Do we know how we know our own minds yet?*

PIERRE JACOB

Introduction: the incompatibilist argument

One feature of the contemporary philosophical situation is puzzling. On the one hand, few if any of the features of the special epistemic authority granted by both the traditional empiricist and the traditional rationalist pictures of introspective self-knowledge have survived recent philosophical scrutiny. On the other hand, several philosophers—the incompatibilists—assume that the alleged special epistemic authority granted to introspective self-knowledge by traditional epistemology can bear the burden of an argument against content externalism. In response, several externalists have tried to argue for the compatibility between content externalism and the alleged special epistemic authority of introspective self-knowledge.[1]

Content externalism is the view that the content of an individual's thought, propositional attitude and perceptual experience does not (always) supervene only upon the internal cognitive resources of the individual.[2] Nor does it (always) supervene only upon the internal physical, chemical and biological properties of the individual's brain. Content externalism comes in two broad varieties: social and non-social externalism. According to the latter, the content of an individual's mental representation may depend upon the individual's non-social environment. According to the former, it may also depend upon what other members of her community think. Notice that social externalism seems tailor-made for the conceptual contents of an individual's thoughts and propositional attitudes, not for the nonconceptual contents of the individual's perceptual experiences.[3]

In a nutshell, the argument put forth for the incompatibility between content externalism and the special epistemic authority of introspective self-knowledge has the general following structure (see e.g., McKinsey, 1991 and Boghossian, 1997). If introspective self-knowledge has special epistemic authority, then content externalism cannot

---

* I am very grateful to Fred Dretske for extensive email exchanges about his views on self-knowledge. I also wish to thank Gabriele Usberti for extensive and penetrating comments on this paper and Max Kistler for a useful conversation.

1 See e.g., Burge (1988) and Davidson (1987).

2 I simply assume without argument both a representational view of the mind and the distinction between the conceptual content of thoughts and propositional attitudes and the nonconceptual content of perceptual experiences.

3 What an individual thinks and believes may depend on what members of his community think and believe. But I assume that what an individual experiences does not depend on what members of his community think, believe or experience.

be true. Introspective self-knowledge has special epistemic authority. Therefore: content externalism cannot be true.

In a little more detail, the incompatibilist argument assumes that one can know with special epistemic authority that one believes that, for example, water is a liquid. Let us say that one knows a priori that one believes that water is a liquid. But one could not believe that water is a liquid unless one had the concept WATER.[4] It follows that one knows a priori that one has the concept WATER. According to content externalism, however, one could not have the concept WATER unless one stood in some appropriate relation to water. It follows that one can know a priori that one stands in relation to water and thus that there is water in one's environment. But this seems incredible: one cannot know a priori that (or have special epistemic authority over whether) one's environment contains water, which is, according to content externalism, necessary for having the concept WATER. Whether one's environment contains water (not something else) is not something one can know a priori.

In summary, two assumptions seem needed for the incompatibilist conclusion that content externalism cannot apply either to the concept WATER or to the belief that water is a liquid. First of all, one must know a priori with special epistemic authority that one believes that water is a liquid. Second of all, the concept WATER must make the same contribution to the simpler content (or truth-conditions) of one's first-order belief that water is a liquid and to the more complex content (or truth-conditions) of one's introspective higher-order belief that one believes that water is a liquid.[5]

The reason I find the contemporary situation perplexing is that I take externalism about the contents of an individual's first-order thoughts about the world to be more plausible—not less plausible—than anything we may think about introspective self-knowledge. On the one hand, content externalism—at least non-social externalism—is a doctrine about first-order human mental representations of the external world. It is a view about the processes—some of which may be common to humans and to non-human animals, e.g., perception and memory—which allow humans to achieve some knowledge of the external world. On the other hand, introspective self-knowledge consists in higher-order representations about first-order mental representations of the world. It is at least conceivable that a creature might have the cognitive resources required for forming reliable beliefs about the external world, even though it lacks the cognitive resources for forming introspective beliefs about its own beliefs about the world.[6] Imposing top down constraints on the *contents* of first-order mental representations about the world from assumptions about the alleged epistemic status of introspective self-knowledge sounds to me like putting the cart before the horse.[7]

---

4 I use words in capital letters to refer to concepts.

5 I shall come back to the second assumption in the conclusion.

6 In order not to beg any question in favor of *epistemological* externalism and against epistemological internalism, I purposefully put the last point in terms of reliable beliefs, not knowledge, of the external world. An epistemological externalist might want to claim, and an epistemological internalist might want to deny, that this is sufficient for knowledge of the external world.

7 It sounds preposterous to impose internalist epistemological constraints, not on knowledge of the world, but on the *contents* of beliefs about the world.

The idea that introspective beliefs about facts involving one's own psychological properties enjoy a unique epistemic authority or privilege has played a different role in traditional rationalist epistemology and in traditional empiricist epistemology. In rationalist epistemology, the primary target of introspective self-knowledge are thoughts. In empiricist epistemology, the primary target of introspective self-knowledge are sense-data or perceptual experiences. Whether one and the same mechanism—introspection—could satisfy both rationalist and empiricist desiderata is far from clear.

On the one hand, in rationalist epistemology, psychological self-knowledge was taken to be the paradigm of both a priori and infallible human knowledge. Rationalist epistemology has three ingredients. First, it is of the essence of the Cartesian mind both that it entertains or forms thoughts (as opposed to having e.g., perceptual experiences). Secondly, thoughts have concepts or ideas as constituents. Thirdly, the mind is transparent to itself: one cannot have a thought of which one is not aware. In rationalist epistemology, what secures the a priority and infallibility of one's introspective awareness of one's psychological properties is that all the psychological properties a mind can exemplify are properties of thoughts or judgments, not experiences. Given that the mind is transparent to itself or that thoughts are reflexive in the sense that one cannot entertain a thought (or make a judgment) unless one is aware that one is, it follows that introspective knowledge of one's own mind is a priori and infallible.

