How I know I am not a Zombie

Thomas Grundmann, Saarbrücken

In a certain sense we are all, or at least most of us are, Cartesians. We are convinced beyond doubt that we *know* what we think, that we *know* how we think it (i.e. in what propositional attitude), and that we *know* we have thoughts at all. Moreover, we are convinced that we know all this not on the basis of studies in empirical psychology or observation of behavior, but in a direct, privileged way that is not available to outside observers. In short, we are sure that we have introspective self-knowledge. I call this the "basic Cartesian intuition." In itself, this intuition implies neither that our self-knowledge is infallible, nor that it is based upon inner observation. The basic Cartesian intuition says only that that we have self-knowledge that is not based on observation from outside.

Fred Dretske and Sven Bernecker have recently argued that this basic Cartesian intuition is not fully compatible with a position that is currently very popular in philosophy of mind – namely with representational externalism. According to this position, all mental states are representational in nature, i.e. they represent objects as having such-and-such a quality. And the representational content of mental states depends upon the relations obtaining between the system and its environment. If this position is correct – Dretske and Bernecker argue – then, although it does not follow that we can have no introspective knowledge at all, it does follow that we cannot have introspective knowledge *that we are creatures with minds and not mindless zombies*. We cannot know via privileged access that we have thoughts and consciousness. For a Cartesian, that is a punch in the face. Bernecker characterizes this consequence as "zombie-scepticism." A zombie, in this context, is someone who is like a person in his or her appearance and behavior, but who has no mental states. What distinguishes him or her from us is not only the absence of phenomenal consciousness (as, for example, David Chalmers suggests), but the absence of any kind of inner mental life. More precisely, the Dretske-Bernecker thesis runs as follows:

(DB) If minds are by nature representational and if the existence of representational properties is dependant upon the relations obtaining between a system and its environment,

then we have introspective knowledge only of the specific contents of our mental representations, not of the fact that we have minds.[1]

Dretske and Bernecker are both representational externalists. And they acknowledge that this position is only partially compatible with the basic Cartesian intuition. They consider externalism defensible, though, because it does not throw out introspective knowledge lock, stock and barrel, but allows it within certain bounds. In my opinion, however, there is another diagnosis to be taken into consideration: if Dretske and Bernecker are right, then content externalism contradicts our basic Cartesian intuition, and should therefore by rejected. Since the sceptical consequences that Dretske and Bernecker derive from externalism can also be understood as a reductio ad absurdum of externalism, I would like to take a closer look at whether content externalism really entails these consequences.

First of all, I shall attempt to demonstrate that the problem of zombie-scepticism does not arise from content externalism taken by itself, but from its conjunction with Dretske's specific information-theoretic concept of knowledge. I shall not question this concept of knowledge, but shall argue that it does not have the sceptical consequences that Dretske and Bernecker take to be unavoidable. Finally I shall show that these consequences can also be avoided if one assumes a different externalist concept of knowledge, namely the one proposed by Robert Nozick. If I am right, representational externalism is perfectly compatible with our basic Cartesian intuition, which in turn speaks clearly in its favor.

I would like to begin by elucidating the Dretske-Bernecker thesis with a few remarks:
(i) The thesis speaks of "introspective knowledge." This is not intended to imply a positive model – like that of inner perception, which of course would raise intractable problems – but only a source of knowledge that is independent of sense perception. It is a purely negative characterization, a minimal concept of introspection. Introspective knowledge is knowledge that is independent of sense perception.
(ii) As we shall see, Dretske's concept of knowledge plays an important role in the argumentation for the Dretske-Bernecker thesis. According to Dretske, a subject S knows p if and only if S has the belief that p and this belief is causally based upon the *information* that p. As for information, however, Dretske has not always defined it uniformly. It is clear that the

[1] Dretske 2003b, 137: „What introspection gives us is the content of our cognitive states (...), not the fact that it is content." Bernecker 2000, 2: „Self-knowledge provides us with knowledge of what is in our minds, but not that we have minds."

