1 Introduction: Putnam's reflections on the brain in a vat

Sanford C. Goldberg

1.1

In 1981, Hilary Putnam published a paper entitled "Brains in a Vat." In it, he reflected on a "science fiction possibility discussed by philosophers," which he describes as follows:

> 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 travelling 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 seem to 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)

This scenario, which will henceforth be called the brain-in-a-vat (or BIV) scenario, has generated a tremendous amount of philosophical commentary in the more than three decades since its publication.

Putnam's reflections on the BIV scenario have a familiar historical precedent, of course, in Descartes's reflections on the Evil Demon scenario. As Descartes had described that scenario in *Meditations on a First Philosophy*, an Evil Demon who is as evil as he is powerful aims to dupe you as much as possible, and so stimulates your mind so as to give you normal-seeming sensations and "experiences," under conditions in which there is no world beyond your mind at all (just the Evil Demon). Descartes employed the thought experiment in the course of employing the method of *systematic doubt*, through which he hoped to discern the "foundations" of all that he knew. Through this method he aimed to rid himself of any belief that was dubitable – any belief that *could* be doubted – until such time as he could certify the truth of the belief (in which case, and only in which case, he would endorse the belief). Accordingly, the Evil Demon scenario served two purposes for Descartes. The first was that it enabled him to determine the extent of what could be doubted: any belief that the Demon could render false is a belief that could be doubted. On reflection, it turned out that this included the belief in a world external to one's own mind, and the corresponding belief in the existence of objects and properties in that world. The second purpose served by the Evil Demon scenario was that of a heuristic. Having the Evil Demon in mind was a constant reminder that Descartes needed to take care not to simply endorse what he found himself naturally believing: here the thought of the possibility of an Evil Demon reminded him of the need for constant vigilance, lest he allow to slip into his system something that could be doubted (and hence might be false).

In contrast to Descartes's primarily epistemological motivations, Putnam's aim in reflecting on the BIV scenario was to defend some points about the nature of intentionality and related matters – in particular, (mental and linguistic) reference and representation. He asked the question, "Could we, if we were brains in a vat in this way, *say* or *think* that we were?" (7; original italics) The burden of his argument was to answer this in the negative. At the same time, he anticipated that his argument would have some implications not only for semantics and the philosophy of mind and language but also for epistemology (the nature and limits of our knowledge of the world; the nature of our knowledge of the meanings of our language and the contents of our thought) and metaphysics (metaphysical realism; the nature of truth).

Putnam's argument for the conclusion that a BIV could not say or think that it was a BIV appealed to the conditions on reference. Starting the paper off by repudiating what he called the "magical theory of reference," Putnam went on to argue that for a subject to be able to refer to an object or a property

in thought or speech requires some sort of *causal contact* with the object or property. With this condition on reference in place, Putnam contends, neither a BIV nor an unenvatted (English-speaking) subject can say or think a truth with "I am a BIV." Consider first the BIV. Because the BIV is not in causal contact with the items of our ordinary environment – no tables, trees, or turtles exist in its environment – the causal contact principle forces us to hold that, when a BIV speaks of "tables" or "trees" or "turtles," we must regard its thought and talk as regarding something *other than* tables or trees or turtles. Perhaps it is referring to internal features of the computer. Alternatively, perhaps it is referring to features of its sensory or "perceptual" experiences: not trees, but whatever it is in the envatted circumstance that causally prompts the BIV's uses of "tree" – say, trees-in-the-mental-image; not tables, but tables-in-the-mental-image; not turtles, but turtles-in-the-mental-image. But the same point holds for a BIV's use of "brain in a vat": since it has not been in causal contact with any such thing, when it thinks "I am a brain in a vat," it should not be construed as referring to (or representing itself as) a brain in a vat, but as something else – in which case its thought is *false*. Of course, if the subject is an ordinary (unenvatted) English-speaking subject, then, while such a subject can use "brain in a vat" to refer to or pick out a brain in a vat, even so, the thought such a subject expresses with "I am a BIV" – namely, the thought that she herself is a BIV – is false. So we see that the thought expressed by "I am a BIV" is false whether the subject is a BIV or an ordinary (English-speaking) subject. Since this argument is perfectly general – the conclusion depends only on generic considerations about the nature of mental and linguistic reference in any natural language – the result is that *no one can truly think of oneself that one is a BIV*. Or so Putnam argued.

