Philosophical Review

Externalism and Knowledge of Content

Author(s): John Gibbons

Source: *The Philosophical Review*, Vol. 105, No. 3 (Jul., 1996), pp. 287-310
Published by: Duke University Press on behalf of Philosophical Review
Stable URL: https://www.jstor.org/stable/2185702
Accessed: 02-11-2018 11:59 UTC

The Philosophical Review, Vol. 105, No. 3 (July 1996)

Externalism and Knowledge of Content

John Gibbons

Many believe that content externalism is inconsistent with commonsense views about our knowledge of the contents of our own thoughts.[1] Content externalism is the view that the propositional contents of an individual's thoughts do not supervene on the intrinsic properties of that individual. Relations between you and your social and physical environment partly determine the contents of your thoughts.[2] But if what determines the content of your thoughts lies partly outside your mind, it might seem that you have to investigate your social and physical environment before you can know the content of your thoughts. If such investigation were necessary, our knowledge of our own mind would be much less direct and much less warranted than we ordinarily believe.

How do we connect content externalism with failures of self-knowledge? I believe that water is wet. This first-order belief about the world is subject to the limitations of ordinary empirical knowledge. But I also believe that I believe that water is wet. Apart from cases of self-deception or conceptual confusion, it is difficult to see how I could be wrong about this second-order belief about my own mind. Even if I am wrong about the world, I know what I am thinking. But if I had grown up on Twin Earth, I would now be confidently asserting that I know that I believe that twin-water is wet. I cannot tell through introspection whether I grew up on Earth or Twin Earth, and there is no qualitative difference between believing that water is wet and believing that twin-water is wet. So how do I know which one I believe?

We typically do not investigate our environment, find that water

[1]Here and in what follows, 'thought' refers to token mental events with propositional content—for example, particular beliefs and desires. I take it that 'the content of a thought' refers to a proposition or a type of thought, though nothing in what follows turns on a particular ontology of contents.

[2]See Hilary Putnam, "The Meaning of 'Meaning'," in *Mind, Language, and Reality* (Cambridge: Cambridge University Press, 1975), 215–71, and Tyler Burge, "Individualism and the Mental," *Midwest Studies in Philosophy,* vol. 4, ed. French et al., (Minneapolis: Minnesota University Press, 1977), 73–121.

rather than something else causes our thoughts, and discover the contents of our thoughts in this way. Nor is such investigation intuitively necessary. The externalist, then, needs to give an independently plausible account of how we know the contents of our thoughts and to explain, using this account, the possibility of self-knowledge, given the fact that external factors partly determine content. This is my project. One of the leading ideas behind my account of self-knowledge is that the warrant of a belief depends heavily on certain features of its causal history. This idea is central to, but not exclusive to, epistemological externalism, the view that the warrant or justification for a belief does not supervene on introspectively accessible properties of the believer. Another leading idea is the familiar content-externalist idea that the content of a thought depends on certain features of its causal history. It is no accident that I put these two ideas together. The kinds of causal relation relevant to knowledge seem to be among the most important kinds of causal relation relevant to the determination of content.

1.

In epistemology, as in action theory, there is a difference between reasons for and reasons for which. You may have good reasons for doing something or for believing something, but they may not be the reasons for which you actually do it or believe it. In action theory, as in epistemology, there is reason to think that the difference between reasons for and reasons for which is a causal difference: the reasons for which you do something or believe something are those reasons for doing it or believing it that actually cause you to do it or believe it.[3]

The idea that the justification relation is a partly causal relation does not apply only to the justification of one belief by another belief. If you infer the belief that p from the belief that q, then your belief that q is clearly a reason for which you believe that p. But if you are in pain, then, at least on some views, the phenomenal experience itself can be part of your justification for believing

[3]For the distinction and reasons for thinking the difference is causal, see Donald Davidson, "Actions, Reasons, and Causes," in *Essays on Actions and Events* (Oxford: Oxford University Press, Clarendon Press, 1980), 3–19.

that you are in pain. Again, the experience helps justify the belief only if it helps cause that belief. If the belief comes about from guessing, conceptual confusion, or another's testimony, the experience of pain is epistemically irrelevant. This is important because the way a first-order thought justifies a second-order belief[4] is much more like the way an experience justifies an experiential belief than it is like the way a premise justifies a conclusion.

So to ask about the warrant for our beliefs about our own propositional attitudes, we need to look at the source of these beliefs. What caused your belief that you are thinking about epistemology? Your thoughts about epistemology caused this higher-order belief. This idea provides the basis for a general account of self-knowledge that shows the compatibility of externalism with self-knowledge.

Consider the following partial functionalist account of conscious thoughts

(F) If a thought of yours is conscious, it must cause a higher-order belief to the effect that you have the thought.

So if it occurs to you that p, this event will cause you to believe that it occurred to you that p. This is analogous to the familiar view that part of what it is to be a pain is to cause beliefs that you are in pain. While having conscious thoughts may involve more than this, a reflective individual cannot typically have conscious thoughts without knowing about those thoughts. This knowledge involves, perhaps among other things, some sort of higher-order belief.[5]

The following considerations motivate the causal requirement in (F). First, suppose that as a result of reading Freud, you come to believe that you have some particular subconscious desire. The mere presence of the higher-order belief does not guarantee that

[4]By 'second-order belief' I mean a belief about any first-order propositional attitude.