In empiricist epistemology, on the other hand, the most primitive and elementary constituents of minds are perceptual experiences or sense-data, not concepts. Concepts (of either psychological or non-psychological properties) are logical constructions out of sense-data. According to much traditional empiricist epistemology from Locke to Russell, knowledge of the external world—knowledge of mind-independent facts—is twice dependent on psychological self-knowledge. First of all, knowledge of mind-independent facts depends on the epistemologically antecedent knowledge of mental or psychological facts about oneself (such as that one is having a particular perceptual experience, sense-datum or idea). Secondly, one's knowledge of mental or psychological facts about oneself derives in turn from one's direct quasi-perceptual acquaintance with some mental entity present to or in one's mind, i.e., the sense-datum or perceptual experience. If one's awareness of one's sense-data consists in being acquainted with them, then one is made aware of one's sense-data by some kind of quasi-perceptual process or peering inside at one's own perceptual experiences.

Both the rationalist and the empiricist pictures of introspective knowledge have come under serious criticism in contemporary philosophy. On the one hand, the empiricist model of a quasi-perceptual process whereby one becomes self-aware of one's own perceptual experiences raises at least three issues. First of all, if one's knowledge of psychological facts about oneself derives from some quasi-perceptual acquaintance with one's own perceptual experiences, then it is questionable whether self-knowledge can still meet the epistemic requirements of a priority and infallibility. Secondly, as Shoemaker (1994) and other philosophers have noticed, whereas vision, audition, olfaction, touch and proprioception can be used to get information about mind-independent facts (involving one's own body and the bodies of others), no inner sense organ provides information about one's own perceptual experiences, let alone about one's thoughts. Finally, as Harman (1990), Tye (1992) and many others have noticed, perceptual ex-

periences are introspectively transparent. In other words, the phenomenology of the introspection of e.g., one's own visual experience of e.g., a bush of blue lavender just is the phenomenology of one's visual experience of a bush of blue lavender. What it is like to introspect and to be aware of one's own visual experience of a bush of blue lavender is nothing but what it is like to have the visual experience of a bush of blue lavender. Presumably, if introspection of one's visual experience of anything involved some quasi-perceptual process, then introspective awareness of one's visual experience would have a phenomenology of its own—in addition to and above that of the visual experience itself.

On the other hand, the Cartesian picture of introspection raises at least two issues. First, the asymmetry between first-person and third-person mindreading that results from a Cartesian picture of introspective self-knowledge raises a genuine puzzle. As Davidson (1984, 1987) recognizes, the asymmetry between first-person and third-person mindreading takes it for granted both that claims to know one's own mind are made independently of any empirical evidence and that they enjoy an epistemic authority of which third-person claims to know the minds of others are deprived. The puzzle is: why should claims *without* evidential support have more epistemic authority than claims based on evidence? The second question is: given that the Cartesian assumption that the mind is transparent to itself has come under heavy attack, what is left of the Cartesian picture of the special epistemic authority of introspective beliefs? Since Freud, it is commonly accepted that one may be blind to some of one's own beliefs and desires. Furthermore, a human mind does not merely entertain thoughts and propositional attitudes; it also has perceptual experiences. On the Cartesian picture, the immunity to error of the mind's introspective beliefs about itself was secured by the joint assumptions that it is of the essence of the mind to entertain thoughts and that one cannot entertain a thought unless one knows that one is doing so. But how could such assumptions entail that one's introspective beliefs are both exhaustive and infallible? How could such assumptions secure infallible introspective beliefs about one's perceptual experiences at all? How could they secure infallible introspective beliefs about each of one's propositional attitudes—both one's occurrent propositional attitudes and one's dispositional propositional attitudes?[8]

In the first section of the paper, I sketch Fred Dretske's (1995) view of a restricted subset of the set of one's introspective beliefs (i.e., one's introspective beliefs about one's own perceptual experiences) based on the model of displaced perceptual knowledge. In the second section of the paper, I examine the question whether the naturalistic view of the contents of first-order mental representations on which it is based has the resources to accommodate the contents of introspective metarepresentations of one's first-order mental representations. Finally, in the third and last section of the paper, I argue that it is a mistake for an externalist to accept wholesale the premisses of the incompatibilist argument and to try to accommodate externalism to these premisses.

---

8 These points are made by Boghossian (1989).

1. The displaced perception model of introspective self-knowledge

In this section, I will sketch what is to my mind a very plausible externalist account of the introspective process whereby one comes to form introspective beliefs about one's own perceptual experiences, i.e., Dretske's (1995) displaced perception model of introspection. What I call a little misleadingly 'the displaced perception model of introspective beliefs' has really two ingredients: the theory of displaced perceptual knowledge proper and the general informationally based teleosemantic (henceforth, IBT) account of the contents of first-order mental representations of the world (from Dretske, 1988, 1995), which I will presently sketch very briefly.

According to IBT, no system can represent anything unless it has a function (a design or a purpose), i.e., the function to indicate or carry information about the presence of some property, e.g., property *F*. A system could not indicate the presence of property *F* unless it were correlated with property *F*. Carrying information about property *F*, however, is necessary but not sufficient for representing property *F*. Unless a system has the function to carry information about *F*, it cannot misrepresent, and hence represent *F*: if a system has the function to carry information about *F*, then it can represent someting as *F* even though it fails to carry information about *F* because what it represents as *F* may fail to be *F*.

Importantly, the IBT account of mental content entails a principle, which I shall dubb "the principle of the reflexivity of content", and which can be formulated thus (see Dretske, 1995: 52):

Reflexivity of content:
A system cannot represent things to be *F* without carrying information about it's being *F* that it is representing.

