Externalism in the philosophy of mind contends that the meaning or content of a thought is partly determined by the environment. The view has garnered attention since it denies the traditional assumption, associated with Descartes, that thought content is fixed independently of the external world. Apparently under this assumption, Descartes also believed that he could know the content of his thoughts while suspending all judgment about his environs. (Indeed, such knowledge was thought indubitable.) Yet if externalism is correct, this may well be a mistake. As we shall see, externalism can suggest that Descartes is unable to know that his own thought represents, say, elm trees (vs. beech trees) without knowing that it is elms (and not beeches) that the thought is connected to in the world. But if such worldly knowledge is a prerequisite, then Descartes could not know the content of this thought just “from the armchair,” so to speak. So there seems to be a conflict between externalism and such armchair knowledge of one’s own thought contents (for short: “armchair self-knowledge”). The question whether this conflict is real is what drives the contemporary debate on externalism and self-knowledge.

Officially, we can put the issue in terms of an apparent tension between the following:

(EXT) Thought content is determined partly by the environment.

(SK) A subject can know from the armchair what content her thoughts have.

The issue is that EXT seemingly implies that knowing about content requires knowing about the environment. And since the latter is empirical, so too would be the former, contra SK. Now it is usually thought that, if EXT is incompatible with SK, this would be a serious problem for EXT (though some think internalism also conflicts with SK; see McLaughlin & Tye 1998a, Farkas 2003, Bar-On 2006, p. 434). Yet some think the incompatibility threatens SK instead of EXT (see section 2.4). Regardless, the interest in the debate goes beyond EXT, for it pertains to many central concerns of philosophers, such as the nature of knowledge and the relation between mind and world. The debate also touches on more specialized topics, including memory, concept acquisition, epistemic responsibility, and transcendental arguments.

In the standard terminology, the dispute is between incompatibilists who affirm the conflict between EXT and SK, and compatibilists who deny it. Incompatibilists have pressed the conflict mainly in two ways. The first way, discussed in section 2, is by a reductio ad absurdum: Incompatibilists argue that EXT plus SK entail the absurdity that one can know just from the armchair contingent facts about the external world. The second way, discussed in section 3, features a thought experiment about “slow switching” between two environments. Incompatibilists argue that such thought experiments show that EXT precludes SK. In section 4, we will end with some lesser-known issues for externalist self-knowledge.
1. Why Externalism?

Given the tension with SK, why should we accept EXT in the first place? One reason is that EXT is more or less entailed by many contemporary theories of content. For instance, several theories say (in short) that a concept has its content in virtue of certain causal relations the concept bears to an environmental referent (Dretske 1981, Millikan 1984; 1989, Fodor 1987; 1990; also, cf. Kripke 1972 and Evans 1982, ch. 6). Others see mental content as posited when interpreting a subject vis-à-vis objects in her environment (Davidson 1973; 1987, Dennett 1975). Still others see content as determined by the inferences that feature the relevant concept. Such an inference, moreover, is often defined partly by the environmental object which normally prompts the inference, or in which the inference normally terminates (Sellars 1954).

Predictably, such theories face various difficulties—so none provide a decisive case for EXT. Yet the foremost arguments for the view are not so parochial. These arguments instead rest on various thought experiments, purporting to show that internally identical subjects can host different thought contents, solely because of environmental differences (Putnam 1973; 1975, McGinn 1977, Stich 1978, Burge 1979; 1982; 1986). If the thought experiments are right about this, then internal states do not wholly determine thought content. The environment would play a part as well.

The most famous experiment imagines a remote planet called “Twin Earth,” a planet which is superficially indiscernible from Earth (so that even its events parallel the events on Earth). But there is one key difference: The compound H2O occurs nowhere on Twin Earth. In its place is a different compound that looks, tastes, etc., exactly the same, which we may call “XYZ”. Now consider Oscar, a normal adult English speaker, along with his Twin Earth doppelgänger, Twin Oscar. (Ignore that people are mostly H2O; that’s an accident of this example.) The externalist intuition is that, even if these “twins” are unschooled in chemistry, Oscar uses ‘water’ to refer to H2O yet Twin Oscar uses ‘water’ to refer to XYZ. After all, when Oscar asserts “water is scrumptious,” this is directed at water, yet Twin Oscar’s parallel assertion is in reference to XYZ, a.k.a. twin water.

Moreover, suppose we accept the Fregean view that the meaning of a term determines what the term refers to (or in Carnap’s words, that “intension determines extension”). Then, given the different referents, the twins’ terms must have different meanings as well. But Oscar and Twin Oscar are physical duplicates; so it appears the meaning of ‘water’ is not wholly fixed by speakers’ internal states. Rather, the environment also figures into it.

This claim about linguistic meaning is then extended to the content of mental representations for natural kinds like water, and for non-natural kinds like sofas (where the latter are defined by social convention) (McGinn 1977, Stich 1978, Burge 1979; 1982; 1986). Consider that when Oscar says “sofas are delightful,” he expresses a thought that references sofas. But suppose that Oscar’s Twin’s community defines a “sofa” slightly differently, so that broad, overstuffed armchairs count as “sofas.” Consequently, when Oscar asserts “sofas are delightful,” he differs from Oscar in expressing a thought that refers to (among other things) broad, overstuffed armchairs. But again, if content determines reference, the difference in reference demonstrates a difference in intension or content, in spite of identical internal states. Thus we arrive at EXT.

Some might find these science fiction stories extravagant, and it may be unclear whether they teach us anything about the real world. Yet there are not so unusual examples which also illustrate the point (Ludlow 1995b; see also Butler 1997, Tye 1998). Consider that ‘chicory’ and ‘endive’ have their denotations switched between Standard American English and British English. The two types of salad green also seem identical to the untrained eye (or palate, as the case may be). Thus a Brit and an American, both untrained, might be internally the same in all relevant respects, yet refer to different greens when asserting “chicory is scrumptious.” So ‘chicory’ will express a concept with different content, despite the subjects being internally alike. (Notably, ‘water’ itself might vary in content among different locations in the actual world. What flows from taps contains different admixtures in different municipalities. Cf. Malt 1994.)

Nevertheless, some philosophers reject these considerations for a variety of reasons (Zemach 1976, Mellor 1977, Searle 1983, ch. 8, Bach 1988; 1998, Crane 1991, Rosenberg 1994, Chomsky 1995; 2000, Segal 2000, Martinich & Stroll 2007, ch. 5, Farkas 2008, Pollock 2015; see also Fine 1975). Other philosophers do not reject EXT as much as restrict it: They posit a kind of content—so called “narrow” content—which is determined independently of the...
Yet incompatibilists reply that the second reading cannot be what is really intended. For it renders EXT into something (1) is not guaranteed here, and the reductio seems in doubt.

implies "H

thought

Even so, the first reading is not the usual interpretation of EXT. On a more standard construal, EXT holds that a water

implications are apriori. So if EXT on this first reading is true, (1) is apriori (and nothing here prejudges what

dilemma, corresponding to two ways that EXT might be interpreted. On the first reading, EXT holds that a water thought

reductio need not assume that (McKinsey 1994; 2002; 2007). One can first see the incompatibilist as presenting a

However, this assumes that E is a condition where water exists (or existed at one time). Yet the incompatibilist's reductio need not assume that (McKinsey 1994; 2002; 2007). One can first see the incompatibilist as presenting a dilemma, corresponding to two ways that EXT might be interpreted. On the first reading, EXT holds that a water thought logically or conceptually implies some environmental condition E. Yet for the incompatibilist, logical or conceptual implications are apriori. So if EXT on this first reading is true, (1) is apriori (and nothing here prejudges what E is).

Even so, the first reading is not the usual interpretation of EXT. On a more standard construal, EXT holds that a water thought metaphysically implies some environmental condition E. Yet metaphysical implications, e.g., that "water is wet" implies "H₂O is wet," are not always apriori. (See Kripke 1972 and the entry on Rigid Designators.) So the apriority of (1) is not guaranteed here, and the reductio seems in doubt.

Yet incompatibilists reply that the second reading cannot be what is really intended. For it renders EXT into something trivial—it is obvious that Oscar's water thoughts metaphysically imply a number of external conditions, e.g., that Oscar

2. The Reductio to Armchair Knowledge of the External World

But is the tension real? One pro argument is a reductio ad absurdum: The claim is that EXT and SK jointly entail an obvious falsehood, namely, that Oscar can know from the armchair contingent facts about the external world (McKinsey 1991; 1994; 2002; 2007, Brown 1995, Boghossian 1997; 1998). If so, then EXT and SK cannot be jointly true, and the incompatibilist wins the debate.

