5 Extended minds in vats

Sven Bernecker

Hilary Putnam has famously argued that "we are brains in a vat" is necessarily false. The argument assumes content externalism (also known as semantic externalism and anti-individualism), that is, the view that the individuation conditions of mental content depend, in part, on external or relational properties of the subject's environment. Recently content externalism has given rise to the hypothesis of the extended mind, whereby mental states are not only externally individuated but also externally located states. This chapter argues that when content externalism is combined with the extended mind hypothesis it is robbed of its anti-skeptical power. Given the extended mind hypothesis, the supercomputer and the envatted brain can be regarded as aspects of the extended mind of the evil scientist. On this view, the thought contents of the coupled brain–computer–scientist system do not differ from those of a normal human. But without a difference in thought contents Putnam's anti-skeptical argument crumbles.

Apart from giving rise to the extended mind hypothesis, content externalism has given rise to the thesis of embedded cognition, that is, the view that cognition depends not just on the brain but also on the body and its interaction with the environment. This chapter argues that when content externalism is combined with the thesis of embedded cognition, the vat has to fill in everything the world provides for the brain to realize the kind of mind we have. But if the vat fills in everything the world provides, then the skeptical problem evaporates.

Section 5.1 explains content externalism and gives a brief account of Putnam's refutation of the brain-in-a-vat scenario. Section 5.2 discusses the hypothesis of the extended mind. Section 5.3 argues that adding the extended mind hypothesis to content externalism undermines Putnam's refutation of the brain-in-a-vat scenario. In Section 5.5 it is shown that the thesis of

For helpful comments on an earlier version of this chapter I am grateful to Peter Baumann and Maura Priest.

embedded cognition takes away the skeptical sting the brain-in-a-vat scenario is said to have. Section 5.5 offers some concluding remarks.

5.1 Putnam on brains in vats

External world skepticism states that we don't know many of the worldly propositions that we take ourselves to know. Skeptical arguments typically take as their starting point skeptical scenarios. A skeptical scenario introduces a possibility concerning how the world really is that is incompatible with how we experience the world and that, given our empirical evidence, we are ostensibly unable to rule out. The skeptical conclusion follows from my inability to rule out this possibility. A classical skeptical scenario is Descartes's Evil Demon scenario:

> I will suppose . . . that . . . some malicious demon of the utmost power and cunning has employed all his energies in order to deceive me. I shall think that the sky, the air, the earth, colors, shapes, sounds and all external things are merely delusions of dreams that he has devised to ensnare my judgment. I shall consider myself as not having hands or eyes, or flesh, or blood or senses, but as falsely believing that I have all these things. I shall stubbornly and firmly persist in this mediation . . . I am like a prisoner who is enjoying an imaginary freedom while asleep. (Descartes 1984: 15)

Hilary Putnam's brain-in-a-vat scenario is a contemporary version of Descartes's Evil Demon story. Here is Putnam's classic passage:

> [I]magine that a human being (you can imagine this to be yourself) has been subjected to an operation by an evil scientist. The person's brain (your brain) has been removed from the body and placed in a vat of nutrients which keeps the brain alive. The nerve endings have been connected to a super-scientific computer which causes the person whose brain it is to have the illusion that everything is perfectly normal. There seem to be people, objects, the sky, etc; but really all the person (you) is experiencing is the result of electronic impulses traveling from the computer to the nerve endings. The computer is so clever that if the person tries to raise his hand, the feedback from the computer will cause him to 'see' and 'feel' the hand being raised. Moreover, by varying the program, the evil scientist can cause the victim to 'experience' (or hallucinate) any situation or environment the evil scientist wishes. He can also obliterate the memory of the brain operation, so that the victim will see himself to have always

> been in this environment. It can even seem to the victim that he is sitting and reading these very words about the amusing but quite absurd supposition that there is an evil scientist who removes people's brains from their bodies and places them in a vat of nutrients which keep the brains alive. The nerve endings are supposed to be connected to a super-scientific computer which causes the person whose brain it is to have the illusion that ... (Putnam 1981b: 5–6)

While the brain-in-a-vat scenario "violates no physical law," and is "perfectly consistent with everything we have experienced," Putnam (1981b: 7) still insists that it "cannot possibly be true, because it is, in a certain way, self-refuting." The statement "we are brains in a vat" is said to be *necessarily* false. If we are not brains in vats, then the statement "we are brains in vats" is obviously false. But Putnam argues that the statement "we are brains in vats" is also false if uttered or thought by a brain in a vat. This part of the argument rests on *content externalism*. Content externalism is the view that the meanings of my words and the contents of my thoughts depend in part on causal relations that I bear to aspects of my social and/or physical environment rather than only on internal properties of my mind and brain. What my words mean and my thoughts refer to is determined by the environmental features their usage or tokening is causally connected to.