System *S* cannot represent something to be *F* unless it has the function to carry information about *F*. If *S* represents correctly something to be *F*, then it does carry information about *F*. Suppose now that *S* represents incorrectly *x* to be *F*. Since it is *S*'s *function* to covary with *F*, in misrepresenting *x* as *F*, *S* is correlated with the property that *would* be instantiated *were S* performing its function according to its design. This property is no other than *F*. Thus, even if *S* incorrectly represents *x* as *F*, still *S* carries information about property *F*, i.e., the property that would be instantiated if *S* were doing its job properly. It follows that by representing something (whether correctly or incorrectly) to be *F*, a device has available information about it's being *F* that it is representing. The principle of the reflexivity of content is, as we shall see in section 3, an important step in Dretske's (2003) reply to the incompatibilist argument. Here, I just note that the principle of the reflexivity of content entails that a representational device has available information about which property it is representing. Now for a device to have such available *information* is not the same thing as *knowing* that it is representing something as *F*. A representational device has available information about the content of its representation, but it does not thereby know what it is doing. To know the latter, it must be able to represent the fact that what it is doing is representing. Unless it has the concept REPRESENTATION, a device cannot represent the fact that what it is doing is representing.

I now turn to displaced perceptual knowledge. 'Displaced perception' is Dretske's (1995) word for what Dretske (1969) called 'secondary epistemic perception'. I shall give a few examples. You hear the dog bark. You thereby come to believe that the dog barks. The dog would not bark unless there was someone at the door. You believe that the dog would not bark unless there was someone at the door. You thereby come to believe that someone is at the door. By hearing the dog bark, you thereby hear that someone is at the door. You see hoof prints in the snow at *t*. You thereby come to believe that there are hoof prints with a particular shape in the snow at *t*. There would not be hoof prints with such a particular shape in the snow at *t* unless a horse had walked on the snow at $t - 1$. You believe that there would not be hoof prints with such a particular shape in the snow at *t* unless a horse had walked on the snow at $t - 1$. You thereby come to believe that a horse walked on the snow at $t - 1$. Although you did not see the horse at $t - 1$, by seeing the hoof prints at *t*, you see that a horse walked on the snow at $t - 1$.

One can represent the general structure of displaced perception in the following sequence of steps. *S* has displaced perceptual knowledge of the fact that object o' is *G* iff

(1) o is *F* (intermediate fact).
(2) *S* has a perceptual experience of o's being *F*.
(3) *S* believes that o is *F* (intermediate belief).
(4) o would not be *F* unless o' were *G* (correlation between facts).
(5) *S* believes (4) (connecting belief).
(6) *S* believes that o' is *G* (from (3) and (5)).

In order to extend the model of displaced perceptual knowledge to introspective knowledge, I shall introduce some of Dretske's own terminology. Step (2) results from step (1) as a matter of perceptual psychology. With Dretske (1969, 1978), we may call nonepistemic or simple perception step (2) and primary epistemic perception step (3). Only a creature with some concept of property *F* could move from step (2) to step (3). In Dretske's (1995) terminology, the fact that o is *F* is the intermediate fact. *S*'s belief that o is *F* is *S*'s intermediate belief. (4) is a correlation between two distinct facts and (5) is *S*'s connecting belief. The fact that *S* comes to believe via displaced perception—the fact that o' is *G*—is the target fact.

Consider now the application of the model of displaced perceptual knowledge to introspective knowledge. Suppose that one has the visual experience of a triangular object: one has the visual experience of o as triangular. *S* has introspective knowledge that *S* has the visual experience of a triangular object iff

(1*) o is *F*
(2*) o looks *F* to *S*.
(3*) *S* believes (2*).
(4*) o would not look *F* to *S* unless *S* had a visual experience of o as *F*.
(5*) *S* believes (4*).
(6*) *S* believes that *S* is having a visual experience of o as *F* (from (3) and (5)).

As Dretske (1995: 60-61) notes, there is one disanalogy between displaced perceptual knowledge and introspective knowledge. One cannot come to believe that o' is *G* by seeing that o is *F* unless o is *F*. In other words, displaced perceptual knowledge that o's is *G* requires that (1) obtains—that o indeed be *F*— and that *S*'s intermediate belief (3) that o is *F* be true. By contrast, *S*'s introspective belief that *S* is having a visual experience of o as triangular might be true even though *S*'s visual experience of o as triangular is non-veridical. *S*'s introspective belief that *S* is visually experiencing *F* does not require either that (1*) obtains or that *S* correctly believes that o is *F*. Whether (1*) obtains is optional, for (2*) may obtain although (1*) does not. *S* may falsely believe that object o is triangular even though o might not be a triangle at all. Object o may look triangular to *S* (as in (2*)) even though either o is not really triangular or there is no object at all: either *S* may misperceive object o as triangular or *S* may have a visual hallucination of a triangular object.

Two features of the displaced perception model of introspective knowledge of one's perceptual experiences are worth emphasizing. First, it is an externalist account since one comes to learn facts about one's own perceptual experiences by having the experiences. Granted, the experiences need not be veridical. But on the assumption that one would not have any visual experiences at all unless natural selection had provided the human visual system with the function to carry information about the visual attributes of objects instantiated in the environment of ancestors of humans, it follows that one's non-veridical visual experiences are parasitic on one's veridical visual experiences. One could not either misperceive a triangle or have a visual hallucination of a triangle unless one could visually perceive triangles. If so, then one learns facts about one's own perceptual experiences by perceiving mind-independent objects, properties and facts in the external world, not by experiencing—or by peering at—one's own perceptual experiences (as the traditional empiricist model of self-knowledge would have it).

Secondly, according to the displaced perception model of introspective knowledge of one's own perceptual experiences, one comes to know that one has perceptual experiences by forming beliefs about oneself. Although it is necessary to perceive mind-independent objects, it is not sufficient for introspective self-knowledge. Presumably, one can have visual experiences whether or not one can form the connecting belief (5*). But unless one can form the connecting belief (5*), one could not form the belief that one is having a visual experience.[9] Arguably, one could not form the connecting belief (5*) unless one had the higher-order concept VISUAL EXPERIENCE. Nor could one come to believe, as in (6*), that one is having a visual experience of a triangular object unless one had this higher-order concept. Similarly, one could not come to believe that one has the belief that o is a triangle unless one had the concept BELIEF.

---

9 On some interpretation, Rosenthal's (1986, 1993) higher-order thought (HOT) theory of conscious mental states would deny that one could have a *conscious* perceptual experience if one could not form, if not the connecting belief (5*), at least some close cousin of (5*)—a higher-order thought to the effect that one is having a perceptual experience.

