thought, then in what sense can it possibly be true that he knows that W at t_2*?

Let's grant for a moment that Oscar's lack of C-knowledge is sufficient for M3. Will the Memory Argument work given this assumption? Unfortunately for the incompatibilist, there are still substantive issues with premise M1. As several commentators have pointed out, Oscar's being told of his history of slow-switching constitutes a doxastic defeater[77] of his belief that W (Brueckner 1997; Burge 1998, n. 18; Kobes 2003). The idea is that Oscar *does* know that W at t_1, but that this knowledge is defeated when he finds out about his history of slow-switching at t_2. Here again, the objection is that M1 is false because it fails to account for the possibility of defeaters.[78]

This response raises a couple of interesting questions. First, notice that on the assumption that Oscar's lacking C-knowledge is sufficient for his failing to know that W, the implication is that Oscar *had* C-knowledge of his thought at t_1 but that this knowledge was defeated when he found out about his history of slow-switching at t_2. But, of course, this presupposes that externalism is compatible with Oscar's having had this kind of C-knowledge in the first place. More to the point: even if it is compatible, might it nonetheless be possible to construct a case where a subject does not know of a thought she is *currently entertaining* whether it is a *water*- or a *twater*-thought? If so, then it should be possible to construct a version of the Memory Argument that doesn't rely on controversial premises about memory at all. Second, are there any good reasons to suppose that knowledge of content does in fact

[77] See footnote 76.
[78] Bernecker offers an independently motivated counterexample to M1 on p. 78 of his 2009.

presuppose C-knowledge of the kind Oscar lacks? Perhaps, despite its initial plausibility, this thesis is wrongheaded after all. In that case, the Memory Argument fails regardless.

I address both of these questions in the next two sections of this chapter. In the next section, I adapt an argument by Sanford Goldberg to show that it *is* possible to construct a case where a subject does not know of a thought she is currently entertaining whether it is a *water*- or a *twater*-thought. The result is in effect a memory-free version of the Memory Argument. In section 4.3, I argue that even this version of the argument fails. To make my case, I consider and reject two reasons to think that a subject in this situation must know which concept figures in her thought in order to know what she is thinking.

4.2 The Argument from Conceptual Omniscience

Consider the following variation on the slow-switching case. It is a revised version of Sanford Goldberg's "argument from conceptual omniscience" (2003b).

Jane is born on Earth where she trained as a chemist. As a result of her training, she both has the concept *water* and knows how to explicate it (i.e. she knows that the concept applies to H_2O). Suppose that at some point t_1 after her training, Jane is switched to Twin Earth. Once there, she is immediately told about the switch. While on Twin Earth, she learns about XYZ. As a result, by time t_2 Jane knows as much about XYZ as she does about H_2O. In particular, she knows how to explicate the concept *twater*. Now suppose that at t_3, and unbeknownst to her, Jane begins a regimen of slow-switching back and forth between Earth and Twin Earth. At some time t_4 later than t_3, she is told (and, we'll suppose, comes to *know*) that at some point after t_2 she became a victim of slow-switching. She is not told, and does not

know, where she is now, where she was at any point after t_2, nor for how long.[79] Now suppose that as she receives this news she is at the same time looking out over the lake by which she is standing and, with the intention of referring to whatever kind of watery-substance happens to fill this lake, thinking to herself a thought T that she would express by using the words "water is refreshing". Let's stipulate that Jane happens to be on Earth at t_4 and is therefore thinking a *water*-thought. Does Jane know which concept (*water* or *twater*) figures in T?

Notice that Jane cannot knowledgably explicate T *despite knowing how to explicate both of the relevant concepts* (*water* and *twater*). Or, more precisely: Jane cannot knowledgably explicate T with respect to the water-concept that T involves. For short, let's just say that Jane cannot explicate T/c_w. What does it mean to say this? Suppose Jane does articulate T by uttering "water is refreshing". Then, to say that she cannot explicate T/c_w means that, for instance, she could not knowledgably answer if asked the question: "What are the application conditions of the concept expressed by the word 'water', as you just used it?" There are two possibilities: Either the water-concept that figures in T applies to H_2O or to XYZ. Since she does not know where she is, it seems that Jane is not in a position to know which.

But this suggests that Jane does not know which concept, *water* or *twater*, figures in T.[80] For if she did know this, then, knowing how to explicate both *water* and *twater*, she could tell

[79] In the original version of this argument, Goldberg simply stipulates that Jane is "conceptually omniscient". In his words, this means that "given any non-logical and non-indexical expression *E* of English [or Twin-English], Jane can correctly and exhaustively explicate the concept expressed by *E*" (2003b, p. 54). This seems to me like a dangerous stipulation to make. Where did Jane acquire these concepts? If we suppose that she has and can explicate the concept *twater*, does this mean that at some point in Jane's history she spent time on Twin Earth? The worry is that by supposing Jane is conceptually omniscient we many inadvertently commit ourselves to certain presuppositions about Jane's history. I think it best to be explicit about these commitments if we can. That is why I try here to construct a case in which Jane's conceptual omniscience (with respect to *water* and *twater*) is explained and her environmental history is clear.

[80] Notice that this is true even if we suppose that Jane is *not* told of her history of slow-switching. In that case, if she is asked at t_4 to explicate T/c_w, then, not knowing she has been slow-switched, she'll confidently answer "XYZ!"

us whether she is thinking an H_2O-thought or an XYZ-thought. The best explanation for the fact that she cannot is that she does not know which concept figures in T – that, in other words, she lacks C-knowledge with respect to T just as Oscar lacks C-knowledge with respect to his earlier thought that W.

The question now is: Does it follow that Jane doesn't know, or is not in a position to know, that W_{Jane}?

(W_{Jane}) I [Jane] am thinking that water is refreshing.

If so, then TWP is false: it is not necessarily true that one is always in a position to know introspectively what one is thinking. Thus, at last, we come to the crux of the problem.

4.3 Does Knowledge of Content Presuppose C-Knowledge?

I now want to consider two reasons to think that Jane doesn't know that W_{Jane}: (i) the argument from cognitive insignificance and (ii) the argument from the Principle of Knowing Identification. I'll explain both of them before moving on to my response.

4.3.1 The Argument from Cognitive Insignificance

The first reason is that knowledge that W_{Jane} *without* the relevant kind of C-knowledge would appear to be a "cognitively insubstantial" kind of self-knowledge.

Some authors have criticized Burgean basic self-knowledge for this reason (Boghossian 1989; Gertler 2000; Farkas 2008, Ch. 6; Wikforss 2008). They compare basic judgments to one's

(assuming she believes she is still on Twin Earth). But, even if she is correct, this would clearly not be a *knowledgeable* explication. It would just be a matter of luck that she gets it right.

judging *I am now here*. Because of the indexical natures of "now" and "here", the latter judgment is bound to be correct. It is self-verifying, and one knows that it is true. But, of course, it doesn't follow that one knows where one is any meaningful sense. For suppose I am kidnapped, blindfolded, and taken to a remote location. In that case, I may know that *I am now here,* but I do not really know where I am.

Likewise, if I sincerely utter, "I am thinking [with this very thought] that water is refreshing", the self-verifying nature of the judgment thereby expressed will guarantee that I am correct. But, it is said, it doesn't follow that I know what I am thinking in any meaningful sense. For suppose that I am an unwitting victim of slow-switching. In that case, it seems that I will not know what I am expressing by my use of the word "water" any more than I know where "here" is in the kidnapping case. Thus, it is alleged, basic self-knowledge is by itself an insubstantial kind of self-knowledge, if it counts as self-knowledge at all.

The same thing can perhaps be said here. The idea would be that Jane's knowing that W_{Jane} without knowing that she is thinking a *water*-thought would be a bit like one's knowing that *I am now here* without knowing where "here" is. But one can know that *I am now here* without knowing where "here" is precisely because the former is a cognitively insubstantial kind of judgment. Thus, one might think, if one can know that W_{Jane} without knowing that one is thinking a *water*-thought, then judging that W_{Jane} must be similarly insubstantial. So, the argument concludes, since propositional knowledge of content is *not* cognitively insubstantial in the manner of *I am now here* (see e.g. Boghossian 1989, p. 19-20), we ought to hold that Jane cannot know that W_{Jane} without knowing that she is thinking a *water*-thought.

4.3.2 The Argument from the Principle of Knowing Identification

In his 1997, Goldberg offers a principle that appears to entail that Jane doesn't know that W_{Jane}. He calls it *The Principle of Knowing Identification*, or PKI (Goldberg 1997, p. 215).

(**PKI**) If S self-ascribes a thought with a form of words W which is such that

(i) there is more than one relevant[81] interpretation that can be attached to W, and

(ii) S herself has no presently available way to select one over the other as the interpretation she intended,

then S's self-ascription does not count as self-knowledge – because it is not a knowledgeable identification – of the thought in question.

Suppose that at t_4 Jane judges that W_{Jane}, a self-ascription that she would express by uttering the words, "I am thinking that water is refreshing". Now, there is certainly more than one relevant interpretation that can be attached to these words (do they express a thought about water or about twater?). And it appears that at t_4 Jane is in no position to select one over the other as the interpretation intended. Therefore, if PKI is true, we apparently get the conclusion that Jane doesn't know that W_{Jane} precisely because she does not know that she is thinking a *water*-thought (as opposed to a *twater*-thought).

[81] In Goldberg's original formulation, condition (i) begins: "*by S's own lights*, there is more than one interpretation…" (emphasis added). Presumably, Goldberg wanted to avoid endorsing a principle that would require (via condition (ii)) one to rule-out all possible alternative interpretations. In that case, the function of "by S's own lights" is to narrow the scope of PKI so that it applies only to cases in which S knows that there are relevant alternative interpretations. However, if PKI is true at all, it seems to me that it should also hold in cases where there are relevant alternative interpretations that S is *not* aware of. For this reason, I have removed the "by S's own lights" qualifier. But to account for Goldberg's (presumed) worry, I have added the word "relevant" (absent from the original formulation) to condition (i).

But is PKI true? Goldberg offers two examples to illustrate its plausibility (Goldberg 1997, p. 215-216). In both cases, we imagine a subject who has simply forgotten which of several interpretations she had attached to an earlier thought.

BANK

Suppose that at t_2 S recalls having at t_1 expressed a thought with the utterance, "The bank is about four blocks from the station," yet does not remember if she had an effluvial embankment or a financial institution in mind. Now suppose that at t_2 S attempts to self-ascribe her earlier thought by sincerely uttering, "I was [at t_1] thinking that the bank is about four blocks from the station," with the intention of using "bank" to mean whatever she had used it to mean at t_1.

MARYS

Suppose that at t_2 S recalls having at t_1 expressed a thought with the utterance, "Mary is in town," but fails to remember which of the several Marys she knows she was thinking about. Now suppose that at t_2 S attempts to self-ascribe her earlier thought by sincerely uttering, "I was [at t_1] thinking that Mary is in town," with the intention of using "Mary" to refer to whomever she had used it to refer at t_1.

Clearly, in neither case does S's self-ascription at t_2 constitute knowledge. And, plausibly, the reason is precisely that at t_2 S has no way to determine which of several relevant alternative interpretations she had attached to the original utterance at t_1. Thus, it appears, a self-ascription's satisfying PKI's antecedent is indeed sufficient for its failing to constitute knowledge.

In short, BANK and MARYS suggest that PKI is true. And PKI appears straightforwardly to entail that Jane doesn't know that W_{Jane}. Hence, it seems that Jane doesn't know that W_{Jane}.

4.3.3 The Response

I now want to address both of the arguments just sketched, taking the argument from PKI first. My response there will help us to see the mistake behind the argument from cognitive insignificance.

The primary issue with the argument from PKI is simply that Jane's case is not analogous to BANK or MARYS. Let's focus on MARYS. In that case, S is unable to provide *any* interpretation or description that would indicate a particular Mary. This is why we want to deny that S's self-ascription at t_2 constitutes knowledge. But this is not true of Jane. Though she cannot interpret her thought as an H_2O-thought, she is in a position to say *something* about what she means by "water".

To illustrate the point, suppose that the particular Mary S was thinking about at t_1 is an old high school friend whom S happened to run into at a party shortly before t_1. Then, there are at least two interpretations that S could to attach to her use of "Mary" such that, were she to select one of them as the interpretation intended, her self-ascription at t_2 ("I was [at t_1] thinking that Mary is in town") would constitute knowledge.[82] The first:

(Int. 1) The woman I went to high school with

The second:

(Int. 2) The woman I met at the party the other night

Notice that S need not attach *both* interpretations to her use of "Mary" in order for her self-ascription to constitute knowledge. Suppose, for example, that S remembers that the Mary she was thinking about at t_1 is the one she met at the party and, as a result, attaches Int. 2 to her

[82] Assuming, of course, that other non-controversial conditions are met.

use of "Mary" at t_2. But suppose also that S did not recognize her old high school friend and, as a result, fails to realize that Mary [from the party] is *Mary* [from high school]. We would still want to say that at t_2 S knows that she was [at t_1] thinking that Mary is in town. Thus, what's important is not that S is unable to attach this or that particular interpretation to her use of "Mary"; it is that she is unable to provide *any* interpretation.

Is the same thing true of Jane? I don't believe that it is. Recall that at t_4 Jane is thinking a thought about the kind of stuff that fills the lake by which she is standing, stuff which happens to be H_2O. Suppose that as she thinks this thought, she correctly judges that W_{Jane}, a self-ascription she expresses with the words, "I am thinking that water is refreshing". Then, there appear to be at least two interpretations that Jane could attach to her use of "water" such that, were she to select one of them as the interpretation intended, her self-ascription would constitute knowledge. The first:

(Int. 3) The kind of stuff that fills this lake

The second:

(Int. 4) The kind of stuff that fills the lakes on Earth, H_2O

Now, it is unclear why Jane would need to attach *both* interpretations to her use of "water" in order for her self-ascription to constitute knowledge. It would seem to be sufficient that she is in a position to say that Int. 3 is correct. It's true that, because Jane is a victim of slow-switching, she cannot know that water [the kind of stuff in the lake] is *water* [H_2O]. But this by itself shouldn't undermine Jane's self-knowledge any more than S's failure to realize that Mary [from the party] is Mary [from high school] should undermine *her* self-knowledge.

In any event, this marks a significant disanalogy between Jane's case and MARYS (and for similar reasons, BANK). In MARYS, S is unable to provide *any* interpretation of her self-ascription. Jane, however, is able to provide *at least one* (viz. Int. 3). It seems to me that this is enough to establish that Jane *does* know that W_{Jane}. But even if one still wants to deny this, it is clear that comparing Jane's case to MARYS will not work.