It is worth noting that, as offering an argument for a causal contact condition on linguistic representation, Putnam's paper "Brains in a Vat" reinforced a conclusion he had already famously advanced in his 1975 paper "The Meaning of 'Meaning.'" In that paper Putnam had sought to show that "meanings ain't in the head," by which he meant that the meanings of our words depend for their individuation on facts that are "external" to the speaker's bodily states. Putnam's preferred formulation of this claim appealed to the possibility of two doppelgängers – subjects who are type-identical as far as the history of their bodily states goes, and whose phenomenological experiences are subjectively indistinguishable. In these terms, the thesis that meanings "ain't in the head" was formulated as the thesis that there could be two doppelgängers who, owing to having grown up in different environments,

think different thoughts when each thinks a thought which he would express with "Water is thirst-quenching": one of them refers to (linguistically represents) *water* (H_2O) as being thirst-quenching, the other refers to (linguistically represents) *another liquid* – call it twin-water (which shares superficial qualities with water, but which is of a different chemical kind) – as thirst-quenching. This difference in the meaning of 'water' in their respective idiolects reflects differences in the liquid kinds with which they have interacted in their respective environments: one has applied 'water' to water, the other has applied 'water' to the other (water-like) liquid kind. These facts about the environmental differences are relevant to determining what each doppelgänger means by 'water'; hence the meaning of 'water' "ain't in the head" (of either one). Such a view came to be known as a *semantic externalist* view about linguistic meaning.

In comparison with Putnam's (1975) argument for semantic externalism, two developments are noteworthy with respect to the (1981) argument in "Brains in a Vat." The first is that the 1981 argument extended the semantically externalist conclusion beyond the case of linguistic meaning, to the contents of thought. Those familiar with the history of these discussions will know that Burge (1979) had already argued for such an extension, albeit by appeal to another consideration first presented in Putnam's 1975 paper: the division of linguistic labor. Putnam's 1981 paper argues for its externalist conclusion about the contents of thought on the basis of cases that have nothing to do with the division of linguistic labor. He asks us to imagine a scenario in which ants trailing across wet send leave tracks that compose an image that is perceptually indistinguishable from a caricature of Winston Churchill (published in some newspaper); Putnam argues that, even so, if some alien from another planet were to view this image, the alien would not be mentally representing or thinking of Churchill, even if the alien's sensory state would be subjectively indistinguishable from a state in which it observed the caricature in the paper. The content of the alien's mental state depends on which item it has perceived – the ant trail or the caricature. Nor is this the only difference between Putnam's 1981 argument involving the BIV and his earlier 1975 arguments from "The Meaning of 'Meaning.'" A second difference is this: in the 1981 paper, Putnam explicitly uses his reflections on reference to develop what appears to be a profoundly new sort of anti-skeptical argument. What is more, as Putnam and others saw, this argument also appears to have implications for metaphysics – in particular, for the (im)-possibility of *metaphysical realism*, and for the lessons to be drawn from the possibility of non-standard models in model theory. That a thesis about the

conditions on reference might have such dramatic epistemological and metaphysical implications perhaps explains why Putnam's reflections on the brain in a vat have attracted a good deal of attention.

I have organized the contributions to this volume to reflect the three main areas where the literature has discussed Putnam's reflections: intentionality and the philosophy of mind and language; epistemology; and metaphysics (including discussions of the implications of the so-called model-theoretic argument). Of course, the issues raised in one ramify to touch topics in the others, and this is something that will be seen throughout the papers in this volume. Still, it is helpful to group the papers according to the main contributions they seek to make. I begin with the papers in the philosophy of mind and language, as these go to the heart of the sort of argument Putnam was trying to make in 1981; and we move from there to epistemology, and then on to metaphysics.

1.2

It is no surprise that Putnam's semantic reflections on the BIV scenario generated discussions in the theory of intentionality (especially in the philosophy of mind and language), and it is with papers addressing these topics that this volume begins. (Other themes from philosophy of mind and language will loom large in several of the papers in the metaphysics section as well.)