A common strategy for motivating content externalism is to describe a situation in which there are two individuals who share all their intrinsic properties, but who differ in respect to some mental properties because they inhabit different environments. In "The Meaning of 'Meaning'" Putnam asks us to consider this thought experiment. Suppose Oscar is an English-speaker who uses the word 'water' in the same way as the other members of his linguistic community. He doesn't have any considerable knowledge of the chemical properties of water. Suppose that there exists somewhere in a nearby galaxy a Twin Earth, that is, a planet that is a molecule-for-molecule duplicate of Earth. The only difference between the two planets is that no water is present on Twin Earth. What is found there instead is a liquid which looks, tastes, and behaves like water but has the chemical composition XYZ and not H_2O. Like each one of us, Oscar has a molecular duplicate on Twin Earth.[1] Twin S has endured experiences that are just like Oscar's, except that where

[1] Here and elsewhere I follow the tradition in neglecting the fact that strictly speaking none of Oscar's molecules are identical with those of his twin, for Oscar's molecules contain H_2O while Twin Oscar's contain XYZ. The thought experiment can be changed to accommodate this fact.

Oscar has encountered H_2O, he has encountered XYZ; and neither of them is aware of the constitution of the liquids they refer to as 'water.'

The reference of Oscar's 'water' expression is different from the reference of his twin's 'water' expression. Oscar's expression refers to H_2O while Twin Oscar's refers to XYZ. Given that reference is part of the meaning, the meaning of Oscar's 'water' expressions is distinct from the meaning of Twin Oscar's 'water' expressions. And if the meaning of an expression is determined by the concepts of the speaker, then the concept expressed by Oscar's term 'water' is different from the concept expressed by Twin Oscar's term 'water.' To translate the concept expressed by the Twin Earthian word 'water' into English we have to coin a new word, perhaps *twater*. Owing to the difference in concepts, Oscar and Twin Oscar express different thoughts when both of them utter, for instance, "Gee, water is wet!" – and this in spite of their molecular identity. Putnam concludes that mental states involving natural kind terms don't supervene on physical states of our brains but on the physical states of our environment.[2] In Putnam's (1975: 227) famous phrase, "meanings ain't in the head."[3]

Let's get back to the brain-in-a-vat argument. Given content externalism, just as Twin Oscar's term 'water' does not have the sorts of causal connections to water needed for it to refer to water, a brain in a vat's usage of "vat" does not refer to vats. As Putnam (1981b: 14) puts it, "'vat' refers to vats in the image in vat-English, or something related (electronic impulses or program features), but certainly not to real vats, since the use of 'vat' in vat-English has no causal connection to real vats." If we were brains in a vat,

[2] Twin Earth cases assume a narrow notion of environment according to which a person's content-determining environment is limited to the planet he resides on. If, however, the notion of an environment is construed broadly, then the environment is made up of everything in the person's universe. Given the broad notion of environment, Twin Earth belongs to the environment of Earth, and vice versa. Oscar's word 'water' then refers to both liquids – H_2O and XYZ.

[3] Putnam's Twin Earth example exploits the fact that in the case of a natural kind term such as 'water,' the nature of the referent plays an essential role in individuating the concept associated with the word. A notorious objection to this thought experiment states, first, that it is limited to natural kind terms and, second, that it rests on the contentious idea that, just as the chemical composition (as opposed to the color, taste, density, or boiling point) is said to be the defining characteristic of water, every natural kind is supposed to have an essential property. While this is still a popular criticism of the Twin Earth example, it seems pretty clear that Putnam's overall position in "The Meaning of 'Meaning'" doesn't presuppose any substantial type of metaphysical essentialism and is intended to apply to most terms in our language (cf. Salmon 1979). Since writing "The Meaning of 'Meaning,'" Putnam (1990b: 70) has become skeptical of the search for necessary and sufficient conditions. "Is it clear that we would call a (hypothetical) substance with quite different behavior water in these circumstances? I now think that the question, 'What is the necessary and sufficient condition for being water in all possible worlds?' makes no sense at all. And this means that I now reject 'metaphysical necessity'."

then our word "vat" would not refer to vats. Consequently, the mere fact that we can raise the possibility that we are brains in a vat shows that we are not. In other words, "[i]f we can consider whether it is true or false, then it is not true . . . Hence it is not true" (Putnam 1981b: 8).

Putnam's anti-skeptical argument crucially depends on the claim that the use of 'vat' in vat-English has no causal connection to real vats. Someone who was just recently subjected to envatment by an evil scientist could have had causal contact with vats before being envatted. Content externalists agree that the conceptual shift caused by an environmental shift takes time. Only if one remains in the vat long enough is one's concept of vat substituted for one's concept of vat-in-the-image. Thus for the brain-in-a-vat argument to work it has to be assumed that the brain has been envatted from "birth."

Another embellishment that has to be added to the brain-in-a-vat scenario for the anti-skeptical argument to work is that brains in a vat don't interact with normal people. For if such an interaction were possible the brain in a vat could think the following true thought: "I am what N.N. means when she says 'Sven is a brain in a vat,'" where N.N. is a normal person whose term 'brain' refers to brains and whose term 'vat' refers to vats. Following Putnam, I use the phrase "brain in a vat" to refer to a brain in a vat from birth living in a community of fellow brains in vats.