2. Displaced perceptual knowledge, IBT and metarepresentations

The displaced perception model of introspection raises an interesting question about the compatibility of IBT and the appeal to metarepresentations. IBT offers a (presumably naturalistic) account of the contents of first-order mental representations of the external world based on the notions of information and function. No doubt, introspective beliefs about one's own perceptual experiences are higher-order mental representations of one's first-order perceptual representations of the world. Unlike the latter, the former are metarepresentations. Does a full naturalistic account require an extension of IBT to the contents of introspective metarepresentations? If so, can it be so extended? This question has been raised as a challenge to a naturalistic approach to the puzzles of mental content by Kemmerling (1999: 323-24), who writes: "How could such a thing—a metarepresentational belief—show up in Dretske's framework? [….] what is lacking, is an account of how a natural system may come to need information which is specifically about the content of its own ground-floor representations […] an account of natural metarepresentational systems [….] a job which is clearly separate from any job of any ground-floor representation". Can the IBT account be extended from the contents of "groundfloor" representations to the contents of introspective higher-order representations? Or should it?

My response to Kemmerling's (1999) challenge will come in two steps. First, I will argue that introspective beliefs notwithstanding, displaced perceptual knowledge itself is already infected by metarepresentations. So the challenge could be directed to displaced perceptual knowledge as well: is displaced perceptual knowledge compatible with a naturalistic semantics? Conversely, if displaced perceptual knowledge is immune to the challenge, so should the displaced perception model of introspective knowledge. Secondly, to parody Jerry Fodor (1994), I will argue that the challenge involves a confusion between semantics and epistemology.

First, in the previous section, I sketched a simple example of displaced perceptual knowledge in which one sees that a horse walked on the snow at t – 1 by seeing hoof prints in the snow at *t*. It does not seem to me out of the question at all that many non-human preys and predators are capable of such displaced perceptual knowledge. Now, consider in more detail some of Dretske's (1995) own examples of displaced perceptual knowledge. One comes to learn that the gas tank of one's car is empty by seeing the pointer of the gas gauge. One comes to learn something about the temperature of a liquid by seeing the level of mercury of a thermometer immersed in the liquid. One comes to learn "what is happening on the other side of the world" by reading a newspaper or watching television (Dretske, 1995: 41). I doubt very much that non-human preys and predators can come to achieve displaced perceptual knowledge of any of these last three kinds. In the sequel, I will distinguish between metarepresentational and first-order displaced perceptual knowledge.

Unless one knows the grammar of the natural language to which the perceived tokens of printed or spoken sentences belong, one will not be able to form the intermediate belief about the propositional content of the linguistic inscriptions or utterances one reads in the newspaper or one hears on television. In other words, unless one knows the grammar of some natural language, one will not grasp what sentences mean or

what they are used to say. Here, I will leave aside the question whether knowledge of the grammar of some natural language or other is necessary for one either to learn that the gas tank in one's car is empty by seeing the gas gauge or to learn what the temperature of a liquid is by seeing the level of mercury on a thermometer. Rather, I want to emphasize the complexity of the cognitive resources imposed by metarepresentational displaced perceptual knowledge of any kind.

In coming to learn that a horse walked on the snow at $t - 1$ from seeing hoof prints in the snow at t, one merey needs to form *first-order beliefs* about mind-independent facts, i.e., the fact that there are hoof prints in the snow and the fact that a horse walked in the snow. In other words, the target belief, the connecting belief and the intermediate belief are first-order mental representations of mind-independent facts. This is why it is plausible that non-human preys and predators can achieve such first-order displaced perceptual knowledge. Not so in the three examples of metarepresentational displaced perceptual knowledge. In all three examples, the target or displaced belief is a first-order mental representation of some fact.[10] But the intermediate belief is *not*. The intermediate fact about which one must form an intermediate belief in each of the three cases is itself a non-mental *representation* of some state of affairs, which need not be mental either.[11] In order to form a target belief about the level of gas in the tank of one's car, one must form an intermediary belief about the representation of the level of gas yielded by the gas gauge. In order to form a target belief about the temperature of the liquid, one must from an intermediary belief about the representation of the temperature yielded by a thermometer. In order to form a target belief about some event happening on the other side of the world, one must form an intermediary belief about the representational content of some linguistic expression. All three intermediary beliefs are about a representation or the state of some representational device. Of course, if some intermediate belief about a representation is metarepresentational, then so is the antecedent of the relevant connecting belief.

Notice that in each three cases, the relevant intermediate belief is *not* about some intrinsic property of the non-mental representation. One must perceive the shape, orientation and color of either the pointer of the gas gauge or the tube containing the mercury in order to achieve belief either about the level of gas in the tank or the temperature of the liquid. Similarly, one must perceive either the shape, orientation and color of the symbols printed in the newspaper or the acoustic properties of the sound structure of the utterances in order to achieve belief about some event on the other side of the world. But the relevant intermediate belief must be about the content, not the intrinsic local properties, of the representation. It must be about what the representation is about or what it stands for. The perceptual experience of the intrinsic properties

10 I do not say 'mind-independent' fact in order not to get entangled in the issue (that is irrelevant to the present discussion) of whether the fact that the gas tank is empty is a mind-independent fact. Of course, the gas tank is part of the car, which is itself an artifact that would not exist if it had not been manufactured by humans with minds.

11 I say purposefully 'state of affairs', not 'fact', for the representation need not be veridical and the state of affairs need not obtain.

of the representation gives rise to what Dretske (1969) calls nonepistemic perception. The intermediary belief about the content of the representation is epistemic perception that the representation means so and so.