[5]A detailed argument connecting consciousness with higher-order thoughts using the notion of reportability appears in David Rosenthal's "Thinking That One Thinks," in *Consciousness: Psychological and Philosophical Essays,* ed. Martin Davies and Glyn W. Humphreys (Oxford: Blackwell, 1993), 197–223. Also see his "Two Concepts of Consciousness," *Philosophical Studies* 49 (1986): 329–59. In the more recent paper, Rosenthal explicitly rejects the requirement that first- and second-order thoughts be causally related (205 n. 16).

the desire is conscious. At the very least, the higher-order belief must be noninferential. But this is not sufficient. If a neuroscientist were to implant the higher-order belief, it would not thereby be inferential. But this belief would not guarantee the consciousness of the desire.

Second, the idea behind (F) is that we know about our conscious states in some sort of immediate way, and this involves higher-order beliefs. But we would not say that your knowledge was of a particular table or chair if the object were not appropriately causally related to your beliefs. Similarly, higher-order beliefs do not count as knowledge of your thoughts unless those thoughts cause the beliefs.

Finally, higher-order beliefs are typically nonconscious. In a case of introspection, we pay attention to one of our own mental states. But this is relatively rare. In the usual case, we pay attention to the world. Also, if every conscious thought caused a higher-order conscious thought, we would be off on a regress. Since the higher-order thought is typically not conscious, we avoid both of these difficulties. In a nonintrospective case, only the first-order thought about the world is conscious, and this determines the focus of attention. Introspective cases differ in that they involve *conscious* second-order thoughts, thus, nonconscious third-order thoughts. There is no regress of states, conscious or otherwise. Since the highest-order belief is not conscious, there is no reason to suppose it causes another, higher-order belief.

The functional role of conscious thoughts involves not only causing the second-order belief but also partly determining the content of that belief. Remember the analogy with pain. States of pain themselves go a long way toward determining the content of the beliefs they cause. If the qualitative features of pain are important to what it is to be a pain, then our beliefs about our pains represent them as states of a certain qualitative sort. The qualitative property exemplified by the pain determines the qualitative property relevant to the content of the belief.

The same goes for our beliefs about our thoughts. Thinking that p not only causes a second-order belief but also partly determines the content of that belief. The second-order belief inherits its content from that of the first. It is important to see what can and what cannot be inherited. The first-order belief that the water is boiling causes a second-order nonconscious belief that you believe that the

water is boiling. The first-order desire that the water be boiling causes a second-order belief that you want the water to be boiling. The first-order thoughts have the same content. Both involve the proposition that the water is boiling. The difference between the thoughts is the different attitude you take toward the proposition.

The second-order beliefs, on the other hand, do differ in content. One belief involves the concept of *belief*, and the other involves the concept of *desire*. While the second-order belief inherits the concepts of *water* and of *boiling* from the first-order causes, it does not inherit the concepts of *belief* and *desire* in this way. The latter concepts cannot be inherited from the first-order thoughts because these concepts are not part of the content of the first-order thoughts. In order for x to inherit an intentional property from y, not only must y cause x, but y must have the relevant property to pass on. So while the proposition ascribed by a second-order belief is inherited from the first-order thought, the attitude ascribed is not.

One clear example of content inheritance comes from intentional action. If you are looking for water (or the Fountain of Youth), your behavior has a certain intentional property. You can be looking for water even if you do not find any and even if there is currently no water in your vicinity to find. Your behavior, for example, your walking around in the kitchen, does not mean *water* or represent water. But the relevant description of your behavior fails the usual tests of existential generalization and substitutivity. In this case, the behavior inherits the intentional property from the beliefs, desires, or intentions that cause it. There is nothing you need to do, over and above looking for water, to ensure that your behavior has this intentional property rather than some other intentional property.

This avoids one problem for introspection-based accounts of self-knowledge. Introspection, whether we understand this on the model of inner vision or on the model of a brain scanner, only affords access to the intrinsic properties of thoughts. So if contents are relational properties of thoughts, then we could know about these contents through introspection only if the syntactic, or phenomenological, or other intrinsic properties of thoughts encode the relational properties. But if the second-order thought inherits its content from that of the first, we do not need to infer the relational properties of thoughts from their intrinsic properties. When be-

havior inherits intentional properties from its mental causes, we have no inclination to think that there must be something over and above the behavior that must first scan those mental causes. Your behavior counts as looking for water because thoughts with the relevant content caused it in a certain way. Your second-order belief counts as knowledge because the relevant thought caused it in a certain way.

Giving a complete account of content inheritance requires saying what kind of causation is involved when one state inherits its content from another. As in the case of intentional action, not just any kind of causation will do. We need an account that will distinguish deviant from nondeviant causal chains. I do not have such an account. But thinking about self-knowledge in terms of content inheritance is still useful. Since our self-knowledge is a matter of the causal history of our higher-order beliefs, not of our knowledge of that causal history, once you are in a position to think the first-order thought, there is nothing further you need to do or find out in order to know the content of that first-order thought. Also, since the kind of causation relevant to the second-order thought is mental causation—that is, causation by the first-order thought—any external factors relevant to the determination of content have already done their job.