The argument is this: Assume EXT and SK for reductio, and let E name the Environmental condition that EXT requires for a concept to have a water content (versus a twin content).[11] Apparently, then, EXT and SK imply that Oscar can know from the armchair—or in the incompatibilist's idiom, know “apriori”—both of the following:

1. If Oscar is thinking that water is wet, then E obtains.
2. Oscar is thinking that water is wet.

In this, p is known apriori (or just: “p is apriori”) iff p is known without a concerted empirical investigation of the external world. (Note that introspective knowledge counts as "apriori" in this context, even though it arises from a kind of introspective experience. For more on the present conception of the apriori, see Miller 1997, Davies 1998, Nuccetelli 1999, and McLaughlin 2000; 2003). On the current line, (1) can be known apriori if EXT is true (though as we shall see, there are different ways to spin this). Whereas, (2) is apriori given SK. Yet if Oscar has apriori knowledge of both (1) and (2), Oscar can then use modus ponens to know apriori that:

3. E obtains.

But E is an environmental condition that presumably can only be known empirically. So EXT and SK apparently lead to an absurd sort of apriori knowledge, and are thus incompatible. (Henceforth, call this argument "the reductio.")

Before we consider the compatibilist's replies, one clarification of the reductio is in order. For it may be unclear whether EXT entails (1) in the first place (whether apriori or not). However, if EXT is derived from one of the causal theories of content (from section 1), then acquiring the water concept requires causally interacting with water (if only via your linguistic community). (See McGinn 1989, ch. 1 passim, McCullough 1999, but contrast with McLaughlin & Tye 1998a, Ball 2007.) Thus possessing the concept requires the existence of water, meaning that the view entails (1). Even so, if the incompatibilist wants to add that (1) is entailed apriori, she would need the contentious premise that some causal theory of content is entirely apriori.

So on a more charitable reading, the incompatibilist's reductio does not target an externalism based on a causal theory. It is rather directed at a view based solely on the Twin Earth thought experiments, given that the experiments are more naturally seen as apriori. (But see Baker 2007.) For the remainder of section 2, 'externalism' and 'EXT' shall refer exclusively to this kind of Twin Earth based externalism.

2.1 Reply 1: Environmental Implications are Empirical

Yet it may remain unclear whether EXT is committed to (1). For the thought experiments alone do not assume that water is necessary for water thoughts (Brueckner 1992). The thought experiments indeed assume that suitable interaction with water is sufficient for Oscar to have water thoughts. They also assume that Twin Oscar's interactions with XYZ are insufficient. But this does not yet imply that interacting with water is the only way that someone can arrive at water thoughts.

However, this assumes that E is a condition where water exists (or existed at one time). Yet the incompatibilist's reductio need not assume that (McKinsey 1994; 2002; 2007). One can first see the incompatibilist as presenting a dilemma, corresponding to two ways that EXT might be interpreted. On the first reading, EXT holds that a water thought logically or conceptually implies some environmental condition E. Yet for the incompatibilist, logical or conceptual implications are apriori. So if EXT on this first reading is true, (1) is apriori (and nothing here prejudges what E is).[12]

Even so, the first reading is not the usual interpretation of EXT. On a more standard construal, EXT holds that a water thought metaphysically implies some environmental condition E. Yet metaphysical implications, e.g., that "water is wet" implies "H₂O is wet," are not always apriori. (See Kripke 1972 and the entry on Rigid Designators.) So the apriority of (1) is not guaranteed here, and the reductio seems in doubt.

Yet incompatibilists reply that the second reading cannot be what is really intended. For it renders EXT into something trivial—it is obvious that Oscar's water thoughts metaphysically imply a number of external conditions, e.g., that Oscar
was born (McKinsey, 1994; 2002; 2007, Rowlands 2003). But EXT is supposed to be more substantive than that. Compatibilists reply, however, that the thought experiments suggest that a water thought metaphysically implies some content specific environmental condition, such as the existence of water (Brueckner 1995).

So the way seems clear for the compatibilist to reject the apriority of (1). But as before, the incompatibilist may respond that we are illicitly assuming E to be the condition simply that water exists. Yet the strongest version of the reductio may not assume that. One might instead understand E as follows:

4. $E = \text{Water or a community using the water concept exists (or, existed)}$ (Brown 1995, Boghossian 1997).

If E is identified as this disjunctive condition, it is more believable that (1) is apriori under EXT. And if that is believed, the reductio then presses that EXT permits apriori knowledge that either water or a community exists. Yet that too should not be apriori.

But assuming (4), why would EXT render (1) apriori? Granted, many externalists find it apriori plausible that besides interacting with water, Oscar could acquire the water concept from other community members. That suggests that if Oscar has acquired the concept, (4) must be true. However, some externalists conjecture that there are other conditions under which the water concept can be acquired. Even if Oscar is socially isolated and lives in a waterless environment, he can conceivably arrive at the concept by hypothesizing the H$_2$O molecule (Burge 1982). If this cannot be ruled out apriori, then EXT does not secure the apriority of (1)—even if (4) is assumed.

Yet incompatibilists reply that (ironically) these very reflections reveal apriori something that should not be apriori (Brown 1995, Boghossian 1997). Namely: If Oscar is aware that he has no such hypotheses (i.e., if he is “agnostic” about chemical essences), then (1) must be knowable apriori. In more detail, if Oscar is agnostic about chemistry and knows that apriori about himself, he can use the reflections above to know apriori that his water thoughts metaphysically require the existence of water or a community. From there, apriori knowledge of E seems imminent.

Nevertheless, the compatibilist’s opening objection lingers: Nothing yet shows that the only ways to acquire water thoughts, under EXT, are in the ways catalogued so far (Ball 2007). For all that has been shown, it remains possible that essence-agnostic Oscar is socially isolated and his water thoughts reference no natural kind. (See also Gallois & Hawthorne 1996, Gallois 2008, and McLaughlin & Tye 1998a, b, c.)

But perhaps additional apriori considerations can rule this out (Brown 1995). Suppose that Oscar has the concept of a sofa and knowledgeably applies the concept in a variety of cases. Yet assume he is agnostic on whether the concept applies to broad over stuffed armchairs. Then, if Oscar has no community to settle the matter, EXT may suggest that Oscar’s concept is indeterminate in what it refers to. And that seems odd.

However, the compatibilist may ask why this is supposed to be a problem (McLaughlin & Tye, 1998a, b, c). The indeterminacy may show only that Oscar’s concept is vague and not that Oscar fails to possess any concept whatsoever. (Brown 2001 replies by defining ‘agnostic’ so it applies only if the concept is non-vague. Yet Brown 2004, ch. 8 concedes that Oscar then can no longer know apriori if he is “agnostic.”) So, here too, it seems that EXT does not reveal apriori that water thoughts metaphysically require a content specific environmental condition.

2.2 Reply 2: Externalism about Empty Concepts

Still, it may remain intuitively compelling that a Twin Earth thought experiment, if cogent, would show (1) apriori—even if E is just the condition where water exists.$^{[13]}$ For the experiments seem to show apriori that, to have the bona fide water concept, the concept must be non-empty, i.e., its referents must exist. However, a compatibilist can counter this by offering a semantics of empty concepts that is consonant with EXT (Stoneham 1999, Sawyer 2003, Goldberg 2006b, Haukioja 2006, Parent 2017, ch. 4; see also Pryor 2007, pp. 184–185). Even so, there are two arguments in the literature suggesting that any externalist view of empty concepts must fail (Boghossian 1997; 1998, Segal 2000, Besson 2012).

On its face, empty-concept externalism can be supported by a Twin Earth thought experiment (Boghossian 1997).$^{[14]}$ Consider “Dry Earth”—a planet exactly like ours, except wherever water occurs on Earth, there are only watery looking holograms (though Dry Earthians mistake these for a real kind). The intuition is that unlike Oscar, Dry Oscar uses ‘water’ to express an empty concept. And if Dry Oscar is not referring to water, then as before, the Frege-Carnap view implies that his concept has a different intension or content. So even for empty concepts, content seems determined partly by facts about the environment (including facts about what is absent from the environment).

However, carrying this further may lead to absurdity, since externalism about empty concepts can suggest that a concept’s form is sensitive to the environment (Boghossian 1997; 1998). Suppose, as is plausible, that empty concepts have a descriptive form—so that (e.g.) the unicorn concept is understood as the concept “horse with a horn.” For most externalists, this contrasts with our concept of water which does not have a descriptive form, but is rather atomic. If that is correct, then Dry Oscar expresses a descriptive concept when using ‘water’, whereas Oscar expresses an atomic one. In which case, the environmental difference creates a difference in form as well as content. But that is strange, since form seems to be a purely syntactic matter, fixed entirely by what’s in the head.