Dretske (1995: 42) notes that "perceptual displacement enlarges the number of facts one perceives without a corresponding enlargment of the number of objects one perceives. [….] One see more facts, not by seeing more objects, but by expanding one's knowledge of what the objects one can already see signify about the objects one cannot see. This is what connecting beliefs (e.g., well-confirmed theories) provide". On the one hand, what Dretske did not emphasize is the extent to which, if I am right, some of his own examples of *non-introspective* displaced perceptual knowledge do presuppose the power to represent representations as such, i.e., metarepresentational resources. On the other hand, the metarepresentational resources involved in the above examples of displaced perceptual knowledge are really the tip of the iceberg of the full battery of human metarepresentational resources. Arguably, unlike other animals, humans derive more, not less, of their first-order representations of the world from the testimony of others (i.e., from verbal communication with their conspecifics) than from perception and memory. For example, an addressee located in Paris would not come to believe that it is raining in San Francisco from his understanding of the speaker's utterance of the sentence 'It is raining' in San Francisco unless the hearer had come to believe that the speaker believes that it is raining in San Francisco and intends to induce in her Parisian hearer the belief that it is raining in San Francisco by means of her utterance.[12]

Thus, human verbal communication requires third-person metarepresentational capacities, i.e., the ability to do third-person mindreading. From an evolutionary standpoint, the adaptive benefits of third-person mindreading among humans seem indeed quite obvious. Not only are the beliefs of one's conspecifics a useful source of information about aspects of the world that one cannot directly perceive (displaced perceptual knowledge), but social cooperation is an important source of mutual (social) benefit. Now, if one individual is contemplating the choice between cooperation and competition with another human agent, then it is useful for the former to be able to detect accurately the latter's goals, intentions, desires and vice-versa. Arguably, if one did not know the content of one's own mind, one could not even contemplate the choice between cooperation and competition. If so, then perhaps the metarepresentational resources necessary for introspective self-knowledge ride piggyback on the metarepresentational resources for knowing the minds of others. Alternatively, given that the adaptive advantages of both third-person and first-person mindreading may well stand or fall together, they may have co-evolved in tandem.[13]

I now turn briefly to my Fodorian claim that semantics is not part of epistemology. The task of IBT, or so I claim, is to provide a semantic account of the contents of

12 This paper is not trying to clarify the complexity of communicative intentions.

13 Notice that here I am considering the relation between the metarepresentational resources required respectively for introspective self-knowledge and for knowing the minds of others. I am not considering the sort of evidence relevant to each kind of knowledge, let alone the sort of epistemic authority that should be associated with each kind of knowledge claims.

first-order (or groundfloor) mental representations of the world. It is neither the task of IBT to provide an epistemic account of their justification, nor to provide an account of how one comes to acquire such higher-order concepts as REPRESENTATION, EXPERIENCE, BELIEF, and so on.

From a semantic standpoint, the content of a metarepresentation includes and depends systematically (or compositionally) upon the content of the representation metarepresented.[14] The content of the latter is a proper part of the former. If and when IBT has provided a naturalistic account of the content of the embedded first-order representation, its job is over. Presumably, one can see a horse and have a visual experience of a horse even though one does not know what horses are—even though one does not possess the concept HORSE. One cannot, however, believe that there is a horse nearby unless one knows what horses are or one has the concept HORSE.

The task of IBT is to offer an account of both the non-conceptual content of a percept of a horse and the content of the concept HORSE. Arguably, one can have either the visual experience of a horse or the belief that there is a horse nearby without believing (or knowing) that one does. One cannot, however, believe that one has either the visual experience of a horse or the belief that there is a horse nearby unless one has the concepts VISUAL EXPERIENCE and BELIEF. Furthermore, one cannot believe that either one has a visual experience of a horse or the belief that there is a horse nearby unless one has the concept HORSE. If one cannot have the belief that one believes that there is a horse nearby unless one believes that there is a horse nearby and if one cannot have the latter first-order belief unless one has the concept HORSE, it follows that one cannot have the former metarepresentational belief unless one has the concept HORSE. Although one can have the visual experience of a horse without having the concept HORSE, one cannot, however, have the metarepresentational belief that one has the visual experience of a horse unless one has the concept HORSE. Given that one may have the visual experience of a horse without having the concept HORSE, how is it that one cannot have the metarepresentational belief that one has the visual experience of a horse unless one has the concept HORSE? Unless one knows what horses are—unless one has the concept HORSE —, one cannot entertain the higher-order concept VISUAL EXPERIENCE OF A HORSE. Similarly, one cannot believe that a piece of paper is a photograph of a horse unless one knows both what horses are and what cameras are—i.e., unless one has the concept HORSE and the concept CAMERA.

The tough epistemological questions are: how does one acquire such higher-order concepts as REPRESENTATION, BELIEF or EXPERIENCE? How does one know that such concepts apply to oneself? To have the concept BELIEF, for example, is to know that beliefs, unlike intentions and desires, have, in Anscombe's (1957) and Searle's (1983) terminology, a mind-to-world direction of fit, not a world-to-mind direction of fit, or that they have truth-conditions. Correlatively, one cannot have the concept

14 I limit myself to second-order metarepresentations of first-order representations of the world. But in verbal communication, the human metarepresentational faculty can ascend to higher levels.

BELIEF unless one knows that, unlike states of knowledge, beliefs can be false. Does a human child learn these features of BELIEF by some ontogenetic process? Has the human brain been endowed by the phylogenetic evolution of the species with mastery of a set of higher-order concepts including the concept BELIEF with such characteristics? In this paper, I certainly do not pretend to offer any response to this question, which is presently the object of much empirical investigation in evolutionary psychology and in developmental psychology.[15]

Rather, I merely wish to divide the content of a metarepresentational belief into two pieces, one of which is the content of the first-order representation of the world that is metarepresented, and the other of which is the conditions for applying the relevant higher-order metarepresentational concept to the content of the first-order representation. I submit that the task of IBT, which deals with semantic mind/world relations, is to offer an account of the first piece. I submit that it is the task of epistemology, which deals with the conditions for believing and knowing something, to offer an account of the second piece.

3. Compatibilism revisited

At the beginning of this paper, I discussed the argument for the incompatibility between content externalism and the special authority of introspective self-knowledge. One can know a priori (with special first-person authority) that one believes that water is a liquid. It follows from externalism that one could not believe that water is a liquid unless one stood in relation to water. If so, then it follows from externalism that one can know a priori (with special first-person authority) that one stands in relation to water. But it is false that one can know a priori (with special first-person authority) that one stands in relation to water. It follows that externalism should be rejected.

