Externalism about Content and McKinsey-style Reasoning

James Pryor
Harvard University
<jpryor@fas.harvard.edu>

Draft 1 — 10/1/01

WARNING: This is a first draft. I want to stress that the footnotes and citations are just off the top of my head at this point, and may very well misrepresent some other philosophers' views.

I

It's widely accepted nowadays that **the contents of some of our thoughts are externalist**: we're only able to have thoughts with those contents because we inhabit environments of certain sorts.[1] For example, to have thoughts about water it may be

---

[1] Different philosophers mean different things when they talk about a thought's **content**. I take a thought's content to be those of its representational properties which are essential to its being *that thought*. These may or may not suffice to determine the thought's truth-conditions. That is a substantive philosophical question. I think that content does determine truth-conditions; but I don't want to assume this *in the very definition* of content. On some views, like Lewis's [cite??], the thought I express by saying "My pants are on fire" and the thought you express by saying "My pants are on fire" are the same thought. Hence, those are views that allow thoughts with different truth-conditions to have the same content, in my sense. Similarly, on some views, the thought I express by saying "Water puts out fires" and the thought my twin on Twin Earth expresses by saying "Water puts out fires" are *the same thought*, thoughts with a single content; it's just that they have different truth-conditions in the different environments. I don't count such views as **externalist** views. As I understand externalism, an externalist has to say that, though my belief and my twin's belief may have interesting properties in common, they are nonetheless *different* beliefs, with *different* contents.

One sometimes hears it said that "one and the same token thought can be 'typed' in different ways." It certainly is true that a single thought can *exemplify* many different types. The thought I express by saying "My pants are on fire" is: a thought about fire, a first-person thought, a thought about Jim Pryor, a false thought, and so on. But not all of these types are relevant to the thought's identity. It is not essential to

necessary to inhabit an environment containing samples of water, which you or other members of your community have causally interacted with.[2] When you think to yourself *Water puts out fires*, the content of your thought is not available to people who live in Twin Earthian environments that contain XYZ instead of water.

On some views, the contents of our demonstrative thoughts, and thoughts we express using proper names, are externalist in the same way. If I see a man walking down the street and I think to myself *He is dangerous*, the content of this thought is not available to someone who lives in an environment that that man never inhabited or left traces in.

If we accept those kinds of externalism, then we have to confront a puzzle that arises about **our ability to know what we're thinking.** This puzzle was first[??] articulated by Michael McKinsey in 1991. It goes as follows. It would seem that you can

---

my thought that it be false, for example. When philosophers say that my thought "can be 'typed' in different ways," they're usually making a claim about the thought's *identity conditions*. According to one way of "typing" my thought that my pants are on fire, it would count as "the same thought" as your thought that your pants are on fire; according to another, they would count as "different thoughts." And both ways of "typing" the thoughts would be correct. (Or perhaps each would be correct in certain settings.) In my view, this ecumenical attitude towards identity conditions has little to recommend it. Certainly the mere fact that it's *controversial* what the identity conditions of our thoughts are does not by itself suffice to make the position plausible. In any event, as I understand externalism, it *does* take our thoughts to have definite identity conditions, conditions which cannot be satisfied unless one's environment is of a certain sort.

[2] It is not easy to specify what the relevant environmental features are, that externalism says are necessary for you to think about water. According to Burge [[Other Bodies]], it would be possible to have thoughts with the content *Water puts out fires*, even if no samples of water ever existed—but only if you or someone else in your community had some beliefs about what water's underlying nature is. [[Cite Mendeleev / Ekaboron story]] If this is right, then assuming you do not yourself have beliefs about water's underlying nature, the way your environment would have to be, for you to think about water, is: it would have to contain *either* samples of water *or* other people who had beliefs of certain sorts. See Brown [[??]] for discussion. These complications will not affect the main points in my discussion, so to keep things simple, I will work with the condition I give in the text.

usually tell the contents of your thoughts, just by introspection. For example, you can know just by introspection that you're thinking that water puts out fires:

McK-1 You're thinking a thought with the content *Water puts out fires*.

And it seems that our knowledge that externalism is true, and hence that you can't have thoughts about water unless you inhabit a certain sort of environment, is a deliverance of *a priori* philosophical reflection on Twin Earth thought-experiments. So it seems that you can know *a priori*:

McK-2 You couldn't have that thought unless your environment is a certain way, e.g., some same samples of water must have sometime existed and you or other members of your community must have causally interacted with them.

But if that's right, then it seems like you should be able to put those two pieces of knowledge together, and conclude:

McK-3 Your environment is the relevant way, e.g., some samples of water have sometime existed…

So it looks like introspection and philosophical reflection would be enough to enable you to know that you inhabit the relevant sort of environment. And that is a puzzling result. *Prima facie*, it seems quite counter-intuitive that you should know those sorts of facts about your environment just on the basis of these kinds of reflection.

Let's be clear about what the source of the puzzle is. On the one hand, we have *a certain argument*: the argument from McK-1 and McK-2 to McK-3. There's nothing wrong with that argument, *per se*. In fact, it's a sound argument. Its conclusion is true and we know it to be true. And *in itself*, there doesn't seem to be anything objectionable about believing that conclusion on the basis of those premises. This only becomes puzzling because of *the kind* of justification you happen to have for those premises. You seem to be able to know McK-1 by introspection, and you seem to be able to know McK-2 *a priori*, so it looks like you're in a position to know the conclusion McK-3 just on the basis of introspection and *a priori* reflection. *That* is what seems puzzling.[3]

---

[3] Some philosophers doubt that philosophy is a purely *a priori* enterprise. This might engender

The puzzle is sometimes formulated as a problem about whether externalism would enable us to have "*a priori* knowledge" of what our environment is like.[4] Authors who speak this way are counting introspection as a kind of "*a priori* knowledge." I think that's quite misleading, and encourages certain confusions. We will be disentangling some of those confusions later in this paper. For clarity, I will *not* count introspection as a kind of *a priori* knowledge. I will use the more general expression "**knowledge by reflection**" to talk about things you know either by introspection or by *a priori* reasoning, or by a combination of the two. So in my terms, the puzzle says that you can know certain facts about your environment *just on the basis of reflection*, and that's what seems incredible.

The puzzle is sometimes formulated as a problem about whether we have certain kinds of *epistemic authority* about our environment, authority which we would have thought we had only about our own mental states. But it is controversial and unclear what kinds of authority we have about our own mental states. And I don't think the current puzzle really depends on any specific assumptions about that. It'd already be very puzzling if we could have knowledge about our environment *from the reflective sources we're discussing*—even if no special epistemic authority came with that knowledge.

If that's right, then it wouldn't be a satisfying resolution of the puzzle merely to say that introspection is *perceptually defeasible*, e.g. by evidence that my environment has never contained any samples of water. Nor do I think the puzzle requires thinking of introspection as being indefeasible. Let's suppose for the sake of argument that my introspective awareness that I'm thinking of water *is* defeasible. How would that help? It might be thought to take the sting out of saying that we can know about our environment

---

doubts about whether our justification for believing McK-2 is purely *a priori*. But even if we accepted that philosophy *isn't a priori*, it would still be puzzling if you could come to know that your environment contains water, just by a combination of introspection *and philosophy*. Even if doing philosophy requires you to have *certain* kinds of empirical evidence, presumably it doesn't require you to have perceived water, or anything like that.

[4] [[Authors who speak this way]]

just on the basis of introspection and *a priori* reasoning. For now that knowledge will be defeasible in the ways we think knowledge about our environment ought to be. But I don't think it really does resolve the puzzle; at least, not by itself. For as I said, it already seems puzzling to say that we can know about our environment *from reflective sources* like introspection and *a priori* reasoning. That is *already puzzling*, without the need for any further assumptions about defeasibility.

There are a number of ways one can respond to this puzzle.

One view *endorses* the conclusion that you can know things like McK-3 by reflection alone. That is one way of construing **Putnam's argument** in Ch. 1 of *Reason, Truth, and History*. Putnam argues that people who have always been brains in vats can't refer to or think about vats. So, in order for us to have thoughts about vats, our environment has to be a certain way: it has to be such that we haven't always been brains in vats. As it happens, we can tell by introspection that *we do* have thoughts about vats. (In fact, such thoughts are *necessary*, to be entertaining the skeptical hypotheses we are entertaining.) So it follows that we haven't always been brains in vats.[5]

Other philosophers regard it as unacceptable to say that we could have knowledge of things like McK-3 by reflection alone. Intuitively, those seem to be facts that can only be known by empirical investigation. So the puzzling McKinsey-style reasoning has to be blocked in some way. One popular way to block it is to be an **incompatibilist** about externalism and our ability to know the contents of our thoughts by reflection alone. The incompatibilist says that it can't be true *both* that a given thought has an externalist

---

[5] Those who take this stance towards McKinsey puzzle: Warfield?? Sawyer??