In the case of direct or noninferential self-knowledge of a conscious thought, the basic account goes like this. The conscious thought causes the second-order belief that the thought occurred. The thought also largely determines the content of that belief. The thought determines the proposition (but not the attitude) ascribed. If we are only concerned with knowledge of content, we can see that a second-order belief that is formed in this way will be correct about that. In virtue of these facts, the thought justifies the belief. For a reliabilist, the belief is justified in virtue of the process that produced it. But the process just is the thought causing and determining the content of the second-order belief. We can know that this process is reliable simply by reflecting on the functional role of conscious thoughts. So if you think that reliability, or known reliability, or known reliability in the absence of defeaters is a sufficient condition for knowledge, you should believe that we can know the contents of our thoughts directly.[6]

[6]Second-order beliefs, even false second-order beliefs, can come about through some other process—for example, through applying a psycholog-

On this account, externalism is clearly compatible with self-knowledge. The fact that the first-order thought determines the content of the second-order belief guarantees the relevant sameness of content. Since the second-order belief inherits its content from the first-order thought, it makes no difference whatsoever what determines the content of the first-order thought. Environmental, social, or neurological factors could play a role here. Only the relation between the first- and second-order thoughts matters to discussions of self-knowledge.

A common theme among many externalist replies to the self-knowledge objection is that just as the environment determines the contents of our first-order thoughts, the environment also determines the contents of our second-order thoughts.[7] I think it is more informative to say that the first-order thought determines the content of the second-order belief. The first way of putting things leaves open the possibility that the environmental determination of the second-order belief is independent of that of the first-order thought. This possibility raises the question of how we know, without investigating the environment, that the environmental determination is the same in both cases. Further, in the case of our knowledge of past thoughts, the distinction does make a difference. If you have a current memory of a past first-order thought, there could be two different environments involved, the past and present environments. To determine which environmental features are relevant, we need to look at the causal history of the current second-order memory. We will return to these issues in sections 3 and 4.

As an account of self-knowledge, this is incomplete, not only in terms of its sketchiness, but in principle. We have here, at most, an account of knowledge of content. We answer the question of

---

ical theory to oneself. But I am not concerned with showing that content externalism is consistent with infallibility with respect to one's own mental states, since we are not infallible. I want to see how we know when we do and whether content externalism is inconsistent with our knowing in this way.

[7]See, for example, Tyler Burge, "Individualism and Self-Knowledge," *Journal of Philosophy* 85 (1988): 654–55; John Heil, "Privileged Access," *Mind* 42 (1988): 238–51; and Kevin Falvey and Joseph Owens, "Externalism, Self-Knowledge, and Skepticism," *Philosophical Review* 103 (1994): 107–37. I discuss the difference between this causal account and Burge's account of self-knowledge in "Externalism and Knowledge of the Attitudes."

how we know *what* we believe, a question about knowledge of content, in terms of content inheritance. We cannot answer the question of how we know that we *believe* something, a question about knowledge of the attitudes, in the same terms. The latter question is more difficult and has received much less attention. An account of our knowledge of the attitudes should specify which features of a thought we use in classifying it in terms of its attitude type, whether they are functional, qualitative, or neurological. Unfortunately, I do not have such an account.

2.

So externalism is compatible with self-knowledge. Or is it? Paul Boghossian presents an argument designed to show that if externalism is true, then under certain circumstances, we do not know what we are thinking unless we investigate the environment.[8] The argument goes essentially like this: Imagine a situation in which my thinking about twin-water is a relevant alternative to my thinking about water. If I were in such a situation, in order for me to know that I am thinking that water is wet, I would have to rule out the possibility that I am thinking that twin-water is wet. But I could only rule out this latter possibility if I knew something about my environment, namely that water rather than twin-water is the dominant causal source of my "water" thoughts, or something of this sort. So, if externalism is true, and if we were in such a situation, we would not know what we are thinking without investigating our environment (Boghossian, 12). Of course, this argument does not show that externalism entails that we do not know the contents of our thoughts. It tries to show that one consequence of externalism is that our knowledge of our own thoughts is more susceptible to empirical contingencies than we may have believed.

What is a situation in which my thinking about twin-water is a relevant alternative to my thinking about water? Suppose that one night when I am sleeping I am transported to Twin Earth. I wake up in a place that looks just like the place where I fell asleep. Since the only difference between Earth and Twin Earth is in the chemical composition of the stuff that flows in streams and comes out

[8]Paul Boghossian, "Content and Self-Knowledge," *Philosophical Topics* 17 (1989): 5–26.

of faucets, and since I do not notice this difference, I have no idea that the switch has occurred. When I first point to some XYZ and say "That's water," what I say is false. My thoughts are still about water, that is, H_2O. But suppose that I stay on Twin Earth for several years. It seems that after sufficient causal contact with XYZ and with members of a language community who use the word 'water' to refer to XYZ, I come to have concepts appropriate to my environment. I come to have the concept of twin-water. Now when I point to some XYZ and say "That's water," what I say is true.

As Boghossian points out, there are two ways of understanding this story. People generally agree that in time I acquire the concept of twin-water. But according to one way of telling the story, I lose the concept of water. According to another way, I end up with access to both concepts. I agree with Boghossian that the second way of telling the story is not only more interesting, it is also more plausible. While it is easy to see how causal contact with a new type of substance can give you a new concept, it is not at all clear how it can take one away. Suppose I say "I remember the first time I went swimming in the ocean as a child. The water was really salty." It seems fairly clear to me that if I went swimming in water as a child, I am now thinking about water, even if I have been switched.

We will return to the issue of the competing interpretations later. But regardless of what we say about this question, even in the imagined case of switching we are right about the contents of our present thoughts. Suppose I say

(6) I believe that there is water in front of me.

If I say this on Earth when my concept of water is operative, then both the first-order thought reported and the second-order thought expressed are about water (H_2O); so my second-order thought is true. If I say it on Twin Earth when my concept of twin-water is operative, then both thoughts are about twin-water (XYZ), and, again, the second-order thought is true. The only way the second-order thought could be false is if the first-order thought involved one content and the second-order thought involved the other. But since the first-order thought determines the content of the second-order thought, it is not possible for the two contents to come apart in this way.