I noted that the incompatibilist argument can go through only on two assumptions. First, it must be assumed that, for any introspective belief one may have, this belief is a priori and/or it has the very special first person epistemic authoritative features that were attributed to it by either the rationalist or the empiricist epistemological traditions or both. Secondly, it must be assumed that one's WATER concept makes one and the same semantic contribution to the content of both the metarepresentation and the first-order representation metarepresented.

In the last section of this paper, I want to examine one compatibilist strategy that is based on the distinction between two sorts of self-knowledge: knowledge of *what* one thinks, believes or experiences—i.e., knowledge of the content of one's mental representations—and knowledge—of the fact—*that* one thinks, believes or experiences whatever it is that one thinks, believes or experiences. As Dretske (forthcoming) has recently put it, "knowing what you think is easy". Knowing that you think is not. Armed with this

15 See e.g., Baron-Cohen (1995).

distinction, the compatibilist strategy argues that what one knows introspectively with special first-person epistemic authority is not the fact that one thinks, believes or desires what one thinks, believes or experiences. Rather, what one knows introspectively with special first-person epistemic authority is what one thinks, believes or experiences. What one may know by introspection is that it is water that one has beliefs about, not that one has beliefs about water. If one's introspective a priori knowledge is not that one believes that water is a liquid, then even if externalism is true of WATER, it still does not follow that one can know a priori—in the way one has introspective knowledge about what one believes—that one stands in relation to water (not to something else).

Although I think that this strategy, which has recently been championed by Dretske (1995, 2003), is a perfectly coherent strategy, I do not approve it. Since I think it is a perfectly coherent strategy, I will not provide any knock-down argument against it. Rather, to show why I do not approve it, I will argue that its cost upsets its utility.

As section 1 of the present paper made clear, Dretske's (1995) model of introspection has two basic ingredients: one is the model of displaced perceptual knowledge, the other is the principle of the reflexivity of content, which in turn derives from the IBT account of the contents of first-order mental representations of the external world. According to the principle of the reflexivity of content, one cannot represent the presence of property *F* unless one has the information that the represented property is *F* (see section 1). Thus, it is a consequence of the principle of the reflexivity of content that by virtue of representing the presence of *F*, one has the information that *F* is the represented property. The combination of the displaced perception model of introspection and the principle of the reflexivity of content yields the following result: one cannot have the visual experience of a triangle unless one has the information that triangularity is what one is experiencing. By visually perceiving a triangular object, one is provided with information about oneself. One gets information about oneself, by turning one's visual attention, not to oneself, but to some triangular object.

This externalist view of introspective knowledge of one's own perceptual experiences seems open to the following objection. First of all, having the information that *F* is the represented property is *not* the same thing as knowing it. According to the principle of the reflexivity of content and the IBT account of first-order representations, from the fact that it is representing the presence of *F*, any representational device will have the information that *F* is the represented property. A thermometer will have the information that temperature is what it represents. From the fact that it is perceiving a cat, a dog will have the information that cathood is the perceived property. Still, neither a thermometer nor a dog can resaonably be credited with introspective metarepresentational knowledge of the contents of their first-order representations. Secondly, one could not know what it is one is experiencing—e.g., triangularity—unless one could apply to oneself the concept EXPERIENCE. One could not know what it is one has beliefs about—e.g., water—unless one could apply to oneself the concept BELIEF. Undoubtedly, thermometers and even dogs lack the mastery of such higher-order concepts as EXPERIENCE and BELIEF. The reason why neither thermometers nor dogs can know what it is that they are representing would be that they cannot know that they are representing—something they are deprived of by the lack of the higher-order concept REPRESENTATION. Does not this show that one could not know what one

experiences unless one knew that one has experiences? Does not this show that one could not know what one believes unless one knew that one has beliefs?

This twofold objection relies on a subtle confusion between semantics and epistemology. It is a semantic truth that one cannot know either what one experiences or what one believes unless one applies to oneself either the concept EXPERIENCE or the concept BELIEF (and hence unless one possesses this concept). This is consistent with the epistemological point that having information about a represented property is different from knowing which property is being represented. One could not move from having the information to knowing that triangularity is the experienced property unless one applied the concept EXPERIENCE to oneself. One could not move from having the information to knowing that water is the property one has a belief about unless one could apply the concept BELIEF to oneself. But now the combination of the displaced perception model of self-knowledge and the principle of the reflexivity of content raises a further epistemological issue. Given that one cannot know either what one experiences (and/or believes) or that one has experiences (and/or beliefs) unless one can apply to oneself the concepts EXPERIENCE (and/or BELIEF), can one's self-knowledge be decomposed into two separable epistemological components: the knowledge of what one experiences (or believes) and the knowledge that one experiences (or believes) it? The above semantic truth provides no answer to this question.

The present epistemological question arises in the context of introspective self-knowledge. But as Dretske (2003) argues convincingly, the very same epistemological question can be raised about perceptual knowledge of the world. Dretske (1969: 93-99) argued that one can know that the water is boiling by seeing it boil. One's grounds for believing that the water is boiling are that one sees it. Before seeing the water boil, one did not believe—let alone know—that it was boiling. On a reliabilist view of what it takes to know the fact that the water is boiling, given that one's visual system is reliable, then by seeing the water boil, one can thereby come to know that it does. However, from the fact that one can see the water boil—and hence know that the water is boiling —, it does not follow that one can see that what is boiling is water. That what is boiling is water may be something one learnt not by seeing the water boil but otherwise. Presumably, one cannot tell by visual perception alone whether something is water, gin or gas. One may believe that what is boiling is water because either one was told that it was or because one tasted it. If so, then presumably one's grounds for knowing that what the water is doing is boiling should not extend to one's grounds for believing that what is boiling is water. Conversely, the question arises whether a (skeptical) challenge directed towards one's grounds for knowing that what is boiling is water could defeat one's claim to know that what the water is doing is boiling.[16]

16 Dretske (2003) has another convincing example to the same effect: Clyde comes to know that Harold told him on the phone that he was going on vacation from hearing Harold's voice, understanding English and retrieving Harold's communicative intention. Clyde may not be able to know by using the same resources that Harold was the person who told him that he was going on vacation. He may not be able to recognize Harold's voice. He may come to know (or form a justified belief) that it was Harold who told him that he was going on vacation by tracing the phone call back to Harold.