This is only one interpretation of Putnam's argument; there are also other interpretations… [[Some interpretations require extra premises about what brains in a vat *do* refer to and think about, when they use the word "vat"; or what *I* would refer to with "vat" if I were a vat.]]

Some versions of Putnam's argument employ, not introspective knowledge of what one is thinking, but rather disquotational knowledge about one's own language. This raises special difficulties that we cannot pursue here.

content *and* that we are able to know our thought has that content, just on the basis of introspection.[6]

Some philosophers take this incompatibility to discredit externalism. Others take it to discredit our capacity for reflective self-knowledge. They say we can have direct, introspective knowledge only of *narrow* features of our thoughts, features that supervene on what we're like intrinsically.

All of these responses strike me as over-reactions. The most sober response would be a story that steered between them. It would *allow* us to have introspective knowledge of the contents of our thoughts, even if those contents are externalist; but it would *deny* that we can know what our external environment is like by introspection and *a priori* reasoning alone. To steer this middle path, we need to find some way, other than the incompatibilist's way, to block the McKinsey-style reasoning.

II

The McKinsey-style reasoning is a form of *modus ponens*. So one way to block that reasoning would be to articulate and defend a limiting principle on when *modus ponens* reasoning is legitimate. Several of the views we'll be considering try to do just this.

One way to limit *modus ponens* reasoning is to deny Closure: that is, to say that we can know McK-1 and McK-2, and know that these entail McK-3, but deny that this entails we're in a position to know McK-3. Some accounts of knowledge do deny Closure in this way. For example, perhaps you're in a position to rule out all the epistemic possibilities that are relevant alternatives to McK-1 and McK-2, but when we're considering McK-3, *more* epistemic possibilities are relevant, and you're not in a position to rule out those additional possibilities.[7]

---

[6] This is McKinsey's own response to the puzzle [cite]. Other people who say this.

There are also other arguments for Incompatibilism besides McKinsey's. The other arguments have to do with "slow switching" thought experiments. I will have to discuss these elsewhere.

[7] [[Dretske]]. Nozick's account of knowledge gives us a different story about why there can be

However, denying Closure is not so popular these days. Even among epistemologists who employ the framework of "relevant alternatives," many nowadays would rather *keep* Closure. They say that in any given context of knowledge-attribution, there is a *single* range of epistemic possibilities that will count as relevant. If we keep that range *fixed*, then whenever a subject's evidence is good enough to know the premises of a *modus ponens* argument, it will also be good enough to know the argument's conclusion.[8]

In any event, I want to focus our attention on ways to limit *modus ponens* reasoning that do *not* require us to deny Closure.

Crispin Wright and Martin Davies have formulated one such limitation. Consider the following argument. You're at the zoo, and you see a striped horse-like animal in the pen in front of you. The sign on the pen says "Zebra." All of this seems to justify you in believing:

ZEBRA-1 That animal is a zebra.

A little reflection convinces you that if the animal is a zebra, it isn't a mule, and *a fortiori* it isn't a cleverly-disguised mule. Hence, you know:

ZEBRA-2 If that animal is a zebra, it isn't a cleverly-disguised mule.

Putting these together, you conclude:

ZEBRA-3 That animal isn't a cleverly-disguised mule.

Now, Wright and Davies do not want to raise any doubts about Closure. They allow that you can know both ZEBRA-1 and ZEBRA-3 to be true. What they doubt, however, is that the reasoning I just sketched could be *what gives you* justification for believing ZEBRA-3. Rather, they think that you would *already need justification for ZEBRA-3 to be in place*, before you could be justified in believing ZEBRA-1 on the grounds I described. This is because those grounds aren't really enough, by themselves, to justify you in believing ZEBRA-1. It's only insofar as they're supplemented by some antecedent or independent justification for believing ZEBRA-3 that they can support ZEBRA-1. Hence, the ZEBRA-

---

failures of Closure.

[8] See [[Stine 1976]], [[Cohen 1988]], [[DeRose 1995]], and [[Lewis 1996]].

argument can not do anything to *enhance* ZEBRA-3's epistemic credentials for you. Wright and Davies put this by saying that the justification you have for the argument's premises does not "transmit" into a reason to believe the argument's conclusion.

This suggests the following limiting principle on *modus ponens* reasoning: a piece of *modus ponens* reasoning cannot give you justification for believing its conclusion when you need to be *antecedently* justified in believing that conclusion, to be justified in believing the premises in the way you do. Reasoning that violates this constraint exhibits a "failure of transmission," even if it doesn't exhibit any failure of Closure. Wright and Davies think that the McKinsey-style reasoning suffers from just this kind of defect.[9]

Their reasons for thinking this are different. Wright thinks it has to do with the specific kinds of thought-contents that McKinsey's argument is dealing with, and the specific kinds of grounds we have for believing we have thoughts with those contents. Davies thinks it has to do with a more general phenomenon. He thinks that *whenever* you argue for a conclusion by appeal to some premise P, you are presupposing that "there is such a thought as P" for you to appeal to, in a sense of "presupposing" that is sufficient to generate failures of transmission.

I agree with Wright and Davies that transmission-failure is a genuine epistemic phenomenon, and that it deserves close study. I think the ZEBRA-reasoning [??] described above provides a good illustration of the phenomenon. However, I also think that Wright and Davies are inclined to see transmission-failure in too many corners.

For example, they've levied that charge against Moore's famous "proof":

MOORE-1 Here is one hand, and here is another.

MOORE-2 If I have hands, then the external world exists.

MOORE-3 So, the external world exists.

But whether one should count Moore's proof as exhibiting transmission-failure depends on what epistemology of perception one accepts. Wright and Davies think you are justified in believing particular perceptual beliefs like MOORE-1 only insofar as you're

---

[9] Cite: [[Wright 1985, Davies 1998, Wright 2000, Davies 2000, Wright Rutgers forthcoming]]

already justified or entitled to accept general background assumptions like MOORE-3. [[Soften their commitment here??]] I've argued elsewhere for an epistemology of perception that denies this. On my view, for your perceptual experiences to justify you in believing things like MOORE-1, you *don't* need to have any antecedent justification for believing that the external world exists, or that you're not a brain in a vat, or that your senses are reliable, or anything else of that sort. It's enough if you *lack* reasons to believe that you *are* a brain in a vat, that your senses are *un*reliable, and so on. Because your perceptual justification for believing MOORE-1 does not require you to have antecedent justification for believing MOORE-3, I do not think that Moore's proof exhibits transmission-failure—or, for that matter, any other epistemic vice.[10]

That dispute in the epistemology of perception is relevant to our present discussion because I think it reveals that certain *kinds of defeasibility* are not enough, by themselves, to make an argument guilty of transmission-failure. Wright and Davies tend to move too quickly from:

(i) Your justification for believing the premises of such-and-such an argument would be defeated by evidence that not-C.

to:

(ii) Your justification for believing the premises rests on the tacit assumption that C, so you need some antecedent justification or entitlement to believe that assumption.

I have argued elsewhere that (i) does not by itself entail (ii). In my view, the MOORE-argument is one important case where (i) is true but (ii) is not; and it is only when we have the kind of epistemic dependence described in (ii) that a charge of transmission-failure will be appropriate.

So if Wright and Davies are going to make the charge of transmission-failure stick against the McKinsey-style reasoning, they'll have to show that our justification for believing the premises of McKinsey's argument requires us to have antecedent

---

[10] It may have various *dialectical* weaknesses, but I think that's a separate issue. See "Skeptic and Dogmatist" and "Is Moore's Argument…?" for discussion.

justification for believing certain assumptions about our environment. Wright and Davies think we will *ordinarily have* that antecedent justification. It need not be justification that we had to *earn* or *acquire*. Rather, we might for some reason have a *default entitlement* to believe those assumptions. We need not be able to articulate any explicit argument in their support. But Wright and Davies will say we do need some antecedent justification or entitlement for those assumptions to be in place, if we're to be justified in believing the premises of McKinsey's argument.[11]

Now, if our justification for the premises of McKinsey's argument *does* require us to have antecedent justification for certain assumptions about our environment, then I agree that Wright and Davies' charge of transmission-failure might be appropriate. But it will take a good deal of argument to show that our justification for believing McK-1 and McK-2 rests upon environmental assumptions. What's more, I think that if we *do* show that, then we'll have shown that our justification for those premises is not purely reflective, in the way the McKinsey-style reasoning requires it to be. That will *already* be enough to block the McKinsey-style reasoning, transmission-failure or no.[12]

So for the time being, I want to set aside Wright and Davies' response to McKinsey's puzzle, and look at other possible solutions. (In the end, we'll see that I do [??] think there is some transmission-failure going on. But not in every case. And this is just something that *falls out* of my solution to McKinsey's puzzle. It doesn't *play any role* in my account of what's wrong with the McKinsey-style reasoning.)