Many formulations and (alleged) improvements on Putnam's anti-skeptical argument have appeared in the literature. For present purposes, the following reconstruction of Putnam's argument will do:[4]

(1) My language disquotes. [From disquotation principle]
(2) In vat-English, 'brain in a vat' does not refer to a brain in a vat. [From content externalism]
(3) In my language, 'brain in a vat' is a meaningful expression.
(4) In my language, 'brain in a vat' refers to a brain in a vat. [From (1) and (3)]
(5) My language is not vat-English. [From (2) and (4)]
(6) If I am a brain in a vat, my language, if any, is vat-English. [Definition of vat-English]
(7) Therefore, I am not a brain in a vat. [From (5) and (6)]

To be sure, the argument does not prove that I am not a brain in a vat; it makes the weaker point that if I were a brain in a vat I could not coherently

[4] This reconstruction is due to Wright (1994: 224). Putnam (1994a: 284) seems to endorse this reconstruction.

think that I were. But from this it does not follow that the skeptical scenario that I might be a brain in a vat could not be true. This would only follow if the brain-in-a-vat scenario violated a physical law or contained a logical contradiction. But this is not what Putnam says. His claim is that the scenario is 'self-refuting' in the sense that the fact of its being entertained makes it false.

Putnam's anti-skeptical argument is clearly valid. What isn't clear is whether we have an *a priori* entitlement to the premises and whether they are true. Some have challenged Putnam's argument by questioning the *a priori* knowability of the fourth premise (see e.g. Brueckner 1986: 103). Since I don't have *a priori* access to the environmental factors that determine what my expression 'brain in a vat' refers to, I cannot know *a priori* that in my language 'brain in a vat' refers to brains in vats. To know that I would have to know whether I am speaking English or vat-English, and I could only know that if I already knew that I was not a brain in a vat.

A different line of attack challenges the *a priori* knowability of premise (4). Some have argued that content externalism has the consequence that I cannot tell, on the basis of reflection, what my words mean and what the contents of my thoughts are. The worry is that I can only know the contents of my thoughts after having investigated the content-determining environment. Elsewhere I have argued that content externalism and privileged self-knowledge are indeed compatible, but that the notion of privileged self-knowledge reconcilable with externalism is weaker and less interesting than the Cartesian notion of self-knowledge (cf. Bernecker 1996, 1998, 2000). Here I will not defend the compatibility of externalism and self-knowledge, but rather assume its truth.

The most fundamental critique of Putnam's anti-skeptical argument concerns the truth of premise (2). Some have argued that, because it seems perfectly coherent to suppose that I am a brain in a vat, premise (2) must be false. The plausibility of skeptical hypotheses swamps the considerations in favor of content externalism. Content externalism is false because skeptical hypotheses are coherent (see, e.g., Falvey and Owens 1994; Stroud 1984).

5.2 Extending content externalism

Some have argued that content externalism doesn't go far enough in repudiating the Cartesian picture of mentality as intrinsic to the subject. Whereas content externalism locates mental states inside the head or body of an individual, the *hypothesis of extended mind* (HEM) claims that the role of the physical or social environment is not restricted to the determination of

mental content. Mental states are not only externally individuated but also externally located states. The reason they are externally located states is that they are realized by cognitive processes that are, in part, constituted by physical or bodily manipulations of environmental states. Cognitive processes are hybrid entities, made up in part of what is going on inside the brain of the creatures who have them, but also made up in part of what is going on in the environment of those creatures. Mental states are externally constituted in the sense that they are composed not only of the internal states of the individual in question but also of objects, properties, or events in his environment together with the appropriate relation connecting the two. "[T]he human organism is linked with an external entity in a two-way interaction, creating a coupled system that can be seen as a cognitive system in its own right" (Clark and Chalmers 1998: 8). In slogan form: the mind is not exclusively in the head.

The hypothesis of extended mind goes under a number of aliases. To distinguish the view expressed by HEM from content externalism it has been labeled *active externalism*, *vehicle externalism*, *enabling externalism*, *locational externalism*, and *environmentalism*. Proponents of HEM tend to embrace content externalism. Yet it is important to see that, though mutually consistent, content externalism is distinct from HEM.[5] It is one thing to say, as content externalism does, that the contents of mental states are individuated by the relations those states bear to certain environmental features. And it is quite another thing to say, as HEM does, that the cognitive processes that realize mental states are constituted in part by an individual's environment. The external individuation of mental contents does not entail the external location of mental states. And just as content externalism doesn't imply HEM, the latter doesn't imply the former. Instead of embracing content externalism, proponents of HEM could endorse some other semantic theory such as conceptual role semantics.

In their seminal paper "The Extended Mind," Andy Clark and David Chalmers provide two arguments in favor of HEM. First, they argue that the mind's cognitive processes can partially consist in processes performed by

[5] Sometimes HEM (or active externalism) is defined so as to include content externalism. Tollefsen (2006: 142n), for instance, writes: "Content externalism holds that the content of a mental state is determined by environmental or causal factors. My twin on Twin Earth does not have the same water beliefs as me because Twin Earth contains XYZ, not H_2O. The content differs. The vehicle, however, [i.e. the pattern of neurological activity] remains ... the same and inside the head. Active externalism argues that the vehicle of content need not be restricted to the inner biological realm. The idea is that both cognitive contents and cognitive operations can be instantiated and supported by both biological and non-biological structures and processes."

external devices. Consider, for instance, the cognitive process of rotating shapes when playing the game Tetris. According to Clark and Chalmers, the computer's rotation of a shape plays the same sort of role, in a person's cognitive economy, as the corresponding internal process of imagining how the shape would appear if it were rotated in various ways (see also Clark 2008: 70–3).