information that p entails that p is true. That is why the information condition of knowledge makes the usual truth condition redundant. In *Knowledge and the Flow of Information*, Dretske espouses a probabilistic concept of information, according to which the signal r carries the information that p iff p, given that r has a conditional probability of 1. Christoph Jäger has recently drawn attention to the fact that conditional probabilities of 1 are closed under logical implication (If the conditional probability of p, given r, is 1, and p logically implies q, then the conditional probability of q, given r, is also 1). Dretske, however, denies the closure of information and knowledge. And in fact he has to for the sake of his argumentation, as we shall see presently. Hence it is advisable to consider a different definition of information, which Dretske presented in an earlier paper (entitled *Conclusive Reasons*) under the label "conclusive reasons." According to this modal definition, a signal r carries the information that p iff r would not have occurred if p had not been the case. It is important to note that the evaluation of this counterfactual sentence would take place in the nearest possible world, not in just any non-p world. This modal definition of information achieves the desired result – information is not closed under logical implication. An example may serve to demonstrate this: my experience of a table carries the information that there is a table in front of me because in the nearest possible world in which there is no table in front of me, I would not experience a table in front of me. *That there is a table in front of me* entails *that I am not a brain in a vat in a world without tables, deceived by an evil scientist*. But my table experience does not carry the information that I am not such a brain in a vat, because if I were such a brain, I would still have the same table experience. Thus the modal definition would lead us to give up the closure principle.

(iii) The Dretske-Bernecker thesis can only be consistent if a particular version of the closure principle – namely the transitivity of reasons – is false. The transitivity principle says:

(T) If S knows that p *on the basis of r*, and q follows logically from p, then S knows q *on the basis of r*.

If this principle were correct, then if S knew *introspectively* that he represented p and not q (the DB-thesis concedes this), then he would automatically also know *introspectively* that he represented (and hence that he were not a mindless zombie), since *S represents that p* entails *S represents*. But this consequence is exactly what the DB-thesis disputes. Thus the closure principle has to be given up; the thesis can only be maintained if the modal concept of information is adopted.

So much for my elucidation of the DB-thesis. Let's have a look now at the argumentation in favor of the thesis. I think it is possible to distinguish two arguments. The first one I would like to call the "argument from absent rationalizing reasons."

ARGUMENT I

(1) Knowledge demands reasons that rationalize the belief.	(Premise)
(2) A reason rationalizes a belief if there is an inferential relationship between the representational content of the reason and the representational content of the belief.	(Analytic)
(3) Only the content of representations and the internal properties of the states of an organism are introspectively accessible.	(Premise)
(4) The content of a representation does not include the fact that it is a state with content.	(Premise)
(5) Since content externalism is true, the presence of representational content cannot be derived from the internal properties of the states of an organism.	(Premise)
(6) Thus the contents of introspection do not rationalize the belief that representational content is present.	(2),(3),(4),(5)
(7) Therefore, there can be no introspective knowledge that representational content is present.	(1), (6)

In his paper *How do you know you are not a zombie?*, Dretske argues above all for premise (4). The quintessence of his reflections can be illustrated with the following example: There is surely a difference, Dretske says, between things that one sees, and things that one does not see. This difference constitutes our perspective on the world. But when one looks at the world around oneself, one simply does not see one's own perspective.[2] Our perspective is not part of the content of our perception. Bernecker makes a complementary argument for premise (5) in *Knowing the World by Knowing One's Mind*. He maintains that there is no internally accessible rationalizing reason: "Just by looking inside, I cannot decide whether or not my so-called thoughts have content or are contentless states." (Bernecker 2000, 12). I

[2] Dretske 2003a, S. 3.

find Dretske's and Bernecker's arguments for premises (4) and (5) convincing. Premises (2) and (3) also have an overwhelming plausibility.

Premise (1), however, is vulnerable. For an epistemological externalist – as Dretske and Bernecker officially profess to be – knowledge does not require rationalizing reasons. If, for example, we assume Dretske's definition of knowledge, then the knowledge that p requires the information that p. But the information that p is not necessarily a rationalizing reason, because the inferential relation necessary for such a rationalizing reason can only obtain between states with representational content. A state can carry information about a certain fact, though, without representing this fact. It is also possible for a state to represent a certain fact without carrying any information about it (most obviously in the case of mis-representations). Hence information and representation are mutually independent of each other. But a representation may also carry the information necessary for knowledge. Dretske's definition is therefore compatible with the existence of rationalizing reasons, although it does not necessarily require such reasons.

Thus the first argument is not persuasive. But there is another one: the "information-theoretical argument," as I would like to call it. This is the main argument for the DB-thesis. It is spelled out most clearly in Dretske's paper *Externalism and Self-Knowledge*. The argument goes as follows:

ARGUMENT II

(8) Knowledge requires beliefs that are supported by corresponding information.