III

Another way in which *modus ponens* reasoning can go awry comes up in a kind of example that Harman discussed in his book *Thought*.[13] [[Is this the same example he uses??]]

---

[11] In addition to Wright and Davies, check also Brewer?? and Sawyer??

[12] [[See Beebee]]

[13] [cite] Harman says the cases were first discussed by Kripke.

Suppose I perform certain mathematical calculations. These calculations justify me in believing some conclusion:

(1) P

Now, P logically entails any claim of the form "___ P." In particular, it logically entails "If my wife says that not-P, then P." Or in other words:

(2) If my wife says that not-P, then my wife is wrong.

So it seems like I ought to be justified in believing (2) on the basis of my calculations, as well. But (2) tells me that if I acquire a certain kind of evidence against P, that evidence will be misleading. And *that* in turn would seem to justify me in adopting a certain *epistemic policy*: the policy of ignoring my wife's testimony about P. (By parallel reasoning, I'd seem to be justified in ignoring *any* evidence I might acquire that not-P.) But the calculations with which I began *shouldn't* license me in adopting any such policy. After all, I may know my wife to be a much better mathematician than me. Suppose *she does* tell me that not-P. So now I have justification for believing:

(3) My wife says that not-P.

Clearly I would not at this point be justified in combining my belief in (3) and my belief in (2), and concluding that my wife is wrong.

Some have claimed that we have a failure of Closure here. They claim that I know (1), and know that (1) entails the conditional (2); but they deny that I know (2). Or they allow that I know (2), and also that I know its antecedent (3); but they deny that I know that my wife is wrong.[14]

But there is a better diagnosis of the example that does not require us to deny Closure. It goes like this. I am *initially* justified in believing (1) and (2), on the basis of my calculations. But this justification is *defeasible*. When my wife's testimony arrives on the scene, and I acquire evidence for believing:

(3) My wife says that not-P.

Where does this math example come from? Harman? Audi?

[14] E.g., Audi [cite book and reply to Feldman] See Feldman for criticism.

this evidence defeats the justification that my own calculations gave me for believing (1). And since my belief in (2) was solely based on the fact that it is a consequence of (1), (3) will also defeat my justification for believing (2). What is interesting about the example is that I have justification for believing a claim of the form A B, but this justification is defeated by the acquisition of evidence that A. The kind of justification I had for believing A B was not what we might call "*modus ponens* robust." It was not of a sort that would survive the acquisition of evidence that A. I was never in a position where acquiring evidence that A would justify me in *concluding* B.[15]

On this solution, I am unable to engage in *modus ponens* reasoning, but there is no failure of Closure. At no one time do I know *both* of A B, and A, but fail to know B. Once I learn that A is true, I'm *no longer* justified in believing that A B.[16]

Like transmission-failure, I think this is a genuine epistemological phenomenon. And the phenomenon isn't confined to reasoning that involves conditionals. I have plenty of evidence that I am in Cambridge, Mass. right now. That justifies me in believing that *I am in Cambridge right now or I am a monkey's uncle*. But of course if I were to acquire evidence that I'm *not* in Cambridge right now, that would *not* justify me in concluding that I am a monkey's uncle.

Is there any reason to think this phenomenon is involved in the McKinsey-style reasoning, though? In the McKinsey-style reasoning, your evidence for believing the conditional:

McK-2 If you're thinking a thought with the content *Water puts out fires*, then your environment is such that…

---

[15] Put in Bayesian terms, it was rational for me to believe A B, but it was not rational for me to assign B a high *conditional probability*, given A.

This is all a function of *the kinds* of grounds I have in the example for believing (2). Another person could have *different* grounds for believing (2), grounds that *do* survive the acquisition of evidence that (3). Such a person *would be* justified in concluding that my wife is wrong.

[16] This is the solution Harman himself favored; see also Feldman.

comes from certain *a priori* thought-experiments. Why should learning that the antecedent of this conditional is true—that you're thinking a thought with a certain content—why should that defeat or undermine your grounds for believing the conditional? There's no evident reason to think it would. So at first glance, the kind of thing that's going on in these Harman-inspired cases doesn't seem to be going on in the McKinsey-style reasoning.

Let's set these Harman-inspired cases aside for now, then. We will come back to them later.

IV

Let's pause at this point and review how the McKinsey-style reasoning goes. Introspection seems to justify you in believing:

McK-1 You are thinking a thought with the content *Water puts out fires*.

and *a priori* philosophical reflection seems to justify you in believing that you could only be thinking a thought with that content if your environment were a certain way:

McK-2 If you're thinking a thought with the content *Water puts out fires*, then your environment is such that…

Putting these together, you seem to be in a position to conclude, on the basis of reflection alone:

McK-3 Therefore, your environment is such that…

Note that for this reasoning to deliver knowledge by reflection alone of McK-3, you have to be in a position to know the conditional McK-2 by reflection alone. It's not enough if, instead:

(iii) You understand an "externalist content" to be a thought-content that works as described by McK-2.

(iv) You know by reflection alone that the thoughts you express with "water" *aim* to have externalist contents.

(v) But you don't know by reflection alone whether they *succeed* in that aim. It's an open possibility, for all you know by reflection alone, that they fail to have externalist contents.

If (iii)–(v) are true, then you would fail to have any *purely reflective* reason to believe McK-2. Even if, as it happens:

(vi) The thoughts you express with "water" *do* have externalist contents, and so McK-2 *is true*.

Remember, the mere *truth* of the McKinsey-style premises is not enough to generate any puzzle. We only get a puzzle if *you can know by reflection alone* that those premises are true.

I think that something like (iii)–(vi) is our actual scenario. Perhaps we can know by reflection alone that the thoughts we express with "water" "aim" in some sense to have externalist contents, of the sort that McK-2 describes. But I don't think we can know by reflection alone whether they *succeed* in that aim, whether they *actually do* have those kinds of contents. So we are not in a position to know by reflection alone whether McK-2 is true.

Of course, this response to McKinsey's puzzle has its work cut out for it. We have to explain the sense in which our "water"-thoughts [[<—?? INTRODUCED HERE]] "aim" to have externalist contents, but might fail in that aim. We have to explain what kinds of contents our "water"-thoughts would have if they *did* fail to have externalist contents. Would we then be thinking *any* contentful thought when we said to ourselves "Water puts out fires"? If we *wouldn't*, then how can we be in a position to know whether *we do* have contentful thoughts, by introspection alone?

These are the issues I want to explore now.

V

Let's look more closely at the premise McK-2. What is the way our environment has to be, for us to be able to entertain or think the content *water*?

One way of construing McK-2 goes like this:

(a) The content you express with "water" is a content that is only available to subject who inhabit environments containing H_2O, which they or other members of their linguistic community have interacted with.

Now this may be *true* of our content *water*, but it's clearly not knowable *a priori*. We can't know *a priori* that the stuff that content represents is H_2O. So (a) could not be appealed to in a piece of McKinsey-style reasoning. Let's consider, instead:

(b) The content you express with "water" is a content that is only available to subject who inhabit environments containing water, which they or other members of their linguistic community have interacted with.

This is a better candidate for being knowable *a priori*.[17] *Prima facie*, it seems like your grounds for believing it might just consist in *a priori* reflection on certain thought-experiments.

Notice that (b) applies only to the thought-contents that *we* express with "water." It does not also apply to the thoughts that *the Twin Earthers* express with "water." There is no water on Twin Earth, so the Twin Earthers don't have any thoughts that are only available in environments containing water. However, the contents of the thoughts *they* express with "water" share an important feature with the contents *we* express with that word. In each case, the content C is a content that is only available to subjects who inhabit environments containing samples of the stuff C represents. The following principle applies to both of our thoughts:

(c) Whatever content C we express with "water," it is a content that is only available to subjects who inhabit environments containing samples of the stuff C represents, which those subjects or other members of their linguistic community have interacted with.

On Earth, this is one content, *water*, representing the stuff H_2O; on Twin Earth, it is a different content, *twater*, representing the stuff XYZ.[18] Both of these contents are "externalist" because they both satisfy principle (c).

---

[17] Though, as I say in fn. 2, above, I am simplifying somewhat.