At least one compatibilist has tried embracing environment-dependent syntax, though he concedes it is an uncomfortable view (Ludlow 2003; 2011). At the least, such a stance would seem ad hoc (Besson 2012). Others question the “supervenience principle” in the argument, viz., that a difference in form necessitates a difference in internal state (Corbi 1998, Brown 1999, Korman 2006, Parent 2015a, 2017, ch. 4). Further argument has been offered...
However this issue unfolds, externalism about empty concepts faces a further challenge (Segal 2000, pp. 54–56). Consider: Either the externalist sees the unicorn-concept as necessarily empty or not. If the former, then the view implies that unicorns are impossible objects, like round squares. And that is rather counterintuitive (yet see Kripke 1972 on unicorns). But taking the other horn (sorry), a contingently empty concept would have its content fixed by descriptors like “horse with a horn.” In this case, EXT seems false. For the constancy of the descriptors guarantees that the concept has the same content in every possible world, whether or not it contains unicorns (Segal 2000; see also Besson 2012). (But for replies, see Korman 2006 and Parent 2015a, 2017, ch. 4.)

2.3 Reply 3: Transmission Failure

If externalism about empty concepts fails, then EXT may well entail (1) apriori, per the incompatibilist’s reductio. But a compatibilist could try a different line of resistance. She might simply grant that (1) and (2) are apriori, but still try to block apriori knowledge of (3). (Here too, E is assumed to be just the condition where water exists, though it is believed that similar points hold if (4) is true instead.) The most prominent strategy of this type is to deny the “transmission of warrant” that supposedly occurs in the deduction of (3) from (1) and (2) (hereafter, “the deduction”). The idea is that, even if the externalist agrees that (1) and (2) are warranted, none of this warrant “transmits” to (3) in the deduction (Davies 1998; 2000; 2003a; 2003b, Wright 2000a; 2000b; 2003; 2011; 2102; 2015; see also Wright 1985). If so, the deduction cannot furnish apriori knowledge of (3), pace the incompatibilist.

For brevity, let us call the advocate of transmission failure “the advocate.” Then, the advocate’s core idea is that even though (1) and (2) entail (3), they do not prove (3). For in brief, the “proof” seems question begging (cf. McLaughlin 2000). ‘Question begging’, however, may not be the right word here (Babee 2001, Prichard 2002, Silins 2011). Still, the idea is that (3) in some sense is presupposed, once the deduction gets started. If such circularity is operative, then the deduction cannot create any warrant for (3), much less prove it. Or in the advocate’s idiom, the warrant for (1) and (2) cannot transmit to (3) in virtue of the deduction.

As a satisficing approximation, the transmission of warrant generally occurs iff:

(TW)

If a subject S is warranted in believing p and is warranted in believing that p entails q (and S further recognizes that q follows deductively from these beliefs)—then S acquires, potentially for the first time, a warrant in believing q. (Cf. Wright 2000a, p. 140–141.)

Yet (TW) should be distinguished from epistemic closure, as discussed by some anti-skeptics (e.g., Dretske 1970). Some compatibilists think that the reductio also exemplifies a failure of epistemic closure (see Bernecker 2000, Hale 2000, Sawyer 2006, although see Sawyer 2015 for a “contrastive” account that upholds closure). Regardless, transmission failure is a different affair, if only because epistemic closure concerns “knowledge” rather than “warrant.” Nevertheless, warrant transmission differs even from the “closure of warrant,” characterized as follows:

(CW)

If S is warranted in believing p and warranted in believing that p entails q, then S is warranted in believing q (or is able to be thus warranted).

The difference between transmission and warrant closure is best seen in cases of question begging. Take an overtly circular argument which concludes (3) using (3) as a premise. Trivially, anyone warranted in believing the premise is warranted in believing the conclusion. So (CW) is satisfied in this case. But the argument does not transmit warrant to its conclusion; it cannot bestow any warrant on (3) that was not already there. From a different angle, a person cannot use the argument to gain knowledge of (3) for the very first time. Hence, warrant fails to transmit onto the conclusion, even though warrant closure holds trivially.

The advocate thus cries “transmission failure” in relation to (1)–(3), on the grounds that (3) is “presupposed.” Yet in this case, the presupposition charge is not as straightforward. For clarity’s sake, let us call (1) and (2) the “premises” of the deduction, to contrast them with the premises of the reductio. (The latter are stronger in claiming that the former are apriori.) Then, the advocate bases the presupposition charge on the following: An externalist can rationally grant that the premises are warranted only if she has certain “background assumptions” which already imply (3). The point is sometimes framed in terms of “information dependence”: An externalist who grants warrant to the premises “rationally requires certain collateral information” that includes (3) (Wright 2000a, p. 149). Consequently, an externalist’s “acceptance of those [background] assumptions cannot be rationally combined with doubt about the truth of [(3)]” (Davies 2003a, p. 43, but contrast Davies 2000, p. 402).

For simplicity (since rational externalists are the only ones we care about), assume henceforth that “the externalist” is rational. Then, the presupposition charge is most readily appreciated via the following approximation: When considering the reductio,

\[ B_{(3)} \]

An externalist grants that the premises are warranted only if (3) is already among her Background assumptions.
ordinarily has a non-empirical warrant. Still, perhaps W. could soften this by reiterating that this is a “thin” warrant whether it is apriori warranted (Davies 2000). But such a response is not available to W., for he clearly thinks that (3) position may suffice for the advocate’s purposes, since transmission failure occurs if (3) is

Advocates can respond in two different ways here. One is to be neutral on whether (3) is apriori warranted. A neutral that (3) can be apriori warranted Sawyer 2006, Brueckner 2006). For the absurdity was not that an apriori warrant transmits; instead, it was in thinking (2) -—then she cannot rationally put (3) in doubt. For if the existence of water were in doubt, then given her warranted metaphysical claim, this rationally compels a further doubt on whether the thought expressed by (2) exists in the first place. But the existence of the thought must be acknowledged before one can concede a warrant to it. So our externalist will rationally allow a warrant to both *premises only if she already assumes (3) as background, which is just what B(3) says.

The second argument we might attribute to a character called “Wrightgenstein” or just “W.” for short, since it originates in Crispin Wright’s interpretive work on Wittgenstein. Qua externalist, W. acknowledges the possibility of error in discerning one’s own mental contents, as when Dry Oscar misjudges his empty concept to have a natural kind content. Even so, he sees no reason to surrender the warrant possessed by ordinary judgments about content. This parallels a natural view of perceptual judgment: Even though perceptual illusions occur, ordinarily we still possess warrant for our perceptual judgments about the world. W. thinks this is because our ordinary perceptual and self-attributing judgments enjoy a kind of “default” warrant. Such a warrant discounts the possibility of illusion, even when one has gathered no supporting empirical evidence. Though if a reasonable suspicion arises, the default no longer applies. But ordinarily our judgments are “presumed innocent” of error until proven (or at least evidenced to be) guilty; this in turn is rooted in our communal practices, and is justified on pragmatic grounds (Wright 2000a, p. 152–153).

All this is important for B(3), since if W. grants (2) a warrant in the absence of empirical investigation, it will be a default warrant (at least in part). And this kind of warrant exists only in *ordinary cases. Thus, he grants this warrant only if he assumes that there is no Dry Earth illusion where (3) merely appears to be true. So in particular, in the context of the reductio, W. allows an apriori warrant to the *premises only if (3) is already assumed as background. QED

Let us now consider some objections to transmission failure. One accuses the advocate of conflating different types of warrant (Raffman 1998, Sainsbury 2000). Note here that, if the deduction fails to prove the conclusion, this does not mean it transmits no warrant at all. (Beebee 2001, Burge 2003). The objector accordingly suggests that the deduction indeed transmits a warrant, though it is a “thinner” warrant, i.e., a warrant that is weaker than what a proof would bestow. But if a thinner warrant transmits, then contra the advocate, the deduction indeed confers some apriori warrant (albeit non-demonstrative warrant) onto (3). This, then, vindicates a version of the reductio—for the intuition is that we cannot warrant apriori a claim about the world at all.

But is this intuition correct? To evaluate it we need further clarity on what a “thin” warrant is. Some understand this as the warrant failing as a warrant for equivalent propositions. For instance, a non-empirical thin warrant for (2) may fail to warrant the extensionally equivalent ‘I am thinking that H₂O is wet’ (Davies 2003b, p. 114 and 118). Alternatively, “thinness” for W. may reflect that the warrant is not something earned by gathering empirical or proof-theoretic evidence. It is instead just a defeasible “entitlement” that holds only in the ordinary cases (Wright 2003).

If the latter is what “thinness” consists in, then an advocate could try to grant the transmission of a thinner warrant—but still emphasize that transmission failure is the rule when the warrants are thicker (Davies 2003b, but compare Wright 2003, p. 72). Yet an advocate might take a stronger stance by doubting whether any thinner warrants transmits. After all, this putative warrant dissipates immediately once any earnest doubt is raised (Wright 2000b, yet contrast with Wright 2003; 2011).

A different objection to the advocate is that, despite appearances, she ends up surrendering SK after all. This is because (3) is empirical and it is conceded that (2) presupposes (3) (cf. Wright 2000, p. 151–152). And so, it is hard to see how (2) can be warranted purely apriori if it presupposes empirical facts about the world.