(9) In a representational system there is no information that is relevant to the introspective knowledge of the fact *that* the system represents something. (The only information in the system that is relevant to introspection is the information about *what* the system represents.)

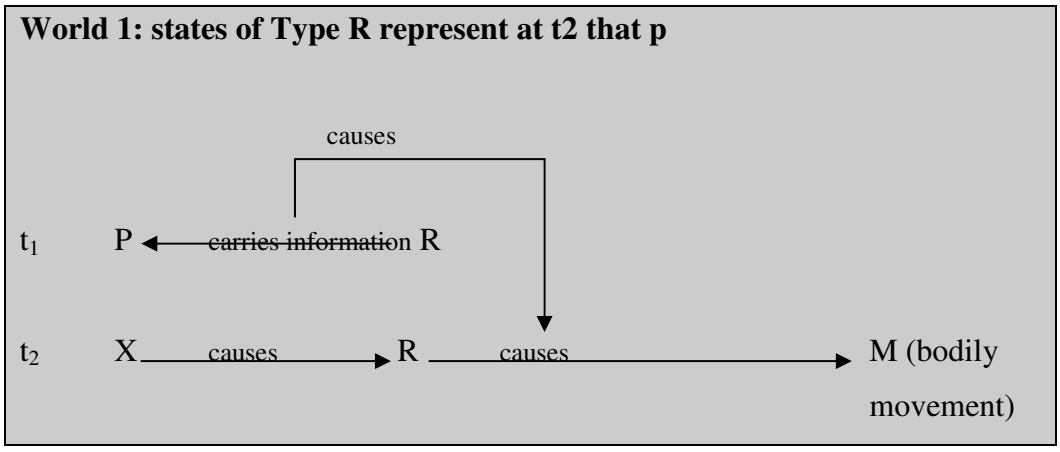
(10) A representational system cannot have introspective knowledge of the fact *that* it is representational.

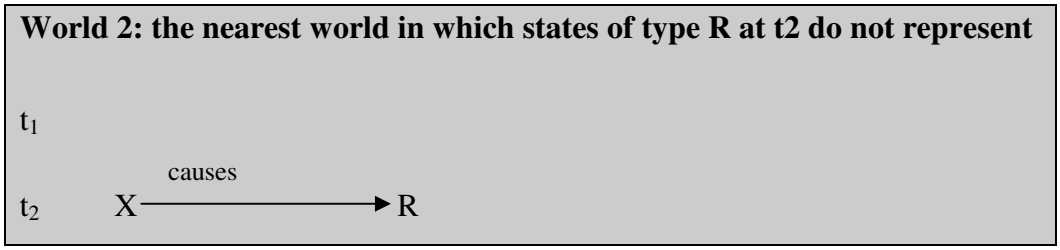
Premise (8) reflects Dretske's information-theoretical concept of knowledge, which I do not want to discuss here. But why should we accept premise (9)? There is internal information about *what* the system represents: if the system represented something else, a different internal realizer of representational content would arise. So the representational state itself

carries information about *what* it represents. But, according to Dretske, there is no internally available information about the fact *that* the system is representational: if content externalism is right, the same internal states that currently have representational content would still arise even if they had no representational content, as long as the external content-individuating factors were not present. Thus a representational state does not carry the information that it is representational.

That becomes clearer when one considers Dretske's explanation of representational content. A system has representational properties if its states have the *function* of carrying information about certain properties. Tokens of type R can take on this function either through the intention of a designer or user (as in the case of instruments) or by natural means (as in the case of cognitive systems). The latter occurs when the informational properties of a state of type R historically cause a successor of type R in the cognitive system to take on a particular functional role. For example, if the neural states of type R in a frog's brain sometime in the past carried the information that a fly was flying by, fly-catching behavior triggered by these states would have been successful. Through evolution, that would have led successors of type R always to trigger this same fly-catching behavior. If that is so, states of type R would have acquired the content that a fly is flying by, and would in fact also have that same content if they happened not to have been triggered by a fly flying by (in the case of mis-representation).

If this is the right explanation of representational content, it is not difficult to see why representational states do not carry the information that they are representational: states of type r might arise even if they did not have any representational content – i.e. if the content-explaining prehistory had not occurred.