[18] Philosophers use the expression "twater" in a variety of ways. Sometimes it's used as our name for

If it is possible for there to be externalist contents of that sort, then it will also be possible for there to be cases where *it only falsely seems* to the subjects that the relevant environmental conditions are satisfied. For instance, consider the subjects on Dry Earth. Dry Earth contains no clear drinkable liquids: no H_2O, no XYZ, nothing. But the subjects there have been hallucinating clear drinkable liquids for some years. In fact, we can stipulate that they've been having courses of experiences that are phenomenally indistinguishable from our own.[19] But the Dry Earthers' environment does not meet the conditions which are necessary for them to be able to entertain or think contents like the ones that we and the Twin Earthers express with "water."

---

the liquid that we're supposing is found on Twin Earth. When it's used in that way, we can be sure that "twater" in our mouths has *the same intension* as "water" has in the Twin Earthers' mouths. But we cannot be sure that the thoughts we express with "twater" *have the same content* as the thoughts the Twin Earthers express with "water." Consider the relation between our words "H_2O" and "water." It is natural to think that there is a certain kind of structure in the content of "H_2O" which is not present in the content of "water." There is also a corresponding *syntactic* structure to the word "H_2O." But we can stipulate that a syntactically simple word, say "hwater," shall have the same semantically complex content as "H_2O" has. Then the thoughts we expressed with "hwater" would not have the same contents, but only the same intensions, as the thoughts we express with "water." If all that's right, then I think we should say that our word "twater" stands to the Twin Earthers' word "water" in the same way that this new word "hwater" stands to our word "water." With "twater" and "hwater," it is part of the very semantics of these terms that they refer to XYZ and H_2O. That is not true of either the Twin Earthers' or our word "water."

A different way to use the word "twater" would be to *stipulate* that it is to have the same content as the Twin Earthers' word "water." But it is not clear that we're in a position to think any thoughts with that content, if we've never been to Twin Earth. We can *name* that content, and talk about it. But it's not clear that we're in a position to *think it*. (To settle that, we'd have to determine *exactly which* environmental conditions need to be satisfied, for the contents the Twin Earthers express using "water" to be available to us. That would mean paying more attention to the issues mentioned in fn. 2.)

I will myself only *talk about* the content *twater* that the Twin Earthers express using "water," and contents containing it. I do not expect the reader to be able to think or entertain those contents.

[19] Here I assume, *contra* Dretske [cite], that a subject's phenomenology is not dependent upon her environment, in the way that contents like *water* are.

The point here is not just that the Dry Earthers aren't able to think *the same* thoughts that we express with "water." That is also true of the Twin Earthers. The Dry Earthers' incapacity is much more radical. When the Dry Earthers say to themselves "Water puts out fires," they aim to be thinking thoughts about a natural kind, just as we and the Twin Earthers do. But there is no natural kind of the right sort for them to be thinking about. The point is not just that there are *no samples* of the natural kind they're trying to think about. There does not seem to even be *a kind*, at all—not even an uninstantiated kind—for their thoughts to be about. When we think the thoughts we express using "water," we rely in part on the world to determine what kind it is we're thinking about. So too do the Dry Earthers. But in the Dry Earthers' case, the world does not cooperate. There's nothing in their environment to make it the case that they're thinking about one natural kind, like H_2O, rather than another, like XYZ.[20]

What, then, shall we say *are* the contents of the thoughts the Dry Earthers express using "water"?

D1 One option would be to say that these thoughts *do* have contents, albeit not the *same kind* of contents that our thoughts and the Twin Earthers' thoughts have. For example, perhaps when a Dry Earther counts something as "being water," he's representing it as having a certain cluster of *observable properties* (being clear, drinkable, liquid, and so on). In fact, nothing in his environment *has* that cluster of properties. But there is a definite cluster of observable properties that his thoughts are representing.

---

[20] The Dry Earthers' situation here is much like the situation that our ancestors were in when they employed concepts like *air* or *phlogiston*. These concepts aimed to be about natural kinds, but they failed, for there were no natural kinds of the right sort for them to be about. (I will explain in a few moments what it means for a thought or concept to "aim" to be about a natural kind.) Compare Kripke's discussion of unicorns, in N&N…

If some of the Dry Earthers had *beliefs about* what the underlying structure was of the liquid their environment seems to contain, that might make a difference. In that case, it could be argued that the thoughts they express with "water" *are* about an uninstantiated natural kind. (Here again the issues mentioned in fn. 2 are relevant.) To keep things simple, let's suppose that the Dry Earthers do not have any such beliefs.

D2 Alternatively, we might say that the Dry Earthers' thoughts would represent, not an *un*instantiated cluster of *objective* properties, but rather a cluster of properties that *really are had by his experiences*. This is the kind of story that many error-theorists about color favor. They don't say that our thoughts and experiences of color are about an objective property that nothing in the world instantiates. Rather, they say that these thoughts and experiences really concern a feature of our experiences, that we erroneously project onto the world. On the proposal we're considering, the thoughts that Dry Earthers express using "water" would turn out to work in just the same way.

D3 Another option, of course, is to say that when the Dry Earthers use the word "water," they're not really thinking or expressing *any* contentful thoughts whatsoever.

It is quite controversial and unclear which of these accounts of the Dry Earthers' thoughts would be best. For now, let's not worry about that. Let's just note these different positions that one can take.

On positions D1 and D2, the Dry Earthers' thoughts do have contents, but they don't have the same kind of contents as the thoughts that we and the Twin Earthers each express using "water." The Dry Earthers' contents are "more qualitative"; they are available to subjects in a broader range of environments. In fact, the Dry Earthers' contents are plausibly available *to us*. When I say that the Dry Earthers' contents are *available* to other subjects, I'm not saying that they are *the contents those other subjects express using "water."* Most of the other subjects will be like us and the Twin Earthers. *We have* the right sorts of clear drinkable natural kinds in our environments; the thoughts we express using "water" have the kinds of externalist contents we expect them to have, not the more qualitative contents that the Dry Earthers' thoughts have. But although the Dry Earthers' contents aren't *the contents we express using "water,"* they might nonetheless be contents *we can think*. We'd just have to choose other words to express them.

So, even if the "qualitative" contents the Dry Earthers express using "water" are available to us, they are not the contents *we* express using "water"; and no content of the kind we express using "water" is available to the Dry Earthers.

This asymmetry gives us a grip on the sense in which the Dry Earthers' thoughts "aim" to have the kind of externalist contents that our thoughts have. We and the Twin Earthers are in the kind of situation that Dry Earthers *would be* in, if there were a natural kind of the appropriate sort for them to be thinking about. In such situations, although the relevant "qualitative" contents would still be *available* to the Dry Earthers, they wouldn't be the contents they expressed using "water." The contents they expressed using "water" would be contents like our *water* or the Twin Earthers' *twater*. (Contrast a different term, which explicitly *sets out* to have the more qualitative kind of content that the Dry Earthers' thoughts in fact have: for instance, the term "clear drinkable liquid." This would continue to have a qualitative content even in worlds like our world and the Twin Earthers' world, where there are natural kinds having those qualities.)

On positions D1 and D2, then, we can say that the thoughts the Dry Earthers express using "water" *aim* to have an externalist content, like our content *water* or the Twin Earthers' content *twater*, but since their environment does not meet the necessary conditions for such a content to be available, the Dry Earthers' thoughts instead have a different, more qualitative **fallback content**.[21]

Let me emphasize: I'm not saying that the Dry Earthers' thoughts *definitely do* have such fallback contents. That is just one sort of position that one might take about the Dry Earthers' thoughts. Another option is position D3, which says that when the Dry Earthers use "water," they fail to think or express any contentful thoughts at all. (Well, there may be other contentful thoughts they're thinking—e.g., general thoughts like *There*

---

[21] These fallback contents need not be *purely* qualitative contents, available to subjects in *all* environments. They may themselves be externalist contents, and they may in turn have other fallbacks, if their own environmental conditions fail to be satisfied. All that's really required, at each stage, is that the fallback contents at that stage be available to subjects in *some* of the environments in which the previous content was unavailable. To keep our discussion simple, though, I will use examples where the relevant fallback contents are available to subjects in *all* of the environments we're considering.

are liquids—but according to D3, those won't be "the real contents" of their "water"-thoughts. D3 says that when the Dry Earthers say to themselves things like "Water puts out fires," they *merely seem* thereby to be thinking a contentful thought.)

Now, I do not think it is possible to determine by reflection alone whether we are in a situation like the one we and the Twin Earthers are in, on the one hand, or a situation like the one the Dry Earthers are in, on the other. So it is not possible to determine by reflection alone whether the thoughts we express using "water" have the kind of externalist content they aim to have, or whether they instead have a more "qualitative" fallback content—or, if we adopt position D3, whether we instead have no contentful thoughts at all. This fact will play a central role in my critique of the McKinsey-style reasoning.