We are right about what we think, but do we *know* what we think?

I am inclined to say that we do know our own thoughts even in this case. According to Boghossian's argument, on the other hand, we do *not* know what we are thinking in these cases unless we investigate our environment. The argument is straightforward. In this situation, I do have to rule out the possibility that I am thinking about twin-water because that is a relevant alternative in this situation. If we assume that I cannot rule out this relevant alternative without investigating my environment, then it follows that in this situation, I do not know what I am thinking unless I investigate the environment (Boghossian, 13).

The problem with the argument is that it is set out in terms of ruling out or excluding certain possibilities. This can give the impression that in order to know that p in a relevant alternative situation, you need to go through a certain process of reasoning by which you rule out the relevant alternative q.[9] According to this picture, noninferential knowledge in a relevant alternative situation is not an option. I think the correct moral of the relevant alternative stories has nothing to do with reasoning. In a knowledge-precluding relevant alternative situation, your true belief that p is just an accident. If things had been just slightly different, you would have had a false belief that p. It is the presence or absence of these counterfactuals, not the presence or absence of reasoning, that is relevant to whether you know.

Let me illustrate. You are sitting by a pond and you see a duck.[10] You are familiar with ducks, the lighting is good, you are paying attention, and so on. You have a justified true belief that there is a duck in front of you. Nevertheless, there are a number of decoy ducks in your vicinity. In order to determine whether your belief counts as knowledge, we need to know the truth of certain counterfactuals. In a knowledge-precluding relevant alternative situation, the following counterfactual is true.

---

[9]Here is how Boghossian concludes the slow argument: "S has to be able to exclude the possibility that his thought involved the concept *arthritis* rather than the concept *tharthritis,* before he can be said to know what his thought is. But this means he has to *reason* his way to a conclusion about his thought" (14; his emphasis).

[10]Both of these cases are modeled on similar cases in Alvin Goldman's "Discrimination and Perceptual Knowledge," *Journal of Philosophy* 73 (1976): 771–91. Goldman discusses the cases in terms of counterfactuals about what you would believe under certain circumstances, not in terms of reasoning.

(P) If a decoy duck had been in front of you, you would have falsely believed that it was a duck.

The contrast case, a knowledge-consistent relevant alternative situation, is one in which you know that there is a duck in front of you, even though there are decoy ducks in your vicinity. How could this be? Well, suppose the relevant decoy ducks were not particularly lifelike. In that case, (P) would be false. In fact, we can strengthen the case so that not only is (P) false, but (C) is true.

(C) If a decoy duck had been in front of you, you would have correctly believed that it was a decoy duck.

In a situation in which (P) is false and (C) is true, it seems that you can know that there is a duck in front of you without doing anything that we would normally call ruling out or excluding the possibility that the thing in front of you is a decoy.

As Boghossian says, the notion of a relevant alternative is an objective notion. If (P) is true under these circumstances, then you do not know. You need not know that (P) is true or have any beliefs about (P) at all for its truth to exclude knowledge. But the objectivity works in the other direction as well. If (P) is false and (C) is true, then under these circumstances you do know. You need not find out that (C) is true or have any beliefs about (C) in order for its truth to guarantee knowledge.

Is the switching case more like the knowledge-precluding situation or the knowledge-consistent one? If the issue is one about the truth or falsity of certain counterfactuals, it seems clear that it is more like the knowledge-consistent situation. Here are the relevant counterfactuals.

(P′) If I had thought about twin-water, I would have falsely believed that I was thinking about water.

(C′) If I had thought about twin-water, I would have correctly believed that I was thinking about twin-water.

Now suppose that first-order conscious thoughts typically cause and determine the content of second-order beliefs to the effect that you are having them. If I have a first-order thought about water, this will produce the correct second-order belief that I am thinking

about water. If I have a thought about twin-water, this will produce the correct second-order belief to that effect. But there is no reason to think that I would also have some other second-order belief that gets the content wrong. So (P′) is false and (C′) is true. This makes the switching case a knowledge-consistent situation.[11]

Intuitively, the difference between the switching case and the knowledge-precluding cases is that it is just not an accident that you are right about your own thoughts. It's just not the case that if things had been slightly different, you would have been wrong. If things had been slightly different, you would have had a different, but still correct, second-order belief. We get the contents of our thoughts right in the switching case not by chance but because there is a systematic connection between the first-order thoughts and the second-order beliefs that are candidates for knowledge. It is this connection, not the mere fact of getting it right, that makes the beliefs knowledge.

3.

So far, we have primarily discussed first- and second-order thoughts that occur at the same time. What about our knowledge of our past thoughts? Boghossian argues that if externalism is true, we do not know our past thoughts. He goes on to argue that if we do not know our past thoughts now, we could not have known them in the first place. As he says, "It is not as if thoughts with widely individuated contents might be easily known but difficult to remember" (Boghossian, 23). I will argue that under certain extreme

[11]Falvey and Owens present a similar objection to Boghossian's argument. We agree that the truth values of the counterfactuals are what is most important and that these truth values come out the way the externalist wants them to. With my account of self-knowledge, I think I can give an explanation of why the relevant counterfactual is true. And I can reply to a question they raise but do not discuss. "How can it be that the subject is always right about the contents of her beliefs, despite the fact that the introspectible evidence in her possession underdetermines their contents?" (Falvey and Owens, 118). If you think of the evidence for p on the model of premises in an argument for p, then on my account, evidence in this sense is not necessary for self-knowledge. For one thing, you must *believe* all of the premises in an argument in order for them to justify the conclusion. But self-knowledge is not like this. A hope, fear, or doubt can cause, determine the content of, and justify a relevant second-order belief.

conditions, there is a sense in which these thoughts *are* difficult to remember.