Thus, the question raised by such examples is this: from the fact that some complex piece of knowledge can be decomposed into two separable components because each can be traced to a particular epistemic pedigree, does it follow that one can know one piece without knowing the other? From the fact that some knowledge claim can be decomposed into two components, one of which is learned in one way and the other of which is learned in another way, does it follow that one can know either without knowing both? This question is perfectly general and it applies to perceptual knowledge as well as to psychological self-knowledge. Arguably, if the answer to the question is positive in the case of perceptual knowledge, then so should the answer be in the case of psychological self-knowledge.

Clearly, Dretske (1969, 2003) thinks that a positive answer can be given to the question in the case of perceptual knowledge and Dretske (2003) concludes that a positive answer can be given to the question in the case of psychological self-knowledge. Dretske (1969) did endorse the view that one can see (and hence know) that the water is boiling even though one does not know—not in the same way—that what is boiling is water. Dretske (2003) endorses the view that Clyde can hear (and hence know) that Harold told him that he was going on vacation even though he does not know—not in the same way—that the person who told him so was Harold. Understandably, Dretske (2003) argues that one can know by introspection what one is representing, even though one may not know—not in the same way—that what one is doing is representing. On this view, although one cannot know either what one thinks or that one thinks it unless one applies to oneself the concept THOUGHT, still one can know what one thinks and fail to know that one thinks it. For example, one knows that the concept WATER applies to the content of one's belief (that water is a liquid). Although one has the concept BELIEF, one merely believes (perhaps justifiably so) that the concept BELIEF applies to oneself (or to what one is doing while one believes that water is a liquid). But conceivably, the conditions for *knowing* that BELIEF applies to oneself might not be met. On this view, direct introspective knowledge is only of what one thinks, not that one thinks it.

As I said above, I think both that this is a coherent picture of introspective self-knowledge and that it undermines the argument for the incompatibility between the special epistemic authority of introspective self-knowledge and externalism. Nonetheless, I want to point out the cost of this picture and why it may be superfluous. The cost of the picture is the denial of closure, i.e., the principle that knowledge is closed under known implication.

Consider how the argument for the denial of closure works. First, it is noticed that in a complex conjunctive belief *K*, one can sort out two components B1 and B2—K = B1 & B2—on the grounds that the epistemic justification of one component is different from the epistemic justification of the other. For example, one learns B1, that the water is boiling by visual perception, and one learns B2, that what is boiling is water, by some other method, e.g., by tasting it. Furthermore, one's grounds for believing one component, B1, meet the requirements for knowledge and one's grounds for believing the other component, B2, do not. It is uncontroversial that one's epistemic justification for believing B1, the component one is in a position to know, should not carry over to one's epistemic justification for believing B2, the component one is *not* in a position to

know. In other words, given the hypothesis, the fact that one knows B1, that the water is boiling, does not entail that one thereby knows B2, that what is boiling is water. It is also uncontroversial, I think, that the epistemic justification of the complex belief K—consisting of the two component beliefs, B1 and B2—should not exceed the epistemic justification of the lowest of the two component beliefs, i.e., B2. What *is* controversial, I think, is the contraposition: namely, that a skeptical challenge directed towards one's justification for the weakest belief, B2, *cannot* ipso facto defeat one's justification for the stronger belief, B1. Certainly, a skeptical challenge so directed should defeat one's justification for the complex belief K. What is controversial is precisely the denial of closure. How loose can one's epistemic standards be for believing that what is boiling is water, B2, consistent with one's knowledge that what the water is doing is boiling, B1? Is skepticism about the fact that one knows that one has beliefs and experiences compatible with one's knowing what one believes and experiences?

Now, given the costs of the two-tiered analysis of psychological self-knowledge, I want to reexamine the strategy an externalist may want to choose in order to avoid the incompatibilist conclusion. In the first section of the paper, I have expressed doubts about one premiss of the incompatibilist argument: namely that we know enough about the process of psychological self-knowledge to accept the various strands of the picture inherited from traditional rationalist epistemology and from traditional empiricist epistemology. Now, I want to question the conditional premiss: if the traditional picture of self-knowledge is right, then externalism cannot be right.

4. Conclusion: how to question the conditional premiss of the incompatibilist argument

Although the conditional premiss does not say so, it does imply, and seems motivated by, the claim that, unlike an externalist view of mental content, an internalist view would be consistent with the traditional claims made on behalf of the special epistemic authority of self-knowledge. This implication of the conditional premiss is puzzling.[17] According to a *physicalist* version of internalism, the content of one's mental states is constituted not by the history of one's brain nor by extrinsic relations between one's brain and properties instantiated in one's environment, but by the current internal physical (chemical and biological) structure of one's brain. If so, then how could one know a priori with special first-personal epistemic authority the physical structure of one's brain any better than one can know a priori with special first-personal epistemic authority the history of one's brain or the nature of the extrinsic relations between one's brain and properties instantiated in one's environment? Indeed, assuming the truth of physicalism, what difference does it make to the traditional picture of self-knowledge

17 As Heil (1992: 174) insightfully put it, "If the contents of one's thoughts depended entirely on the state of one's brain, for instance, why should that fact alone render our access to those contents any less direct or problematical?".

whether one is an externalist or an internalist? Perhaps ontological dualism was a better match for the traditional special epistemic authority of self-knowledge than physicalism of either an externalist or an internalist variety.[18]