Before we proceed, though, I want to make sure it's clear what fallback contents are and what they are not. When we talk about the Dry Earthers' thoughts having a fallback content, this is meant to be *a different content* than is had by the thoughts that *we* express using "water," or the thoughts that the Twin Earthers express using "water." We're *not* discussing a view that says that all our thoughts have a single content, which in our world has H_2O in its extension, in other worlds XYZ, and in yet other worlds a certain cluster of observable properties. That is *a* possible view about the content of our thoughts, but it is not an externalist view. So it is not one of the views I want to discuss in this paper. On the views we're discussing here, we have *three different* contents. For a given subject, at most *one* of them can be the content she expresses using "water." If her environment is like ours, then her "water"-thoughts will have one externalist content, *water*, that represents the natural kind H_2O in every world. No one can think thoughts with *that* content unless they inhabit an environment containing H_2O. If her environment is like Twin Earth, then her "water"-thoughts will have a *different* externalist content, *twater*, that represents the natural kind XYZ in every world. And if her environment is like Dry Earth, then her "water"-thoughts will represent something other than a natural kind, e.g. a cluster of observable properties.[22] It can be an open question, for all the

---

[22] Some may be tempted here to construe "water" as an indexical term, like "I" or "here," which in different contexts expresses different rigid contents. I *don't* think we should construe "water" in that way;

subject knows by reflection, which of the three contents we just described is identical to the content she expresses by saying "water." But don't confuse that *epistemological* fact with the *semantic* claim that the content she expresses with "water" is one that really does include H_2O, XYZ, and arbitrary clear drinkable liquids in its extension, with respect to different worlds.[23]

If our demonstrative thoughts, and thoughts we express using proper names, also have externalist contents, then much of what I've said here about natural-kind thoughts will also apply to them. For example, perhaps the name "Homer" *aims* to have a *de re* content, which is available only to subjects in environments that contain a particular

---

but I will have to discuss that elsewhere.

[23] In the previous section, I allowed that we can know by reflection alone what kinds of contents our "water"-thoughts *aim* to have. I'm also inclined to think that we can know by reflection alone whether there are fallback contents for our "water"-thoughts to have, if they fail to have the kinds of contents they aim to have. And if there are such fallback contents, I think we ought to be able to tell by reflection alone what they are. These things may not be *easy* to know; but I do think they can be known by reflection alone. Block and Stalnaker deny this… [[1999, pp. 21ff., pp. 36ff, p. 43. They deny that we can know, just in virtue of knowing the language, what to say in every case where there is no natural kind of the right sort. (Well, I don't say that either. In some cases there may be no determinate answer. And it may take more than just knowledge of the language: it takes *a priori* reflection as well. Still, I think that B&S would deny even my more moderate view.)]]

Two-dimensionalists like Chalmers [[cite??]] also have a view where it's an open question, for all we know by reflection alone, whether our "water"-thoughts refer to H_2O, or XYZ, or a cluster of observable properties; but whichever it turns out to be, our thoughts refer to that same thing with respect to every world. Two-dimensionalists also think that we can determine by reflection alone what fallback contents our "water"-thoughts would have, if our environment turned out to be like Dry Earth. So to those extents we are agreed. However, that is not enough to make me a two-dimensionalist. To be a two-dimensionalist, one also has to accept a number of other theses, theses that I reject. I discuss these other theses in "Bad Intensions" and elsewhere. (To take one example, I don't think we will generally be able to know by reflection alone purely qualitative necessary and sufficient conditions for our thoughts to have externalist contents. In this paper, I'm only assuming that we know *necessary* conditions for our "water"-thoughts to have externalist contents, and I've allowed myself to use the word "water" in stating those conditions.)

historical figure. But if it turns out that there is no historical person of the right sort to be the referent of "Homer," then this name may have a more qualitative fallback content, instead. Perhaps it represents a fictional character.[24] Or perhaps it represents a historical role, that in fact nobody fills. Similarly, perhaps our empty demonstrative thoughts represent a cluster of observable properties, such that there falsely seems to be some object we're demonstrating which has those properties. Or perhaps when we use empty names or demonstratives, we fail to have any contentful thoughts at all.

As before, it is a thorny question which of these different positions is best. But for our purposes, I think it is neither necessary, nor desirable, to take a stand on that. It is better if we keep the full range of positions in mind. To simplify our discussion, I will make some stipulative assumptions. I will assume that *there is* a fallback content for our "water"-thoughts to have, if it turns out that they can't have the kind of externalist content they aim to have. We can let this fallback content be a cluster of observable properties. I will also assume that there is *no* fallback content for the thoughts we express with "Homer." Either those thoughts have a *de re* content representing a real historical person, or else we have no contentful thoughts at all when we use the word "Homer."

You may not like these assumptions. Perhaps you think that "Homer" works more the way I'm saying "water" works. That is, perhaps there are also fallback contents for our "Homer" thoughts to have, should no *de re* content of the right sort be available. Or perhaps you think that our "Homer"-thoughts *don't even aim* to have externalist contents.

---

[24] If so, then its content would again be an externalist one, since no one can be thinking about a particular fictional character in worlds where the relevant fiction was never constructed. But the content wouldn't be externalist in the same way that names of flesh-and-blood Greeks are. It would plausibly be *more* qualitative, available in a broader range of environments, than a *de re* content about a particular person would be.

Even if our historical evidence about Homer turns out to be fabricated, I do think that the name "Homer" *aims* to represent a flesh-and-blood person. It would only represent a fictional character as a fallback, because the kind of content it aims to have turns out to be unavailable. Contrast the name "Sherlock Holmes," which *always aimed* to represent a fictional character. (Even if, unbeknownst to us and to Conan Doyle, there *was* a person who did the things described in the Holmes stories, he wouldn't be Holmes.)

It doesn't really matter. I'm just using these examples to illustrate the range of different positions one might hold. I want to argue that the McKinsey-style reasoning fails to go through, *no matter which position one holds*.

I said that it's not possible to determine by reflection alone whether one is in a situation like the one we and the Twin Earthers are in—where our thoughts have the externalist contents they aim to have—or whether one is in a situation like the one the Dry Earthers are in. Now, I *don't* take that to entail that one can't know what the contents of one's own thoughts are. I think we *do* know what the contents of our "water"-thoughts are; we just can't know *everything about* those contents by reflection alone. So long as *there is* some content, *water*, that we're thinking when we say "water," introspection enables us to be aware of this content and to know that it is the content we're thinking. What we can't know by reflection alone is whether this content is an externalist content, of the sort described by principle (c) earlier:

(c) Whatever content C we express with "water," it is a content that is only available to subjects who inhabit environments containing samples of the stuff C represents, which those subjects or other members of their linguistic community have interacted with.

or whether this content is instead a more qualitative fallback content. For all we know by reflection alone, we might be Dry Earthers. If we were, then (c) and McK-2 would no longer be true. Hence, when we're dealing with thoughts like our "water"-thoughts, where *they might* have more qualitative fallback contents, we won't be in a position to know McK-2 by reflection alone.[25]

It might for all that be that our environment *is* cooperative, and our thoughts *do* have the externalist contents they aim to have. In that case, McK-2 would be *true*. We're just not in a position to know by reflection alone whether that's so. Whether our thoughts have externalist contents or not will depend on whether our environment *is* cooperative, not on whether we know that it is.

---

[25] Cite McLaughlin & Tye's papers. [[Also Raffman thinks we can't simultaneously know all the premises of the McKinsey-argument by reflection alone.]]

These observations are an important start to a diagnosis and defusal of the McKinsey-style reasoning. But many problems remain. What I've said may be fine for thoughts that work the way I'm assuming our "water"-thoughts work. These are thoughts where, if the kind of externalist content they aim to have is unavailable, they instead have a more qualitative fallback content. But what about cases where one of the fallback options is *having no content*? What about thoughts that work the way I'm assuming our "Homer"-thoughts work, that is, such that they if they have *any* content at all, then it is the externalist content they aim to have? For these thoughts, we *should* be able to know, just by understanding them, that if they have any content, then (c) and McK-2 are true of that content. So we need to figure out what our epistemic position is when we're dealing with those sorts of cases.

VI

To do that, we need to get clearer about *a priori* justification, and the differences between it and certain kinds of introspective and perceptual justification.