Boghossian's argument from a later lack of knowledge goes like this (Boghossian, 22–23). Suppose that after I am on Twin Earth long enough to acquire the concept of twin-water, you tell me that the switch has taken place, but not when this switch occurred. Now you ask me, "Last year, were you thinking about water or twin-water?" There is a clear sense in which I do not know the answer to this question. But, the argument continues, if I do not know now what I was thinking then, there are two possibilities. Either I have forgotten something I once knew, or I never knew it. But *straightforward* memory failures are extraneous to the discussion, and we can exclude them by stipulation.

If you know something at one time and fail to know it at another, we need some explanation for this change in your cognitive state. Memory failure and the acquisition of misleading information are possible explanations.[12] But neither of these possibilities explains what happens in the switching case. Fortunately, there is a further possibility. A change in your conceptual repertoire, however this comes about, is another possible explanation for a loss of knowledge. Knowing that p requires being able to think that p. Since a process of conceptual revision can take away this ability, it can take away knowledge as well. In order to see when and where this conceptual revision takes place, we need to look more closely at the semantics of switching.

As I have said, there are two interpretations of the switching story. One interpretation allows that I have access to both contents after the switch. The other interpretation does not allow this. On the second version of the story, the process of acquiring the concept of twin-water is a process of *replacing* one concept for another. I will focus primarily on the view I take to be more plausible, the interpretation that allows access to both contents. I will begin with some justification and discussion of this interpretation and then proceed to the discussion of our knowledge of past thoughts in two stages. I will first discuss the situation on Twin Earth before

[12]For the second possibility, see Carl Ginet, "Knowing Less by Knowing More," in *Midwest Studies in Philosophy*, vol.5, ed. French et al. (Minneapolis: Minnesota University Press, 1980), 151–61.

you inform me of the switch, and then, in the next section, see what happens when you tell me about the switch.

According to the usual externalist intuitions, it is possible for two state tokens of distinct individuals to differ in content even though those tokens are intrinsically, functionally, qualitatively, and syntactically indistinguishable. Furthermore, according to both interpretations of the switching stories, it is possible for two intrinsically indistinguishable states of the same person at different times to differ in content. In each case, a difference in the causal or historical features of the token states explains the difference in content. Why would an externalist believe that no two intrinsically indistinguishable states of the same individual at roughly the same time could differ in content? Presumably, the only reason to believe this is that you believe, for some reason, that no two such states can differ in the relevant causal or historical features. For if this pair differed in the relevant causal features, then they, like the other pairs, would differ in content.

Is it possible for two intrinsically indistinguishable states of the same individual at roughly the same time to differ in the relevant causal or historical features? Consider a case of *de re* conflation of individuals. Suppose that you have two look-a-like cousins, from different sides of the family and both named "Vinnie." You have met them and know what they look like, but you have never seen them together. You believe that you have only one cousin named "Vinnie," and you have no view about what side of the family he's from. To use a popular metaphor, you have one file where you store information (and misinformation) about both Vinnies. Since the file metaphor is presumably to be cashed out in functional or conceptual role terms, this comes to the idea that, everything else being equal, your thoughts about one Vinnie have the same functional role as your thoughts about the other.

Clearly, different states in your "Vinnie"-file have different causal histories. Some of them are beliefs caused by perceptions of the Vinnie on your mother's side, and some are beliefs caused by perceptions of the Vinnie on your father's side. Are these differences in causal features semantically relevant? Perhaps all of your "Vinnie"-thoughts refer to whichever individual is the dominant causal source of the information in that file.[13] Perhaps there is just no

[13]Gareth Evans, "The Causal Theory of Names," *Aristotelian Society*, supp. vol. 47 (1975): 187–208.

fact of the matter about which individual any of the thoughts in that file are about. Perhaps every thought in the file is about both Vinnies.[14] If any of these views are correct, then this difference in causal history does not make a semantic difference. I think that there are intuitively clear cases in which two state tokens with the same functional role at roughly the same time do refer to distinct individuals.

There are at least some cases where you can determinately think about and refer to one individual despite the conflation. If you are three feet away from one of the Vinnies while the other is nowhere nearby and you say, "Hey Vinnie, what time is it?" you single out an individual in language and thought. If, five minutes later, someone asks you what time it is, and you say, "Vinnie just said it was around ten," the intuition is still very strong that you are referring to just one person. Memory preserves reference.

So, consider the following two cases. Suppose you are at dinner with a number of people from your mother's side of the family, discussing your Aunt Clara's wedding. No one in the conversation (with the possible exception of you) has thought about or mentioned the Vinnie from your father's side of the family, but you discuss the other Vinnie a great deal. You were at the wedding and you remember seeing your cousin Vinnie (from your mother's side) dancing at the reception. You consciously judge, on the basis of this memory, that Vinnie is a good dancer. In an attempt to express this belief, you say, "Vinnie sure can dance." To whom are you referring? The conversation, your memories, the causal ancestry of the belief, your desire to be relevant, and the natural interpretation of your audience all point to one man, your mother's sister's son.