Clark and Chalmers's second argument in favor of HEM takes the form of a thought experiment that is meant to show that standing beliefs (such as memories and dispositional beliefs) can be partially constituted by factors external to the skin. The thought experiment involves two characters, Inga and Otto. Inga wants to visit the Museum of Modern Art in New York. At first she doesn't remember where the museum is located. She thinks for a moment and recalls that it is on 53rd Street. Clark and Chalmers note that it is uncontroversial to assume that Inga has had this belief about the museum's location all along, even if it has not always been occurrent. The belief was stored in memory, waiting to be accessed. Next consider Otto, who suffers from Alzheimer's disease. Owing to his failing memory, Otto always carries around a notebook in which he jots down information. He too decides to go to the same museum as Inga, and he too doesn't remember the location of the museum. Unlike Inga, Otto cannot retrieve the desired piece of information from his memory. Instead he consults his notebook, which says that the museum is on 53rd Street.

Clark and Chalmers argue that the information in the notebook plays the same functional role for Otto that an ordinary non-occurrent, but explicitly encoded, belief plays for Inga. We should therefore count the notebook as part of Otto's mind, and the location of the museum as one of Otto's beliefs. This inference rests on the *parity principle*:

> If, as we confront some task, a part of the world functions as a process which, were it to go on in the head, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world is (so we claim) part of the cognitive process. (Clark and Chalmers 1998: 8)

If external devices act as memory traces, then they can be treated as memory traces, notwithstanding the difference in their location. Given the parity principle, Otto has the (dispositional) belief that the Museum of Modern Art is on 53rd Street even before he consults his notebook because the notebook plays the same role for Otto that memory traces play for Inga and because any alternative explanation involving intermediate beliefs about the contents of the notebook introduces "one step too many" (Clark and Chalmers 1998: 13) to the description of Otto's mental life.

Clark and Chalmers are careful not to trivialize the proposed extension of the boundaries of the mind. The external aids that count as part of an individual's mental processing must meet the following requirements (Clark and Chalmers 1998: 17; Clark 2010: 46): they must be (i) consistently available, (ii) readily accessible, (iii) automatically endorsed, and (iv) present (internally or externally) because they were consciously endorsed in the past. Clark and Chalmers are confident that constancy, direct accessibility, and present endorsement are constitutive of believing but they concede that it is debatable whether past endorsement belongs to the necessary conditions for belief. Without the past-endorsement condition there doesn't seem to be a difference between remembering and relearning. Yet adding this condition seems to rob HEM of its attraction. For if an extended belief requires conscious endorsement and if conscious endorsement is ultimately an internal process, then HEM is not really that different from the received picture whereupon the mind does not extend beyond the skin.

A problem for HEM is that it seems to take very little to satisfy the four criteria for inclusion of external props into an individual's cognitive system. In response to the worry that, given HEM, there is no stopping the 'leakage' of mind into the world Clark and Chalmers (1998: 8) also require that "the human organism is linked with an external entity in a two-way interaction, creating a coupled system." Clark dubs this two-way interaction continuous reciprocal causation (CRC), and Menary (2007) calls it cognitive integration. "CRC occurs when some system S is both continuously affecting and simultaneously being affected by activity in some other system O" (Clark 2008: 24). Unlike conditions (i)–(iv) above, CRC is intended only as a sufficient, not a necessary condition on HEM.

5.3 Outsourcing the mind

Clark and Chalmers's Otto is not so different from a brain in a vat. Otto relies on a notebook to acquire the belief that the Museum of Modern Art is on 53rd Street. If a brain in a vat comes to believe that the Museum of Modern Art is on 53rd Street it does so because it relies on the supercomputer. In both cases there is a systematic coupling between a brain, on the one hand, and on the other hand an external device that feeds the brain information. The coupling conditions are met in both cases: (i) the notebook and the computer are reliably available and typically invoked, (ii) information retrieved from the notebook and the computer is automatically endorsed, and (iii) information provided by the notebook and the computer is easily available as and when

required. What distinguishes the case of Otto from that of the brain in a vat is the kind of external device they rely on. The brain in a vat relies on a computer while Otto relies on a notebook. We can further approximate the case of Otto to that of a brain in a vat by imagining a Twin Earth where Otto's brain has been envatted and where a supercomputer plays the role of the notebook.

How does embodied Otto differ from his envatted twin? One difference is that envatted Otto uses the laptop not only for selective tasks such as figuring out the whereabouts of the Museum of Modern Art but for acquiring all kinds of beliefs. But surely this isn't a relevant difference. It doesn't matter how often a brain forms a coupled system with an external device. Besides relying more heavily on external devices, envatted Otto differs from embodied Otto in that he is not aware of his reliance on external devices. Embodied Otto can tell whether and when he is using an external device (notebook) to form a belief. Envatted Otto, by contrast, would be very surprised if he found out about the physical and computational operations underlying his cognitive processes. But, once again, this is not a relevant difference (cf. Clark 2008: 164). HEM allows for cognitive integration to be opaque. A cognitively integrated system need not conceive of itself as such but may instead take itself to be wholly intra-cranial.