One of the earliest influential assessments of the success of Putnam's argument was by Tony Brueckner. It is fitting to lead off this volume with a chapter from him, both to honor his memory – he passed away more than a year before this volume made it to press – and to acknowledge his central role in leading the assessment of the BIV-based arguments for anti-skepticism. In a series of very influential papers starting in the 1980s (see especially Brueckner (1986) and (1992a)), Brueckner sought to clarify the nature of the argument, and hoped to identify the (philosophy of mind and language) assumptions that underlie Putnam's anti-skeptical use of the BIV thought experiment. Brueckner originally and frequently did so in an attempt to *criticize* Putnam's argument; the charge of his seminal (1986) paper was that the argument was question-begging, and he often concluded that the argument failed to establish a serious anti-skeptical conclusion. However, in his contribution to this volume, "Putnam on brains in a vat," Brueckner limits himself to a re-assessment of the argument and various objections that he (and others) have levelled against it. (Or rather he re-assesses the *two* arguments through

which the BIV scenario can be used to support an anti-skeptical conclusion, and he raises and replies to various objections to these ways.) As this re-assessment focuses on objections to assumptions in the philosophy of language, Brueckner's chapter serves to introduce the reader to a variety of topics in the philosophy of language, as these arise in connection with an assessment of the success of Putnam's anti-skeptical argument. Brueckner concludes his contribution by noting, as he had argued in his (1986) paper, that Putnam's anti-skeptical argument has certain limitations: it cannot be used to show that you are not a *recently envatted* BIV.

In his contribution, "How to think about whether we are brains in vats," Gary Ebbs aims to identify precisely Putnam's objective in his reflections on the BIV scenario. Ebbs begins by noting that in "Brains in a Vat" (Chapter 1 of *Reason, Truth, and History*), Putnam argues that, even if it is physically possible that there be brains that are always in a vat, it cannot actually be true that we are always brains in a vat. But Ebbs thinks that Putnam's readers have misunderstood the purport of this argument. The best-known reconstructions of the argument assume that its goal is to rule out *a priori* the supposedly coherent possibility that we are always brains in a vat.[1] While Ebbs agrees that this sort of argument cannot succeed, he thinks this was not the purport of the argument. On the contrary, the purport of that argument was "to dissolve a problem that apparently arises from *within* [one's] own *non*-skeptical account of the methodology of inquiry." Ebbs concludes that, so interpreted, Putnam's argument is both illuminating and successful.

Many philosophers (including several in this volume) suppose that Putnam's BIV-based anti-skeptical argument is motived by appeal to semantic externalism in the philosophy of mind and language. Those who endorse this supposition might well think that anyone who rejects the doctrine of externalism about the contents of thought cannot exploit the anti-skeptical conclusion Putnam sought to establish through his reflection on the BIV. In his contribution, "Brains in vats, causal constraints on reference and semantic externalism," Jesper Kallestrup takes aim at this impression. Kallestrup argues that we ought to distinguish the claim that there are causal *constraints* on reference from the doctrine known as the *causal theory of reference*; and he goes on to argue that, whereas Putnam's BIV-based anti-skeptical argument relies only on the former, externalism

[1] Although Brueckner's chapter in this volume does not weigh in on whether Putnam's argument succeeds in establishing a conclusion against the most *radical* forms of skepticism (according to which you are *and have always been* a BIV), many of his other papers had done just that. For a recent review by Brueckner himself, see Brueckner (2012).

about the mental relies on the latter. The result, according to Kallestrup, is that one need not be a semantic externalist to endorse Putnam's anti-skeptical argument. Kallestrup goes on to point out that there are theorists who embrace a causal constraint on reference, yet who reject the causal theory of reference; so if his argument from this chapter is correct, such theorists can embrace Putnam's anti-skeptical conclusions even as they reject semantic externalism about the mind.