In response, at least one externalist has relinquished apriori knowledge of (2), but emphasizes that some self-knowledge still remains apriori (Rowlands 2003, ch. 8). In particular, an externalist could still know apriori that:

5. If E obtains, then I am thinking that water is wet.

But for his part, W. thinks apriori knowledge of (2) should not be surrendered. Instead, he claims we should not assume that the warrant for (3) is inevitably empirical. For (3) ordinarily has a non-empirical “default” warrant. So given that (2) presupposes (3), a non-empirical warrant for (3) makes possible an apriori warrant for (2). Even so, W. is quick to remind us that a default warrant for (3) is too “thin” to prove (3).

However, if the advocate grants an apriori warrant to (3), this may succumb already to the reductio (Brown 2003; 2004, Sawyer 2006, Brueckner 2006). For the absurdity was not that an apriori warrant transmits; instead, it was in thinking that (3) can be apriori warranted at all.

Advocates can respond in two different ways here. One is to be neutral on whether (3) is apriori warranted. A neutral position may suffice for the advocate’s purposes, since transmission failure occurs if (3) is presupposed, regardless of whether it is apriori warranted (Davies 2000). But such a response is not available to W., for he clearly thinks that (3) ordinarily has a non-empirical warrant. Still, perhaps W. could soften this by reiterating that this is a “thin” warrant.
knowledge might be possible in that, once the subject has acquired the relevant concepts, she can use them to gain
In response, a defender of SK may concede that there is a “training period” for acquiring a concept. Still, apriori
acquisition depends on empirical knowledge. Yet if that’s right, then even knowledge of (2) has an empirical basis,
Indeed, it is typically assumed that a person acquires a concept from experience—which already portends that
argument against SK can be formulated as long as
In the acquisition process, however, the role of demonstrative identification may be unclear. Nevertheless, a similar
knowledge of (3). So one can hardly learn (3) for the first time via knowledge of (2)—the learning proceeds in the other
It comes to the reductio, the absurdity stems from thinking that (2) is apriori, since knowledge of (2) already rests on
That in turn normally suffices for knowing that water exists. So when
kind by perceptual means (Brewer 1999; 2000a; 2000b). If so, then to possess the water concept is already to know
perceptually where (instances of) the stuff resides. That in turn normally suffices for knowing that water exists. So when
attributions of thought, which cuts into so called “first-person authority.”
There is a more recent incarnation of this sort of view, supported by the latest from cognitive science (Carruthers 2011).
There is a more recent incarnation of this sort of view, supported by the latest from cognitive science (Carruthers 2011).
The claim is that one interprets sensory input when attributing thoughts to oneself, just as in attributing thoughts to
A rather different line against SK derives from an externalist view about concept possession. This view hypothesizes
A final worry for W. is that he is conflating the psychological phenomenon of SK with something merely linguistic or
grammatical” (Snowdon 2012). If an externalist wishes to explain how knowledge of content is possible, why would she appeal to a linguistic convention to illuminate the matter? It is clear, however, that W. never meant to forbid psychology from having its say (Wright 2015, p.55). But it is possible that our linguistic practices for crediting self-knowledge have mislead us about the psychological facts. An analogy with ethical expressivism is helpful on this score (cf. Wright 2015, p. 55) Instead of investigating the “real moral facts,” ethical expressivists try to account for moral discourse, precisely due to the suspicion that the discourse may not be grounded in bona fide facts of the relevant sort. Similarly, there may be a suspicion in W. that a robust psychological phenomenon of “armchair self-knowledge” simply does not exist.
Nonetheless, it may be desirable for us to speak as if there was such knowledge, given our aims and interests more broadly.

2.4 Reply 4: Self-Knowledge is Empirical

So far, the replies to the reductio have been “combative” in that they seek to undermine some tacit or explicit premise therein. Yet other responses are “concessive” in that they admit that the reductio has a point. One concessive response is to reject EXT in light of the reductio, though as we saw in section 1, EXT has its own arguments to contend with. But a different concession, considered in this section, is to forgo SK. (There is also a “quasi-concessive” response, considered in section 2.5, where the apriority of (3) is simply embraced rather than rejected as absurd.)

Some incompatibilists indeed take the lesson of the reductio to be that SK is false (see, e.g., McKinsey 2002). The sentiment is often that SK should have been highly suspect from the start. Jacob (2004) once expressed this well as follows:

[T]he contemporary philosophical situation is puzzling. On the one hand, few if any features of the special epistemic authority granted by both the traditional empiricist and the traditional rationalist pictures of introspective self-knowledge have survived recent philosophical scrutiny. On the other hand, several philosophers...assume that the alleged special epistemic authority granted to introspective self-knowledge by traditional epistemology can bear the burden of an argument against content externalism...I take externalism about the contents of an individual’s thoughts...to be more plausible...than anything we may think about introspective self-knowledge...Imposing top down constraints...from the assumptions about the alleged epistemic status of introspective self-knowledge sounds to me like putting the cart before the horse. (pp. 401, 402)

This attitude is encouraged by research from empirical psychology, which suggests that self-knowledge is largely an illusion. (See for instance Nisbett & Wilson 1977, Nisbett & Ross 1980, Gopnik 1983.) However, even if the research shows we are often ignorant of our own minds, it does not follow that this is most often the case. In fact, it has been argued that the current research does not even preclude infallible self-directed judgments, such as those described in section 3.1 (Parent 2016; 2017, ch. 2).

Still, there are other reasons to surrender SK, and this would be one way to understand the lesson of the reductio. For instance, well-known Quinean arguments exist against apriori knowledge, though this is not the place to discuss these. (But see section 3 of the entry on Quine.) Ryle 1963, ch. 6 independently argued that self-knowledge is apriori; such knowledge was instead achieved via empirical observation of one’s own behavior. Yet many contemporary philosophers balk at this, since it at least underestimate the “authority” that Oscar usually has in judging his own mental states (compared to others’ judgments) (Boghossian 1989, pp. 7–8). On the other hand, given the psychological research mentioned above, a neo-Rylean view may provide just the right amount of first-person authority, to the extent that Oscar observes more of his behavior than anyone else (Ludlow & Martin 1993; Martin 1994; McGeer 1996). There is a more recent incarnation of this sort of view, supported by the latest from cognitive science (Carruthers 2011). The claim is that one interprets sensory input when attributing thoughts to oneself, just as in attributing thoughts to others. However, the “sensory input” behind self-attributions may include input from interoception, proprioception, and the like. Nevertheless, the view maintains that interpretation of “input” is central in both first-person and third-person attributions of thought, which cuts into so called “first-person authority.”

A rather different line against SK derives from an externalist view about concept possession. This view hypothesizes that possession of a (non-empty) kind concept requires the ability to ostend or demonstratively identify instances of the kind by perceptual means (Brewer 1999; 2000a; 2000b). If so, then to possess the water concept is already to know perceptually where (instances of) the stuff resides. That in turn normally suffices for knowing that water exists. So when it comes to the reductio, the absurdity stems from thinking that (2) is apriori, since knowledge of (2) already rests on knowledge of (3). So one can hardly learn (3) for the first time via knowledge of (2)—the learning proceeds in the other direction.

In the acquisition process, however, the role of demonstrative identification may be unclear. Nevertheless, a similar argument against SK can be formulated as long as some kind of empirical knowledge is needed to acquire the water concept. Indeed, it is typically assumed that a person acquires a concept from experience—which already portends that acquisition depends on empirical knowledge. Yet if that’s right, then even knowledge of (2) has an empirical basis, since possession of the water concept does. Hence, knowledge of (2) does not qualify as purely apriori.
In response, a defender of SK may concede that there is a “training period” for acquiring a concept. Still, apriori knowledge might be possible in that, once the subject has acquired the relevant concepts, she can use them to gain...
new knowledge without further empirical inquiry (Boghossian & Peacocke 2000, p. 2). But one problem is that EXT usually portrays concept possession in a way that blurs the distinction between the "training phase" and the post-acquisition phase (Hawthorne 2007, p. 213). For the externalist thought experiments allow a subject to possess (e.g.) the sofa concept even prior to being "trained" that no armchair is a sofa.¹⁹

A final line against SK is that, if (1) is not apriori (as is plausible), it follows that apriori knowledge of (2) is impossible (Gertler 2004). After all, if (1) is not apriori, then Oscar apparently cannot discriminate apriori between his water concept and the Dry Earthian concept. So it seems incorrect to say that (2) is something known. Now in fact, this raises the extremely important question of whether such discriminations are necessary for self-knowledge, but further discussion will be delayed until section 3. The discrimination sub-literature usually addresses the "slow switching" arguments rather than the incompatibilist's reductio.

2.5 Reply 5: A Transcendental Argument?

The last response to the reductio is to deny that the argument is a reductio at all. Instead, this compatibilist embraces that EXT and SK facilitate armchair knowledge of the world. Some externalists have even offered independent arguments to that effect (Putnam 1981). But these independent arguments are significantly more complex, having to do with model theory and Skolem's "paradox." Such things shall not be discussed here (but see the entry on skepticism and content externalism).