As I said in §I, I don't count the introspective justification we have for beliefs about our own experiences, intentions, and thought-processes as *a priori* justification. I regard the belief that I am in pain, the belief that I'm thinking of a number greater than 12, and the like, as a justified *a posteriori*. Now, Shoemaker and Burge and others [cite] have argued that there are important differences between introspection and perception, and that it's wrong to think of the awareness we have our own experiences and thought-processes as a kind of "inner perception." I agree. However, it still seems to me that the kind of justification we have for beliefs about our own experiences and thought-processes is much *more* like the justification we get from perception than it is like the justification we rely on when doing things like math, metaethics, and analytic metaphysics. The boundaries between *a priori* and *a posteriori* justification are controversial; and they may be vague. And it may be difficult to say what everything we regard as *a priori* has in common. Nonetheless, I think there is a natural and robust category here, that includes the justification we get from mathematical reasoning, conceptual analysis, reflecting on

philosophical thought-experiments, and the like, and that excludes our awareness of what's currently going on in our minds, or in our perceptual environments.[26]

Now, I said that we could introduce a composite notion, "reflective justification," that combines *a priori* and introspective justification. And it is true that the epistemic notion used in McKinsey's argument is this composite notion. However, we will learn some valuable lessons by thinking about the ways in which *a priori* and introspective justification are *distinct*. These lessons will enable us to use the composite notion of reflective justification more carefully. They will also help us to answer the questions I raised at the end of the preceding section.

There are four lessons I want to stress about *a priori* justification.

The first lesson is that experiences can be necessary *for you to entertain* a thought, without thereby playing a role in *your justification* for believing that thought. For instance, assume that Homer did in fact exist, and consider:

(4) the *de re* thought, of Homer, that if he exists then he is self-identical

One needs to have had certain kinds of experiences to be in a position to entertain this *de re* thought. But those experiences aren't the source of your justification for believing (4). Your justification for believing it comes from your understanding of what it is to be self-identical. Hence, I would class (4) as something you can know *a priori*.[27]

The second lesson is that we shouldn't apply the first lesson indiscriminately. Sometimes the experiences that enable you to entertain a belief *do* play an essential role in your justification for that belief. Here are some examples.

Suppose you look out the window and see a canary. You form the thought *That canary exists*. Or you walk into an office building and hear a distant phone ringing. You

---

[26] Like me, BonJour 1998, pp. 7ff. counts introspective justification as *a posteriori*. Contrast Boghossian 1989. Kitcher discusses whether introspectively-based knowledge should be counted as *a priori* knowledge in Kitcher 1980?? §V. [[Check his book.]]

[27] The following authors also argue that the experiences necessary to entertain a thought need not be playing a justifying role: BonJour 1995, p. 53; Kitcher 1980??, pp. 4-5; Burge 1993, p. 460; Audi 1983??; Alston 1976, p. 293; Alston 1983, pp. 62-3; Plantinga 1993, pp. 103-4; BonJour 1998, pp. 9-10.

form the thought *That phone is ringing*. Or, stumbling about in a dark room, you place out your hand and touch a wall. You form the thought *This wall is perceived by me*. In each of these cases, you require certain experiences just to be able to entertain your demonstrative thought. And in each case, the experiences that enable you to entertain the thought also give you all the *a posteriori* justification you need, to be justified in believing that thought. Your visual experiences of the canary both enable you to think the demonstrative thought *That canary exists* and justify you in believing it. Your auditory experiences of the ringing phone both enable you to think the demonstrative thought *That phone is ringing* and justify you in believing it. And so on. We can call this **presentational justification**, because it's the experiences that "present" the thought to you, or enable you to entertain it, which justify you in believing it. In these cases, your justification is still coming from your experiences; so I count them as cases of *a posteriori* justification rather than *a priori* justification. Suppose a friend had all the same experiences you have; but instead of forming demonstrative thoughts, he formed the general thoughts that *A canary exists*, *Some phone is ringing*, and *I am perceiving a wall*. Your friend's justification for his beliefs is acknowledged on all sides to be *a posteriori*. Your justification for your beliefs seems to me to be cut from the same cloth as your friend's. The fact that you can only entertain your beliefs by having certain experiences, experiences which are enough to justify that belief, does not seem to me to make your justification any less experiential or *a posteriori*.

With your belief *That canary exists*, then, your justification may be presentational but it's still *a posteriori*. With thoughts like (4), on the other hand, I think that experiences only play a role in enabling you to entertain the thought. Your justification for believing it comes from your understanding of the relation of self-identity, and seems to be fully *a priori*.[28]

I would not say the same thing about:

---

[28] This distinction between presentational justification and *a priori* justification hasn't played any large role in discussions of *a priori* knowledge, though I think it should. A few other philosophers have noted the distinction, or related distinctions. See BonJour 1998, p. 10; Salmon [where?]; and Soames on trivial knowledge [where?]

(5) the *de re* thought, of Homer, that he is self-identical

Notice the difference between (4) and (5). (4) is of the form "If he exists, then he is self-identical." (5) is of the form "He is self-identical." I don't think we can have purely *a priori* justification for believing (5). This is because (5) entails that Homer exists. And intuitively, that's not something we should be able to establish on *a priori* grounds. (That's just the kind of McKinsey-style result that we find so puzzling.) Hence, unless we find compelling reason to think otherwise, I think we should be reluctant to count (5) as being justifiable by purely *a priori* considerations. You may have *presentational* justification for believing (5); but I'm inclined to think that all you can have *a priori* justification for believing is (4).[29]

Now consider *the sentences* S(4) and S(5) that we use to express (4) and (5):

S(4) "If Homer exists, he is self-identical."

S(5) "Homer is self-identical."

If you don't know whether Homer exists, then you won't be able to determine, by *a priori* reasoning alone, whether either of those *sentences* succeeds in expressing a contentful thought. For both sentences contain the name "Homer"; and, given the way I'm assuming our "Homer"-thoughts work, if that name has no referent, then sentences containing it will not express any contentful thought.[30] But this brings us to our third

---

[29] [[Alternatively, one might try to bring in a limiting principle on *modus ponens* reasoning like the one we formulated when discussing Wright and Davies. But a limiting principle of that sort would apply here only if, to be justified in believing (5), you needed to be antecedently justified in believing that Homer exists. And since, in my view, justification for believing that Homer exists can only be *a posteriori*, this means that the limiting principle would apply only if your grounds for (5) needed to be supplemented by some antecedent *a posteriori* justification for believing that Homer exists. And to my ears, that sounds tantamount to conceding that you can't have purely *a priori* justification for believing (5), after all. Just as I was saying.]]

I think it is intuitively clear that (4) does not entail that Homer exists. It is a tricky matter coming up with an adequate semantics and logic that preserves that intuition. But it is one we should preserve.

[30] Evans 1979 stresses this point, that if "N" doesn't refer, then neither "N is so-and-so" nor "If N exists, N is so-and-so" will express a contentful thought. See also Oppy 1994, fn. 14.

lesson, which is that there is a difference between knowing that some sentence expresses a contentful thought, and having justification for believing the thought it expresses.

Suppose that Homer does exist, and hence the sentence S(4) does express a contentful thought, viz., the thought (4), which you are entertaining. As I've already said, I think you can have *a priori* justification for believing that thought (4). Your justification for believing it comes from your understanding of the relation of self-identity. But I think you won't have *a priori* justification for believing either of the following:

(vii) S(4) expresses a contentful thought.

(viii) If there is a contentful thought that S(4) expresses, I am entertaining it.

You won't have *a priori* justification for believing (vii), because given your knowledge of how the word "Homer" works, having justification for believing (vii) seems to presuppose having justification for believing that Homer exists. If he didn't exist, then the word "Homer" would have no content. You won't have *a priori* justification for believing (viii) because your justification for it will come from your introspective awareness of your current cognitive activities, and I'm not counting introspective justification as *a priori*.

Nonetheless, in the case we're considering, *it is true* that S(4) expresses a contentful thought, which you are entertaining; and I think it is also true that you can have *a priori* justification for believing that thought. I don't think it's a requirement, for you to be justified in believing a thought P, that you be justified in believing that the sentence you use to express P does in fact express P or any other content.[31] To be sure, it would be peculiar to go ahead and believe P, if you had doubts about whether the sentence you used to express P did in fact express anything. Perhaps it would be not

---

[31] [[Here I part company with Davies. He thinks that in order to be justified in believing P, you need to be antecedently entitled to the belief that there is such a belief as P. Hence, if, as I believe, we can only have *a posteriori* justification for our belief that there is such a belief as P, it follows that in order to be justified in believing P, we need to have antecedent *a posteriori* justification for believing something else. It seems like this would make it impossible to be *a priori* justified in believing P. I reject Davies' claim that in order to be justified in believing P, you need to be antecedently entitled to the belief that there is such a belief as P.]]

merely peculiar, but irrational. Perhaps you could never be rational in believing P, while having doubts about whether the sentence you used to express P really succeeded in expressing a contentful thought. But what if you *had no view about* whether the sentence you used to express P expressed a contentful thought? Or what if you just (truly) took it for granted that the sentence expressed a contentful thought, without having any justification for believing so? Would it still be irrational for you to believe P, in those cases? I don't see why it should. For one thing, it takes a certain amount of philosophical sophistication to think about one's sentences, and what they express. One can certainly *use* sentences to express thoughts, like (4), long before one has achieved that level of sophistication. And one can also have the understanding of self-identity that seems to justify belief in (4), before one has achieved that level of sophistication. So I think it is possible to be justified in believing (4), even if one has no view about, and lacks any justification for believing, the claim that the sentence one uses to express (4) does express a contentful thought. (Cases where you do believe that the sentence expresses a contentful thought, without having justification for believing so, are more complicated. But I think in those cases, too, you can be justified in believing the thought the sentence expresses, without needing to be justified in believing that the sentence expresses it.)