Sven Bernecker's contribution to this volume, "Extended minds in vats," seeks to resist the anti-skeptical thrust of Putnam's argument. However, where many others do so by criticizing one or another of the argument's assumptions in philosophy of language or mind (see e.g. the contributions to this volume by Folina, Douven, Sher, and Sundell), Bernecker does so by appeal to an auxiliary thesis in the philosophy of mind. The auxiliary thesis in question is the *extended mind hypothesis*, according to which the mind as a cognitive system sometimes "extends" to include features of the environment. Such a view is often defended on the grounds that the manipulation of these environmental features is itself part of the information-processing that is done by the mind *qua* cognitive system. Bernecker motivates his appeal to this auxiliary claim by arguing, first, that Putnam's BIV-based anti-skeptical argument trades on semantic externalism about the mind, and second, that at least some of those who embrace semantic externalism also embrace the doctrine of the extended mind.[2] At the very least, then, Bernecker's argument targets such folks; it is with them in mind that he writes that "Given the extended mind hypothesis, the supercomputer and the envatted brain can be regarded as aspects of the extended mind of the evil scientist." The burden of Bernecker's chapter is to show that under these assumptions the distinctly anti-skeptical thrust of Putnam's reflections on the BIV "is lost."

1.3

Another part of the vast literature generated by Putnam's reflections on the BIV scenario concerns the anti-skeptical purport of these reflections. Here, several issues were discussed in the literature in epistemology.

In Chapter 6, "Putnam on BIVs and radical skepticism," Duncan Pritchard and Chris Ranalli argue that Putnam's reflections on the BIV fail in their anti-skeptical ambitions. However, unlike those attempts to show this by

[2] The doctrine of the extended mind goes beyond standard semantic externalism about the mind in that, where the latter is a thesis about the individuation of mental states, the former is a thesis about the mind *qua* cognitive system itself.

appeal to the falsity of one or another assumption about the nature of language or thought, Pritchard and Ranalli argue that it is the very transcendental pretensions of the argument that get it in trouble. In particular, they argue, Putnam's "proof" that we are not BIVs appears to fall victim to one of two influential criticisms of so-called transcendental anti-skeptical arguments. After revisiting Brueckner's (1986) reconstruction of Putnam's BIV-based anti-skeptical argument, Pritchard and Ranalli follow Crispin Wright in maintaining that Brueckner's reconstruction of the argument had misidentified the problematic nature of the argument. Whereas Brueckner (1986) thought that Putnam's "proof" was problematic for depending on an assumption pertaining to the identity of the language in which the argument was framed,[3] Pritchard and Ranalli argue that the problematic nature of the "proof" has to do with its transcendental nature. In particular, Pritchard and Ranalli argue that, as a transcendental argument, Putnam's BIV-based anti-skeptical argument appears to fall victim to remarks that Barry Stroud and Tom Nagel have directed against transcendental anti-skeptical arguments.

Another epistemological issue which has been much discussed in connection with the BIV scenario, and which touches on issues of skepticism, concerns the nature of epistemic justification itself. Admittedly, the discussion in question (to be described in what follows) is intelligible only if we can make sense of the radical skeptical possibility in which one oneself is a BIV, and so only if Putnam's "proof" to the contrary fails. Even so, it is worth considering this line of reflection on the BIV, if only because it has played such a central role in thinking about the nature of epistemic justification.

Consider Descartes's version of the scenario involving the Evil Demon. As noted above, Descartes had used this scenario to make vivid the idea that your perceptual beliefs regarding particular features of the external world, as well as your general belief in a world of objects and properties "external" to your mind, are more susceptible to skeptical doubts than you might have realized. In making this plain, Descartes's agenda was to render intuitive the claim that if any of these beliefs of yours are to be justified, you must first be justified in believing that the Evil Demon scenario itself is non-actual. But Stew Cohen (1984) came up with an alternative use for the Evil Demon scenario. Comparing you to your envatted twin doppelgänger, Cohen highlighted how intuitive it is to suppose that, given the subjective indistinguishability of your and your doppelgänger's internal history and perceptual experiences, your and your doppelgänger's perceptual beliefs are *equally*

[3] Brueckner's contribution to this volume acknowledges that this diagnosis was faulty.

well justified. If this is so, then whether or not you are a BIV envatted by the Demon *does not matter to how well justified your empirical (perceptual) beliefs are*. This intuition, which has been dubbed the "new" Evil Demon intuition, was used by Cohen to criticize reliabilist accounts of epistemic justification, according to which a belief is justified only if it was formed through a process that reliably produces true beliefs. Cohen's own favored view of epistemic justification was an "internalist" one, according to which the facts that determine one's degree of justifiedness supervene on one's "internal" (non-factive) mental states.