Yet as concerns the incompatibilist's reductio, it is unclear if any externalist accepts it as an anti-skeptical argument (though see Peacocke 1996, p. 152). Even if one believes that some transcendental argument can succeed against skepticism, it is hard to believe that the incompatibilist's reductio is a case in point (cf. Davies 1998, p. 353).

Be that as it may, Warfield (1998) and Sawyer (1998) are often interpreted as embracing the reductio qua anti-skeptical argument. But the fact is that Warfield is mostly addressing Putnam's (1981) argument. (Warfield mentions the reductio in passing, but he endorses Brueckner's 1992 compatibilism from section 2.1, instead of a transcendentalist reply.)

Sawyer, moreover, explicitly discourages those who interpret her (1998) as embracing the reductio as an anti-skeptical argument: "Remarks by others indicate that in general the question of whether externalist arguments yield apriori warrants for beliefs about environmental conditions is seen as synonymous with the question whether we have an apriori refutation of external world skepticism. I do not see the questions as synonymous" (2006, p. 150). Instead, Sawyer's view seems similar to that of W., where speakers are given a prima facie (although quite defeasible) warrant when ascribing their own mental contents. Granted, Sawyer allows that given EXT, such a non-empirical warrant can be transmitted to (3).²⁰ But it is not the kind of warrant that suffices to undermine skepticism; the warrant is rather of the "thin" variety, noted at the end of section 2.3. Nevertheless, although Sawyer's view is not as radical as many have thought, she would face similar obstacles as W.

3. Slow Switching Arguments

We have explored one argument against the joint truth of EXT and SK, namely, the reductio to armchair knowledge of the world. But there is a further group of arguments to the same incompatibilist end. Such arguments rest on a different thought experiment, the "slow switching" experiment, to bolster the incompatibility of EXT and SK. In brief, the new thought experiment suggests that under EXT, self-attributions of thought content are not sufficiently discriminating to vindicate SK.

The slow switch thought experiment again features Oscar on Earth—yet this time, we suppose he is unwittingly switched to Twin Earth (e.g., during his sleep by a secret government operative) (Burge 1988, Boghossian 1989). Upon his arrival, he then takes up his usual activities, being none the wiser. Now the externalist's intuition is that although Oscar's use of 'water' on Twin Earth initially refers to water on Earth, over time it comes to denote XYZ. (The switch is thus a "slow" one.) After all, Oscar will acquire the habit of using 'water' in the presence of XYZ, and of using 'water' to converse with Twin Earthians about the stuff. Yet if 'water' changes its referent, then (by the earlier Frege-Carnap view) this signals a shift in intension or content. And the change in content can occur without Oscar noticing. Indeed it seems Oscar cannot detect the switch just from the armchair.

What this indicates is that Oscar is unable to discriminate from the armchair whether assertions using 'water' have a water content versus a twin water content. As with the original Twin Earth example, the lesson here is extended to mental content as well. Hence, if a "water thought" is one that Oscar expresses with help from the term 'water', the slow switch experiment indicates that:

A. Oscar cannot discriminate from the armchair that his water thought has a water content rather than a twin water content.

His inability to discriminate, moreover, seems pertinent to SK. As Burge (1988, p. 653) puts it, (A) can easily suggest that SK is false in Oscar's case, i.e.:

B. Oscar cannot from the armchair what content his water thought has.

Generalizing from this, (B) suggests that SK ends up false if EXT is true.

In section 3.1, we shall see that Burge denies that (A) entails (B). Yet in light of his resistance, Boghossian (1989)
formulates a different version of the slow switch argument. This second version, known as the memory argument, runs as follows. As before, suppose Oscar unwittingly is subject to a slow-switch, so that the concept he expresses with ‘water’ at $t_1$ denotes water, and the concept he expresses at $t_2$ denotes twin water. Then, if $p$ is a thought that Oscar uses ‘water’ in expressing, the memory argument runs as follows (adapted from Ludlow 1993a):

a. If Oscar forgets nothing, then what Oscar knows at $t_1$, Oscar knows at $t_2$.

b. Oscar forgot nothing.

c. Oscar does not know at $t_2$ that he thinks $p$.

d. So, Oscar does not know at $t_1$ that he thinks $p$. [from (a)-(c)]

Premise (a) is seen as a platitude about memory (though some deny this; see section 3.3). Premise (b) is regarded as stipulative, and premise (c) is thought to be the lesson of the slow switch experiment. The thinking behind (c) is this: Even if Oscar knows his water thoughts at $t_1$, yet these morph into twin water thoughts at $t_2$, he no longer knows at $t_2$ about any water thoughts. For Oscar no longer has water thoughts at $t_2$.

There are further variations on these slow switch arguments. What’s more, there ends up being multiple interpretations of the initial thought experiment. (See the supplementary document, Variations on Slow Switching, for more details.) Nonetheless, all variations embody the same basic thought: If EXT is true, then in a slow switch Oscar cannot distinguish from the armchair a water content from a twin content. This then forms the basis for why EXT precludes SK. Given the two types of slow switch argument, note that sections 3.3 addresses only the second type, the memory argument. The remaining sections, however, plausibly have some bearing on both types of argument.

3.1 Reply 1: Self-Verifying Judgments

As concerns the (A) to (B) maneuver, Burge protests that this glosses over an especially secure type of armchair self-knowledge. The claim is that some judgments about one’s own thought contents (a.k.a., “second order” judgments) are self-verifying. Roughly, these are judgments where “thinking makes it so”—where judging “I am now thinking that $p$” is enough to make it true that one is now thinking that $p$. Importantly, however, this self-verifying feature is limited to one’s current thinking of $p$ (in contrast to any standing or dispositional thoughts).

To illustrate the self-verifying feature, consider the judgment:

(W) I am thinking, with this very thought, that water is wet.

Burge’s idea is that in executing this judgment, one “runs through” the very thought that the judgment is about. That is, judging (W) necessitates thinking the thought that water is wet—and thinking this thought is precisely what (W) contends. So the very act of judging (W) suffices to render the judgment true. It is in that sense that the judgment is self-fulfilling or self-verifying. Besides Burge, self-verifying judgments were touted around the same time by Davidson (1987; 1988) and Heil (1988; 1992), though others have subsequently embraced them as well (e.g., Falvey and Owens 1994, Gibbons 1996, Bar-On 2004, Parent 2007, 2017, chs. 3 and 7).

Neverthelese, the suggestion that certain judgments are infallible can seem exceedingly strong. But for one, infallibility does not imply that the judgment is indubitable. A judgment could de facto be perfectly reliable, yet the subject might still have some unanswered questions about it. Second, it is crucial that ‘thinking’ is construed minimally, in that to “think” that $p$ is not necessarily to believe that $p$. “Thinking” that $p$ just means having some mental state or other with the content $p$—it does not imply any particular attitude toward that content.

In light of things like Freudian repression, it can be easy to dismiss Burge’s view. But it is important that it pertains only to a limited class of rather atypical judgments. For instance, it applies only when the second-order judgment contains a “self-referential” mechanism, expressed in (W) by the phrase ‘with this very thought’. Moreover, the infallibility concerns occurrent judgments that attribute the co-occurrent first-order thought. And the judgment that (W) can be true even if the subject’s mind also harbors, consciously or unconsciously, a variety of other thoughts.

Although self-verifying judgments are atypical, Burge still thinks they are sufficient to block the inference from (A) to (B). For apparently, they provide a kind of Cartesian access to one’s own thought-contents—meaning that (B) is false even if (A) is true. Granted, self-verifying judgments are a rarefied kind of judgment, so the view is not satisfying as a general view of self-knowledge (Boghossian 1989, pp. 19–22). Still, the main goal was to block the inference from (A) to (B), and self-verifying judgments seem to do that.

Quite a detailed literature has ensued concerning self-verifying judgments; the interested reader is directed to the supplemental documents, The Epistemic Status of Self-Verifying Judgments, and A Problem with Critical Reasoning.

3.2 Reply 2: Relevant Alternatives

A different reply to slow switch arguments may have seemed obvious from the start. It is that the slow switch thought experiment is a bizarre piece of science fiction, and is simply irrelevant to whether we know our own thoughts from the armchair. After all, unlike Oscar, we are not the victims of covert switches. So how could the experiment bear on our capacity for armchair self-knowledge?

The point is often framed in terms of a “relevant alternatives” epistemology (cf. Goldman 1976; 1986). The suggestion...
is that in order to know that $p$, one does not need to rule out every possible scenario where $\neg p$. Rather, it suffices just to rule out the "relevant alternatives" where $\neg p$. And ordinarily, deviant skeptical possibilities are not among the relevant alternatives that a knower needs to exclude. Thus, while driving along the countryside, one can ordinarly know if one sees a barn—even without excluding the possibility of a mere barn façade. For barn façades are ordinarily irrelevant. In the same way, slow switch possibilities ordinarily are irrelevant to knowing what content a thought has (Warfield 1992, Falvey & Owens 1994, Gibbons 1996, Brown 2004, ch. 2). Accordingly, it is normally unnecessary to rule out such a possibility in order to know from the armchair one’s own thought.