Now for my objection. I think Dretske is right that representational states do not themselves carry the information *that* they (or the system to which they belong) are representational. *But it does not follow* that this information is nowhere to be found in the system. Indeed, there are other facts in the system that carry precisely this information – namely the fact that the representational states of the system have a particular functional role. The figure above should make this clear. If the states that currently have representational content did not have representational content, the historical factors that constitute the content would also have been absent. But if these factors had been absent, the structuring cause of the functional role of these states would have been absent, too (as in world 2). Hence it is fair to say that, if the states that currently represent had no representational content, they would not have the functional role that they have. So the fact that they have this functional role carries information that the states are representational. Thus the system does have – contra Dretske and Bernecker – internal states that carry the information that the system is representational.

One additional remark in passing: in his paper *Die Grenzen des Selbstwissens*, Bernecker expresses doubt that we can know introspectively in what propositional attitude we represent – whether we have a belief, a perception or a wish, or are merely considering a representational content. But in my view there are also internal facts that carry information about *the attitude in which* we represent. The fact that a representational state of a particular type arises in a particular module – let's say the belief module and not, for example, the perception module – carries the information that the corresponding representational content is a belief and not a perception. For if it had been a perception rather than a belief, it would have appeared in the perception module rather than in the belief module.

At first glance, it may seem there is a reply that could torpedo my suggestion that the functional role of a representational state carries the information that this state is representational. Couldn't there be a system with states playing the same functional roles as the states of my representational system, but without its functional architecture having been caused by content-providing factors? In short: couldn't there be a functional duplicate of me that did not have a mind? Content externalism appears to admit just such a possibility. Davidson's swamp man dramatizes this possibility in a vivid way. Why should it not be possible in principle for a system with the same internal makeup and functional structure as mine to arise suddenly in the morass of a swamp struck by lightning? From the perspective of externalism, this system would not have the prehistory necessary for representational content. Thus a functional duplicate without a mind seems possible. And Dretske and Bernecker appeal to precisely this possibility in their argumentation.

My response is to point out that, although such scenarios are of course possible, they are only realized in worlds that are *irrelevant* to the evaluation of informational properties. A signal r carries the information that p if it would not have arisen in the *nearest* possible world in which p and all facts logically and causally relevant to p had not been the case. And precisely this condition is fulfilled by the functional role. In the *nearest* possible world in which the causes that provide the functional role with its content are absent, there are no alternative causes for the genesis of the functional role. In other words, although swampman-worlds are possible worlds, they are very distant, exotic worlds. The world would have to be a lot different for a functional duplicate of me to arise from the morass of a swamp struck by lightning. According to Dretske's theory of information, the informational content of a signal would not be endangered if the signal were also to arise in a very distant world where the fact about which the signal carries information were absent. Otherwise perceptual states could not carry information about the material external world either, because in very distant worlds they could also be brought about by an evil demon. And Dretske rejects precisely this consequence. So only the nearest non-p worlds are relevant; and in those worlds, the functional role would not exist in the absence of the content-providing factors. So much for my response to the objection.

There is however another objection to my suggestion. It may well be that the functional role of the representational states of a system carries the information that the states are representational (presumably it also carries information about what content the states

represent and how they represent it), but why should beliefs based upon this information be instances of *introspective* knowledge? The following consideration appears to imply that they are not: if content externalism is right, the information that the states are representational contains information about external content-providing facts. Can such knowledge about the external world really be introspective?

In order to answer this question, it is important to be clear about the fact that sources of knowledge – like perception, introspection and memory – are not individuated by their domains of objects. One cannot simply define perceptual knowledge by saying that it applies to the external world; introspective knowledge by saying that it applies to the inner world; and memory knowledge by saying that it applies to the past. It is obviously possible for the same objects to be accessible with the help of different sources. One may, for example, introspect that one is nervous about a test, but one may also become aware of one's nervousness by observation of one's own behavior, or by having it pointed out by someone else. So the introspective character of my knowledge that I have thoughts and consciousness is not undermined by the implications my knowledge has for the external world.

More is required, however, to defend the introspective character of my knowledge that I have representational states: I shall also have to explain what distinguishes the unique character of introspection. In my view, introspective knowledge differs from perceptual knowledge in that it is not based upon rationalizing reasons. In the case of perceptual knowledge – my visual knowledge that there is a rectangle in front of me, for example – the representational content of my visual experience is a reason that inferentially rationalizes my belief. I see that there is something rectangular in front of me, and this reason rationalizes my belief that there is something rectangular in front of me. The case of introspection is completely different. In order for there to be knowledge, there has to be information about what is represented, about how it is represented, and that something is represented. This information is not given, however, in the form of a representation, but through non-representational and unconscious facts like functional roles. Introspective knowledge is not based upon rationalizing reasons; it is non-inferential. That is what distinguishes its unique character.