At roughly the same time, the next week, the next day, or later that evening, after having completely forgotten the conversation about Clara's wedding, you are discussing a party at your uncle Tony's with your father's side of the family. No one (with the possible exception of you) has mentioned the other Vinnie, but your father's side Vinnie has come up in conversation. You were at the party, and you remember seeing the Vinnie from your father's side,

[14]Igal Kvart, "Divided Reference," *Midwest Studies in Philosophy*, vol. 14, ed. French et al. (Notre Dame: University of Notre Dame Press, 1989), 140–79.

who also happens to enjoy dancing. On the basis of this (qualitatively identical, if you like) memory, you judge that Vinnie is a good dancer. Once again, you express this belief with the words, "Vinnie sure can dance." In this case, everything points to the Vinnie on your father's side.[15]

Here we have two utterances of the same sentence that refer to distinct individuals. Since the beliefs have the same functional role, we cannot account for the difference in reference in those terms. According to Donnellan's discussion of this sort of case,[16] to find out the referent on a particular occasion, we need to ask about the point of the utterance. But to ask about the point is to ask about the reasons for which the sentence was uttered. And to ask about the reasons for which someone does something is to ask about the beliefs, desires, and intentions that caused it. Asking about the point is one way of asking about the causal history. If two intrinsically indistinguishable states are produced for different reasons and these reasons determine the reference or content of the states, the states will differ in reference or content.[17]

So, on this interpretation of the switching story, do I know my past thoughts? After the switch, but before I learn about the switch, I say

(7) Last year I thought that there was water in front of me.

---

[15]Many thoughts in these conflation cases will not be determinately about one candidate or the other. For example, you may believe that your cousin Vinnie is a lawyer and base this belief on some evidence that is determinately about one cousin and some which is about the other. Hartry Field has developed a semantics for referentially indeterminate expressions that extends quite naturally to this sort of case. See "Theory Change and the Indeterminacy of Reference," *Journal of Philosophy* 70 (1973): 462–81, and "Quine and the Correspondence Theory," *Philosophical Review* 83 (1974): 200–28.

[16]"Proper Names and Identifying Descriptions," in *Semantics of Natural Language,* ed. Davidson and Harman (Dordrecht: D. Reidel, 1972), 356–79.

[17]In order to move from this discussion of reference back to the level of sense, or content, where we began, all we need is the Fregean view that sense determines reference: if x and y have the same sense or content, then they have the same reference or extension with respect to the same possible world and time. It follows easily enough from this that if two of your "Vinnie"-thoughts have different referents or extensions (with respect to the same world and time), they have different contents, despite the fact that the members of each pair are intrinsically indistinguishable.

This expresses a second-order belief about a past conscious thought. According to the usual externalist intuitions, the substance causally responsible for the utterance determines the content of the second-order belief. Now, (7) will be false if twin-water is causally responsible *in the content-determining way* for the utterance of and second-order belief expressed by (7) while twin-water was not responsible for any first-order belief of mine last year. Suppose that last year I thought about water and had not been to Twin Earth yet. So the first-order thought that (7) reports is about water. But surely the fact that I utter (7) on Twin Earth does not automatically guarantee that the second-order belief is about twin-water. We need to know where that belief came from, that is, the reasons that produced it.

Well, last year, I had a conscious thought that there was some water in front of me. This thought caused a second-order belief to the effect that I had that thought. This second-order belief inherited its content from its first-order cause and so involved the concept of water. It seems that (7) is an expression of this very same dispositional belief. If going to Twin Earth does not deprive me of any concepts I once had, there is no reason to think that the second-order dispositional belief changes its content. So if (7) is an expression of the most common sort of second-order belief, then it is about water, and I do know what I thought last year even after the switch.

In discussing dispositional beliefs in general and memories in particular, I am assuming that there is a difference between what is stored in memory and what is inferred, on a particular occasion, from what is stored. For example, it is possible that you have your birth date stored in memory but do not have the season of your birth stored. Since the state that represents your birthday is usually not conscious (that is, not until the question arises), we say that you dispositionally believe that you were born on such and such a day. Since the inference from the date of your birth to the season of your birth is so simple and need not even be available to consciousness, we say that you dispositionally believe that you were born in such and such a season.

Though we call each of these attitudes toward the two propositions dispositional beliefs, the difference between them is important. Since the inferential process involved in the second sort of dispositional belief need not be conscious, the distinction is not

introspectively discernible.[18] But given the distinction, we can describe the switching situation like this: Merely switching to Twin Earth does not change the content of what is stored in memory. The content of what is stored is determined by the causal history of the stored state. The content of a state that is inferred (consciously or otherwise) exclusively from stored states is determined by the content of the stored states. The content of a state inferred partly from stored states and partly from current information is determined by the content of both sorts of states, and so on.

So if (7) is an expression of a belief stored in memory, that belief counts as knowledge. Of course, the belief expressed by (7) could have a different source. I may infer this belief from other thoughts that only involve the notion of twin-water. So I may think that there is some twin-water in front of me, and believe that I thought the same thing last year, and utter (7) on the basis of an inference from these beliefs. In this case, the second-order thought would also be about twin-water, and so it would be false. But the kind of self-knowledge that we are interested in saving is *direct* self-knowledge. Evidential knowledge is subject to the limitations of the evidence on which it is based. In the case described above, I make a mistake about *sameness* of content. But surely I can know that x is water and that y is H_2O without knowing that x is y or that water is H_2O.[19]

Given the sameness of functional role, I will take my beliefs about water as evidence for and against my beliefs about twin-water, and vice versa. This may lead to false beliefs about the world, but it does not by itself lead to misidentification of thoughts. Consider the following case.[20] Suppose that while on Twin Earth, I see some twin-water with a purplish glint. This leads me to say, "I was wrong last year to think that water never has a purplish glint." This is equivalent to the following

---

[18]I assume that there is nothing peculiar or even particularly Freudian about the notion of a nonconscious inferential process. According to most cognitive scientists, your visual ability to detect edges depends on such a process.