Given HEM, embodied Otto remembers the whereabouts of the Museum of Modern Art (even before he consults the notebook) because the notebook plays the same kind of functional role vis-à-vis information storage and retrieval as Inga's biological memory. Given the parity principle, if Inga's biological memory counts as part of the cognitive process *she* employs, then Otto's notebook counts as part of the cognitive process *he* employs. But just as there is a functional parity between Inga and embodied Otto, there is a functional parity between embodied Otto and his envatted twin. The vat and the connected supercomputer play for envatted Otto the same kind of functional role that the notebook and the rest of the external world play for embodied Oscar. The parity principle does away not only with unwarranted 'metaphysical bio-prejudice' (Clark 2008: xxvi) but also with the kind of techno-prejudice that makes us treat brains in vats like second-class citizens. "Otto in a vat," Clark claims, "shares all the standing beliefs of our worldly Otto. This should not come as much of a surprise because envatted Otto just *is* (functionally speaking) Otto the extended mind" (Clark 2008: 164).[6]

[6] "It would seem, then, that Otto and his duplicate will share their unquestionably mental lives, and may or may not share the aspect of their life that is anyway questionably mental" (Horgan and Kriegel 2008: 362).

To say that envatted Otto has beliefs if embodied Otto does is one thing; to say that the two Ottos hold the very same beliefs is quite another. HEM states conditions for when a system is minded; it does not specify conditions for the individuation of thought content. HEM leaves it open whether envatted Otto's beliefs have the same content as those of his embodied twin. As was explained before, HEM is compatible with different theories of mental content. When HEM is combined with a position whereby the meaning-bestowing relations are intra-linguistic or intra-cognitive,[7] then embodied Otto comes out as having the very same thought contents as envatted Otto. On this theory of mental content Putnam's semantic critique of external-world skepticism would not get off the ground.

But what happens when HEM is combined with content externalism? As was explained in Section 5.1, content externalism has it that the contents of an agent's thoughts depend on causal relations she bears to aspects of her physical and social environment. Whether envatted Otto has the same thought contents as embodied Otto crucially depends on what counts as envatted Otto's content-determining environment. Given Putnam's interpretation of content externalism, the content-determining environment of envatted Otto differs from that of embodied Oscar. This is why Putnam claims that a brain in a vat lacks the conceptual tools to entertain the thought that it is a brain in a vat. But when content externalism is combined with HEM the difference between the content-determining environment of the brain in a vat and that of a regular person disappears. For, just as the brain in a vat forms a coupled system with the supercomputer that feeds it all of its sensory-input signals, the supercomputer forms a coupled system with the evil scientist who programs it.[8] So even though the brain in a vat lacks *direct* causal contact with real brains and real vats, given HEM, it has *indirect* causal contact with real brains and real vats, for it has direct contact with the evil scientist who has direct contact with real brains and real vats. So, given HEM, the brain in a vat is sufficiently connected to the external world to 'speak' English and to have the same thoughts as the evil scientist.

To see that the combination of content externalism and HEM has the consequence of leveling the semantic difference between envatted and

[7] Examples of internalist semantics are meaning holism, functional role semantics, conceptual role semantics, causal role semantics, and the like.

[8] "Envatted-brain scenarios usually involve a systematic coupling between the brain, on the one hand, and on the other hand the external device (usually envisioned as a super-computer) that monitors the brain's motor-output signals and feeds it sensory-input signals. Arguably, the states of the external device could be regarded as aspects of an extended mind that is physically realized by the coupled brain–computer system as a whole" (Horgan and Kriegel 2008: 370–1, fn 30).

embodied brains it is helpful to compare the evil scientist in Putnam's skeptical scenario with Clark and Chalmers's Otto who is suffering from Alzheimer's disease. Just as the notebook is a record of Otto's thoughts, the brain in a vat is a record of the evil scientist's thoughts. And just as Otto can consult the notebook to find out what he used to think, we can consult the brain in a vat to find out what the evil scientist is thinking right now and can consult the brain in a vat's memories to find out what the evil scientist used to think. But if the notebook counts as an extension of Otto's mind, as Clark and Chalmers claim, then the brain in a vat should count as an extension of the evil scientist's mind. For the brain in a vat is like a mirror website which republishes information verbatim from another (originating) site without exercising independent editorial control. And if the envatted brain's mind is an extension of the evil scientist's mind, then the envatted brain's thought contents don't differ from those of the evil scientist. For what makes it the *same* mind (or the same mental process) that is located both in the evil scientist's head and in the brain in a vat? Intuitively it is the same mind if it thinks the same thoughts. And two thoughts are the same if they have the same content. Thus when applied to the brain-in-a-vat scenario HEM suggests that the evil scientist and the brain in a vat think the same thoughts. But if there is no difference in thought contents between the brain in a vat, on the one hand, and an ordinary person who has causal contact with brains and vats, on the other, then Putnam's semantic critique of external-world skepticism crumbles.

Let's go over the argument once again. We have started from the observation that the brain in a vat forms a coupled system not only with the supercomputer but also with the evil scientist operating the computer. The second step of the argument is to realize that coupling or cognitive integration is taken to be a sufficient condition on HEM. This was explained in Section 5.2. The upshot of the first two steps is that the evil scientist's mind extends to the brain in a vat. The third step of the argument is to say that the evil scientist's mind can only extend to the brain in a vat if they think the same thoughts. Sameness of mind requires sameness of thoughts. Two thought tokens belong to the same type if they have the same content. Thus, for the mind of the evil scientist to extend to the brain in a vat the scientist's brain and the envatted brain have to entertain the same thought contents. The fourth step says that since the evil scientist has the required causal contact with brains in vats to think the thought *I am a brain in a vat* and since the brain in a vat thinks the same thoughts as the evil scientist it too thinks the thought *I am a brain in a vat*. And if we assume that the indexical 'I' refers to the local

realization of the thought in question, then the brain in a vat is having a true thought when it thinks that it is a brain in a vat. So, contrary to Putnam's contention, it is not the case that the thought *I am a brain in a vat* is necessarily false. When content externalism is combined with HEM it loses its anti-skeptical punch.