But what determines relevance exactly? A popular answer comes from the epistemic contextualists (DeRose 1995; 2009, Lewis 1998; see also the entry on epistemic contextualism). Contextualism in the first instance is a view about knowledge, yet it implies that whether an alternative is relevant is determined by the evidential standards in the context. Consider that in order to know whether it is raining, normally one can just take a look out the window. When engaging a Cartesian skeptic, however, that is insufficient. For the skeptic “raises the bar” on the evidence needed to know by imagining bizarre alternatives that he wants ruled out, e.g., the possibility that you are dreaming. Similarly, some compatibilists regard slow switching as a bizarre skeptical possibility which makes exceedingly high the evidential standards for knowledge (Hohwy 2002; Neta 2003 is also pertinent). So if context makes that possibility relevant, one may lose the ability to self-know from the armchair. Still, if switching possibilities are irrelevant, contextualism says that even an externalist can know her own thoughts from the armchair.

When first championing the memory argument, Boghossian (1989) noted the reply from relevant alternatives, yet maintained that the slow switch experiment still shows a conflict between EXT and SK. For one can imagine cases where slow switching is a relevant alternative. And in those cases, one could not rule out this possibility just from the armchair. Even the relevant alternatives account denies the externalist armchair self-knowledge in such a scenario. Still, why does this reveal a problem with our self-knowledge? If slow switching actually occurred on a regular basis, it would be a relevant alternative that could not be excluded from the armchair, and the externalist would lose SK. But to acknowledge a possible scenario of this sort does not mean that EXT and SK are incompatible (Warfield 1992). It shows merely that if EXT is true, self-knowledge is not necessarily within reach from the armchair.

In reply, some incompatibilists subversively suggest that switches do occur on a regular basis (Ludlow 1995b; Butler 1997). In section 1, we saw that one “Twin Earth” scenario might occur in the actual world: Recall the superficially indiscernible extensions of ‘chicory’ and ‘endive’, which are swapped between British and Standard American English. An American ex-patriot thus could be unwittingly fooled into denoting chicory when speaking or thinking about “endive.” Still, this seems like an atypical case. But to the contrary, the potential for an undetected switch might be as common as the potential for polysemy in the language (Ludlow 1995b). Consider that between different philosophical circles, the term ‘pragmatism’ denotes different yet superficially similar philosophical theses. So thanks to unnoticed switches between subgroups, one might be fooled into saying or believing things about different kinds of pragmatism. Such variation among linguistic subgroups seems entirely commonplace. It is the prevalence of these switches, moreover, that makes them relevant even if no actual switch occurs in a given instance.

In her defense, a compatibilist can observe that these cases are not cases of slow switching, and the usual view is that contents do not change in a “quick switch.” (Though the individuation of some cognitive processes may switch immediately; see Clark & Chalmers 1998). Yet even if slow switches regularly occur, this still does not show that EXT and SK are incompatible (Warfield 1997). It might mean that EXT and SK are not jointly true in our world. Nevertheless, they still might hold in some other possible world; so no incompatibility strictly follows.

Regardless, it may be troubling enough if an externalist cannot actually have armchair self-knowledge (Ludlow 1997). It is standardly agreed, moreover, that the externalist should concede one type of armchair self-knowledge—namely, knowledge that is discriminatory between water contents and twin contents, a.k.a. “comparative knowledge” of content (Falvey & Owens 1994; see also Mendola 2008). But this may not be a huge concession, since one may remain capable of knowing a self-justifying judgment that (W) from the armchair, absent any comparisons with twin contents. Besides, there is independent reason to reject such comparative knowledge anyway (see Owens 1990; 1992).

A different issue is that if a compatibilist concedes SK when switches are relevant, she may be conceding too much (Goldberg 2006a). For there is a persistent Cartesian intuition that armchair self-knowledge should be unaffected by skeptical hypotheses about the external world. Following Meditation Two, even if an evil demon is deceiving me, it still seems I can know what I am currently thinking. Yet the slow switch experiment suggests that this Cartesian intuition is mistaken. For if a slow switch hypothesis is relevant, then even a relevant alternatives epistemology demands that we discriminate water thoughts from twin water thoughts. Yet one is unable to do so from the armchair. So apparently, some deviant hypotheses about the external world can undermine armchair knowledge of one’s own current thoughts, contra the Cartesian intuition. Something important about armchair self-knowledge may be lost.

But a relevant alternatives compatibilist might still rescue the Cartesian intuition. First, note that in many slow switch arguments, the issue is not whether Oscar knows that he thinks that $p$ from the armchair; rather, it is whether Oscar knows what he thinks. And unlike ascriptions of “knowledge that,” ascriptions of “knowledge what” are appropriate only relative to one’s goals or purposes in a context (Boër & Lycan 1986, Braun 2006, DeRose 2009, ch. 2, appendix).[22] Hence, whether it is apt to say “Oscar knows what he thinks” partly depends on contextually salient goals or purposes. Granted, if Oscar’s goal is to discriminate between a content and a twin content, then he cannot “know what” he thinks from the armchair. However, it may be that for other anti-skeptical purposes, Oscar can indeed “know what” (Parent 2015b, 2017, chs. 5 and 10). For instance, suppose the skeptic challenges Oscar on whether he knows the truth of
Before closing this section, it is worth noting that compatibilists sometimes prefer not to insist on the irrelevancy of slow switches, but rather to contest whether their relevance undercuts armchair self-knowledge (Falvey & Owens 1994, Gibbons 1996, Butler 1997, McLaughlin & Tye 1998a, Vahid 2003, Brown 2004, Morvarid 2015). A variety of inferential principles might lead to the incompatibilist’s conclusion from the relevance of slow switching. As a more noteworthy example, the principle here might be some kind of “truth tracking” condition; cf. Nozick 1981 (Falvey & Owens 1994). On this approach, the slow switch argument assumes that (W) does not express knowledge, since the subject S fails to track the truth in the following sense:

(TC)

S’s justification for the belief that p is such that, if some relevant alternative were true instead, S would still believe that p.

According to (TC), S fails to know since her justification would not divert her judgment, in a relevant scenario where the judgment is false. However, if this is the principle operative in the arguments, it seems they can be resisted (Falvey & Owens 1994). Plausibly, we do not satisfy (TC) when it comes to self-verifying judgments (see section 3.1). For if we consider a counterfactual where I have twin water thoughts in lieu of water thoughts, I cannot believe I have water thoughts. A self-attribution expressed as “I am thinking that water is wet” would inevitably attribute a twin water thought instead. My self-attribution “tracks the truth” perfectly in the counterfactual circumstances.

### 3.3 Reply 3: Externalism about Memory

Let us now focus our attention on the second version of the slow switch argument, a.k.a. the memory argument, which is repeated here for the reader’s convenience. Where t2 is a time after a slow switch, and t1 is a time before, then:

a. If Oscar forgets nothing, then what Oscar knows at t1, Oscar knows at t2.

b. Oscar forgot nothing.

c. Oscar does not know at t2 that he thinks p.

d. So, Oscar does not know at t1 that he thinks p. [from (a)-(c)]

(Slow-switching is a relevant alternative here, since it is assumed that Oscar actually is slow-switched.) Now as one might expect, each of the premises have been challenged by various compatibilists. Regarding (a), some reject the suggestion that even if p was known at t1, remembering p does not imply knowing it at t2 (Bernecker 2010, ch. 3). But even if remembering implies knowing, there are other ways Oscar can lose knowledge besides forgetting (Brueckner 1997a; Burge 1998, n. 18). For instance, suppose Oscar knows he is driving through a countryside with at least one barn, but some defeating condition is introduced (e.g., he enters Barn Façade County). Then, he may no longer know at t2 that there is a barn in his environs, even if he truly believes it. That's because, although nothing has been forgotten, a defeating condition exists at t2 that was absent at t1.

An incompatibilist might try to stipulate that Oscar meets all the conditions needed to know p on the basis of memory. However, the compatibilist can reply that this is inconsistent with Oscar being slow switched (Brueckner 1997a). For a slow switch is a defeating condition, akin to entering Barn Façade County. In that case, the incompatibilist’s argument does not even get started.

Alternatively, a compatibilist can attack the portrayal of memory suggested by the argument. Here, one might claim that Oscar’s memory content is partly determined not by the Earthly environment in which the memory was formed. Rather, memory content is conditioned by the environment in which recall occurs (Ludlow 1995a; 1996; 1999; Gibbons 1996; Tye 1998; Bernecker 1998). This means that, even if knowledge of a water thought was initially stored in memory, what is “recalled” on Twin Earth is a twin water thought. If so, then a slow switch causes a kind of memory failure, since the knowledge initially stored in memory is not what is recalled. (N.B., memories of second-order thoughts are the most dialectically relevant here, though the literature often uses first-order memories as examples, cf. Kraay 2002.)