In Chapter 7, "New lessons from old demons: the case for reliabilism," Thomas Grundmann takes aim at this use of the BIV scenario. To do so he explicitly targets both Descartes's use of the Evil Demon scenario as well as Cohen's "new" use of that scenario. Grundmann argues that, contrary to what many epistemologists appear to suppose, considerations like those that motivated Descartes can be used to motivate reliabilism; and he argues further that reliabilism can accommodate something very much in the spirit of the intuition elicited by Cohen's "new" Evil Demon scenario.

One of the more robust discussions generated by Putnam's reflections on the BIV concerned the interplay of issues regarding external world skepticism and authoritative self-knowledge of the contents of one's thoughts and the meanings of one's terms. Admittedly, this discussion was generated even prior to Putnam's reflections on the BIV scenario; it began as early as his arguments for semantic externalism in Putnam (1975).[4] But it is clear that the worries discussed there were seriously reinforced by the would-be anti-skeptical proof offered in Putnam (1981b). What is more, the (1981) argument also suggested a renewed formulation of the problem. The worry itself can be stated, at least to a first approximation, as follows. Suppose that Putnam's "proof" succeeds at showing that anyone who utters (or thinks the thought expressed by the English sentence) "I am a BIV" says or thinks something false. Suppose further that, by reflecting on this perfectly general result, one can acquire the first-personal knowledge to the effect that one oneself is not a brain in a vat. Then it would seem that, given the nature of the considerations involved, one who reasons through the argument can come to

[4] I should note that *Putnam himself* did not think that his BIV-based argument raised problems for authoritative self-knowledge of one's thoughts or meanings. On the contrary, in his reflections on the BIV scenario, he assumed that subjects have such authoritataive self-knowledge, and used this assumption to draw conclusions about the nature of thought and language in relation to the world. But many other authors worried about this; and it is a topic that shows up in many of the contributions to this volume (for which see footnote 6).

know *a priori* (or at least from the armchair) that one oneself is not a BIV. But the impression remains that one *cannot* know this *a priori* (nor can one know it from the armchair). If this impression is correct, it underwrites a *modus tollens* inference. In particular, we can appeal to the impossibility of *a priori* or armchair knowledge that one oneself is not a BIV, in order to conclude that one of the two suppositions above must be false: either Putnam's "proof" is no such thing; or else what one knows, when one knows that "I am a BIV" is false, *is not tantamount to knowing that one oneself is not a BIV*. The latter option holds if (and perhaps only if) we call into question the subject's knowledge of what she herself has expressed with "I am not a BIV," and correspondingly if we call into question the subject's knowledge of the meaning of "BIV" in her mouth, as well as the thought content she herself expresses when she utters "I am not a BIV." Hence it appears that the "proof" succeeds only at the cost of calling into question the subject's authoritative knowledge of the meanings of her words and the contents of her thoughts.

As I say, this sort of worry has a history that goes back to Putnam's original argument for semantic externalism (Putnam 1975). For example, if the meaning of a natural kind term such as 'water' depends for its individuation on the scientific nature of the kind to which it applies, as Putnam (1975) had argued, then it would seem that a thinker who knew the philosophical argument for this conclusion would be in a position to reason, from her thought that water is thirst quenching, to the conclusion that her environment contains water. But then it seems that Putnam's (1975) argument has discovered an unexpected route to *a priori* (or at least armchair) knowledge of (some of) the features of one's environment! Insofar as one regards this as absurd, one will reject either the supposition that Putnam's argument succeeds at establishing semantic externalism, or else that the subject has authoritative knowledge of what she is thinking when she thinks a thought that she would express with such a natural kind term like "water." This worry was famously turned into an objection to externalist views of the mind by Michael McKinsey (1991), and it has been much discussed in the literature.[5]

In Chapter 8, "BIVs, sensitivity, discrimination, and relevant alternatives," Kelly Becker addresses this dialectic as it arises in connection with Putnam's (1981) "proof."[6] He considers how this dialectic looks if we assume a

[5] For one of the most influential monographs on this and related topics, see Brown (2004). See also Brown (1995, 2001) and Sawyer (2001) for earlier discussions.

[6] This issue is also touched on in this volume in the chapters by Folina (at the very end of her chapter) and Douven (at the end of Section 11.3).