Note that it doesn't help to change the skeptical scenario by increasing the distance between the evil scientist and the brain in a vat. We can imagine a version of Putnam's famous Twin Earth where 'brain' refers to something other than brain, say, computer programs, and where vat refers to something other than vat, say, vat holograms. We can further suppose that the evil scientist and the supercomputer are located on Earth while the brain in a vat is located on Twin Earth. The brain in a vat on Twin Earth is connected to the computer on Earth by a very long cord. Would changing the skeptical scenario in this way make a difference regarding the brain in a vat's thought contents? The answer is "no." To the extent that the brain in a vat forms a coupled system with the evil scientist, both of them think the same thoughts, regardless of the distance between them.

Given HEM, for the brain in a vat to lack the conceptual resources necessary to think the thought *I am a brain in a vat* the evil scientist who is coupled with the brain in a vat via the supercomputer would also have to be unable to think the thought *I am a brain in a vat*.[9] But under what circumstances would the evil scientist be unable to think this thought? One possibility is that the evil scientist who programs and operates the computer is Davidson's Swampman. This is Davidson's Swampman thought experiment:

> Suppose lightning strikes a dead tree in a swamp; I am standing nearby. My body is reduced to its elements, while entirely by coincidence (and out of different molecules) the tree is turned into my physical replica. My replica, the Swampman, moves exactly as I did; according to its nature it departs the swamp, encounters and seems to recognize my friends, and appears to return their greetings in English. It moves into my house and seems to write articles on radical interpretation. No one can tell the difference. (Davidson 2001: 19)

[9] A critic might object that the referents of 'I' are different when the evil scientists thinks *I am a brain in a vat* and when the brain in a vat thinks *I am a brain in a vat*. This would mean that the evil scientist and the brain in a vat cannot share indexical thoughts. I disagree. Following Fost (2013), I think that we sometimes think of our identities as being extended, in more or less the sense of HEM. When the evil scientist talks about 'I' he is referring not only to himself but also to the coupled system consisting of the supercomputer and the brain in a vat.

But, Davidson argues, there is a difference because the Swampman cannot "be said to mean anything by the sounds it makes, nor to have any thoughts." The reason is that Swampman lacks the proper causal-historical connections to the world which underpin meaning and thought content.[10]

We can combine the brain-in-a-vat scenario with the Swampman scenario by imagining that the supercomputer has been built and is operated by Swampman. The brain in a vat has now turned into a *brain in a swamp*. If Swampman does not have thoughts, neither does the envatted brain that is 'cognitively' integrated with Swampman. So if we assume that the supercomputer is built and operated by a mindless creature like Swampman, the hypothesis of the extended mind doesn't get a foothold. For if there is no mind in the first place, then there isn't anything that can be extended to the envatted brain.

Putnam too considers a version of the brain-in-a-swamp scenario, for he worries that a brain in a vat still has indirect causal connections to real vats and real brains through the evil scientist who programs the supercomputer that feeds the brain in a vat sensory information. To sever this last connection a brain in a vat has with the real world Putnam modifies the scenario in such a way that the vat and automated machinery are no longer designed by an intelligent scientist, but rather are "supposed to have come into existence by some kind of cosmic coincidence" so that they "have no intelligent creator–designers" (Putnam 1981b: 12). Putnam's point is that that if brains in *vats* don't have our concepts neither do brains in *swamps*.

> For there is no connection between the *word* 'tree' as used by these brains and actual trees. They would still use the word 'tree' just as they do, think just the thoughts they do, have just the images they have, even if there were no actual trees. Their images, words, etc., are qualitatively identical with images, words, etc., which do represent trees in *our* world.[11] (Putnam 1981b: 12–13)

The skeptical worry raised by the brain-in-a-swamp scenario differs from that raised by the brain-in-a-vat scenario. The brain-in-a-vat scenario illustrates

[10] According to Hawking (1993: 112–13), Swampman is a physical possibility. Millikan (1984: 93) says of her Swampwoman: "that being would have no ideas, no beliefs, no intentions, no aspirations, no fears and no hopes . . . this because the evolutionary history of the being would be wrong."

[11] Horgan and Kriegel (2008: 371) imagine a scenario similar to the brain-in-a-swamp scenario discussed here: "Let the pertinent scenario be an envatted brain that receives randomly-generated inputs from a long-term surrounding electrical storm that just happens, via cosmic coincidence, to give the brain an ongoing phenomenal mental life that exactly matches yours."

the possibility that my thoughts have determinate contents but that these contents differ from those of its embodied duplicate thereby making it impossible for the brain in a vat to even consider the possibility of external-world skepticism. The brain-in-a-swamp scenario, by contrast, illustrates the possibility that I am a propositional zombie that lacks a mind and lacks thoughts. As I have argued elsewhere (Bernecker 2000, 2004) content externalism makes it impossible to rule out zombie skepticism. For if mental content consists in the nomic dependence of brain states on environmental conditions, a type of configuration in my brain is a mental state only if it is normally tokened by some environmental condition. But I cannot know *a priori* whether there is any systematic relationship between my internal states and states of an external kind. I am unable to know *a priori* whether my so-called thoughts are indeed thoughts as opposed to some other kind of states lacking mental content. Content externalism implies that I am unable to know *a priori* that I am a minded being rather than a propositional zombie.