This view of memory as “shifty” between different environments may be a consequence of EXT. Apparently, we can just as easily construct a Putnam-style thought experiment about Oscar’s and Twin Oscar’s memories (Ludlow 1995a; 1996; 1999). Moreover, in the slow switch experiment, the presumption is that Oscar’s water thoughts generally change over to twin water thoughts—including whatever (first- or second-order) thoughts are stored in memory.

It may be quite unintuitive to see memory in this way, and we shall attend to objections shortly. But it is worth noting that the view allows an externalist to resist the memory argument. A compatibilist might first grant that Oscar does not forget strictly speaking—after all, we can assume no neurological impairment has occurred in him. Yet she can add that Oscar loses some memory content regardless, thanks to the slow switch. Even though Oscar does not truly “forget,” he has still lost knowledge by a different route; premise (a) of the memory argument is thus false. (The “shifty” view of memory has waned in recent years; Bernecker 2004; 2010 relinquishes it in favor of something like Burge’s 1998 view, described below. Ludlow 2004 trades it for a view where a memory is a temporally extended object, one which has “stages” that hosts different contents at different times. Yet see Burge 1998 n. 2 for remarks against such a view.)
When spelled out, this “memory free” slow switch argument assumes the following “principle of knowing identification:”

absent any premises about memory.

armchair, this might express a water thought or a twin water thought. Such “underdetermination” in the interpretation is relevant alternatives, and he asserts ‘I am thinking that water is scrumptious’. Then, for all Oscar can tell from the that Oscar’s self-attributions can be open to more than one interpretation: Suppose Oscar recognizes slow switches as memory seem inessential to the so-called “memory argument” (Goldberg 1997; 1999a; 2000a). After all, the key idea is As a final thought, memory may ultimately be a red herring in defending incompatibilism. For suppositions about found in the supplementary document, acquainted with twin water (Boghossian 1989; 1992a, Gibbons 1996, Burge 1998). More discussion of the issue is repertoire. The default view is that the slow switch causes Oscar’s water concept to be of the slow switch thought experiment. For it complicates what effect the slow switch has on Oscar’s conceptual Whether one accepts the “shifty” view of memory or not, the debate is important to opening up different interpretations see Schroeter 2007 for a different issue in cases where Burge rejects (c).) Yet the objection may beg the question of how memory should function (Ludlow 1996). For the issue is precisely whether memory is supposed to preserve past contents across different environments. The counter-suggestion is that memory instead offers information about past events, in terms pertinent to the present environment (Ludlow 1996; 1999, Bernecker 1998). Ludlow (1999) provides an illustration: “If at t₁ I believed it was possible to drown in water, memory will deliver a t₂ belief that it is possible to drown in twater—and good thing, too! Twater is no less …dangerous than water” (p. 167).

As a second worry, if Oscar’s memory contents shift, has he really forgotten nothing? A case can be made that the memory shifts indeed constitute forgetting (Brueckner 1997a). For even though memory offers up a similar substitute content, it seems clear that Oscar fails to recall something about water. Yet a failed attempt at recall just means that Oscar “forgot.” If so, then it is premise (b) rather than (a) which is the culprit in the memory argument. (Even so, there may remain a different sense in which Oscar does not forget: Brueckner 1997a, n. 21; see also Kraay 2002).

Nevertheless, it can be debated whether EXT entails the “shifty” theory of memory. When Burge first presented his Twin Earth experiment, he conceded that it does not show that all mental contents satisfy EXT (though it suggests that that is typically so) (1979, p. 92, passim). The contents of memory thus might be an exception. And indeed, Burge (1993; 1998) endorses the view that memory should “preserve” past thought contents. So in a slow switch case, Burge holds that Oscar’s properly functioning memory preserves an Earthian thought content, so that the very same content is available later for recall.

Accordingly, Burge rejects premise (c) of the memory argument: If Oscar truly forgot nothing, then preservative memory means he knows at t₂ what he knew at t₁. Per Falvey & Owens (1994), Oscar’s memory may not provide comparative knowledge of content, i.e., knowledge that enables armchair discriminations between water and twin water contents. But memory still preserves knowledge of his thought at t₁—meaning that (c) is false. (The incompatibilist might rejoin that (c) is stipulative in the thought experiment; but in that case, Burge sides with Brueckner 1997a in rejecting (b). Yet see Schroeter 2007 for a different issue in cases where Burge rejects (c).)

Whether one accepts the “shifty” view of memory or not, the debate is important to opening up different interpretations of the slow switch thought experiment. For it complicates what effect the slow switch has on Oscar’s conceptual repertoire. The default view is that the slow switch causes Oscar’s water concept to be wholly replaced by the twin water concept. But it may be unclear why a slow switch destroys Oscar’s water concept, even as he becomes acquainted with twin water (Boghossian 1989; 1992a, Gibbons 1996, Burge 1998). More discussion of the issue is found in the supplementary document, Variations on Slow-Switching.

As a final thought, memory may ultimately be a red herring in defending incompatibilism. For suppositions about memory seem inessential to the so-called “memory argument” (Goldberg 1997; 1999a; 2000a). After all, the key idea is that Oscar’s self-attributions can be open to more than one interpretation: Suppose Oscar recognizes slow switches as relevant alternatives, and he asserts ‘I am thinking that water is scrumptious’. Then, for all Oscar can tell from the armchair, this might express a water thought or a twin water thought. Such “underdetermination” in the interpretation is what Oscar cannot resolve without empirical investigation. So already there is a problem with armchair self-knowledge, absent any premises about memory.

When spelled out, this “memory free” slow switch argument assumes the following “principle of knowing identification:”

(PKI)
If a subject S recognizes that there is more than one relevant interpretation of her assertions/beliefs—and S has nothing to decide between them—then S lacks knowledge of the content of her assertion/belief.
If Oscar satisfies the antecedent of (PKI), it may seem that he is missing only *comparative* knowledge of content—i.e., knowledge that allows armchair discrimination between a content and a twin content (Brueckner 1999; cf. Falvey & Owens 1994). And we saw that many compatibilists are already willing to surrender that type of knowledge. That’s because Oscar’s assertion “I am thinking that water is scrumptious” might still count as one kind of armchair self-knowledge (cf. section 3.1), even if it is not of the comparative type.

However, (PKI) is applied to a specific kind of case, namely, where (i) there is more than one relevant interpretation, and (ii) the subject recognizes this (Goldberg 1999a). There, even a relevant alternatives epistemology requires Oscar to rule out the Twin Earth interpretation. Toward making this clearer, suppose that Oscar is informed of the switches at some later time t₂. In that sort of case, (i) and (ii) will be satisfied. Thus if he asserts “I am thinking water is scrumptious,” he will be aware of relevant alternative interpretations that he cannot rule out from the armchair.

Notably, since the “memory free” argument is restricted to a special sort of case, it does not show that EXT and SK are entirely *incompatible* (Goldberg 1999a, p. 217; Goldberg 2000a). It still suggests, however, that EXT precludes SK when certain skeptical hypotheses are entertained—contrary to, e.g., the view in Meditation Two.

### 3.4 Reply 4: The Argument is Self-Undermining

As should be clear, slow switch arguments can be seen as skeptical arguments, where the skeptic raises the possibility that we are in Oscar’s predicament. The externalist is thus asked to rule out deviant thought contents from the armchair. Moreover, since the skeptic thinks she cannot do so, she claims that the externalist lacks armchair self-knowledge. But one sort of reply is that such skepticism is *self-undermining* (Ebbs 1996; 2001; 2005; Brueckner & Ebbs, 2012, ch. 12).

The contention is that, assuming EXT, the skeptic cannot both deny SK and be justified in accepting the premises of his own argument. If so, skeptical slow switch arguments end up being incoherent.

Consider that, since the skeptic wants to make a general point about knowledge, he should see slow switch possibilities as relevant to his own situation. The skeptic should not make an exception of himself, re: his skeptical thesis. Hence, when the skeptic levies the slow switch argument, his premises force him to say of himself:

1. If I am on Twin Earth, then my saying ‘water is scrumptious’ expresses the thought that twin water is scrumptious.
2. If I am on Earth, then my saying ‘water is scrumptious’ expresses the thought that water is scrumptious.

His premises also require him to say that:

3. I cannot know from the armchair which of the two thoughts is expressed by ‘water is scrumptious’ (given that switching possibilities are relevant).

The problem, however, is that (III) applies to ‘water is scrumptious’ whenever the skeptic says it. Therefore, it applies to the utterance as it occurs in the consequent of (II). So if (III) is assumed, it follows that the skeptic does not know from the armchair what thought is expressed by the consequent of (II), hence, by (II) itself. Yet if it is unknown what thought is expressed by (II), one can hardly be justified in accepting the thought. Hence, the skeptic cannot simultaneously hold (III) and be justified in accepting (II). (Ebbs also makes the analogous point about the justification for (I), followed by similar objections and replies. But for brevity’s sake, we shall just consider (II).)