[19]Falvey and Owens also distinguish the claim that we have introspective knowledge of content from the claim that we have introspective knowledge of sameness and difference of content. They argue that the latter claim is implausible independently of externalism.

[20]I owe this example to an anonymous reader for the *Philosophical Review.*

(8) Last year, I thought water never has a purplish glint, and that belief was false.

We do not have to worry about the second conjunct, since this involves knowledge of truth value. Only the first conjunct expresses a possible item of knowledge of content. Presumably, there is a fact of the matter as to where this second-order belief came from. If this is an expression of a belief stored in memory, the belief is noninferential and counts as knowledge. If, on the other hand, I infer the second-order belief from a (possibly nonconscious) belief about sameness of content, it may well be false. But this failure of self-knowledge is a result of the inferential nature of the second-order belief.

On the other interpretation of the switching story, there seems to be no way of saving knowledge of past thoughts. On the interpretation according to which one content replaces another, after the switch, all of my second-order beliefs will involve the concept of twin-water. Any of these beliefs that are about thoughts that occurred on Earth before the switch will be false. Perhaps this is a further reason to prefer the first interpretation.

4.

I take it as shown that after the switch, but before I learn about the switch, I do know my past thoughts in an authoritative way, at least as long as moving to Twin Earth does not deprive me of any concepts I had when I got there. But what happens when you inform me of the switch? I learn that two different substances have been causally responsible for my utterances of 'water' and that my utterances of sentences involving the word 'water' have expressed different propositions at different times. Now you want to ask me the following question.

(9) Last year, did you think about water or twin-water?

In order for me to understand this question, I must understand the terms 'water' and 'twin-water', and I must understand them in such a way that I designate, or at least intend to designate, distinct substances with each.

In fact, it seems that just informing me of the switch alters my

situation in this way. After I'm informed of the switch, my thoughts about water have a different functional role from my thoughts about twin-water. For example, I do not take my beliefs about water as evidence for and against my beliefs about twin-water. But before I'm so informed, these types of thought did have the same functional role. So the functional role changed for at least one type of thought. But given the similarity of cases, there are no grounds for choosing which type of thought remained functionally unchanged. So the functional role of both types must have changed.

But what about (9)—do I know the answer? Of course, (9) is ambiguous in a familiar way. We use 'about' in an extensional and in an intensional sense. We might say that the sentence 'Water is wet' is about, or refers to, the substance that covers three-quarters of the Earth's surface. But clearly, if we use 'about' in this extensional sense, I can know that I think that water is wet without knowing that my thought is about the F, where 'the F' is any description that differs in content from 'water'. Knowledge of this latter fact clearly involves knowledge of the external world.

So 'about' in (9) must be used in an intensional sense. We must be asking about notions or concepts. But did my thought last year involve either of the notions—*water* and *twin-water*—involved in (9)? The answer to this depends on how we individuate "notions" or "concepts." So let's review the facts. The following should help keep things organized.

t_1 I think that there is water [C1] in front of me
- I am switched to Twin Earth
- I acquire a concept of twin-water

t_2 I believe I thought that there was water [C2] in front of me
- I find out about the switch

t_3 At t_1 did I think about water [C3] or about twin-water [C4]?

Now suppose that C1 is the concept I express with the word 'water' at t_1, C2 is the concept expressed at t_2, and so on. I am switched between t_1 and t_2, and enough time goes by after the switch for me to acquire the concept of twin-water by t_2. I have argued that at t_2, I am right about what I thought at t_1 because C1 is C2. Not only do C1 and C2 both pick out water (H_2O), they have the same

cognitive significance or functional role, whatever exactly this turns out to be. But, on or about t_2, I could have used the word 'water' to express a concept distinct from C2, one that picked out or referred to XYZ. This concept might have the same functional role as C2, but it picks out a distinct substance. And this, for the externalist, is a sufficient condition for the concepts to be distinct.

So when you ask me what I was thinking about last year, we must read the 'about' intensionally. This must be a question about the concepts involved. Here is one way to reformulate the question.

(9′) Is C1 C3, or is it C4?

As we have seen, the functional role of C1 is not the same as that of either C3 or C4. So on any conception of concepts that takes functional role to be relevant to their individuation, C1 is neither C3 nor C4. But can an externalist hold that functional role is relevant to the individuation of concepts? Externalism is the view that the propositional attitudes of an individual do not supervene on the intrinsic properties of that individual. This does not mean that intrinsic properties are not relevant. It simply means that those properties are not sufficient to determine content.

So it is consistent with externalism to require some similarity of functional role as a necessary condition for sameness of content. But similarity in what respects? It clearly goes against the spirit of much of the externalist literature to say that any time you learn something new about an individual or kind you end up with a different concept. But overall similarity of functional role is not the decisive feature in the individuation of concepts. Consider the overall functional similarity between C3 and C4. These are the concepts of water and twin-water I have after I learn about the switch. From an application of each concept I am disposed to infer *clear, potable, covers three-quarters of the planet,* etc. These two concepts may be more functionally similar than two different individuals' concepts of water are. But one difference outweighs all of the similarities. I believe that the kinds picked out by the concepts are distinct. In functional terms, the concepts do not inferentially interact in the right way for C3 to be C4. Thoughts involving one concept do not count as evidence for and against thoughts involving the other without the mediation of other thoughts.