Elsewhere (Bernecker 2000, 2004) I have argued that even though I cannot know *a priori* I am having thoughts as opposed to contentless states I *can* know *a priori* what my thoughts are about, provided they have content. I can know what I am thinking – the specific content of the thought – without knowing that I am thinking – the fact that my so-called thoughts are not contentless states. *A priori* knowledge of what I am thinking about is consistent with me lacking the ability to rule out, on the basis of *a priori* reasoning, the possibility that I don't have any propositional attitudes. That's why the inability to rule out the brain in a swamp scenario doesn't undermine the kind of self-knowledge presupposed by the fourth premise of Putnam's brain-in-a-vat argument.

Let's take stock. We saw that adding HEM to content externalism has the consequence of undermining Putnam's refutation of the brain-in-a-vat scenario. For given HEM, the brain in a vat is sufficiently connected to the external world to 'speak' English and to entertain the thought *I am a brain in a vat*. But if the brain in a vat can think the thought that it is a brain in a vat then Putnam is wrong in claiming that external-world skepticism is self-refuting. For the brain in a vat to not have the ability to think the thought that it is a brain in a vat no one may be interacting with the supercomputer who is able to think this thought. One way of spelling out this possibility is to imagine a brain in a swamp. Content externalism is unable to rule out the brain-in-a-swamp scenario.

The upshot is that content externalism loses its anti-skeptical force when it is combined with HEM. And if it is a motivation for content

externalism that it refutes the brain-in-a-vat scenario, then adding HEM makes content externalism less appealing. So contrary to Clark's (2008: 78) contention, HEM is not "orthogonal to the more familiar Putnam–Burge style externalism"; instead HEM and content externalism are in tension with one another.

5.4 Embedding the vat

The hypothesis of the extended mind (HEM) states that the boundaries of the mind extend beyond the boundaries of individual organisms. There are a number of theses in the vicinity of HEM that are easily mixed up with HEM. The thesis of *embodied cognition* states that cognition depends not just on the brain but also on the body. The thesis of *enactive cognition* proposes that cognition is not something that occurs inside of an agent, but is a product of the interaction between agents and their environment. The thesis of *distributed cognition* states that cognition and knowledge are not confined to an individual but are distributed across objects, individuals, artefacts, and tools in the environment. Each of these theses contributes to a picture of mental activity as dependent on the context in which it occurs, whether that context is relatively local (as in the case of enactivism and embodiment) or relatively global (as in the case of distribution and extension). According to my usage, *embedded cognition* is the genus, and embodied, enactive, embedded, and distributed cognition and their ilk are species. This usage isn't standard but it seems as good as any.

In the previous section we have seen that, assuming the extended mind thesis, the brain-in-a-vat scenario is coherent but Putnam's argument against this scenario fails. The goal of this section is to show that, assuming the thesis of embedded cognition, the brain-in-a-vat scenario is incoherent.

In Putnam's version of the brain-in-a-vat story, one is asked to imagine that one's brain has been removed from one's body and placed in a vat of nutrient fluids, and that all of its normal neural inputs and outputs are being simulated by a supercomputer. The brain has no way of knowing whether it is in a skull or in a vat. Can we be sure that this is not our current situation? How do we know that anything beyond our brains is real rather than virtual? The moral of the brain in a vat story is that the neural representations of body and world are only indirectly related to real external things. But for the brain-in-a-vat story to make sense we have to assume that a suitably working human brain is sufficient all on its own for the instantiation or realization of

our mental life. And this very assumption is challenged by the thesis of embedded cognition. Clark writes:

> The mistake, then, is to infer that the sufficient mechanism is the biological stuff alone, just because the biological stuff, in the special vat-context, helps support thinking and experience. At the limit of this thought experiment we have the single neuron in a dizzyingly complex vat ... We would not conclude that experience and thought constitutively depend only on the activity of that single neuron. (Clark 2009: 980–1; see also Clark 2008: 163–4; Shapiro 2004: 218; 2011: 162)

Assuming the thesis of embedded cognition, the mind is not realized simply in the brain but its realization is integrated throughout the body. Mental processes are not the deliverances of only the brain, but a brain that is intimately enmeshed with a body. Damasio explains:

> It might be argued that if it were possible to mimic, at the level of the dangling nerves, realistic configurations of inputs as if they were coming from the body, then the disembodied brain would have a normal mind. Well, that might be a nice and interesting experiment 'to do' and I suspect the brain might indeed have *some* mind under those conditions. But what that more elaborate experiment would have done is create a body surrogate and thus confirm that 'body-type inputs' are required for a normally minded brain after all. And what it would be unlikely to do is make the 'body inputs' match in realistic fashion the variety of configurations which body states assume when those states are triggered by a brain engaged in making evaluations. (Damasio 1994: 228; see also Heylighen 2012)

If the mind depends on the body, then envatment is impossible. And if it were possible for an isolated brain to have our mental life, then the entire body and local environment would have to be envatted as well. Thompson and Cosmelli remark:

> Careful examination of this thought experiment indicates that ... any adequately functional 'vat' would be a surrogate body, that is, that the so-called vat would be no vat at all, but rather an embodied agent in the world. Thus, what the thought experiment actually shows is that the brain and body are so deeply entangled, structurally and dynamically, that they are explanatorily inseparable. Such entanglement implies that we cannot understand consciousness [and cognition] by considering only the activity of neurons apart from the body, and hence we have good explanatory

> grounds for supposing that the minimal realizing system for consciousness [and cognition] includes the body and not just the brain.[12] (Thompson and Cosmelli 2011: 163)

So given the thesis of embedded cognition, the vat and the supercomputer have to do all the complex work of body, action, and world. But the vat can only do the work of the body, action, and world if the vat contains the body and the world. And if the vat contains the world, the brain in vat is not cut off from the world in the way suggested by the skeptic. Envatting the world thus goes hand in hand with disarming the skeptic. So in the end we are left with a kind of dilemma: If the vat *does not* fill in everything the world provides, then the brain-in-a-vat scenario is incoherent; and if the vat *does* fill in everything the world provides, then the skeptical threat evaporates.

5.5 Conclusion

I have argued for two conditional claims. If content externalism is combined with the extended mind thesis, Putnam's refutation of the brain-in-a-vat scenario collapses. And if content externalism is combined with the embedded mind thesis, the brain-in-a-vat scenario ceases to pose a skeptical problem. Provided these claims can be believed there are a number of conclusions we can draw. First, we can either reject the embedded mind thesis because it renders the brain-in-a-vat scenario incoherent or reject the brain-in-a-vat scenario because it violates the embedded mind thesis. Second, we can either reject the extended mind thesis and hold on to Putnam's anti-skeptical argument or embrace the extended mind thesis and reject Putnam's anti-skepticism. Third, we can simply abandon the brain-in-a-vat story and work with a skeptical scenario that does not violate extension, embedding, and content externalism. A case in point is Descartes's waking dream scenario. In this scenario my words and thoughts do refer to objects in my

[12] "When theorists invoke the notion of a brain in a vat, they invariably take a unidirectional control perspective and view the brain as a kind of reflexive machine whose activity is externally controllable. Yet numerous neurobiological considerations count against this viewpoint and indicate that the brain needs to be seen as a complex and self-organized dynamical system that is tightly coupled to the body at multiple levels" (Cosmelli and Thompson 2010: 362; see also Thompson and Stapleton 2009: 27). "So I hold that the supposed consciousness of a causally detached brain – say, a living brain floating listlessly in a vat, as in Hilary Putnam's famous thought-experiment – even though it seems both conceivable and logically possible, just would not be a consciousness like ours. On our view, a consciousness necessarily involves a brain that is causally-dynamically coupled with all the other vital systems, organs, and processes of our living body" (Hanna 2011: 21).

environment and my mind is entirely within the physical boundaries of the brain and skull. The third option strikes me as the most promising. But this is not the point of the paper. The point of the paper is that buying into the content externalist's critique of brain-in-a-vat skepticism comes at the price of rejecting popular theories of the mind that have grown out of content externalism.

I have assumed, and not argued for, the truth of the extended mind thesis.[13] But even if the extended mind thesis turns out to be indefensible, the problems for Putnam's anti-skepticism don't go away. For on some versions of content externalism one can possess the concept X even if one has never had causal contact with the referent of X. Burge (1982), for instance, claims that the inhabitants of Putnam's waterless Twin Earth *can* have water thoughts provided that there exists sufficient knowledge of chemistry among the Twin Earthian experts to distinguish water from various twin-concepts such as twater.[14] These experts qualify for possession of the concept *water*. And if dry-Oscar defers to the experts when it comes to his use of 'water,' then he too qualifies for possession of the concept *water* despite not knowing anything about chemistry and never having interacted with water. Now assuming Burge's version of content externalism, we can imagine a community of brains in vats theorizing about brains (as opposed to brains-in-the-image) and vats (as opposed to vats-in-the-image) and thereby coming to possess the concepts *brain* and the concept *vat* despite never having had (direct) contact with brains and vats. Thus we don't even need to buy into the extended mind thesis to conclude that a brain in a vat can think the true thought *I am a brain in a vat*.

[13] In Bernecker (2014) I argue for the following conditional claims: if the extended mind debate is a substantive dispute, then we have only superficial understanding of the extended mind hypothesis. And if we have deep understanding of the extended mind hypothesis, then the debate over this hypothesis is nothing but a verbal dispute.

[14] Burge is not alone. See also Ball (2007), Goldberg (2006a), Korman (2006), and McLaughlin and Tye (1998a).

The Brain in a Vat

Edited by

Sanford C. Goldberg

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107069671

First published 2016

Printed in the United Kingdom by Clays, St Ives plc

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication Data
The brain in a vat / edited by Sanford C. Goldberg.
pages cm. – (Classic philosophical arguments)
ISBN 978-1-107-64338-3
1. Philosophy of mind. 2. Metaphysics. 3. Knowledge, Theory of. I. Goldberg, Sanford, 1967– editor.
BD418.3.B723 2016
128'.2–dc23 2015026793

ISBN 978-1-107-06967-1 Hardback
ISBN 978-1-107-64338-3 Paperback

Contents