The argument can be put another way. MARYS and BANK provide inductive support for PKI only if we read condition (ii) as saying

(ii*) S has no presently available way to select *any* of the relevant possible interpretations as the one she intended.

They do not support a version PKI any stronger than that. For both examples show only that a subject S's self-ascription will fail to constitute knowledge when both (i) and (ii*) are satisfied. But, as we have just seen, Jane does not satisfy condition (ii*). Therefore, it does not follow from PKI that Jane doesn't know that W_{Jane}.

What we ought to say in response to the argument from cognitive insignificance should now be reasonably clear. The concern there was that Jane's knowing that W_{Jane} without knowing that she is thinking a *water*-thought would be like knowing that *I am now here* without knowing where "here" is. But if not knowing where "here" is means being unable to say anything non-trivial about one's location, then the comparison fails. For although Jane does not know that she is thinking a *water*-thought, she is nonetheless in a position to say something substantive about what she is thinking (cf. Kobes 2003, last full paragraph on p. 216). It is not as if, for all she knows, her use of the word "water" might refer to a type of flower or to racecars. (Compare: One can know that *I am now here* despite its being the case that, for all

one knows, "here" might be Omaha, Nebraska; the North Pole; the kitchen; under the oak tree; or any number of other places.) She knows that she is thinking specifically about the kind of stuff that fills the lake by which she is standing and can interpret her use of "water" accordingly.[83] Therefore, we can maintain that Jane can know that W_{Jane} without thereby committing ourselves to saying that this knowledge (or propositional knowledge of content more generally) is cognitively insubstantial.

In sum, it looks like the Memory Argument (whether the original or the "memory-free" version) fails. Regarding the 1989 original, regardless of how exactly we fill in the details of Oscar's case, at least one of M1-M3 ends up being false. Regarding the memory-free version of the argument, we have considered two reasons to think that Jane doesn't know that W_{Jane}: (i) the argument from cognitive insignificance and (ii) the argument from the Principle of Knowing Identification. I have argued that neither succeed. On the contrary, Jane's ability to provide a substantive interpretation of her use of "water" suggests that she *does* know that W_{Jane}.

This concludes the main argument of this chapter. In the final section, I argue that, unlike Jane, one will typically be in a good position to know which concepts figure in one's thoughts. I concede, however, that this knowledge is empirically defeasible. This sets the stage for the next chapter on McKinsey's reductio.

[83] Perhaps Jane's knowing that W_{Jane} without knowing that she is thinking a *water*-thought is like one's knowing that *I am standing beside City Hall* without knowing which city one is in.

4.4 The Defeasibility of C-Knowledge

Goldberg's argument from conceptual omniscience is useful because it gives us a way to understand what exactly it means to say that S knows that c "figures" in her thought. Or, at least, it suggests a test we can use to determine whether S has this knowledge. In particular, it suggests that

(**CK**) S knows that concept c figures in T *iff* [S knows how to explicate c □→ S is in a position to knowledgably explicate T/c].[84]

In this section, I want to argue that if CK is true, then one is normally in a good position to know which concepts figure in one's thoughts. To demonstrate this, I'll consider a series of variations on Jane's case, each closer than its predecessor to the epistemic situation of those of us in the actual world.

Here's the second version of Jane's case. Let's refer to this new Jane as "Jane$_2$". Like Jane, Jane$_2$ is born on Earth where she trained as a chemist (so she too knows how to explicate *water*). On her fortieth birthday (t_1) she begins, unbeknownst to her, a regimen of slow-switching between Earth and Twin Earth. Suppose she is apprised of this fact (say) twenty years later, on her sixtieth birthday.[85] She is not told, and does not know, where she is now, where she was at any point after t_1, nor for how long. After learning of her situation, Jane$_2$ begins reminiscing about a beach vacation she took before t_1 and, prompted by this memory, thinks to

[84] The conditional in the consequent is a Lewis (counterfactual) conditional. Also, CK's antecedent should be read in the de dicto sense. For as the argument in the previous section (4.3.3) suggests, it may be that Jane knows *of* the concept *water* that it figures in T but that she cannot recognize it as such (as evidenced by her inability to explicate T/c_w).

[85] Notice that she is not told anything about her environmental history *before* t_1. In particular, she is not told that she was then on Earth. Thus, the evidence she has for believing that she was on Earth before t_1 *has not changed one way or the other*.

herself a thought T_2 that she would express by using the words "water is refreshing". Does Jane know which concept figures in T_2?

Elsewhere, Goldberg (2005, p. 110-116) suggests that in this kind of case, we must ascribe to $Jane_2$ a *water*-thought *even if* she is currently embedded in Twin Earth's linguistic community. This is because (i) $Jane_2$ clearly intends to think about the kind of watery substance in which she swam while on vacation and because (ii) that substance was water.[86] In this case, because of $Jane_2$'s intention, and because she lacks reason to doubt that she was still on Earth prior to t_1, my intuition is that $Jane_2$ can know that her thought involves *water*. This intuition is supported in part by the fact that $Jane_2$ *would* be able to explicate T_2/c_w. If prompted to do so, she could confidently assert that she is thinking an H_2O-thought. Thus, given CK, Jane knows that *water* figures in T_2.

Now, let's change the case again (version three). $Jane_3$ is never slow-switched. She resides on Earth her whole life. On her sixtieth birthday, $Jane_3$ begins reminiscing about a beach vacation she took before she turned forty (at t_1) and, prompted by this memory, thinks to herself a thought T_3 that she would express by using the words "water is refreshing". Does $Jane_3$ know which concept figures in T_3? It would seem so. First, she could easily and knowledgably explicate T_3/c_w. Second, and perhaps more importantly, if we say that $Jane_2$ knows that she is thinking a *water*-thought, then we must say the same thing of $Jane_3$. The only difference between their two cases is that now we are supposing that no slow-switching has taken place. And it would be absurd to suggest that $Jane_2$'s being (and subsequently discovering that she has been) slow-switched somehow puts her in a *better* position to know

[86] See Appendix A.

her thought contents than she would otherwise be. But this is what we would be forced to say if we think that $Jane_2$, but not $Jane_3$, knows the contents of her water-thought.

Version four: Suppose that everything is as it was in version three, except that $Jane_4$ is *not* familiar with the application conditions of the concept *water* and so does not know how to explicate it. Thus, $Jane_4$ cannot knowledgably explicate her water-thought, T_4 (which is token distinct from, but otherwise identical to, T_3). However, as $Jane_3$'s case makes clear, $Jane_4$ *would* be in a position to knowledgably explicate T_4/c_w if she knew *how* to explicate *water*. Given CK, then, it follows that $Jane_4$ knows that *water* figures in T_4.

Finally, suppose that $Jane_4$, tired of reminiscing, looks out over the lake by which she is standing and, inspired by the sight, thinks to herself a thought T* that she would express by using the words "water is beautiful". It seems to me that if $Jane_4$ knows which water-concept figures in T_4, then she knows which water-concept figures in T*. For there is no obvious epistemic difference between T_4 and T*.

Now, notice that by the time we get to $Jane_4$, we are more or less describing the situation in the actual world – a world in which slow-switching does not often happen.[87] Thus, I think the considerations above provide good reason to think that in the actual world one is generally in a good position to introspectively know which concepts figure in one's conscious,

[87] Ludlow (1995a) argues that slow-switching *does* happen in the actual world (and happens quite frequently). In his telling, we switch between linguistic communities all the time. A good example is when English speakers move between the US and Britain. With respect to certain concepts (e.g. *chicory*), it is a bit like moving between Earth and Twin Earth. Thus, someone switching between the US and Britain may very well find herself in a position much like Jane's. Brown (2004, p. 138-142) argues, however, that while slow-switching may indeed happen in the actual world, it must necessarily be rare. For if people moved between different linguistic communities often enough, the linguistic differences between those communities would break down such that there would no longer *be* two distinct linguistic communities. As Brown puts it, "[w]ords that previously had different meanings in the two languages would tend to settle on a single meaning" as "speakers would tend to consult common authorities for correct linguistic practice" (Brown 2004, p. 141). I think that Brown is probably correct about this, and will assume that she is for the remainder of this work.

occurrent thoughts – even when those thoughts involve externally individuated contents (and even when one does not know the precise application conditions of the concepts involved in those thoughts). Anyone wishing to deny this must either (i) explain away the intuition that $Jane_2$'s situation is epistemically different than Jane's or (ii) find some reason to deny that if $Jane_n$ knows which water-concept figures in T_n, then $Jane_{n+1}$ knows which water-concept figures in T_{n+1} (where n is either 2 or 3). For as I have tried to show, there is no epistemically relevant difference between any two successive cases. This is such that if we admit that $Jane_2$ introspectively knows the contents of T_2, then we must admit that $Jane_4$ introspectively knows the contents of T*. And since $Jane_4$'s case is typical, we should accept that one is typically in a position to know by introspection which concepts figure in one's thoughts.

But, importantly, the argument from conceptual omniscience makes clear that this knowledge is empirically defeasible. If one has reason to believe that one is in a situation like Jane's, then whatever justification one may have otherwise had for believing certain things about certain of one's thoughts is defeated. Let's adopt, then, the following thesis, which we'll call *weak C-knowledge*:

(**WCK**) For any person S and any true proposition r of the form

(r) S's thought that p involves concept c

It will typically[88] be the case that: If S knows *a priori* that she is thinking that p, then S is in a position to have *weak a priori* knowledge that r.

S's knowledge that p is *weakly a priori* just in case (i) it is not based on empirical evidence (including perceptual observation) but (ii) is empirically defeasible (see Field 1996). And S's

[88] See previous footnote.

knowledge that p is *empirically defeasible* just in case it is possible for the belief that constitutes it to be rendered unjustified by S's acquisition of empirical evidence.

What follows from TWP and WCK? Suppose I am thinking that water is refreshing and, introspecting on this thought, judge that

(p) I am thinking that water is refreshing.

Supposing that my faculty of introspection is functioning properly and that my belief that p is grounded in the right way, it follows from TWP that I know *a priori* that p. Then, assuming I am not a victim of slow-switching, it follows from WCK that I am in a position to know weakly *a priori* that I am thinking a thought that involves the concept *water*. That is, I am in a position to know weakly *a priori* that

(p*) I am thinking a thought involving *water*.

But, since it is only *weakly a priori*, my knowledge that p* will be empirically defeasible.

Notice that it will be empirically defeasible in both the *undercutting* and the *rebutting* sense (see Pollock 1986). Suppose that Jane thinks her thought T just before she is told that she is a victim of slow-switching. Having no reason to suppose she not still on Twin Earth (and has been since t_1), she believes she is thinking a *twater*-thought. Plausibly, at this point, she is justified in so believing – and, we'll suppose, justified on the basis of introspection. But, once she learns that she has been slow-switched, this justification disappears. She will have acquired an *undercutting defeater* – one that undercuts the evidential connection between introspection and her belief. Now suppose that she is told, not that she has been slow-switched, but that at some point shortly after t_2 she was returned to Earth where she has been ever since. Here, justification for her original belief is defeated by her acquiring reason to believe an

incompatible proposition: that, actually, she is thinking a *water*-thought. In this case, she has acquired a *rebutting defeater*.

We are now ready to turn our attention to McKinsey's reductio. Do TWP and WCK imply, absurdly, that one can typically know substantive propositions about one's environment *a priori*? To explore this question is the task of our final chapter.

Chapter 5: The McKinsey Reductio

Consider the following propositions, where E is a contingent proposition about S's environment or environmental history.

(MK1) S is thinking that water is refreshing.

(MK2) If MK1, then E.

Suppose that MK1 is true and that S's faculty of introspection is functioning properly. Then, it follows from TWP that S is in a position to have *a priori* knowledge that MK1. And it is generally agreed that if externalism is true, then there is some E such that MK2 expresses a conceptual truth (Putnam 1981, McGinn 1989, Brown 1995 & 2004, McLaughlin & Tye 1998a & 1998b, Sawyer 1998, Warfield 1998, Nuccetelli 2003).

Michael McKinsey (1991, 2002, 2007) argues, however, that it cannot both be true that MK1 is knowable *a priori and* that MK2 is a conceptual truth.[89] For consider the following closure principle.

> *Closure of Apriority under Conceptual Implication (CA)*
> Necessarily, for any person x, and any propositions P and Q, if x can know a priori that P, and P conceptually implies Q, then x can know a priori that Q (McKinsey: 2002, p. 207; 2007, p. 55).[90]

If MK1 can be known *a priori*, and MK2 is a conceptual truth, then it follows from CA that E can be known *a priori*. But, McKinsey argues, the notion that one could have *a priori* knowledge

[89] The original (and therefore more frequently cited) version of his argument can be found in his 1991. The 2002/2007 version is a revision based on the reaction to the 1991 version. Brown (1995) and Boghossian (1998) also develop incompatibilist arguments along these lines.

[90] McKinsey articulates CA in terms of "logical implication". But he makes clear that he means "'logically implies' in a broad sense that includes what I have elsewhere called 'conceptual implication'" (McKinsey 2007, p. 53).

that E (*whatever* E ends up being) is absurd. Therefore, if externalism is true, then TWP must be false.

Taking the truth of externalism for granted, we could respond in one of three ways: we could concede that TWP is false, reject CA, or simply accept that it is possible to know that E *a priori*.[91] In this chapter, I defend the latter two options. In section 5.1, I show that if externalism is true, then there are counterexamples to CA. This means that the externalist cannot be expected to accept it. After that, I evaluate the claim that *a priori* knowledge that E is absurd. Whether it is seems to me to depend on what exactly E is.[92] If, for instance, E is the proposition that *water exists*, then certainly one cannot know that E *a priori*. I want to show, however, that MK2 constitutes a conceptual truth only if E is a relatively modest proposition – modest enough that it is not implausible that it can be known *a priori*. That is the task of section 5.2. Then, in section 5.3, I address the worry that even a modest version of MK2 absurdly implies that it is possible to know *a priori* that one is not a brain in a vat in an otherwise empty world. Again, my strategy here is to simply accept that one can know this *a priori*. Indeed, I think it is possible to deduce it from *a priori* premises about one's thought contents. This is sometimes called "McKinsey-style" reasoning (e.g. Pryor 2007). McLaughlin (2003) has argued that McKinsey-style reasoning is necessarily question-begging. I structure my own argument around McLaughlin's objections, using them as a vehicle for launching a positive case to think that McKinsey-style reasoning can be perfectly cogent.

---

[91] Ball (2007) proposes a fourth option: deny that externalism entails that *any* version of MK2 expresses a conceptual truth. Part of his argument is discussed below in section 5.2.