Perhaps an example will help. Suppose that on Monday, you

believe that water never comes in bottles. On Tuesday, you go to the store and see bottled water for the first time. There are at least two very different possibilities to consider. In the most obvious case, you simply change your mind. You used to believe p. Now you believe not-p. The content of your concept of water does not change even though the functional role of that concept does. Contrast that case with the following possibility. You might think that 'water' is ambiguous and so introduce a distinct concept that, as it happens, also refers to water. In this case, the belief you acquire on Tuesday does not count as evidence against the belief you had on Monday, which, we can assume, is still stored in memory. This fact alone about the functional roles of the two tokens is enough for us to say that the two tokens differ in content. But the claim that direct inferential interaction is a necessary condition for sameness of content does not commit us to the holistic conclusion that any change in functional role involves a change in content. When you change your mind, the necessary condition is met despite the change in functional role.[21]

Anyone who thinks that proper names are directly referential but that someone might believe that Cicero was bald without believing that Tully was bald must give some account of the behavior of proper names in propositional attitude contexts.[22] But any such account that did not also apply to predicates, especially natural kind terms, would be inadequate. If you fail to believe that Cicero is Tully, then you will not take your beliefs about Cicero as evidence for and against your beliefs about Tully. If I fail to believe that something is of *that* kind [thinking of the kind I called "water"

[21]Of course, any two beliefs can inferentially interact. Your belief that there are brick houses on Elm Street might lead you to believe that Jones is a liar. But if this is a case of inference rather than association, you need mediating beliefs. For example, you might believe that Jones said that there are no brick houses on Elm Street and that Jones would not be mistaken about this. But the addition of mediating beliefs must come to an end somewhere on pain of a regress. Where it comes to an end, you have *direct* inferential interaction.

[22]If you believe that your concept of Cicero is distinct from your concept of Tully when you do not know that Cicero is Tully, then we have another case where gaining information involves losing a concept. If, after being informed of the identity, you express one concept with both terms, then the new concept cannot be identical to both of the old ones. But we can easily devise cases where there are no grounds for choosing among the two. So the new concept must be distinct from both.

last year] just in case it is of *this* kind [thinking of the kind before me now], I am in the analogous situation with respect to kinds or types. Since I do not know when I was switched, I fail to believe the relevant biconditional. This is exactly the sort of situation where we take differences in functional role as relevant to the individuation of concepts for the intrapersonal case. So, many externalists already know that they need an account that will distinguish between C1 and C3 or C4. If problems of self-knowledge reduce to this problem, then self-knowledge does not pose an independent threat to externalism.

So, if C1 is neither C3 nor C4, then I do know the answer to (9): *neither*. To the extent that we think that my utterances must be about one of the two substances, we are using 'about' in the extensional sense. Clearly, my utterances must have referred to either water or twin-water. But this knowledge of reference involves knowledge of the external world. If (9) is really a question about how I conceived of things last year, I did not think about water in either of the ways involved in (9). The difference between the concepts involved in the question rests on a distinction that I did not make. Nevertheless, it follows from this solution that I do not know what I was thinking last year. Since at t_3 I employ concepts distinct from those I employ at t_1, I cannot say or think what I was thinking then. But if I cannot think it, I cannot know it. Still, even if we do take a fine-grained approach to the individuation of concepts, we do know our present thoughts, and for familiar reasons. Since the first-order thought determines the content of the second-order belief, it makes no difference whatsoever what determines the content of that first-order thought.

So let's return to Boghossian's argument from a later lack of knowledge. According to this argument, if at t_3 I do not know what I was thinking at t_1, then I could not have known at t_1 what I was thinking then. The argument depends on the claim that we can exclude memory failure from the discussion by stipulation. Now I grant that there is a sense in which at t_3, I do not know what I was thinking at t_1. But while straightforward memory failures are extraneous to the discussion, some cases where you are unable to remember something are relevant. According to the more common interpretation of switching stories, the process of acquiring the concept of twin-water is a process of replacement. So at t_2, I am no longer able to think thoughts about water (H_2O). But if I

cannot think about water, I cannot think that I thought that there was water in front of me. In a case of conceptual change, you cannot think your past thoughts. But if you cannot think it, you cannot remember it. So cases of conceptual change are cases where you are unable to remember something, but they are clearly not irrelevant to the discussion, and we cannot exclude them by stipulation. On the other interpretation of the switching stories, moving to Twin Earth by itself does not deprive me of any concepts. It just gives me access to another. But when you tell me about the switch, I must introduce two new concepts into my repertoire. This does seem like a conceptual change of the relevant sort. I no longer have access to my past thoughts because of the conceptual change.

So, on either interpretation of the switching story, it is possible to lack knowledge of your past thoughts because of the conceptual change. We have seen that this does not threaten our authoritative knowledge of our present thoughts. But given that we do not know about our past thoughts in the hypothetical situation, what does this say about our *actual* knowledge of our past thoughts? On the assumption that being switched to Twin Earth is not a relevant alternative, according to most contemporary accounts, the fact that we would not know our past thoughts is consistent with the fact that we do know those thoughts. Furthermore, according to the preferred interpretation of the switching story, you only lose knowledge when you become convinced that you have been switched to Twin Earth. Since we know our present thoughts, we know that we do not believe that we have been switched. So we can rule out this alternative, relevant or not.[23]

New York University

[23]I would like to thank Paul Boghossian, Jaegwon Kim, Roy Sorensen, Ernie Sosa, Ed Stein, Peter Unger, Ed Witherspoon, and two anonymous readers for the *Philosophical Review* for comments on earlier versions of this paper.