The skeptic might offer two different replies to the self-undermining charge. The first is that, although the exact content of (II) may be unknown, the skeptic can still be justified in accepting (II) (Brueckner 1997b). For it seems that (II) can be shown true, no matter whether the skeptic speaks English or Twin English. After all, regardless of what ‘water’ denotes, it is true to say “‘water is scrumptious’ expresses the thought that water is scrumptious.” Moreover, since the skeptic can recognize this from the armchair, it seems he can be justified in accepting (II), even while assuming (III).

But to the contrary, it is possible to imagine “weird worlds” where (II) ends up false (Ebbs 2001). Consider a world much like this one, where H₂O fills the lakes and rivers, etc., and where ‘water’ refers to H₂O. The difference is that ‘Earth’ in this weird world does not denote the planet that the skeptic currently occupies. Rather, it denotes Twin Earth. In that case, when the skeptic states (II), the proposition expressed is that if he is on Twin Earth, then asserting ‘water is scrumptious’ expresses the thought that water is scrumptious (i.e. H₂O) is scrumptious. (Remember that ‘water’ in this scenario still denotes H₂O.) Yet if that is what (II) expresses, then the skeptic says something false. And so, since the skeptic apparently cannot exclude a weird world from the armchair, he cannot prove to himself that (II) is true. So once more, the skeptic cannot justifiably adopt (II) while assuming (III).

The skeptic has a second line of reply to the self-undermining charge, which takes the form of a dilemma. Observe that ‘water is scrumptious’ either expresses the thought that the skeptic thinks it expresses, or not. If it does, then (II) is true—and so, in conjunction with (I) and (III), the slow switch argument remains standing (Brueckner 2003). On the other hand, suppose that ‘water is scrumptious’ does not express the content that the skeptic thinks it does. That just means he has a false belief about the content of his first order thought. So he lacks knowledge of its content (since knowledge requires truth). And thus SK is false in this case too. Consequently, on either horn of the dilemma, SK is threatened (Brueckner 2003).

Nevertheless, the skeptic could not justifiably claim that (I)-(III) figure into a sound argument (Ebbs 2005). For at this stage, the skeptic has conceded that (II) might not be true. Even so, it seems that he can justifiably identify a different
The debate on externalism and self-knowledge should continue to see new developments. After all, as one writer purposes.

Many of the issues here turn on what kind of armchair knowledge one's own thoughts would too easily refute eliminativism about the mental (Dretske 2003a; 2003b; 2004, 2005). Briefly, if a quale is partly individuated by the subject's environment, then it seems one could not know that one had a quale of fire engine red unless one antecedently knew the actual color of fire engines. But since the latter is not knowable from the armchair, then it seems neither is the former (Levine 2003, Ellis 2007).

A variety of other anti-recognitionalist views exist. E.g., there is a neo-expressivist view which is gaining prominence (Bar-On & Long 2001, Finkelstein 2003, Bar-On 2004; 2006; 2011; 2012a, b; 2103, Parent 2017, ch. 8). There are also views which utilize agency-theoretic ideas to explain self-knowledge (Moran 2001, Heal 2002, Bilgrami 2006, O'Brien 2007, Coliva 2009). The interested reader is encouraged to visit the entry on self-knowledge for more.

There is a surprising number of further issues regarding EXT and SK. Some of these consist in lesser-known incompatibilism arguments. For instance, there is a kind of “inverted” memory argument suggesting that, if Burge’s self-verifying judgments count as bona fide self-knowledge, then one’s memory would be infallible in ways that it is obviously not (Goldberg 2000b). In addition, it is sometimes argued that EXT and SK preclude the standard analysis of epistemic possibility. (Ebbs 2003; 2011. But see McLaughlin 2004.)

Questions about compatibility have been extended to more than just knowledge of content. There is also a case to be made that EXT precludes armchair knowledge of one’s attitudes toward those contents (Bernecker 1996, Gibbons 2001). Thus, even if it is known that you are currently thinking that water is scrumptious, EXT may preclude armchair knowledge that you believe this. For what counts as “belief” might vary according to one’s linguistic community.

In addition, new issues have arisen in light of the Language of Thought Hypothesis falsifies SK, assuming EXT (Boghossian 1989, p. 6, Bonjour 1991). Further, some writers have worried that if one can know what one thinks, then one can know that one thinks. If so, then the ability to armchair knowledge one’s own thoughts would too easily refute eliminativism about the mental (Dretske 2003a; 2003b; 2004, Bernecker 1998; 2004b, Jacob 2004, Parent 2017, ch. 9; some papers in Smithies & Stoljar 2012 are also relevant).

In addition, there is a burgeoning literature on whether EXT is compatible with epistemic internalism, a view where knowledge or justification supervenes on what’s in the head (Pritchard & Kallestrup 2004, Brown 2007, Gerken 2008; also cf. Chase 2001, Brueckner 2002. Goldberg 2007a is germane too). Many of the issues here turn on what kind of self-knowledge is allowed by EXT, and in particular, whether such knowledge is discriminating enough for justificatory purposes.

The debate on externalism and self-knowledge should continue to see new developments. After all, as one writer observes:

in the last two decades several developments in the philosophy of mind, the philosophy of language, and epistemology make it worthwhile to reconsider semantic externalism’s implications for self-knowledge...[T]hese developments include various versions of 2D semantics, the emerging popularity of so-called transparency views of self-knowledge...various doctrines pertaining to the semantics of ‘knows’ (such as contextualism, contrastivism, and...
pragmatic encroachment), a renewed focus on anti-luminosity arguments (deriving from the work of Williamson (2000)), and developments in the epistemology of testimony as well as in the epistemology of understanding. (Goldberg 2015b, pp. 7-8)

Moreover, the externalism/self-knowledge debate seems not to be just a short-lived fancy; instead, it is arguably a significant plot-point in the history of Western philosophy. This is because:

[T]he issue reflects an important clash between philosophical paradigms before and after the linguistic (or conceptual) turn. On one hand, it is characteristic of Cartesianism that we can know contents introspectively, independently of knowing the external world. On the other hand, Frege taught us that content determines which external objects we refer to (or in Carnap's terminology, that intension determines extension). And as Putnam (1975, p. 218ff.) argues, Frege's idea seems to result in externalism: If content determines reference, then a difference in reference between Oscar and Twin Oscar shows a difference in content. That seems to hold, even if their narrow psychological states are the same. So if Frege's view leads to externalism, yet this precludes Cartesian knowledge of content, then apparently one of these paradigm-defining ideas has to go. Either Cartesians were wrong to say we can know contents just by introspection, or Fregeans were misguided to think that content is what determines reference. (Parent 2015b, p. 1346)

The debate on whether EXT is compatible with SK thus deserves our continued, serious attention.

Bibliography

Works Cited


——, 2004, “We Can’t Know A Priori that H2O Exists, but Can We Know that Water Does?,” Analysis, 64(1): 44–47.


Collections


Academic Tools

How to cite this entry.
Preview the PDF version of this entry at the Friends of the SEP Society.
Look up this entry topic at the Indiana Philosophy Ontology Project (InPhO).
Enhanced bibliography for this entry at PhilPapers, with links to its database.

Other Internet Resources

• Externalism and Self-Knowledge, category at PhilPapers.org.

Related Entries

introspection | mental content: externalism about | self-knowledge | skepticism: and content externalism

Copyright © 2017 by T. Parent <parentt@vt.edu>

Open access to the SEP is made possible by a world-wide funding initiative. The Encyclopedia Now Needs Your Support Please Read How You Can Help Keep the Encyclopedia Free
Self-knowledge and externalism bill brewer. (E1) I believe that p. (E2) If x believes that p, then x’s environment is thus-and-so. ∴ (E3) My environment is thus-and-so. (w1) I believe that water is wet. (w2) If x believes that water is wet, then x’s environment contains (or did contain) water. ∴ (w3) My environment contains (or did contain) water.

1. Adequate acct. of self-knowledge → (E1) NEK (non-empirically knowable).
2. Content externalism → (E2) NEK.
3. So content externalism plus an adequate acct. of self-knowledge → (E3) NEK.
4. Non-empirical knowledge of empirical facts like this is in A person’s authoritative self-knowledge about the contents of his or her own beliefs is thought to cause problems for content externalism, for it appears to yield arguments constituting a wholly non-empirical source of empirical knowledge: knowledge that certain particular objects or kinds exist in the environment. Thus, a person’s knowledge that he or she has an externalist belief, since this depends upon actually having that belief, and therefore upon possessing the relevant externist concept, presupposes knowledge about the object/kind in question, which provides a direct and perfectly empirical source of knowledge that this object/kind exists. Hence, the putatively problematic arguments do not constitute a non-empirical source of such empirical knowledge. (shrink).