[92] For reasons I don't quite understand, McKinsey denies this latter claim. He says that "it doesn't matter at all to my argument which empirical, 'external' proposition we choose as our instance of 'E', since it will be quite clear, for any such choice, that the proposition cannot possibly be known a priori" (McKinsey 2002, p. 201).

5.1 Is Knowledge Closed Under Conceptual Implication?

If externalism is true, then knowledge is not in general closed under conceptual implication. The reason comes down to the fact that, on the externalist picture, one can possess and entertain thoughts involving a concept despite having an incomplete understanding of that concept.

Consider an example. Suppose that Oscar tells Sally that his best friend, whom Sally has never met and about whom Sally otherwise knows nothing, is a bachelor. And suppose that Sally thereby comes to know that

(1) Oscar's best friend is a bachelor.

Now, (1) conceptually implies the proposition

(2) Oscar's best friend is an unmarried man.

So, it seems that as long as Sally understands the concept *bachelor*, she is also in a position to know that (2). Assuming, then, that possessing a concept requires understanding it, Sally's knowing that (1) puts her in a position to know that (2) (for of course Sally must possess the concept *bachelor* in order to know that (1)).

Externalists, however, deny that possessing a concept requires *fully* understanding it. They hold that one may possess and entertain thoughts involving a concept despite having a partial or incomplete understanding of that concept. To see the significance of this, suppose that Sally correctly applies "bachelor" to unmarried men, but is unsure whether it also applies to unmarried women. The externalist will say that, despite her incomplete understanding, Sally may still possess the concept *bachelor* if she belongs to a community that has the concept. Let's suppose, then, that Sally does belong to such a community. Then she can believe, and

know, propositions involving the concept *bachelor* despite her incomplete understanding. Thus, she can come to know on the basis of Oscar's testimony that (1). Again, suppose Sally does know that (1). Then, if knowledge is closed under conceptual implication, Sally is also in a position to know that (2). But, since Sally is unsure whether "bachelor" also applies to unmarried women, she is *not* in a position to know just on the basis of Oscar's testimony that (2). For all she knows, Oscar's best friend is an unmarried woman, in which case (2) is of course false.

This shows that if externalism is true, then knowledge is not in general closed under conceptual implication. A similar counterexample shows that apriority in particular is not closed under conceptual implication: Sally can know *a priori* that

(3) All bachelors are bachelors.

All this requires is that Sally has the concept *bachelor* and basic logical competence. The proposition (3) conceptually implies the proposition

(4) All bachelors are unmarried men.

But, given her incomplete understanding of *bachelor*, Sally is not in a position to know *a priori* that (4). Thus, given externalism, apriority is not closed under conceptual implication. That is, CA is false.

One might wonder if this is ultimately a problem for McKinsey. After all, Sally *could* know *a priori* that (4) if she had a comprehensive enough understanding of *bachelor* – indeed, it is reasonable to suppose that Sally must know *a priori* at least some analytic truths concerning *bachelor* if she has a firm enough grasp of the concept to possess it at all (Brown 2001, p. 221-224). In other words, it is certainly *sometimes* the case that if a subject S knows *a priori* that p,

and p conceptually implies q, then S is in a position to know *a priori* that q. It might therefore be possible to amend CA (by coming up with a third conjunct to attach to CA's antecedent) so that it is immune to counterexamples while remaining friendly to the reductio.

Perhaps this can be done, but there is reason to be doubtful. As Anthony Brueckner (2010, p. 250-255) argues, it is probably the case that instances of CA are true only when the relevant conceptual implication is *recognized* as such in an *a priori* way. For in those cases the corresponding conditional can be known *a priori*. If that's right, then CA fails in Sally's case because her incomplete understanding of *bachelor* prevents her from seeing that *if (3) then (4)*. This suggests something like the following amendment to CA.

> *Closure of Apriority under Known Conceptual Implication (CA*)*
> Necessarily, for any person x, and any propositions P and Q, if x can know a priori that P, P conceptually implies Q, *and x is thereby in a position to know a priori that if P then Q*, then x can know a priori that Q.

CA* is immune to the counterexample described above since Sally is not in a position to know *a priori* that the relevant conditional holds. But it cannot obviously be used to facilitate McKinsey's reductio. For not only would MK2 have to constitute a conceptual truth, but S would have to be in a position to know MK2 *a priori*. As compatibilists like to point out, however, for all one can tell *a priori* "water" is on a par with "phlogiston" (McLaughlin & Tye 1998b; Brueckner 2001 & 2010; Brown 2004, Ch. 8). In other words, one might turn out to be in a Dry Earth-type scenario where "water" fails to refer to a natural kind.[93] But if so, then, at

[93] Here again is Boghossian's description of Dry Earth: It is "a planet just like ours in which, although it very much seems to its inhabitants that there is a clear, tasteless and colorless liquid flowing in their rivers and taps and to which they confidently take themselves to be applying the word 'water', these appearances are systematically false and constitute a sort of pervasive collective mirage" (Boghossian 1998, p. 206). It's important to note that Dry Earth is not simply a scenario in which water does not exist. As we will see below, one can possess the concept of water in its absence. Rather, it is a scenario in which the baptismal event where "water" is introduced is based

least as far as externalism is concerned, no particular environmental proposition is presupposed by one's thinking that water is refreshing. In that case, MK2 may very well be false.

The upshot is that one cannot know *a priori* that MK2 unless one can know *a priori* that one is not in a Dry Earth-type scenario. But this is an empirical matter that cannot be known *a priori*. Hence, one cannot know *a priori* that MK2.

There is one response to this kind of worry that I think is promising. To keep the discussion manageable, I will not evaluate it at length here. But it is worth mentioning. Boghossian (1998) argues, very roughly, that if S can know *a priori* that MK1, then she can know *a priori* that the relevant thought has determinate content. (The basic idea is that if one knows *which* content one's thought has, then one knows one's thought has *some* determinate content or other.) But, he argues, externalism implies that if one is in a Dry Earth-type scenario, then one's "water"-thoughts have no determinate content. Thus, he concludes, if S can know *a priori* that MK1, then she is also in a position to rule out *a priori* the possibility that she is in a Dry Earth-type scenario.[94]

---

on a mirage. We imagine, for example, a scenario where one attempts to introduce "water" by saying, "Water is the substance instantiated by the bits of clear liquid over there", where *nothing at all* is "over there". The community then continues to apply "water" to such mirages.

[94] Notice that Boghossian does not claim that one can know *a priori* that one is not in a Dry Earth-type scenario, just that one *could* if one could know MK1 *a priori*. McKinsey himself actually employs a similar strategy, arguing that if MK1 is *a priori*, then so is MK2 (McKinsey 2002, p. 206-210). McKinsey, however, relies on the false closure principle

> *Partial Closure Under Conceptual Implication (PCC)*
> Necessarily, for any person x and any propositions P and Q, if x can know a priori that P, and P conceptually implies Q, then x can know a priori that if P then Q (McKinsey 2002, p. 209).

PCC is susceptible to precisely the same kinds of counterexample as CA. Consider: Sally knows *a priori* that (3); (3) conceptually implies (4); nonetheless, given her incomplete understanding of *bachelor*, Sally is not in a position to know *a priori* that *if (3) then (4)*. Notice also that CA* and PCC jointly entail CA. Therefore, since CA is false the conjunction of CA* and PCC is also false.

Both steps of this argument have been criticized. Jessica Brown (2004, p. 284-289) criticizes the first step, endorsing Nozick's (1981) sensitivity condition to argue that it is false that if S knows *a priori* that MK1, then she can know *a priori* that the relevant thought has determinate content. For, she argues, we can suppose that S's introspectively grounded belief that MK1 is sensitive. Given the redeployment thesis (see section 2.2 above) and that S's introspective faculties are functioning properly: if MK1 were false, S wouldn't believe that MK1. If, however, S were in a Dry Earth-type scenario and her "water"-thought had no determinate content, she would continue to believe otherwise. Therefore, introspectively grounded belief to the effect that one's thought has determinate content is *in*sensitive. So, assuming that only sensitive belief constitutes knowledge, Boghossian's argument fails.[95]

The second step of Boghossian's argument presupposes what Jessica Brown (2004) calls the "illusion version" of externalism (see Ch. 3). The illusion version holds that in no-reference cases, such as when a Dry Earther attempts to think a thought she would express by uttering "water is refreshing", one fails to think a thought with any determinate content. But the illusion version is controversial. Korman (2006), for instance, argues that externalists can consistently hold that on Dry Earth, "water" expresses a descriptive concept – that it means something like *the watery stuff*.[96] But if that's right, then "water"-thoughts *do* have determinate content even on Dry Earth. Then, even if one can tell *a priori* that one's "water"-

[95] Of course, Brown's response here is only as plausible as the sensitivity condition on which it relies. For criticism of the latter see Kripke (2011). See Becker & Black (2012) for a more in depth discussion.

[96] Gerken 2007 also offers a persuasive critique of the illusion version. Ludlow (2003) is also relevant, though he is concerned more with the semantics of proper names than of natural kind terms. The descriptive interpretation of no-reference cases is not without its critics. Häggqvist & Wikforss (2007) argue that it commits us to an implausibly strong version of externalism. For on this version, not only would the meaning of a natural kind term depend on the external environment, but its *semantics* (i.e. whether it is referential or descriptive) would as well.

thought has determinate content, one would not thereby be in a position to know that one is not in a Dry Earth-type scenario.

Here is where this leaves us. Since CA is false, it seems that McKinsey is forced to rely on CA*. But, even if CA* is true, it cannot be used to facilitate McKinsey's reductio since that would require that S can know MK2 *a priori*. But MK2 cannot be known *a priori* since one cannot rule out *a priori* the possibility that one is in a Dry Earth-type scenario. One might respond that if S can know *a priori* that MK1, then she can know *a priori* that the relevant thought has determinate content, which would put her in a position to rule out *a priori* the possibility that she is in a Dry Earth-type scenario. But, as we have just seen, this response will not work unless (i) sensitive belief is *not* necessary for knowledge and (ii) the illusion version of externalism is true (or is the most plausible version of externalism).

Successfully defending (i) and (ii) is certainly not an insurmountable task. On the contrary, both (i) and (ii) are reasonably plausible. Let us then grant for the sake of argument that they are true. What does this mean for McKinsey's reductio? Earlier I said that whether it is absurd to suppose that E can be known *a priori* depends of what exactly E is. In the next section, I argue that MK2 constitutes a conceptual truth only if E is a modest enough proposition that it is not implausible that it can be known *a priori*. Thus, McKinsey's reductio is on shaky ground even if we ignore the problems with CA.

5.2 What It Takes to Think About Water

Recall Putnam's Twin Earth thought experiment, which I briefly outlined in Chapter 1. Despite being intrinsically identical, Oscar and his doppelgänger Toscar differ with respect to

their propositional attitude contents. While Oscar thinks that *water is wet*, Toscar thinks that *twater is wet*. But if what goes on inside their heads is exactly the same, then what accounts for this difference? The answer is that their environments are different. Oscar, a resident of Earth, lives in an H_2O environment; Toscar, on the other hand, lives in an XYZ environment. This is important because it suggests that Oscar's "water"-thoughts are causally connected to contact with instances of H_2O whereas Toscar's are causally connected to instances of XYZ. Thus we have an explanation for the difference in content.

It is natural to go a step further and to say that causal interaction with instances of H_2O is *necessary* for thinking *water*-thoughts.[97] For Toscar lacks the concept *water*, and the reason is apparently that there is no H_2O on Twin Earth for him to interact with. This has led McGinn (1989, p. 30-36 & 47-48) to posit the following condition on concept possession.

(**M**) If the concept of k is an atomic natural kind concept, then one possesses it only if one has causally interacted with instances of k.[98]

The qualifier "atomic" is included to head off counterexamples like H_2O (McGinn 1989, p. 35). The concept of H_2O is a natural kind concept, but one need not interact with instances of H_2O in order to possess the concept of it. One need only possess the concepts of hydrogen and

[97] Notice, however, that this does not strictly follow from the Twin Earth thought experiment (Bilgrami 1992, Ball 2007). The thought experiment shows that if two token-distinct thoughts have different causal histories, then they may also have different contents. What kind of casual history might be *necessary* for entertaining certain thought contents is another question.

[98] M is plausible only if we interpret "causally interacted" liberally. For instance, we should not read M as saying that one must *directly* interact with instances of k in order to have the concept of k. I have (probably) never directly interacted with yttrium, but I possess the concept because I am a member of a linguistic community that possesses the concept. It may be necessary that *someone* in my linguistic community has directly interacted with instances of yttrium in the appropriate way. I could then acquire the concept of yttrium by interacting with this person (or with someone who has interacted with this person, etc.). In that case, I will have casually interacted with instances of yttrium in the sense that there is a causal chain linking me to instances of yttrium. This kind of causal-chain interaction is plausibly necessary, but direct interaction is not. We should interpret M (and the amendments to M proposed below) accordingly.

oxygen, and it is possible to possess these concepts in the absence of H_2O. McGinn maintains that this is not a counterexample to M because the concept of H_2O is not atomic, but molecular. That is, it is made up of other concepts (viz. the concepts of hydrogen and oxygen). The concept of water, on the other hand, *is* atomic.[99] Thus, it follows from M that causal interaction with instances of H_2O is necessary for possessing the concept *water*.

Despite its initial plausibility, M is false. As McLaughlin and Tye (1998b, p. 300-302) have argued, where it is possible to possess the concept of H_2O, it is possible to possess the concept of water (see also Burge 2007, p. 96-98). Imagine, for example, a possible world in which H_2O does not exist, but hydrogen and oxygen do. The scientists in this world could theorize about H_2O in its absence. Suppose they do, dubbing the hypothetical molecule "water". Laypeople might then see mention of "water" in a newspaper, thereby acquiring the concept of water despite never having interacted with instances of it. Indeed, it is conceivable that someone could acquire the concept of water in the absence of both H_2O *and* other people (McLaughlin & Tye 1998b, p. 302). For it is conceivable that someone could develop a theory of chemistry in isolation and then theorize their way to the concept of H_2O. Plausibly, this person would then also have the concept of water. And they would have this concept without having had to acquire it from a linguistic community.

Notice, however, that the residents of the waterless world just described are able to theorize their way to the concept of water because they have the concepts of hydrogen and oxygen. This fact suggests the following revision of M.

[99] Of course, the concept of water is distinct from the concept of H_2O. For instance, one can possess the former without possessing the latter.

(**M'**) If the concept of k is an atomic natural kind concept, then one possesses it only if one has either causally interacted with instances of k or one possesses the concepts of the kinds that make up the kind k.