Now, one noteworthy fact about S(4) and S(5) is that, whenever these sentences do express a contentful thought, the thought they express has to be true. And that's something we can know *a priori*.[32] But—and here is our fourth and final lesson—knowing that:

(ix) A sentence S is true whenever it *does* express a contentful thought.

is different from knowing:

(x) the thought that S in fact expresses.

One difference is that you can know (ix) without *knowing which* thought S expresses. But even when you *do* know which thought S expresses, it could be that what justifies you in

---

[32] This assumes that our knowledge of these sentences' semantic properties is *a priori*. In fact, I do not believe that knowledge of semantic properties ever is *a priori*. But the current discussion is already complicated enough; so I will not make any fuss about this here.

believing that thought, (x), is different from what justifies you in believing (ix). Here's an example. If you understand what the following sentence means, then that's enough to know that whenever it expresses a contentful thought, that thought is true:

S(6) "I am uttering a sentence."

S(6) can only express a definite content if there is some agent to be the referent of "I." So it can only express a contentful thought when it's uttered. (For present purposes, count saying the sentence to yourself privately as a kind of utterance.) So given what the sentence means, it follows that whenever it expresses a contentful thought, the thought it expresses is true. Anyone who understands the sentence is in a position to know that. However, suppose you *do* utter the sentence (either privately or aloud). What then justifies you in believing that you are uttering it, or any sentence? It can't be your understanding of the sentence. That would only justify you in having beliefs about what's true *when the sentence is being uttered*. It doesn't help you determine when that condition is fulfilled. The natural thing to say is that what justifies you in believing that you are uttering the sentence is your introspective or perceptual awareness of uttering it. Hence, your justification for believing the thought that S(6) expresses is *a posteriori*—despite the fact that you know, just in virtue of understanding S(6), that whenever it expresses a contentful thought, that thought is true.

I think the same holds for sentences like:

S(7) "I exist."

Your understanding of that sentence is enough to justify you in believing that whenever it expresses a contentful thought, that thought is true. We can count *that* justification as *a priori*. But your justification for believing the thought that S(7) expresses is not itself *a priori*. What justifies you in believing that you exist would be something like your introspective awareness of your own thought-processes—including perhaps the act of entertaining the thought that you exist.[33]

---

[33] I would say the same thing about other *cogito*-type thoughts, like *I am thinking* and *I am hereby thinking about the number zero*. I can know *a priori* that such thoughts are true whenever they are entertained, but my justification for believing them rests on my introspective awareness that I am entertaining them, and is at best presentational. It is not *a priori*. Audi 1999, pp. 212-13 also argues that our

Similarly, although you may know *a priori* that whenever:

S(5) "Homer is self-identical."

expresses a contentful thought, the thought it expresses it true, I do not think your knowledge of the thought it expresses will ever itself be *a priori*. To be justified in believing that Homer is self-identical, I think you need to have some perceptual or historical evidence that Homer exists.

VII

Armed with these lessons, we're now in a position to complete our diagnosis and defusal of the McKinsey-style reasoning. Let's recall once again how that reasoning goes.

I will begin by splitting the original premise McK-1 into two more cautious sub-premises. Introspection is supposed to justify you in believing both:

McK-1a If you are thinking *any* contentful thought right now, you are thinking a thought with the content C.

and:

McK-1b You are thinking a contentful thought right now.

And *a priori* philosophical reflection is supposed to justify you in believing:

McK-2 If you are thinking a thought with the content C, then your environment is such that…

Armed with the lessons of the previous sections, we're now in a position to be more critical about these claims.

We've already discussed thoughts like the ones we express with "water," where, if the kind of externalist content they aim to have is unavailable, they instead have a more qualitative fallback content. In such cases, I argued, you wouldn't be able to know anything like McK-2 by reflection alone.

---

justification for such *cogito*-type thoughts is not genuinely *a priori*.

But what now of thoughts like the ones we express with "Homer," where if the thought has any content at all, it has to be an externalist content? There it looks like you *could* know McK-2 by reflection alone. At least McK-2 would have the same status as:

(4) the *de re* thought, of Homer, that if he exists then he is self-identical

You might not be able to know by reflection alone whether the sentences you use to express (4) or McK-2 succeed in expressing contentful thoughts. But when they do succeed, you are able to entertain the thoughts (4) and McK-2, and I think it is plausible that your justification for believing those thoughts would be *a priori*.[34]

But what about McK-1a and McK-1b?

I think you *can* have purely reflective justification for believing McK-1a. If there is a contentful thought you're thinking, then you can be introspectively aware of that thought, and introspection will enable you to know that—if you can set aside the possibility of illusions of content—*it* is the thought you're thinking. In other words, if there is *any* contentful thought you're thinking, it is that one. Of course, when you're trying to express a thought with "Homer," there is a possibility that you might *fail* to be thinking a contentful thought at all. But, just as you can have *a priori* justification for believing McK-2, when the sentence you use to express it *does* express it, so too do I think you can have introspective justification for believing McK-1a, when the sentence you use to express it succeeds in expressing a contentful thought.

With McK-1b, on the other hand, I think the McKinsey-style reasoning will strike out. So long as it's an open possibility that your attempts to express thoughts with "Homer" fail to express contentful thoughts, I don't see how you could be in a position to know McK-1b by reflection alone. In such a case, to be justified in believing McK-1b, I think you'd need to have some perceptual or historical evidence that Homer exists. Now, the way in which you acquired the name "Homer" may have *brought along with it*

---

[34] [[This is something *I myself believe* about McK-2, and that *I am prepared to grant* to the defender of McKinsey-style reasoning. It is not a claim that *I want to defend*. You might think instead that all we can know *a priori*, when we're dealing with thoughts like the ones we express with "Homer," is that whenever the sentence S(McK2) [??] expresses a contentful thought, that thought will be true. I'm just granting, though, for the sake of argument, that we could know the thought McK-2 itself *a priori*.]]

empirical justification for believing that Homer exists. Or you may have subsequently acquired such justification. But neither of those will constitute having reflective justification for believing that Homer exists. So I don't think you'll be able to have *purely reflective* justification for believing McK-1b.

Hence, for neither kind of thought are you in a position to have the justification that the McKinsey-style reasoning requires, for all the premises simultaneously. With thoughts you express using "water," which have qualitative fallback contents, you won't have the right kind of justification for McK-2. With thoughts you express using "Homer," which are externalist-or-bust, you won't have the right kind of justification for McK-1b. And "water" and "Homer" are meant to illustrate the whole range of positions one might take about what happens to our thoughts when the environment is uncooperative. I think that no matter what position one ends up adopting, subjects will never have the kind of justification that's needed, to derive claims about their environment by reflection alone.

VIII

We'll be able to understand my solution to McKinsey's puzzle better, and add some refinements, by comparing it to the various solutions that we set aside earlier.

I've argued that you can't be justified in believing all the premises of McKinsey's argument, in the way that's needed to generate the puzzling result. Isn't this just the same as the incompatibilist response I set aside in §I?

No, it is not. The incompatibilist says that it can't be true both that a given thought has an externalist content, and that you are able to know by reflection alone that you're thinking a thought with that content. In other words, if McK-2 is *true*—if your thought *in fact* has an externalist content, available only to subjects in certain sorts of environments—then you can't know by reflection alone that you're thinking that thought. That's not my view. My view says that if McK-2 is *known by reflection alone*—if you know by reflection alone that your thought has an externalist content—then you can't know by reflection alone that you're thinking that thought. But McK-2 might very well be *true*, without being knowable by reflection alone. For example, McK-2 is true of our

"water"-thoughts, since we inhabit a cooperative environment, not a place like Dry Earth. But we can't know by reflection alone that that is so.