On the one hand, the account of introspective self-knowledge of one's perceptual experiences based on the principle of the reflexivity of content and the displaced perceptual knowledge model is an elegant externalist account. First, it clearly shows that only if one has metarepresentational resources can one become aware of the fact that one is having a perceptual experience. Secondly, it shows that one becomes aware that one is having a perceptual experience, not by peering inside at one's perceptual experience, but by having the perceptual experience itself, i.e., by experiencing the world. Thirdly, if the displaced perception model of self-knowledge is right, then so is the Cartesian thesis that self-knowledge involves higher-order (or metarepresentational) thought, not some quasi-perceptual mechanism. But its vindication of the Cartesian model of introspective self-knowledge is not unmitigated. Whereas the Cartesian model of introspective self-knowledge applies to one's knowledge that one has thoughts, the displaced perception model explains how one knows that one has perceptual experiences, not that one has beliefs or desires.[19]

On the other hand, Dretske (2003) takes both premisses of the incompatibilist argument seriously enough to choose to deny closure in order to block the incompatibilist conclusion. Given the costs incurred by the denial of closure, the externalist might consider questioning both premisses of the incompatibility argument. One option is to question the conditional premiss on the grounds that physicalist internalism does not seem easier to accommodate with special epistemic authority of self-knowledge than externalism. Another option is to reject the assumption that the concept WATER makes the same contribution to the content (or truth-conditions) of one's first-order belief that water is a liquid and to the content (or truth-conditions) of one's metarepresentational belief that one believes that water is a liquid.

On the second option, Oscar's concept WATER on Earth and Twoscar's concept TWATER on Twin-Earth would have, in Kaplan's (1989) terminology, different *contents*, but they would have one and the same *character*. Oscar's concept WATER would contribute its content to the truth-conditions of his belief that water is a liquid. Twoscar's concept TWATER would contribute its content to the truth-conditions of his belief that twater is a liquid. Thus, Oscar's and Twoscar's first-order beliefs would have different truth-conditions. Suppose now that Oscar's concept WATER contributes its character, not its content, to the truth-conditions of Oscar's introspective belief that he believes that water is a liquid. Suppose that Twoscar's concept TWATER contributes its character, not its content, to the truth-conditions of Twoscar's introspective belief that he believes that twater is a liquid. Suppose that the character of WATER is the same as the character of TWATER. Then Oscar's introspective belief and Twoscar's introspective belief would have the same-truth-conditions. If so, then externalism would

18 Something Heil (1992: 174) expresses doubts about.

19 This seems to me vindicated by higher-order introspection: one's belief that one has such or such a perceptual experience is generally more reliable than one's belief that e.g., one has a particular desire.

be true of one's first-order beliefs about the world, not of one's introspective beliefs about one's own beliefs. Whether the resulting picture is still externalist is a topic for another paper.

References

Anscombe, G.E. (1957) *Intention*, Oxford: Blackwell.

Baron-Cohen, S. (1995) *Mindblindness*, Cambridge, Mass.: MIT Press.

Boghossian, P.A. (1989) "Content and self-knowledge", in Bernecker, S. & Dretske, F. (eds.) (2000) *Knowledge, Readings in Contemporary Epistemology*, Oxford: Oxford University Press.

Boghossian, P. A. (1997) "What the externalist can know a priori", in Wright, C., Smith, B. C. & Macdonald, C. (eds.) *Knowing Our Own Minds*, Oxford: Oxford University Press.

Burge, T. (1988) "Individualism and self-knowledge", in Bernecker, S. & Dretske, F. (2000) *Knowledge, Readings in Contemporary Epistemology*, Oxford: Oxford University Press.

Davidson, D. (1984) "First-person authority", in Davidson, D. (2001) *Subjective, Intersubjective, Objective*, Oxford: Oxford University Press.

Davidson, D. (1987) "Knowing one's own mind", in Davidson, D. (2001) *Subjective, Intersubjective, Objective*, Oxford: Oxford University Press.

Dretske, F. (1969) *Seeing and Knowing*, Chicago: Chicago University Press.

Dretske, F. (1988) *Explaining Behavior*, Cambridge, Mass.: MIT Press.

Dretske, F. (1995) *Naturalizing the Mind*, Cambridge, Mass.: MIT Press.

Dretske, F. (1999) "The mind's awareness of itself", in Dretske, F. (2000) *Perception, Knowledge and Belief*, Cambridge: Cambridge University Press.

Dretske, F. (2003) "Knowing what you think vs knowing that you think it", in Schantz, R. (ed.) *The Externalist Challenge: New Studies on Cognition and Intentionality*, Berlin: Walter de Gruyter.

Dretske, F. (forthcoming) "How do you know you are not a zombie" in Gertler, B. (ed.) *Privileged Access and First-Person Authority*, Ashgate Publishing Co.

Fodor, J.A. (1994) *The Elm and the Expert*, Cambridge, Mass.: MIT Press.

Harman, G. (1990) "The intrinsic quality of experience", in Block, N., Flanagan, O. & Güzeldere, G. (eds.) *The Nature of Consciousness*, Cambridge, Mass.: MIT Press.

Heil, J. (1992) *The Nature of True Minds*, Cambridge: Cambridge University Press.

Kaplan, D. (1989) "Demonstratives", in Almog, J., Perry, J. & Wettstein, H. (eds.)(1989) *Themes from Kaplan*, New York: Blackwell.

Kemmerling, A. (1999) "How self-knowledge can't be naturalized (some remarks on a proposal by Dretske" *Philosophical Studies*, 95, 311-328.

McKinsey, M. (1991) "Anti-individualism and privileged access" in Chalmers, D.J. (ed.)(2002) *Philosophy of Mind, Classical and Contemporary Readings*, Oxford: Oxford University Press.

Rosenthal, D.M. (1986) "Two concepts of consciousness", *Philosophical Studies*, 99, 3, 329-59.

Rosenthal, D.M. (1993) "Thinking that one thinks", in Davies, M. and Humphreys, G.W. (eds.) *Consciousness: Psychological and Philosophical Essays*, Oxford: Blackwell.

Searle, J. (1983) *Intentionality. An essay in the philosophy of mind*, Cambridge University press, Cambridge.

Shoemaker, S. (1996) *The First-Person Perspective and Other Essays*, Cambridge: Cambridge University Press.

Tye, M. (1992) "Visual qualia and visual content", in Crane, T. (ed.)(1992) *The Contents of Experience*, Cambridge: Cambridge University Press.