Given M', one possesses the concept of water only if one has either casually interacted with instances of H_2O or one has the concepts of hydrogen and oxygen. But what might it take to acquire the concepts of hydrogen and oxygen? Perhaps there is hydrogen and oxygen in one's environment, and one acquires the concepts via causal interaction with instances of those kinds. Alternatively, one could theorize one's way to those concepts in the same way that the scientists in our example theorized their way to the concept of water. This would be possible if one possessed the concepts of the kinds that make up hydrogen and oxygen. Those would be the concepts of the relevant subatomic particles (electron, proton, neutron). How might one acquire *those* concepts? Again, it looks like either by casual interaction or by theoretical reconstruction. And so on.

Following McLaughlin and Tye (1998b, p. 301), then, we might adopt the following interpretation of M'.

(**M+**) If the concept of k is an atomic natural kind concept, then one possesses it only if one has either causally interacted with instances of k or one has causally interacted with instances of the kinds that make up the kind k^{100}, or the kinds that make up these kinds, or the kinds that make up these kinds, and so forth.

[100] Following Ball (2007, p. 462), I have slightly amended McLaughlin and Tye's version of M+. Their version ends where I have inserted this footnote.

According to M+, one possesses the concept of water only if one has casually interacted either with instances of H_2O; instances of hydrogen and oxygen; instances of electrons, protons, and neutrons; or with other, more basic subatomic particles.

I think that M+ gets us pretty close to the truth. But, Derek Ball (2007) describes a potential counterexample. Natural kind terms, he reminds us, are typically introduced via an initial baptism. And this can be done by use of a definite description. Thus, to borrow an example from Kripke, I may introduce the natural kind term "gold" by saying, "Gold is the substance instantiated by the items over there, or at any rate, by almost all of them" (Kripke 1980, p. 135). But, as Ball points out, we can use definite descriptions to refer to objects with which we have never causally interacted. This suggests that the baptismal event need not involve casual interaction with instances of the kind thereby introduced or, for that matter, with instances of the kinds that make up that kind (etc.).[101] Hence the following counterexample.

UNEARTH

Imagine that in a distant part of the universe, causally isolated from Earth, there is a planet called "Unearth" on which the laws of physics are very different. On this planet, matter is not made of molecules. There are no protons, neutrons, or electrons, nor are there smaller subatomic particles or even energy as we know it. In fact, this planet has no natural kinds in common with Earth. There is, however, a substance that looks, smells, and tastes very much like water. We may call this substance "unwater". The inhabitants of Unearth are relatively scientifically ignorant: they have no knowledge of chemistry or physics. They do, however, have a well-developed astronomy that

[101] From here on, I will drop the "etc." when discussing issues related to M+ (or variations thereof).

> includes a coordinate system with which they can pick out the location of other stars and planets. Now suppose that an inhabitant of Unearth fixes the reference of the term "schmwater" with the following definite description: "the substance that has the phenomenal properties of unwater on the planet at <X,Y,Z>", where Earth is the planet at <X,Y,Z>.
>
> (Ball 2007, p. 465)

As a result of this initial baptism, the residents of Unearth plausibly now possess the natural kind concept of schmwater.[102] By hypothesis, however, they have never causally interacted with instances of schmwater or with instances of the kinds that make up schmwater. For instances of schmwater are just H_2O, and neither H_2O nor any of the kinds that make it up exist on Unearth.

We therefore appear to have a counterexample to M+. If so, then one may possess a natural kind concept despite being in no way causally connected either to instances of that kind or to instances of the kinds that make it up. Still, it looks as though that kind (or the kinds that make it up) must exist *somewhere* in one's universe. For example, if neither H_2O nor any of the kinds that make it up existed at <X,Y,Z> or anywhere else, then it is hard to see how the residents of Unearth could end up with the concept of schmwater. For then no definite description they might use to fix the referent of "schmwater" would succeed in picking out either H_2O or any of the kinds that make it up. Therefore, if it succeeded in picking out anything at all, the baptismal event would result in the introduction of a concept whose referent is something *other* than H_2O or any of the kinds that make up H_2O. So, the concept they end up with could not be that of schmwater (or of water or H_2O, for that matter) or of

[102] Ball maintains that "schmwater" would express the concept *water* (Ibid., p. 466). But the argument doesn't hinge on this.

anything that might help them to theoretically reconstruct the concept of schmwater (e.g. the concept of hydrogen or oxygen).

Given this, I think we can with some confidence adopt the following, weaker version of M+.

(**M-**) If the concept of k is an atomic natural kind concept, then one possesses it only if there exist either instances of k or instances of the kinds that make up the kind k, or the kinds that make up these kinds, or the kinds that make up these kinds, and so forth, somewhere in one's world.

What does this mean for McKinsey's reductio? Given that the concept of water is an atomic natural kind concept, M- entails the following version of MK2.

(MK2*) If MK1, then there exist either instances of water or instances of the kinds that make up the kind water, or the kinds that make up these kinds, and so forth, somewhere in S's world.

Granting that it is an *a priori* conceptual truth, MK2*, in conjunction with CA* and the apriority of MK1, entails that S can know *a priori* that

(MK3) There exist either instances of water or instances of the kinds that make up the kind water, or the kinds that make up these kinds, and so forth, somewhere in my [S's] world.

Is it absurd to suppose that MK3 could be known *a priori*?

I submit that it is not. Notice that MK3 entails nothing *specific* about S's world.[103] It does not entail that water exists. And though it does entail that *some* natural kind or kinds exist, it is silent on what those kinds might be. Insight into that would require knowledge of the chemical composition of water. But that remains an empirical matter regardless of the

[103] Noordhof (2004, p. 54-55) makes a similar point in the context of a discussion on Jessica Brown's (1995) version of the reductio.

apriority of MK3. Thus, even if we suppose that S can know MK3 *a priori*, all this puts her in a position to know for sure is that there exists *some* natural kind or other somewhere in her world.[104] But this will be true almost regardless of what S's world is like. For instance, it will be true so long as at least one kind of fundamental particle (*any* kind) exists. It may even be the case, though I imagine this is more controversial, that it will be true so long as anything *at all* exists in S's world.

Given the modesty of MK3, it is not obviously absurd to suppose that it might be known *a priori*. An argument is needed to establish this. McKinsey cannot simply assume it to be the case (cf. Brueckner 1992, p. 117-118).

One such argument might go as follows. If MK3 is true, then the following skeptical hypothesis must be false.

(A) I am a BIV in an otherwise empty world (henceforth, a "lonely BIV"). All that exists is a brain (namely, me) and whatever is necessary to keep it alive.

For by hypothesis, if one is a lonely BIV, then neither water nor any of the kinds that make it up (whatever they happen to be) exist anywhere in one's world. Thus, if S can know MK3 *a priori*, it appears she is also in a position to know *a priori* that ~A. But, one might think, it is absurd to suppose that one could know this *a priori*. So the reductio stands.

I am inclined to embrace the conclusion that ~A is knowable *a priori*. Before we get to that, however, we should note that the route to ~A is not as straightforward as my imagined interlocutor has made it out to be. For instance, S cannot infer ~A from MK3 unless she knows

[104] Compare with Brueckner's P (Brueckner 1992, p. 117) and E3 (Brueckner 2010, p. 247) and with Ball's W (Ball 2007, p. 469).

that the concept expressed by "water" as it appears in MK3 is *not* the same concept a lonely BIV would express by "water". But S may not be in a position to know this, at least not *a priori*. For she could be in a situation much like Jane's (see section 4.2 of the previous chapter), where, instead of having been slow-switched between Earth and Twin Earth, she has been slow-switched between Earth and (perhaps) Vat-Earth. Suppose, for example, S learns that a mad scientist repeatedly (and seamlessly) envats and then replaces her brain, the switches happening at random intervals but typically far enough apart for S's thought contents to change. When envatted, her *water*-thoughts become *vat-water*-thoughts, the process reversing itself sometime after S's brain is replaced. In these circumstances, S will be unable to say which concept (*water* or *vat-water*) is expressed by her use of the word "water". In particular, she will be unable to say which concept is expressed by "water" as it appears in MK3. But then, as far as S can tell *a priori*, MK3 may be perfectly consistent with the truth of A.

But what if S is not a victim of the kind of slow-switching just described? In that case, if S is thinking a thought that does in fact involve our concept *water*, she will be in a position to have weak (i.e. empirically defeasible) *a priori* knowledge to that effect.[105] This is a consequence of the conjunction of TWP and the thesis I referred to in section 4.4 as *weak C-knowledge* (or WCK), which, recall, states that one is typically in a position to knowledgably identify the concepts that figure in one's conscious, occurrent thoughts. But if S can know *a priori* that her thought involves our concept *water* (as opposed to *vat-water*), then it seems she can also know *a priori* that her use of "water" expresses a concept inaccessible to a lonely BIV.

[105] This knowledge would be defeated if e.g. S acquired reason to believe that she was subject to the kind of slow-switching described in the previous paragraph.

For given M-, a lonely BIV by definition cannot possess *water*. So, as long as S understands skeptical hypothesis A, she can know *a priori* that *water* is inaccessible to a lonely BIV.[106]

Given this, I believe the incompatibilist can offer the following revised version of McKinsey's reductio. Consider the propositions below.

(SA1) I am thinking a thought involving *water*.

(SA2) If SA1, then ~A.

Given the conjunction of TWP and WCK, it follows that I can have empirically defeasible *a priori* knowledge that SA1. Then, granting that SA2 is an *a priori* conceptual truth[107], it follows from CA* that I can know *a priori* that

(SA3) ~A.

Thus, assuming that ~A cannot be known *a priori*, the reductio stands.[108]

I agree that if the foregoing is correct then we arrive at the conclusion that one can know *a priori* that one is not a lonely BIV. Indeed, I believe that given [TWP ^ WCK] and the apriority of SA2, it should be possible to *reason* one's way to the conclusion that one is not a lonely BIV from *a priori* premises about one's thought contents. What's more, I believe that it should be possible to do so in such a way as to resolve certain skeptical doubts.

---

[106] Of course, there is still the issue of whether it can be known *a priori* that our concept *water* is non-empty, as it would be in a Dry Earth-type scenario. If it cannot, then S cannot know *a priori* that the concept *water* is inaccessible to a lonely BIV (for if *water* were an empty concept, then, at least as far as externalism is concerned, there is no reason to suppose that a lonely BIV could not possess it). Recall, however, that for the purposes of sections 5.2 and 5.3 we are granting for the sake of argument that one can rule out *a priori* the possibility that one is in a Dry Earth-type scenario. See section 5.1 above. So, there are two issues here. First, (i) can S know *a priori* that she is thinking a *water*-thought (as opposed to a *vat-water*-thought)? Second, (ii) if she can know this *a priori*, can she also know *a priori* that our concept *water* is non-empty? We are granting for the sake of argument that the answer to (ii) is "yes"; what I am doing now is defending the claim that, so long as she is not a victim of slow-switching, the answer to (i) is also "yes".

[107] See previous footnote.

[108] Notice that this version of the reductio does not directly implicate TWP. If sound, it shows only that ~[TWP ^ WCK]. It is therefore open to the compatibilist to save TWP by rejecting WCK.

In what remains of this chapter, I aim to defend these theses. It may seem incredible that one could defeat (even modest forms of) skepticism simply by reflecting on and reasoning about one's thoughts à la Descartes. To the extent that that is true, I risk undermining the thesis of this chapter – that, *pace* McKinsey, the conjunction of externalism and privileged access does not reduce to absurdity. My hope, however, is that like a magic trick it will seem less incredible once it is revealed how it is possible. If so, then the dialectical force of McKinsey's reductio will dissipate and the externalist will be left with a promising means of addressing certain modest forms of skepticism.

5.3 The Externalist and the Skeptic

The cogency of an anti-skeptical argument is often a matter of whether it begs the question against the skeptic. That informs the approach I take here. Let's use "SA" to refer to the bit of reasoning whereby one attempts to deduce SA3 from SA1&2. And let's refer to the kind of skeptic who *believes* that A as "Skeptic A". Then, the question I'm concerned to answer in this final section is this: *Does SA beg the question against Skeptic A*?[109]

Now, belief that A is an unusually strong kind of skepticism. Given this, it may not be obvious why I would build a discussion around it. However, I am assuming that if SA does not beg the question against those who believe A, then it neither does it beg the question against those who merely *doubt* that ~A (and certainly not against those who have no antecedent doxastic attitudes whatsoever about whether or not A). Thus, if I succeed in showing the former, then I will have succeeded in showing the latter.

[109] I am assuming that one can beg the question against oneself.

At the end of the previous section, I argued that the first premise of SA (viz. SA1) is empirically defeasible. One might think that this by itself is reason to suppose that SA must beg the question against someone like Skeptic A. Brian McLaughlin (2003) argues that if either premise of a McKinsey-style argument (i.e. an argument of the kind exemplified by SA) is empirically defeasible, then that argument must necessarily beg the question. This follows, he thinks, from the fact that if either premise is empirically defeasible, then the negation of any proposition jointly entailed by those premises will be an empirical defeater of their conjunction. Hence, he reasons, the McKinsey-style reasoner must already be "epistemically entitled to presuppose that E", where E is the conclusion of the argument in question (McLaughlin 2003, p. 91).

Applied to this case, McLaughlin's reasoning suggests that A is an empirical defeater of the conjunction of SA1&2 and that therefore anyone attempting to deduce ~A from SA1&2 must already be "epistemically entitled to presuppose" ~A. So, McLaughlin's argument suggests, SA must beg the question. I intend to structure my own argument around McLaughlin's objections, using them as a vehicle for launching a positive case to think that McKinsey-style anti-skeptical reasoning can be perfectly cogent.[110]

The account of question-begging I adopt for the purposes of this section is derived from Pryor's (2004) notion of "dialectical power". Given this, our question will be intimately related to whether SA has the power resolve the doubts, or change the mind, of someone like Skeptic A.

[110] There is a preexisting debate in the literature over whether McKinsey-style anti-skeptical reasoning is cogent. For arguments suggesting that it is, see Tymoczko 1989 and Warfield 1998. For those who argue that it is question-begging, see McLaughlin 2000 & 2003, Steup 2003, and McKinsey 2018 (section 8).

Pryor's account begins with a distinction between S's having *justification* to believe p and S's being *rationally committed* to believing p. He gives the following example. Suppose I believe that I can fly. In that case, I am rationally committed to believing that someone can fly. It would be irrational for me to lack the belief that someone can fly while at the same time believing that *I* can fly. But notice that this is true even if I lack justification for either proposition. I may even have justification to believe that no one can fly. Hence, one's being rationally committed to believing p is compatible with one's lacking justification to believe p.