Incompatibilism strikes me as an *over-reaction* to McKinsey's puzzle. McKinsey's puzzle gets going in the following way:

> You allegedly know McK-1 by reflection alone, and know McK-2 by reflection alone; so it seems like you can derive claims about your environment by reflection alone.

My solution says:

> You *can't* know McK-1 by reflection alone, if McK-2 is really also *known by reflection alone.*

The incompatibilist says:

> You can't know McK-1 by reflection alone, if McK-2 is really *true*.

The incompatibilist is saying more than is needed, to block the McKinsey-style reasoning. That is why I call incompatibilism "an over-reaction." I think it is a stronger reaction than McKinsey's puzzle warrants.[35]

Now, what of Wright and Davies' charge that the McKinsey-style reasoning is guilty [??] of transmission-failure?

With thoughts like the ones we express with "water," there is no threat of an illusion of content. I think you *can* know by introspection alone that you're thinking a thought of that sort. And here I see no threat of transmission-failure.[36]

---

[35] There is one thing that the incompatibilist and I have in common. We both deny that *all* the relevant properties of our thoughts are knowable by reflection alone. The incompatibilist says: we can't know by reflection alone what the thought's content is. I say: depending on the thought, either we can't know whether it's a genuinely contentful thought by reflection alone (though when it is contentful, I think we can know what content it has); or we can't know by reflection alone whether its content is externalist or qualitative. In every case, there is *some* interesting property of our thought which is not knowable by reflection alone. To that extent, the incompatibilist and I agree. But we disagree about the details; and the details matter.

[36] Nor does Wright; see C&QB pp. 144-45, 151-52.

With thoughts like the ones we express with "Homer," on the other hand, there *is* a threat that your putative thoughts really have no content. Ordinarily, you'll be in a position to rule that threat out. ("Homer" is an unusual case; ordinarily we *do* know whether the referents of our thoughts exist.) But I don't think that *introspection* will be what enables you to rule it out. *Introspection alone* won't be enough to justify you in believing that your "Homer"-thoughts are genuinely contentful.

Wright and Davies claim that we have a *default entitlement* to believe that no illusions of content are taking place.[37] That may very well be so. But Wright and Davies think it enables us to count your justification for believing things like *I am thinking that Homer was blind* as still being "introspective" (or perhaps, "reflective"). I suppose that what's motivating them to say this is that your belief could be justified, even if it was *formed* solely by a process of introspection, without being *inferentially based on* any evidence about your environment.

But on their view, your belief does still *epistemically rest on* the assumption that no illusions of content are taking place—or, as Davies would put it [??], the assumption that "there is such a belief" as *Homer was blind*. Wright and Davies think you *do* need to be antecedently justified or entitled to believe that, before you can be justified in believing *I am thinking that Homer was blind*. And given your knowledge of how "Homer" works, you know that there *will be* an illusion of content unless Homer exists. Hence, it seems to me that you can be *justified* in believing *I am thinking that Homer was blind* only if you have antecedent justification for believing that Homer exists.[38] It doesn't

---

[37] See, e.g., Wright C&QB, pp. 152-3, 156-7.

[38] I *don't* mean to be relying on any general principle here, to the effect that *if* having justification for believing A requires you to have antecedent justification for believing B, *and* you know that B will be false unless C, *then* having justification for believing A would also require you to have antecedent justification for believing C. I don't think that general principle would be *true*. Rather, it's the particular details of this case that persuade me that you can be justified in believing *I am thinking that Homer was blind* only if you're antecedently justified in believing that Homer exists.

Notice: I'm *not* saying that you can *have* "Homer"-thoughts only if you have antecedent justification for believing Homer exists. Nor am I saying you can be *justified in having* "Homer"-thoughts only if you have antecedent justification for believing Homer exists. As I've already pointed out, I think

seem plausible that *introspection* could be the source of that justification. Even if your belief *I am thinking that Homer was blind* was formed spontaneously and non-inferentially, it seems like *your justification* for it would require more than introspection itself is capable of giving you. Introspection would only justify you in believing things like McK-1a:

McK-1a If you are thinking *any* contentful thought right now, you are thinking a thought with the content *Homer was blind*.

To know the antecedent of that conditional (that is, to know McK-1b), you would have to supplement your introspective justification with perceptual or historical evidence that Homer exists.

Ordinarily *you will have* such evidence. And you won't need to explicitly call it to mind, to be justified in believing *I am thinking that Homer was blind*. But what matters for our purposes is that it won't be introspection that gives it to you. Once the possibility of an illusion of content is open, it won't be *introspection* that enables you to close it.

Hence, even if Wright and Davies' charge of transmission-failure turns out to stick in these cases—even if your justification for believing the premises of McKinsey's argument requires you to be antecedently justified in believing its conclusion—I don't think we need to *appeal to* that fact to block the McKinsey-style reasoning. We've *already* blocked the reasoning, once we've shown that you don't have purely reflective justification for believing all the premises. That's enough to show that you're in no position to derive conclusions about your environment from reflection alone.[39] If the McKinsey-style reasoning *also* exhibits transmission-failure, that's merely an interesting corollary.

---

you can have purely *a priori* justification for believing *If Homer exists, he is self-identical*. All I'm saying here is that to be *justified in believing that you have* contentful "Homer"-thoughts, you need to be antecedently justified in believing that Homer exists.

[39] [[See Beebee here??]

Let's go back to Wright and Davies' claim that we have a default entitlement to believe that no illusions of content are taking place. I'm inclined to believe this. But the entitlement is of course *defeasible*. And I think that Wright and Davies have *misidentified* the kind of defeat it is susceptible to. I think that what defeats it is not *perceptual evidence* that your environment fails to meet certain requirements, e.g., evidence that Homer does not exist. Rather, I think that what defeats it are *the kinds of a priori considerations* that justify you in believing that your "Homer"-thoughts have no fallback contents, and hence that they are contentful only if your environment meets the relevant requirements. That is, when *a priori* reasoning convinces you of McK-2, *that* is what defeats your ability to know by introspection alone that you're thinking contentful thoughts about Homer.

On this view, you *start off* entitled to believe that your thoughts are contentful no matter what your environment is like. When you're in that epistemic situation, evidence that Homer doesn't exist will have no tendency to defeat your introspective justification for believing you're thinking contentful thoughts. Once you acquire justification for believing McK-2, though, that's enough to open the possibility of illusions of content; and this is a possibility that I said *introspection* would not enable you to rule out. Hence, I think that, *regardless* of how much evidence you have for or against Homer's existence, once you know McK-2, you're no longer able to know McK-1b by introspection alone.

Now how could you acquire justification for believing McK-2? In cases where McK-2 is knowable only *a posteriori*, as with our "water"-thoughts, you can acquire justification for believing McK-2 only by acquiring evidence that you inhabit a cooperative environment like Earth, rather than an uncooperative one like Dry Earth. That evidence will tell you that your thoughts have the externalist contents they aim to have. That will be a case where you have evidence for McK-2, but, precisely because it's evidence that your environment is cooperative, it wouldn't be evidence that leaves you in the dark about *whether your thoughts are contentful*. I think it will only be when McK-2 is knowable *a priori*, as with our "Homer"-thoughts, that you'll be able to come to know McK-2 in a way that leaves it an open question whether the relevant thoughts are contentful. (That doesn't mean *your own* knowledge of McK-2 has to be *a priori*; you might have *a posteriori* evidence that *someone else* has good *a priori* reasons to believe

McK-2.) And in those cases, I think your knowledge of McK-2 would already be enough, by itself, to render you incapable of knowing McK-1b by introspection alone.

So evidence that Homer doesn't exist *isn't sufficient* to defeat your introspective justification for believing McK-1b, if you haven't yet encountered some reason to believe McK-2; and evidence that Homer doesn't exist *isn't necessary*, if McK-2 is knowable by reflection, and *you do have* some reasons to think it true.

If our epistemic situation concerning McK-2 can change in the ways I've described, then this can generate effects that are interestingly akin to the Harman-style instabilities we discussed in §III. There we discussed cases where you started out with justification for believing $A \supset B$, but when you acquired evidence for A, that defeated the justification you had for believing $A \supset B$. In the present context, we're discussing cases where you start out with introspective justification for believing *I am thinking that Homer is bald*, but when you do some philosophy and realize *If I am thinking that Homer is bald, then my environment has to be such-and-such ways*, that defeats your ability to be justified in believing *I am thinking that Homer is bald* by introspection alone. In the Harman cases, it was your justification for believing the antecedent that defeated your justification for believing the conditional; here it is your justification for believing the conditional that defeats your justification for believing the antecedent. So the cases are not quite the same. But I find the parallel quite intriguing.