If the information necessary for the knowledge that I represent something is not given by a perceptual representation (but by a functional role), then the case can be made that it is introspective knowledge. Of course this does not exclude the possibility of knowing this fact

by perceptual-empirical means. Such empirical knowledge of my mind would be based upon a theory, which would be supported by observation of my interaction with the outside world.

If my account of introspective knowledge is correct, it cannot be objected that the specific functional role of our representational states is (at least introspectively) inaccessible. The information relevant to our introspective knowledge would only have to be present; it would not itself have to be introspectively accessible to consciousness. Such a requirement, in any case, cannot be derived from Dretske's definition of knowledge.

Thus it seems that someone who is an externalist about mental content can claim to know introspectively that he or she has a mind and is not a zombie. Assuming Dretske's information-theoretic definition would not change that. In the remaining space available to me, I would like to see how the issue appears from Nozick's perspective. In the 1980's, Nozick presented the much-respected "truth-tracking" definition of knowledge. According to this definition, a subject S knows that p if and only if the subject has the belief that p, the belief is true and two further conditions are fulfilled: the sensitivity condition and the stability condition. The sensitivity condition (the actual "truth-tracking" aspect) demands that S would not believe that p if p were not the case. So, for Nozick – in contrast to Dretske – there do not have to be sensitive reasons for the belief. It is sufficient for the belief itself to be sensitive. The stability condition demands further that S would still believe that p under slightly different conditions.

Let's assume that Nozick's definition of knowledge is correct and that content externalism is also valid. Can S's belief (b) that he has a mind (i.e. that he has representational states) fulfil the conditions of knowledge laid out by Nozick? Well, if this belief is present, it is automatically true, since S has a mind if he has any beliefs at all. So the belief is self-verifying. That is enough to fulfil the first two conditions. The sensitivity condition is also fulfilled automatically, since if belief (b) were false (i.e. if the external factors that constitute the mental meaning-content were generally absent), it would also be absent. A zombie has no representational states – not even the belief that is has such representations. As for the stability condition, it can be fulfilled, too – at least if our mental states are self-representational and automatically bring about the belief that we represent content. Thus, given Nozick's definition, there is nothing difficult about knowing one has a mind.

What about the second-order belief (c) that I am presently having a first-order thought? This belief is obviously not self-verifying, since it is not self-referential. But if it arises in a cognitive system with first-order representational states, Nozick's first two conditions of knowledge are fulfilled. And the other conditions? Well, let's assume that there is a mechanism in the cognitive system that always causes the realizers of first-order thoughts to bring about the belief (c). Let's also assume that Dretske's explanation of representational content is correct. Then the belief (c) that I am presently having some first-order thought or other is only possible if the type-identical predecessors of the realizers of this belief carried information about the presence of first-order thoughts. So the realizers of the first-order thoughts must at least in the past have had representational content. But if historical externalism is correct, they have this content permanently. That is a consequence of the factors that constitute the content of (c). Thus, if the realizers of my current first-order thoughts did not currently have representational content (i.e. because the external conditions were unfulfilled), the realizer of my belief (c) would not have the semantic content that I am presently having some first-order thoughts or other. That seems to fulfil the sensitivity condition. But if, as we have assumed, there is a mechanism that always causes the realizers of first-order thoughts to bring about the belief (c), the stability condition is also fulfilled.

The picture that arises from my considerations is as follows. As long as we assume that knowledge involves internalist conditions (like rationalizing reasons), externalism about mental content has far-reaching sceptical consequences for the possibility of introspective self-knowledge. Epistemological externalists can however continue to believe that we know introspectively what we think, how we think it , and that we think at all – even if externalism about mental content is correct. From the perspective of epistemological externalism, content externalism is therefore completely compatible with the basic Cartesian intuition. I hope that my discussion of the Dretske-Bernecker thesis against the background of various externalist conceptions of knowledge has demonstrated this.

So a thoroughgoing externalism about content and knowledge seems to be completely compatible with the basic Cartesian intuition. If being compatible with our basic intuitions speaks in favor of a position, then my conclusion strengthens content externalism to no small degree. Especially if one assumes that epistemological externalism is plausible for other reasons. I think it is, but that is another story.