But, as Pryor points out, there is an interesting modal relationship between the two concepts. For instance, it seems that my belief that p rationally commits me to believing q if, were I to gain justification to believe p, I would thereby gain justification to believe q. That seems to be what's happening in the case above. Since I would gain justification to believe that someone can fly if I were to gain justification that I can fly, my belief that I can fly rationally commits me to believing that someone can fly. The modal relationship that interests us here, though, is the following:

(**RO**) S's belief that p **rationally obstructs** her from believing q on the basis of w *iff* empirical justification for p would indicate to S that w isn't a reliable basis for believing q (thus preventing S from justifiedly believing q on the basis of w).

A rationally obstructed belief is an *irrational* belief. For example, suppose I believe that some very realistic fake dimes have been planted nearby. If I were justified in so believing, then I wouldn't be in a position to justifiedly believe I'm holding a dime on the basis of my visual experiences as of a dime in my hand. In these epistemic circumstances, it would be irrational of

me to believe I am holding a dime on the basis of my visual experience as of a dime in my hand (whether I am retaining this belief or adopting it for the first time).

Let's apply this to the idea of dialectical power. Roughly, an argument is dialectically powerful to the extent that its intended audience could rationally accept it. The more dialectical power an argument has, the more *effective* it is. One way for an argument to be dialectically *in*effective is if its intended audience is rationally obstructed from believing one of its premises. Sometimes one is rationally obstructed from believing an argument's premises *because* of one's doxastic attitude toward the conclusion. When this happens, we'll say that the argument is *question-begging*:

(**QB**) A deductive argument from p to q **begs the question** against S *iff* S's belief that ~q rationally obstructs her from believing p.

Given QB, we can reframe our initial question: *Does Skeptic A's believing A rationally obstruct her from believing SA1*? The problem is clear. If Skeptic A is rationally obstructed from believing SA1, then she cannot rationally believe SA1. But then, of course, SA would be quite useless to her.

Now, it turns out that the question of empirical defeasibility is closely related to our primary question – i.e. whether SA begs the question against Skeptic A. Let's be precise about what we mean when we talk about defeat by empirical evidence. Earlier (in section 4.4) we mentioned that there are two distinct senses of *defeasibility*: undercutting and rebutting. By an undercutting empirical defeater, we understand the following:

(**UD**) p is an **empirical defeater$_U$** of q relative to w *iff* empirical justification for p would indicate to S that w isn't a reliable basis for believing q.

By a rebutting empirical defeater, we understand:

(**RD**) p is an **empirical defeater$_R$** of q *iff* empirical justification for p would undermine one's justification for q by directly supporting a proposition r incompatible with q.

For example, suppose that Ted is looking over at a hill he sees in the distance. He seems to perceive that there are sheep on the hill and comes to believe that there are on the basis of this perception. Suppose the conditions are such that Ted's perceptions justify this belief. Now, the owner of the land Ted is admiring comes by and informs him that, in fact, there no sheep on the hill. Plausibly, this testimony justifies him in believing that there are no sheep on the hill. This defeats his justification for believing that there are by directly supporting an incompatible proposition. Thus, Ted's belief is empirically defeated$_R$. If, on the other hand, Ted had instead acquired justification to believe that his vision was impaired in the relevant way, his belief that there are sheep on the hill would have been defeated$_U$. In that case, he won't have acquired any reason to believe anything incompatible with what he now believes. But he *would* have acquired reason to doubt the reliability of the connection between his perceptions *as of sheep* and his belief that he is seeing sheep. Hence the former would fail to justify the latter. (Notice the intimate relationship between defeasibility$_U$ and rational obstruction. We'll return to that momentarily.)

Corresponding to each of these notions, two sets of questions present themselves. The first set is this:

Q1 Is A an empirical defeater$_R$ of SA1?

Q2 If the answer to Q1 is "yes", does it follow that Skeptic A is rationally obstructed from believing SA1?

The second set:

Q3 Is A an empirical defeater$_U$ of SA1?

Q4 If the answer to Q3 is "yes", does it follow that Skeptic A is rationally obstructed from believing SA1?

The answer to Q4 is obvious enough. From RO and UD it follows that:

(5) p is an empirical defeater$_U$ of q relative to w *iff* S's believing p would rationally obstruct her from believing q on the basis of w.

Hence:

(5*) A is an empirical defeater$_U$ of SA1 relative to introspection *iff* S's believing A would rationally obstruct her from believing SA1 on the basis of introspection.

Then, from QB and 5*:

(6) SA begs the question against Skeptic A *iff* A is an empirical defeater$_U$ of SA1 relative to introspection.[111]

As we're understanding things, the idea of rational obstruction is intimately related to empirical defeasibility$_U$. Hence, (from 5*) it straightforwardly follows that the answer to Q4 is "yes". It further follows (from 6) that Q3 is just another way of asking whether Skeptic A is rationally obstructed from believing SA1. Q3 is therefore the target of our inquiry. Before answering this question directly, however, I want to take a look at the first set of questions. McLaughlin (2003) has more or less directly addressed Q1 in connection with the cogency of McKinsey-style

[111] There is a suppressed premise here: *S's belief that A rationally obstructs her from believing SA1 on the basis of introspection* iff *S's belief that A rationally obstructs her from believing SA1*. Essentially, we're ignoring the possibility that Skeptic A might base a belief in SA1 on something other than introspection.

reasoning. I think that is a good place to start. This discussion will prove instructive, leading us directly to an answer for Q3.

It is not a leap to think that our admitting that SA1 is empirically defeasible$_R$ commits us to saying that it is defeasible$_R$ *by A*. After all, ~A is an environmental proposition entailed by the conjunction of SA1&2. Thus, if we suppose that the negation of any environmental proposition entailed by p must be an empirical defeater$_R$ of p, then it follows that A is an empirical defeater$_R$ of the conjunction of SA1&2. But, supposing that (being a conceptual truth) SA2 is *not* empirically defeasible$_R$, A is an empirical defeater$_R$ of the conjunction of SA1&2 only because it is an empirical defeater$_R$ of SA1.

In his own effort to show that McKinsey-style reasoning is question-begging, McLaughlin (2003) adopts precisely this line of reasoning:

> [SA1&2] jointly entail *E*, a contingent environmental proposition. […] If one of these premises were strongly a priori and the other merely weakly a priori, […] then not-*E* would be an empirical defeater of the merely weakly a priori premise [that is, the defeasible premise].[112]
>
> (McLaughlin 2003, p. 90-91)

Let's assume that SA2 can be known strongly *a priori*. Then, McLaughlin's argument entails

(7) SA1 is empirically defeasible$_R$ → [SA1&2 jointly entail E → ~E is an empirical defeater$_R$ of SA1]

Then, given that introspective knowledge of SA1 is empirically defeasible$_R$, we can infer

(8) SA1&2 jointly entail E → ~E is an empirical defeater$_R$ of SA1 relative to introspection

And, since SA1&2 jointly entail ~A, it follows from (8) that

[112] McLaughlin makes clear that he is talking specifically about *rebutting* defeaters (2003, p. 89). S's knowledge that p is *strongly a priori* iff it is both (i) weakly *a priori* and (ii) empirically indefeasible.

(9) A is an empirical defeater$_R$ of SA1.

Therefore, the answer to Q1 is "yes". McLaughlin argues that this conclusion commits us to saying that SA is question-begging. Given QB, this is correct only if Skeptic A is rationally obstructed from believing SA1. This brings us to Q2: *Does the truth of (9) mean that Skeptic A is rationally obstructed from believing SA1*?

We would be committed to answering in the affirmative if something like the following principle were true:

(ϕ) p is an empirical defeater$_R$ of q → S's believing p would rationally obstruct her from believing q

I think ϕ is false. Before evaluating ϕ, however, we should note that (7) presupposes a general principle as well. We'll call it *McLaughlin's Principle*:

(**MP**) p is empirically defeasible$_R$ → [p entails E → ~E is an empirical defeater$_R$ of p]

MP allows us to infer (7), which, together with ϕ, leads to the conclusion that Skeptic A is in fact rationally obstructed from believing SA1. We can reconstruct the reasoning as follows:

(I) MP (Pr.)
(II) ϕ (Pr.)
(III) SA1&2 is empirically defeasible$_R$ (Pr.)
(IV) SA1&2 entails ~A (Pr.)
(V) A is an empirical defeater$_R$ of SA1&2 (I,III,IV)
(VI) S's believing A would rationally obstruct her from believing SA1&2 (II,V)
(VII) Skeptic A's belief rationally obstructs her from believing SA1&2 (VI)

Of course, maintaining the assumption that SA2 is a non-defeasible conceptual truth, then Skeptic A would be rationally obstructed from believing the conjunction SA1&2 because she would be rationally obstructed from believing the first conjunct, SA1. Either way, from VII and

QB it follows that SA begs the question against Skeptic A. To avoid this conclusion, we must reject either MP or ϕ. MP can be derived directly from RD. If empirically defeasible$_R$ proposition p entails E, then ~E is incompatible with p. Thus, justification for ~E would directly support a proposition incompatible with p. Hence (by RD) ~E is an empirical defeater$_R$ of p. The problem, I think, lies instead with ϕ. To see this, consider the following counterexample to the reasoning just employed.

Suppose Mary is at an arcade and in need to some pocket change (having already spent all of hers on Pac-Man). Let's say you need 70¢ to get one more game in. Mary asks a friend if he has 70¢ to spare, and he hands her a small plastic bag with some loose change in it. He tells Mary, "I brought $5, but spent $4 on Space Invaders. So there should be at least 70¢ left." On the basis of this testimony, Mary both believes and is justified in believing that she is holding at least 70¢**.** Now she opens the bag, pours the change out into her hand, and counts it. She sees what looks just like one dime, two quarters, and a nickel. Then, she reasons as follows:

CHANGE

(w_{C1}) *This* perception as of one dime…etc.

(C1) I'm holding one dime, two quarters, and a nickel.

(C2) If C1, then C3.

(C3) I'm holding less than 70¢.

There are three things to note about this case. The first is that C1&2 is empirically defeasible$_R$ because C1 is empirically defeasible$_R$. (To simplify matters, let's assume for a moment that C2 is a conceptual truth and is therefore empirically indefeasible.) It is certainly possible to acquire evidence that would directly support a proposition incompatible with C1. Second, C1&2 entails

C3. Finally, before Mary opens the bag, she both *believes* and is *justified in believing* ~C3. Given this, we should be able to reason as follows:

(I) MP (Pr.)

(II) ϕ (Pr.)

(III*) C1&2 is empirically defeasible$_R$ (Pr.)

(IV*) C1&2 entails C3 (Pr.)

(V*) ~C3 is an empirical defeater$_R$ of C1&2 (I,III*,IV*)

(VI*) S's believing ~C3 would rationally obstruct her from believing C1&2 (II,V*)

(VII*) Mary's belief that ~C3 rationally obstructs her from believing C1&2 (VI*)

Is it true that Mary's believing ~C3 rationally obstructs her from believing C1&2? Again, we're supposing C2 to be a conceptual truth. Assuming that Mary is not rationally obstructed from believing conceptual truths, then she is rationally obstructed from believing the conjunction of C1&2 only if she's rationally obstructed from believing C1. Now, to say that Mary's belief in ~C3 rationally obstructs her from believing C1 trivially implies that it rationally obstructs her from believing C1 on the basis of w_{C1}. From RO it follows that Mary is rationally obstructed from believing C1 on the basis of w_{C1} *only if* justification for ~C3 would sever the evidential connection that would otherwise exist between w_{C1} and C1. But this condition is not met. As we described the case, Mary *is* justified in believing ~C3 before she opens the bag. Nonetheless, it is clear that this justification does *not* prevent w_{C1} from warranting her belief in C1. The evidential connection remains intact despite Mary's prior justification to believe ~C3. Furthermore, rationally obstructed beliefs are beliefs that it would be *irrational* to hold. But Mary's belief in C1 is certainly not irrational, regardless of her prior belief that ~C3.[113]

[113] Of course, it would be irrational of Mary to believe C1 *while still believing* ~C3. But it doesn't follow that belief in ~C3 rationally obstructs belief in C1. There are other sources of irrationality. In this case, the irrationality would simply come from the fact that belief in C1 is inconsistent with belief in ~C3.

The weak link here is clearly ϕ. A proposition p is an empirical defeater$_R$ of q when justification for p directly supports some proposition (perhaps p itself, perhaps some other proposition) incompatible with q. One consequence of ϕ, then, is that one is always rationally obstructed from believing *any* proposition q incompatible with one's current beliefs. For justification for one's current beliefs would directly support a proposition (namely, the conjunction of the propositions one currently believes) incompatible with q. It would follow (from RD and ϕ) that one's current beliefs rationally obstruct one from believing anything incompatible with them. But in that case empirical defeasibility$_R$ itself would be impossible. For one's being rationally obstructed from believing any q incompatible with one's current beliefs is (by RO) for it to be the case that one's being justified in holding one's current beliefs would prevent one from justifiedly believing q. Thus, one could never acquire justification to believe any proposition incompatible with any of one's current, justified beliefs B. But this is precisely what is required if B is to be defeated$_R$.

Take, for instance, the sheep on the hill mentioned earlier. Ted believes, and is warranted in so believing, that there are sheep on the hill on the basis of perception. He is then told that, in fact, there are no sheep on the hill. If ϕ is true, then neither this nor any other evidence could defeat$_R$ Ted's warranted belief that there are sheep on the hill. Let p = *There are sheep on the hill*; and let q = *There are no sheep on the hill*. Notice that just as q is an empirical defeater$_R$ of p, so too is p an empirical defeater$_R$ of q.[114] Trivially, justification to believe that there are sheep on the hill would directly support a proposition incompatible with

[114] This is obvious if we tell the story the other way. That is, suppose that Ted is told that there are no sheep on the hill only to later see them as plain as day. In that case, his warrant to believe that there are no sheep is defeated by evidence to the contrary. Which belief "wins out" is just a function of how we tell the story.

the proposition that there are none. Thus, if ϕ is true, then Ted's belief that p rationally obstructs him from believing q. But this is the case only if his being justified in believing p prevents his being justified in believing q. And q is not special, for we could apply the same reasoning to *any* proposition r incompatible with p. Thus, it follows that Ted could never acquire justification to believe any proposition incompatible with p. But then it is impossible that his justification for p will ever be defeated$_R$. This is obviously an absurd conclusion. Therefore, we must reject ϕ.

Now, even if the *reductio* sketched in the previous two paragraphs is sound, it doesn't follow that ϕ *always* fails. There may yet be confirming instances. Thus, we still need to know *when* it fails and why it fails when indeed it does. CHANGE is useful for this purpose. ~C3 is an empirical defeater$_R$ of C1. Yet Mary's believing ~C3 does not rationally obstruct her from believing C1. Why? The answer seems simply to be that Mary has no reason to distrust the perceptual experiences on which she bases her belief in C1. In particular, her friend's testimony to the effect that ~C3 does not give her any reason to distrust her experiences *as of* a dime, etc. Because ~C3 is not a reason to doubt one's perceptual experiences, Mary's justification for ~C3 does not keep w_{C1} from warranting her belief in C1. The justificatory connection remains intact. Hence, Mary is not rationally obstructed from believing C1.

Is Skeptic A's epistemic situation relevantly like Mary's? That is, is reason to believe A reason to mistrust one's introspective faculties? If so, then justification for A would sever the justificatory connection between introspection and SA1. Hence, Skeptic A is rationally obstructed from believing SA1 on the basis of introspection. But if A is *not* a reason to mistrust one's introspective faculties, then it seems that Skeptic A can rationally believe SA1 just as Mary

can rationally believe C1. This leads us directly back to Q3. My contention is that Skeptic A's epistemic situation is relevantly similar to that of Mary. In particular, since A is *not* a reason to mistrust one's introspective faculties, Skeptic A is not rationally obstructed from believing SA1 on the basis of introspection. Or, what comes to the same thing, A is not an empirical defeater$_U$ of SA1 with respect to introspection.

To see this, let's look at some typical cases of rational obstruction. The first is derived from Dretske's (1970) famous zebra case. Suppose you're at the zoo and want to see the zebras. You have some trouble finding them, so you ask a zoo employee where they are. She gives you directions to the zebras but warns you that "a good deal" of them are actually cleverly disguised mules. Apparently, the zoo couldn't afford too many actual zebras. Let's say that you believe the employee. You then go to see the zebras for yourself. You see what looks to you just like a zebra. Here, your belief that many of the zebras are actually cleverly disguised mules rationally obstructs you from believing that the animal you are now seeing is a zebra. At least, it rationally obstructs you from so believing on the basis of your zebra-like perceptual experiences. For reason to believe that many of the zebras are actually cleverly disguised mules is reason to believe that the conclusions one would *normally* arrive at on the basis of one's perceptual experiences could now *easily* be wrong. For if many of the zebras *are* actually cleverly disguised mules, then one's perceptual experiences *as of* a zebra could easily be *of* mules. Thus justification to believe the employee would undermine the reliability of the evidential connection between experiences *as of* a zebra and the belief that the animal one is seeing is a zebra.

Consider another example. A friend warns you that your roommate is attempting to prank you: "He replaced all the coins on your desk with very realistic fakes!" Having put these coins in your pocket earlier, you now believe that all the coins in your pocket are very realistic fakes. You take one out to inspect it. You're now holding what looks to you to be a real dime. Again, however, you are rationally obstructed from believing that you are now holding a real dime. Reason to believe that one is a victim of this kind of prank is not only reason to believe that the conclusions one would normally arrive at on the basis of one's perceptual experiences could now easily be wrong, but that they *would* be wrong. For if one *is* a victim of this kind of prank, then one's perceptual experiences *as of dime* would actually be *of* counterfeits. Thus justification to believe one's friend would undermine the reliability of the evidential connection between experiences *as of dime* and the belief that one is holding a real dime.

That brings us to Skeptic A. Skeptic A believes that A – i.e. that she is a lonely BIV. Does this belief rationally obstruct her from believing SA1 on the basis of introspection? The answer, I submit, must be "no". Notice that reason to believe that one is a lonely BIV is *not* reason to believe that the conclusions one would normally arrive at on the basis of introspection either would, or could easily be, wrong. For it is not the case that if one were a lonely BIV then introspection would mislead one about the contents of one's thoughts. In the cases above, one's perceptual experiences are rendered unreliable by one's circumstances. But the introspective faculty of a lonely BIV is no less reliable than that of any unenvatted person. Thus, being a lonely BIV is no reason to mistrust the deliverances of introspection.

Critics of arguments like SA must imagine there to be a radical mismatch between what a lonely BIV *thinks* it is thinking and what it *is* thinking. They must imagine a lonely BIV

believing that it is thinking a *water*-thought when, unbeknownst to it, it is actually thinking a thought involving the twin concept *vat-water*. But this is a misunderstanding. As Burge says, "if background conditions are different enough so that I am thinking different thoughts, then they will be different enough so that the objects of […] self-ascription will also be different" (Burge 1996, p. 96; cf. Burge 1988, p. 659). If a lonely BIV cannot think *water*-thoughts, then neither can it ascribe to itself *water*-thoughts. Instead, it would ascribe to itself *vat-water*-thoughts and would be correct in doing so.

Now, recalling the case mentioned at the end of the previous section, if a skeptic were to believe that she has recently been envatted or that she is repeatedly envatted and restored to normal life, then the epistemic situation would be different. Let's call this scenario "B", and let's refer to the kind of skeptic who believes it "Skeptic B". I think it is clear that Skeptic B *would* be rationally obstructed from believing SA1. For the possibility that one is thinking some twin concept (like *vat-water*) would have the same epistemic effect as the possibility that one is seeing a fake zebra. Thus, a victim of scenario B could *easily* be wrong about which concept figures in her water-thoughts. She would be in a situation much like Jane's. For scenario B is essentially an example of slow switching – not between Earth and Twin Earth, but between Earth and Vat-Earth. But Skeptic A doesn't have to worry about any of this. There are no twin concepts a lonely BIV might mistakenly attribute to itself. Where would it even acquire the concepts necessary for this to be a possibility? Unlike Jane or a victim of scenario B, a lonely BIV (living in a world in which nothing at all exists other than what is necessary to keep it alive) cannot be a victim of slow-switching.

Given these considerations, I do not see any reason to think that Skeptic A's situation is any different than Mary's. Mary believes that ~C3. But even though this belief is justified, it does not give her any reason to mistrust her perceptual experiences. Therefore it is reasonable for her to trust them *even where they conflict with ~C3*. So, when she has an experience w_{C1} incompatible with ~C3, she revises her beliefs accordingly. She realizes that her belief that ~C3 was mistaken, adopting C1 in its place. I think we can tell the same story about Skeptic A. Skeptic A believes that A. But even if this belief were justified, it would not give her any reason to mistrust her introspections. Therefore it is reasonable for her to trust them *even where they conflict with A*. So, when introspection reveals to her that she is thinking a thought incompatible with her being a lonely BIV, she ought to revise her beliefs accordingly. She should realize that her belief that A was mistaken and adopt SA1 in its place. And just as it is rational for Mary to conclude C3 on the basis of C1, so it is reasonable for Skeptic A to adopt ~A on the basis of SA1. Hence, SA does not beg the question against Skeptic A.

If the foregoing argument is correct, then externalists committed to the apriority of SA1 are also committed to the conclusion that it is possible to defeat certain forms of skepticism simply by reflecting on and reasoning about one's thoughts. I do not think that this is an unpalatable conclusion. In fact, I consider externalism's anti-skeptical potential to be an attractive feature.[115] In any case, ~A is itself quite modest. It does not entail that one is not a BIV, only that one isn't a *lonely* BIV (or hasn't *always* been a BIV). A recently envatted BIV can certainly think *water*-thoughts, so there is no *a priori* path from SA1 to the conclusion that one

[115] I say "potential" because, remember, we have had to assume the apriority of SA2 to get to this point. But, as we saw in section 5.1, that the relevant instantiation of MK2 (including SA2) is *a priori* is far from clear.

hasn't been envatted or isn't being radically deceived in some other way. Given this, it is not obvious to me why the compatibilist should be expected to dismiss out of hand the possibility that ~A might be *a priori* or that externalism might have anti-skeptical implications. Again, an argument is needed to establish this. McKinsey cannot simply assume it to be the case.

5.4 Conclusion

This concludes my defense of compatibilism. First, I addressed the Discrimination Argument, which appeals to the fact that, if content externalism is true, then we will not always be able to discriminate one thought-type from another. I argued that this is not a problem for the compatibilist because knowledge does not require such an ability. Then, I addressed Jessica Brown's Illusion Argument, arguing that its implications do not extend beyond singular externalism. Next, I evaluated Boghossian's Memory Argument for incompatibilism. I argued, first, that it relies on false premises about memory and that, second, it cannot be reconstructed without these false premises. In this chapter, I discussed McKinsey's reductio. I argued that even if we ignore the problems facing the closure principle on which it relies (viz. CA), it is far from clear that compatibilism reduces to absurdity. In particular, I argued that though the compatibilist may be committed to saying that certain environmental proposition are *a priori* (including e.g. ~A), these propositions are modest enough that it is not obviously absurd to suppose that they might be *a priori*.

References

Aasen, S. (2017). Object-Dependent Thought Without Illusion. *European Journal of Philosophy, 25*(1), 68–84.

Alston, W. (1995). How to Think about Reliability. *Philosophical Topics, 23*(1), 1-29.

Audi, R. (1997). The Place of Testimony in the Fabric of Knowledge and Justification. *American Philosophical Quarterly, 34*(4), 405-422.

Ball, D. (2007). Twin-Earth Externalism and Concept Possession. *Australasian Journal of Philosophy, 85*(3), 457-472.

Bartlett, G. (2018). Occurrent states. *Canadian Journal of Philosophy, 48*(1), 1–17.

Becker, K. (2008). Epistemic Luck and the Generality Problem. *Philosophical Studies, 139*(3), 353-366.

Becker, K., & Black, T. (Eds.). (2012). *The Sensitivity Principle in Epistemology.* New York: Cambridge University Press.

Bernecker, S. (2009). *Memory: A Philosophical Study.* New York: Oxford University Press.

Bilgrami, A. (1992). Can Externalism Be Reconciled with Self-Knowledge? *Philosophical Topics, 20*(1), 233-267.

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

Boghossian, P. (1989). Content and Self-Knowledge. *Philosophical Topics, 17*, 5-26.

Boghossian, P. (1994). The Transparency of Mental Content. *Philosophical Perspectives, 8*, 33-50.

Boghossian, P. (1998). What the Externalist Can Know A Priori. *Philosophical Issues, 9*, 197-211.

Brown, J. (1995). The Incompatibility of Anti-Individualism and Privileged Access. *Analysis, 55*(3), 149-156.

Brown, J. (2001). Anti-individualism and agnosticism. *Analysis, 61*(3), 213-224.

Brown, J. (2004). *Anti-individualism and Knowledge.* Cambridge, MA: MIT Press.

Brueckner, A. (1992). What an Anti-Individualist Knows A Priori. *Analysis, 52*(2), 111-118.

Brueckner, A. (1997). Externalism and Memory. *Pacific Philosophical Quarterly, 78*(1), 1-12.

Brueckner, A. (2001). A priori knowledge of the world not easily available. *Philosophical Studies, 104*, 109–114.

Brueckner, A. (2010). Externalism and Privileged Access Are Consistent. In A. Brueckner, *Essays on Skepticism* (pp. 243-258). Oxford: Oxford University Press.

Buchak, L. (2014). Belief, credence, and norms. *Philosophical Studies, 169*(2), 285–311.

Burge, T. (1979). Individualism and the Mental. *Midwest Studies in Philosophy, 4*(1), 73-122.

Burge, T. (1988). Individualism and Self-Knowledge. *The Journal of Philosophy, 85*(11), 649-663.

Burge, T. (1996). Our Entitlement to Self-Knowledge. *Proceedings of the Aristotelian Society, 96*, 91-116.

Burge, T. (1998). Memory and Self-Knowledge. In P. Ludlow, & N. Martin (Eds.), *Externalism and Self-Knowledge* (pp. 351-370). Stanford: CSLI Publications.

Burge, T. (2003). Mental Agency in Authoritative Self-Knowledge: Reply to Kobes. In M. Hahn, & B. Ramberg (Eds.), *Reflections and Replies: Essays on the Philosophy of Tyler Burge* (pp. 417-433). Cambridge, MA: MIT Press.

Burge, T. (2007). Other Bodies. In T. Burge, *Foundations of Mind* (pp. 82-99). Oxford: Oxford University Press.

Butler, K. (1997). Externalism, Internalism, and Knowledge of Content. *Philosophy and Phenomenological Research, 57*(4), 773-800.

Cohen, S. (1988). How to Be a Fallibilist. *Philosophical Perspectives, 2*, 91–123.

Comesaña, J. (2005). Unsafe Knowledge. *Synthese, 146*(3), 395-404.

Conee, E., & Feldman, R. (1998). The Generality Problem for Reliabilism. *Philosophical Studies, 89*(1), 1–29.

Davidson, D. (1987). Knowing One's Own Mind. *Proceedings and Addresses of the American Philosophical Association, 60*(3), 441-458.

Dierig, S. (2010a). The Discrimination Argument Revisited. *Erkenntnis, 72*(1), 73-92.

Dierig, S. (2010b). Goldmans Scheunen-Beispiel und das Problem der Vereinbarkeit von Externalismus und Selbstkenntnis. *Grazer Philosophische Studien, 80*(1), 179-207.

Dierig, S. (2014). The Discrimination Argument and the Standard Strategy. *Grazer Philosophische Studien, 90*, 213-230.

Dierig, S. (2018). Against Boghossian's Case for Incompatibilism. *Logos & Episteme, 9*(3), 285-306.

Dretske, F. (1970). Epistemic Operators. *The Journal of Philosophy, 67*(24), 1007-1023.

Dretske, F. (1981). The Pragmatic Dimension of Knowledge. *Philosophical Studies, 40*(3), 363-378.

Dummett, M. (1993). Testimony and Memory. In *The Seas of Language* (pp. 411-428). Oxford: Oxford University Press.

Evans, G. (1982). *The Varieties of Reference.* (J. McDowell, Ed.) Oxford: Oxford University Press.

Falvey, K. (2000). The compatibility of anti-individualism and privileged access. *Analysis, 60*(1), 137–142.

Falvey, K., & Owens, J. (1994). Externalism, Self-Knowledge, and Skepticism. *The Philosophical Review, 103*(1), 107-137.

Farkas, K. (2008). *The Subject's Point of View.* New York: Oxford University Press.

Feldman, R. (1985). Reliability and Justification. *The Monist, 68*(2), 159–174.

Field, H. (1996). The Aprioricity of Logic. *Proceedings of the Aristotelian Society, 96*, 359-376.

Forbes, G. (1995). Realism and Skepticism: Brains in a Vat Revisited. *The Journal of Philosophy, 92*(4), 205-222.

Frise, M. (2018). The Reliability Problem for Reliabilism. *Philosophical Studies, 175*(4), 923-945.

Gardiner, G. (2020). Profiling and Proof: Are Statistics Safe? *Philosophy, 95*(2), 161-183.

Gerken, M. (2007). A False Dilemma for Anti-Individualism. *American Philosophical Quarterly, 44*(4), 329-342.

Gertler, B. (2000). The Mechanics of Self-Knowledge. *Philosophical Topics, 28*(2), 125-146.

Gibbons, J. (1996). Externalism and Knowledge of Content. *Philosophical Review, 105*(3), 287-310.

Gibbons, J. (2001). Externalism and Knowledge of the Attitudes. *The Philosophical Quarterly, 51*(202), 13-28.

Goldberg, S. (1997). Self-Ascription, Self-Knowledge, and the Memory Argument. *Analysis, 57*(3), 211-219.

Goldberg, S. (2000). Externalism and Authoritative Knowledge of Content: A New Incompatibilist Strategy. *Philosophical Studies, 100*(1), 51-79.

Goldberg, S. (2003a). What Do You Know When You Know Your Own Thoughts? In S. Nuccetelli (Ed.), *New Essays on Semantic Externalism and Self-Knowledge* (pp. 241-256). Cambridge, MA: MIT Press.

Goldberg, S. (2003b). Anti-Individualism, Conceptual Omniscience, and Skepticism. *Philosophical Studies, 116*(1), 53-78.

Goldberg, S. (2003c). On our alleged a priori knowledge that water exists. *Analysis, 63*(1), 38-41.

Goldberg, S. (2005a). The Dialectical Context of Boghossian's Memory Argument. *Canadian Journal of Philosophy, 35*(1), 135-148.

Goldberg, S. (2005b). (Nonstandard) Lessons of World-Switching Cases. *Philosophia, 32*(1), 93-129.

Goldberg, S. (2006). Brown on self-knowledge and discriminability. *Pacific Philosophical Quarterly, 87*(3), 301-314.

Goldman, A. (1976). Discrimination and Perceptual Knowledge. *The Journal of Philosophy, 73*(20), 771-791.

Goldman, A. (1986). *Epistemology and Cognition.* Cambridge, MA: Harvard University Press.

Goldman, A. (2008). Immediate Justification and Process Reliabilism. In Q. Smith (Ed.), *Epistemology: New Essays* (pp. 63-82). Oxford: Oxford University Press.

Häggqvist, S., & Wikforss, Å. (2007). Externalism and a posteriori semantics. *Erkenntnis, 67*(3), 373-386.

Heil, J. (1988). Privileged Access. *Mind, 97*(386), 238-251.

Henderson, D., & Horgan, T. (2010). *The Epistemological Spectrum.* Oxford: Oxford University Press.

Hudson, R. (2014). Saving Pritchard's anti-luck virtue epistemology: the case of Temp. *Synthese, 191*(5), 801–815.

Kelp, C. (2009). Knowledge and Safety. *Journal of Philosophical Research, 34*, 21–31.

Kelp, C. (2013). Knowledge: The Safe-Apt View. *Australasian Journal of Philosophy, 91*(2), 265-278.

Kobes, B. (2003). Mental Content and Hot Self-Knowledge. In M. Hahn, & B. Ramberg (Eds.), *Reflections and Replies: Essays on the Philosophy of Tyler Burge* (pp. 201-227). Cambridge, MA: MIT Press.

Korman, D. (2006). What Externalists Should Say about Dry Earth. *The Journal of Philosophy, 103*(10), 503-520.

Kripke, S. (1980). *Naming and Necessity.* Oxford: Blackwell.

Kripke, S. (2011). Nozick on Knowledge. In S. Kripke, *Philosophical Troubles: Collected Papers, Volume 1* (pp. 162-224). New York: Oxford University Press.

Lehrer, K. (1965). Knowledge, Truth and Evidence. *Analysis, 25*(5), 168-175.

Lehrer, K., & Paxton, T. (1969). Knowledge: Undefeated Justified True Belief. *The Journal of Philosophy, 66*(8), 225-237.

Loar, B. (1988). Social Content and Psychological Content. In R. Grimm, & D. Merrill (Eds.), *Contents of Thought* (pp. 99-110). Tucson: University of Arizona Press.

Loar, B. (2003). Phenomenal Intentionality as the Basis of Mental Content. In M. Hahn, & B. Ramberg (Eds.), *Reflections and Replies: Essays on the Philosophy of Tyler Burge* (pp. 229-258). Cambridge, MA: MIT Press.

Ludlow, P. (1995a). Externalism, Self-Knowledge, and the Prevalence of Slow-Switching. *Analysis, 55*(1), 45-49.

Ludlow, P. (1995b). Social Externalism, Self-Knowledge, and Memory. *Analysis, 55*(3), 157-159.

Ludlow, P. (1998). Social Externalism and Memory: A Problem? In P. Ludlow, & N. Martin (Eds.). Stanford: CSLI Publications.

Ludlow, P. (2003). Externalism, logical form, and linguistic intentions. In A. Barber (Ed.), *Epistemology of Language* (pp. 399–414). Oxford: Oxford University Press.

Luper, S. (2006). Dretske on Knowledge Closure. *Australasian Journal of Philosophy, 84*(3), 379-394.

Majors, B., & Sawyer, S. (2005). The Epistemological Argument for Content Externalism. *Philosophical Perspectives, 19*(1), 257-280.

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

McDowell, J. (1986). Singular Thought and the Extent of Inner Space. In P. Pettit, & J. McDowell (Eds.), *Subject, Thought, and Context.* New York: Oxford University Press.

McGinn, C. (1984). The Concept of Knowledge. *Midwest Studies in Philosophy, 9*, 529-554.

McGinn, C. (1989). *Mental Content.* Oxford: Basil Blackwell.

McKinsey, M. (1991). Anti-Individualism and Privileged Access. *Analysis, 51*(1), 9-16.

McKinsey, M. (2002). Forms of Externalism and Privileged Access. *Nous, 36*(16), 199-224.

McKinsey, M. (2007). Externalism and Privileged Access are Inconsistent. In B. McLaughlin, & J. Cohen (Eds.), *Contemporary Debates in Philosophy of Mind* (pp. 53–66). Oxford: Blackwell.

McKinsey, M. (2018). *Skepticism and Content Externalism*. (E. N. Zalta, Editor) Retrieved from The Stanford Encyclopedia of Philosophy: <https://plato.stanford.edu/archives/sum2018/entries/skepticism-content-externalism/>

McLaughlin, B. (2000). Skepticism, Externalism, and Self-Knowledge. *The Aristotelian Society, Supplementary Volumes, 74*, 93-117.

McLaughlin, B. (2003). McKinsey's Challenge, Warrant Transmission, and Skepticism. In S. Nuccetelli (Ed.), *New Essays on Semantic Externalism and Self-Knowledge* (pp. 79-96). Cambridge, MA: MIT Press.

McLaughlin, B., & Tye, M. (1998a). Is Content-Externalism Compatible with Privileged Access? *The Philosophical Review, 107*(3), 349-380.

McLaughlin, B., & Tye, M. (1998b). Externalism, Twin Earth, and Self-Knowledge. In C. Wright, B. Smith, & C. Macdonald (Eds.), *Knowing Our Own Minds* (pp. 285-320). New York: Oxford University Press.

Morvarid, M. (2012). The Epistemological Bases of the Slow Switching Argument. *European Journal of Philosophy, 23*(1), 17-38.

Morvarid, M. (2013). Reference Failure, Illusion of Thought and Self-Knowledge. *dialectica, 67*(3), 303–323.

Nagasawa, Y. (2002). Externalism and the Memory Argument. *dialectica, 56*(4), 335-346.

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

Noordhof, P. (2004). Outsmarting the McKinsey-Brown argument? *Analysis, 64*(1), 48-56.

Nozick, R. (1981). *Philosophical Explanations.* Cambridge, MA: Harvard University Press.

Nuccetelli, S. (2003). Knowing That One Knows What One Is Talking About. In S. Nuccetelli (Ed.), *New Essays on Semantic Externalism and Self-Knowledge* (pp. 169-184). Cambridge, MA: MIT Press.

Owens, J. (1992). Psychophysical Supervenience: Its Epistemological Foundation. *Synthese, 19*, 89-117.

Peacocke, C. (1996). Entitlement, Self-Knowledge and Conceptual Redeployment. *Proceedings of the Aristotelian Society, 96*, 117-158.

Pollock, J. (1986). *Contemporary Theories of Knowledge.* Totowa, NJ: Rowman & Littlefield.

Pritchard, D. (2007). Anti-Luck Epistemology. *Synthese, 158*(3), 277-297.

Pritchard, D. (2012). Anti-Luck Virtue Epistemology. *The Journal of Philosophy, 109*(3), 247-279.

Pritchard, D. (2016). *Epistemology.* New York: Palgrave Macmillan.

Pritchard, D., Millar, A., & Haddock, A. (2010). *The Nature and Value of Knowledge: Three Investigations.* Oxford: Oxford University Press.

Pryor, J. (2004). What's Wrong with Moore's Argument? *Philosophical Issues, 14*, 349-378.

Pryor, J. (2007). What's Wrong with McKinsey-Style Reasoning? In S. C. Goldberg (Ed.), *Internalism and Externalism in Semantics and Epistemology* (pp. 177-200). Oxford: Oxford University Press.

Putnam, H. (1975). The Meaning of 'Meaning'. In H. Putnam, *Mind, Language, and Reality* (pp. 215-271). Cambridge: Cambridge University Press.

Putnam, H. (1981). Brains in a Vat. In H. Putnam, *Reason, Truth and History* (pp. 1-21). Cambridge: Cambridge University Press.

Sawyer, S. (1998). Privileged Access to the World. *Australasian Journal of Philosophy, 76*(4), 523-533.

Sawyer, S. (2002). In Defence of Burge's Thesis. *Philosophical Studies, 107*(2), 109-128.

Schroeter, L. (2007). The Illusion of Transparency. *Australasian Journal of Philosophy, 85*(4), 597 – 618.

Sosa, E. (1999). How to Defeat Opposition to Moore. *Philosophical Perspectives, 13*, 141-153.

Sosa, E. (2007). *Apt Belief and Reflective Knowledge, Volume 1: A Virtue Epistemology.* Oxford: Oxford University Press.

Stalnaker, R. (1991/1999). Narrow Content. In *Context and Content: Essays on Intentionality and in Speech and Thought* (pp. 194-209). New York: Oxford University Press.

Steup, M. (2003). Two Forms of Antiskepticism. In S. Nuccetelli (Ed.), *New Essays on Semantic Externalism and Self-Knowledge* (pp. 277-293). Cambridge: MIT Press.

Tolly, J. (2019). Does Reliabilism Have a Temporality Problem? *Philosophical Studies, 176*(8), 2203–2220.

Tye, M. (1998). Externalism and Memory. *Proceedings of the Aristotelian Society Supplementary Volume, 72*(1), 77-94.

Tymoczko, T. (1989). In Defense of Putnam's Brains. *Philosophical Studies, 57*(3), 281--297.

Vahid, H. (2003). Externalism, Slow-Switching and Privileged Self-Knowledge. *Philosophy and Phenomenological Research, 66*(2), 370-388.

Warfield, T. (1992). Privileged Self-Knowledge and Externalism are Compatible. *Analysis, 52*(4), 232-237.

Warfield, T. (1997). Externalism, privileged self-knowledge, and the irrelevance of slow-switching. *Analysis, 57*(4), 282-284.

Warfield, T. (1998). A Priori Knowledge of the World: Knowing the World by Knowing Our Minds. *Philosophical Studies, 92*, 127-147.

Wikforss, Å. M. (2008). Self-Knowledge and Knowledge of Content. *Canadian Journal of Philosophy, 38*(3), 399-424.

Appendices

Appendix A

Let's flesh this objection out in more detail to better highlight what's at stake. Recall that Oscar has been slow-switched, but is now back on Earth. He has been back on Earth long enough to have reintegrated into his old, English-speaking linguistic community – one where the utterance "water" refers to water. Now suppose that Oscar retains the ability to think about twater. In this situation he will express both his *water*- and *twater*-thoughts via utterances that involve the word "water". So, how can we tell the difference? When should we interpret him as thereby expressing a thought about twater instead of water?

The most systematic attempt to answer this question is due to Goldberg (2005b). First, Goldberg endorses what he calls the Current Face Value Presumption.

(**CFP**) S's being a member of a linguistic community C is sufficient to generate the *presumption* that, as long as S remains a member of C, S's utterances are to be interpreted according to the semantic norms associated with the language of C.

Thus, given the semantic norms of Oscar's linguistic community, the *presumption* is that when Oscar utters e.g. "water is refreshing" he thereby expresses a thought about water. However, Goldberg argues, this presumption can be defeated. One way it can be defeated is if Oscar *intends* to be referring to a substance that happens to be twater (whether he knows it or not). For example, let's suppose that while on Twin Earth, Oscar had a swim to cool down after a strenuous hike. Sometime after having been reintegrated into his old English-speaking linguistic community, he begins reminiscing about that swim and how pleasant it was. He then thinks to himself a thought that he expresses by uttering U_1.

(**U_1**) "*That* water is refreshing"

Intuitively, "water" in this case does not refer to water, but to twater. For if Oscar were pressed to say what he intends to refer to by his use of the word "water", he would say something like SI_1.

(**SI_1**) "The stuff I swam in after my hike"[116]

Since that "stuff" is not what "water" normally refers to when uttered by members of Oscar's linguistic community, the presumption that we should interpret Oscar according to the semantic norms of that community is defeated.[117]

In the example above, I follow Goldberg in attributing to the subject of the thought experiment an attitude that involves a *de re* reference to a specific bit of watery stuff. But I think the point can be illustrated just as well if we imagine that Oscar, after reminiscing about his swim, thinks to himself a thought that he would express by uttering U_2,

(**U_2**) "Water is refreshing"

where "water" is intended to refer to a kind. For given the fact that his utterance is prompted by his reminiscing about his swim, it is plausible to interpret him as intending to say something about the *kind* of stuff he was swimming in. If he were pressed to say what he intends his use of the word "water" to refer to, he would probably offer something like SI_2.[118]

---

[116] "SI" is short "speaker intention". I follow Goldberg here in understanding speaker intention to "designate the words S uses [or would use] to express her speaker intentions" (2005b, p. 112).

[117] In Goldberg's analysis, two other conditions must also be met. In particular, Oscar must react to the news that his intended use of the word "water" in this case runs counter to how the word would normally be interpreted given the semantic norms of his linguistic community both "(a) by disavowing any intent to be expressing with [his utterance] W the proposition that would be expressed by members of [his linguistic community] were they to utter W […]; *and* (b) by acquiring the disposition to correct any false beliefs she has regarding the subject-matter of her utterance…" (p. 114). Let us suppose these conditions are met (in this and in the following example). Goldberg's full account of the conditions under which CFP is defeated can be found in section 3 of his 2005b (p. 110-116).

[118] If pressed further, he might (given that he is now on Earth) start pointing to instances of water as examples of the kind of stuff to which he intends to refer. This might be taken as evidence that he intends by U_2 to be saying something roughly equivalent (in truth value, if not in meaning) to: *For all x, if x is either water or twater, then x is refreshing*. But, I think, a more plausible explanation for the fact that he would be disposed to indicate instances

(**SI$_2$**) "The kind of stuff I swam in after my hike"

So, again, since that "stuff" is not what "water" normally refers to when uttered by members of Oscar's linguistic community, the presumption that we should interpret Oscar as saying something about water is defeated.

For ease of reference, let's refer to the Oscar of the foregoing story as "Oscar$_2$", retaining the name "Oscar" for the subject of the original thought experiment outlined in section 2.1. Thus, Oscar$_2$, after being reintegrated into his old English-speaking community, thinks to himself a *twater*-thought that he would express by uttering U_2. So, he is thinking a first-order thought about twater (despite the semantic norms of his linguistic community). The main difference between Oscar and Oscar$_2$, then, is that CFP is defeated in the latter's case but not in the former's.

The question now is: *What does Oscar$_2$ believe about his thought in this case*? Notice here that it is not obviously impossible for Oscar$_2$ to *think* that he is thinking about water (in the same way that, for example, it would be impossible for a non-slow switched Toscar to think he was thinking about water). And if he *does* mistakenly believe that he is thinking about water in this case, then it would follow that the original Oscar does not know that he is thinking about water.

---

of water is simply that he would *think* (mistakenly) that he was pointing to instances of the same kind of stuff he swam in on Twin Earth. Indeed, we could easily imagine him correcting himself after learning that there is a difference between the two kinds. Suppose, for example, that he is apprised of his history of slow-switching and of the fact that water ≠ twater. He then considers the fact that, for all he has been told, water might not be refreshing in the same way (perhaps, for some reason, he's never gone swimming while on Earth). In this case, he will probably not want to commit himself to the claim that *water* is refreshing. Here, it is particularly clear that Oscar specifically intends to be saying something *only* about the kind of stuff he swam in on Twin Earth.

Why? Recall that in the original version, Oscar is thinking about water (CFP is not defeated). We wanted to know whether he was in a position to *know* this. In particular, we wanted to know whether his second-order belief that he is thinking about water qualifies as knowledge. The standard strategist's position is that it does as long as the reliability requirement is satisfied – or, to be more precise, that Oscar's history of slow-switching doesn't give us any special reason to doubt that Oscar's second-order belief counts as knowledge because it doesn't give us any special reason to doubt that it satisfies the reliability requirement. Now, given his history of slow-switching, the possibility that Oscar is thinking about twater is a relevant alternative. So, according to the reliability requirement, his second-order belief counts as knowledge only if it is false that *if* he were thinking about twater, he would still believe that he was thinking about water.

Now, if Oscar were thinking about twater, then he would be in $Oscar_2$'s situation – not one in which he is in some radically different environment, but simply one in which CFP is defeated. So, to determine what Oscar would believe if he were thinking about twater, we need to examine $Oscar_2$'s second-order belief. If it turns out that $Oscar_2$ believes he is thinking about water, it would follow that Oscar's second-order belief to the effect that *he* is thinking about water fails to satisfy the reliability requirement. And that is simply because then the following counterfactual would be true: *If Oscar were thinking about twater, he would still believe that he was thinking about water*.

It should be clear that the original strategy of pointing to the fact that both first- and second-order thought contents are externally individuated will not by itself allow us to rule out this possibility.

Appendix B

Another problem for GP_S is posed by a case originally due to Buchak (2014, p. 292). Here's a slightly amended version of it:

> **IPHONE**
>
> Fiona steps out of the office to get a drink, and she comes back to find that her iPhone has been stolen. There were only two people in the office, Jake and Barbara. Fiona has no evidence about who stole the phone, and she doesn't know either party very well, but she does know that nine out of ten iPhone thefts are committed by men. On this evidence Fiona believes that (**J**) Jake stole the phone. Fiona's belief is true.

Plausibly, Fiona does not *know* that Jake is the thief. As Buchak points out, that most iPhone thefts are committed by men does not by itself constitute "evidence that [Jake] in particular stole the phone" (Buchak 2014, p. 292). But Georgi Gardiner argues that Fiona's belief is probably safe nonetheless (Gardiner 2020). This is because, she says, "theft is not typically random" (Ibid., p. 172). The thought, I take it, is that people who are not otherwise inclined to steal do not randomly decide to do so. On the contrary, when someone steals something, their doing so is likely the result of a modally stable predisposition to steal[119], a predisposition that Barbara (given the fictitious statistic cited in IPHONE) probably lacks. If that is the case here, then "a lot would need to change for Barbara, rather than Jake, to be culpable" (Ibid., p. 172). It would follow that there is no nearby possible world in which Fiona forms her belief in the way

[119] Gardiner doesn't say this explicitly, but something like it is suggested by the discussion in section 4 of her 2020. Perhaps a predisposition to steal could be modeled after Gardiner's fictional "Disease C", where "Habit H" is replaced by some social force.

that she actually does – that is, on the basis of her knowledge that nine out of ten iPhone thefts are committed by men – and ends up with a false belief.[120]

Let's suppose that Jake's stealing the phone was in fact the result of a modally stable predisposition (one that Barbara lacks). Then we have a counterexample to the claim that safety is sufficient for knowledge. But, in order to turn this into a counterexample to GP_S, it must be possible to amend the case so that Fiona justifiedly believes that her belief that J is safe *but still doesn't know* that J. At first blush, this may seem straightforwardly impossible. For what would it take for Fiona to justifiedly believe that her belief that J is safe? She would have to be justified in believing that there is no nearby possible world in which Jake is innocent (for then there would be a nearby world in which she falsely believes that he is guilty). Presumably, this would involve justification to believe, among other things, something about Jake's predisposition for stealing. But if we add that Fiona has enough insight into Jake's character to be justified in believing that he has (or may have) a stable predisposition to steal, then it is no longer obvious that Fiona *doesn't* know that J.

One might counter, however, that where Fiona is justified in believing that J, she will typically also be justified in believing that J is true in all nearby possible worlds in which her phone is stolen (and hence that coming to believe J in the way that she does is in fact safe).[121]

---

[120] This is because her belief-formation method will reliably lead her to blame the man – that is, Jake. If she had blamed Jake only because, say, he was the one sitting closest to the door (and therefore had a better chance at making a quick escape), her belief would not have been safe. For Barbara could easily have been the one sitting closest to the door. In that case, Fiona would have falsely believed that Barbara was guilty. What if Barbara *had* done it? Suppose that she did, and that Fiona correctly believes so on the basis of a distrust of other women. It is interesting to note that we have just as much reason to suppose that this belief is safe as we do to suppose that her belief in IPHONE is safe. If Barbara stole the phone, her doing so was probably the result of a modally stable predisposition to steal; and a distrust of women would reliably lead Fiona to blame Barbara.

[121] Thanks to Georgi Gardiner for suggesting this to me.

That's because if J is true at all, then it is very probably because Jake has the relevant modally stable predisposition to steal. So, one might reason, evidence that Jake stole the phone will typically also constitute evidence that Jake is the kind of person who steals phones – that is, that he is a *thief*.[122] If that's right, then justification to believe the latter will typically accompany justification to believe the former. And it is plausible that if Fiona is justified in believing that Jake is a thief, then she is justified in believing that any nearby world in which her phone is stolen is one in which Jake has stolen it. The upshot of this line of reasoning, if it is sound, is that it may be much easier to get a counterexample out of IPHONE than one might have thought. For, apparently, it would require only that we amend the case so that Fiona is justified in believing, but still does not know, that J. And that does not seem impossible.

In fact, it suggests that IPHONE may constitute a counterexample to GP_S more or less as is. For one might hold, pace Buchak, that Fiona's belief that J *is* justified by her evidence – namely, that nine out of ten iPhone thefts are committed by men (and that her phone is gone). Then, given the reasoning outlined in the previous paragraph, Fiona's evidence also justifies her in believing that Jake is a thief and, therefore, that any nearby world in which her phone is stolen is one in which Jake has stolen it. Now we need only add that Fiona does justifiedly

[122] Plausibly, it will do so except in cases where the evidence points to some other explanation for the theft. For example, suppose that Fiona finds out that Jake was coerced into stealing the phone. In that case, she will have acquired reason to believe that Jake stole the phone but *not* reason to believe that Jake is a thief, if by this we mean that he has a stable predisposition to steal. To illustrate a confirming instance: Suppose Fiona had earlier noticed Jake eyeing her phone and behaving in a shifty manner. If this, in conjunction with the fact that her phone is now missing, constitutes evidence that Jake stole it, then it is reasonable to think that it also constitutes evidence that Jake is a thief.

believe this and that, as a result, her belief that J is safe. Assuming that she still doesn't *know* that J, we have a counterexample to GP_S.[123]

My response is that however plausible it might otherwise be to think that Fiona's belief that J is justified by her evidence, it becomes clearly implausible when conjoined with the reasoning just sketched. For, according to that reasoning, if Fiona's evidence justifies her in believing that J, then it also justifies her in believing that any nearby world in which her phone is stolen is one in which Jake has stolen it. But, I maintain that if Fiona is justified in believing that any nearby world in which her phone is stolen is one in which Jake has stolen it, then she is justified in believing that Barbara is not *also* a thief. Why? Consider that if Barbara is also a thief, then not a lot would need to change for Barbara, rather than Jake, to be the culprit. This means that there is a nearby possible world in which Barbara, rather than Jake, *is* the culprit.[124] But then it is false that any nearby world in which Fiona's phone is stolen is one in which Jake has stolen it. It follows that if any nearby world in which Fiona's phone is stolen is one in which Jake has stolen it, then Barbara is not also a thief. Thus, if Fiona is justified in believing the former, then she must also be justified in believing the latter.[125]

My sense, however, is that Fiona is *not* justified in believing that Barbara is not a thief. For consider the epistemic situation at the point in time just before Fiona discovers that her phone has been stolen. At this point, either Fiona is already justified in believing that Barbara

[123] As Gardiner (2020, p. 172) notes, Fiona's belief that J is insensitive. It is reasonable to think that it does not constitute knowledge for this reason, even if it is justified.
[124] Perhaps Jake is the culprit in half of all of the nearby possible worlds in which Fiona's phone is stolen, Barbara the culprit in the other half. Who is the culprit in which of these possible worlds may have to do with minute differences affecting which of the two, Jake or Barbara, is presented with a more promising opportunity to get away with it.
[125] Again, this assumes that justification is closed under entailment. However, the objection to GP_S under discussion clearly assumes a closure principle for justification *at least* as strong as the one I'm presupposing here.

isn't a thief, or she is not. If she is not, then neither will she be *after* discovering that her phone has been stolen. That's because the discovery will constitute *positive* evidence (E) that Barbara is a thief (clearly, the probability that Barbara is a thief given E is greater than the prior probability that Barbara is a thief).[126] Thus, while E may not be sufficient to justify Fiona in believing that Barbara is a thief, it will certainly not justify her in believing that Barbara is *not* a thief. Positive evidence for a proposition will typically not justify one in believing its negation.

One may hold, however, that Fiona *is* already justified in believing that Barbara is not a thief (for perhaps we are generally justified in believing of others that they are not thieves). But, even if that is correct, it seems to me that this justification is defeated by E. As stipulated in IPHONE, Fiona knows that nine out of ten iPhone thefts are committed by men. But this suggests that a full 10% are *not* committed by men. Together with E, this fact indicates a non-negligible chance that Barbara stole the phone (and, therefore, that she is a thief).[127] Given this, it would seem that Fiona is justified in believing neither that Barbara is a thief *nor* that she isn't.

[126] The prior probability here is the probability given Fiona's background knowledge – that is, all that she knew *before* learning that her phone had been stolen.

[127] One might respond by amending the case so that Fiona knows that, say, 99/100 iPhone thefts are committed by men. In that case, prior justification to believe that Barbara isn't a thief plausibly would *not* be defeated by E. For, even given E, Barbara's being the culprit would still be exceedingly unlikely. The problem with this is that the more unlikely it is that Barbara is the culprit, the harder it becomes to insist that Fiona doesn't know that J. It may be that if we make the statistic lopsided enough to ensure that Fiona will remain justified in believing that Barbara isn't a thief, then it will no longer be obvious that Fiona doesn't know that J. Indeed, if Fiona knows that virtually *all* iPhone thefts are committed by men *and* safely forms the belief that J (in part) on the basis of this knowledge, then I am inclined to say that Fiona does know that J. (Notice that it does not follow that it would be appropriate to hold Jake legally liable for the theft just on the basis of this kind of purely statistical evidence. I'm sure it often happens that the relevant people – judges, juries – *know* that S committed crime c without it being appropriate for them to convict or otherwise hold S liable for having done c. This will happen whenever the evidence underlying that knowledge is, or ought to be, legally inadmissible. Perhaps IPHONE is one such case.)

This, of course, does not prove that IPHONE cannot be made into a successful counterexample. But I think the prospects are dim. For we would need to posit (a) some new piece of evidence E that, together with everything else that she knows, would justify Fiona in believing that J and (b) either that (b1) E would also justify Fiona in believing that Barbara is not a thief or else that (b2) Fiona has prior justification to believe that Barbara is not a thief and that E is consistent with (i.e. doesn't defeat) this justification. Not only that, but the addition of E would also need to be consistent with the intuition that Fiona doesn't know that J. As we have just seen, purely statistical evidence is unlikely to do the trick. And I suspect that the addition of any evidence that directly incriminates Jake will be inconsistent with the intuition that Fiona doesn't know that J.

Vita

Though originally from Ohio, Donnie Barnett has spent most of his life in East Tennessee. He attended East Tennessee State University where he completed a Bachelor of Science in Engineering Technology. It was during this time that he discovered philosophy. He liked it so much that he decided to pursue a PhD in the subject. He ended up at the University of Tennessee doing research at the intersection of epistemology and philosophy of mind.