How (not) to respond to Putnam’s brain in a vat argument

Tim Kraft
Institut für Philosophie, Universität Regensburg
tim.kraft@ur.de
May 24, 2017

Abstract
Putnam’s argument that we are not brains in a vat has recently seen a resurgence in interest. Although objections to it are legion, an emerging consensus seems to be that even if it successfully refutes one version of the scenario, lifelong envatment, it is powerless against recent envatment. Although initially appealing, this strategy generates problems of its own. In this paper I argue that merely switching to recent envatment is indeed a bad response to Putnam’s argument. Yet there is a different version, recent memory-altering envatment, that Putnam’s argument does not refute and is also sufficiently radical. The crucial issue turns out to be which epistemic sources sceptical scenarios may attack. I argue that there is no convincing reason for exempting memory from the sceptical attack. In the end Putnam’s argument does not fail because of some ‘deep’ philosophical mistake, but because it overlooks how flexible and adjustable sceptical scenarios are.

Keywords
Putnam · brains in a vat · Cartesian scepticism · constraints on sceptical scenarios · scepticism and content externalism · memory
1 Putnam’s argument: The state of the debate

Putnam’s argument that we are not brains in a vat (BIV) has recently seen a resurgence in interest (cf. Button 2013, Madden 2013, Goldberg 2016, Thorpe 2017). Objections to this argument are legion: It has been accused of being question-begging or a faulty transcendental argument, of confusing claims about language with claims about reality, of taking a kind of self-knowledge for granted that is inconsistent with semantic externalism and of being pointless because a BIV can repeat it verbatim (cf. Brueckner 2012, Goldberg 2016). Yet another prominent strategy is to concede that Putnam’s argument successfully refutes some version of the BIV scenario, lifelong envatment, but to reply that it is powerless against a scenario in which one was only recently envatted. Though prima facie convincing this strategy also generates problems of its own: It is doubtful that recent envatment is a genuine sceptical scenario. In this paper I argue that merely switching to recent envatment is indeed a bad response to Putnam’s argument, yet there is a different version of the BIV scenario that his argument does not refute: The classical BIV scenario is restricted to questioning sense perception as a source of knowledge, but there is no convincing reason why this restriction should not be lifted to include memory among the epistemic sources under attack. The lesson of this result is twofold: For one, even if successful against metaphysical realism, Putnam’s argument is an unconvincing argument against Cartesian external world scepticism. For another, it does not fail because of
some ‘deep’ philosophical mistake, but because it overlooks how flexible and adjustable sceptical scenarios are. After briefly summarising Putnam’s argument (sec. 2), I discuss why switching to recent envatment is a bad response (sec. 3), what a better response looks like (sec. 4) and why the latter is indeed a good response (sec. 5).

2 A sketch of Putnam’s argument

There is no such thing as the BIV scenario; instead a shared template is varied and embellished in myriad ways: Where are the BIV and the supercomputer located? What else exists in the universe? How long has the BIV been envatted? Why was the BIV created? If the envatment did not happen by chance, how and why was it covered up? And these are just the basic questions. Additional questions can be raised about what happened to other sentient beings, the laws of physics etc.

In discussions of Putnam’s argument the scenario under consideration is usually lifelong envatment in its most radical form:¹

**Lifelong envatment.** By sheer chance the whole universe consists of nothing but the supercomputer and a brain in a vat attached to it. All sensory experiences of the envatted brain are the result of the supercomputer stimulating it in such a way that it has the same experiences that I actually have.

Putnam’s argument works well with this version of the BIV scenario: Against ‘magical’ theories of reference he defends a causal constraint and points out that by hypothesis BIV’s do not meet this necessary condition for brains, hands etc.: Since there are no hands or other ordinary objects, there is no causal connection to them. And although there is a brain, a vat and a computer, the causal connection to them is not of the kind required for reference. Since a BIV cannot refer to brains, vats etc., it cannot think that it is a BIV.

¹ Putnam mentions both lifelong envatment (1981: 6, 12, 50) and recent envatment by an evil scientist (1981: 5f.).
But, if a BIV cannot think that it is a BIV whereas I can, I am not a BIV.\(^2\)

3 A bad response to Putnam’s argument

As mentioned in the introduction, a popular reply to Putnam’s argument (cf. the list of references in Thorpe 2017: 11) is to concede all of the last section, but to point out that there are versions of the BIV scenario whose victims can entertain the thought that they are in that scenario, e.g. because of past causal connections. Let us call this strategy ‘Putnam-prooﬁng’: A sceptical scenario is Putnam-proof iﬀ its victim can entertain the thought that she is in that scenario. A natural way of Putnam-prooﬁng the BIV scenario is to switch from lifelong to recent envatment. If the envatment happened yesterday, last week or last year, its victim can exploit past causal connections to entertain whatever thoughts she was able to entertain before envatment.

**Recent envatment.** Last year someone was kidnapped and envatted. Since then their sense experiences are the result of a supercomputer stimulating their brain so that they have the same experiences that I actually have.

However, recent envatment by itself cannot be used to challenge all or even most of my empirical beliefs. Beliefs about the past and inductive beliefs based on past observations are outside the scope of the resulting sceptical argument. This restriction has been noted quite often in the literature, but disagreement kicks in as to whether and why this is a problem for the sceptical argument. I think there are two separate reasons why the restriction compromises recent envatment as a sceptical scenario.

The first problem is the *distinction without a difference problem*. The restriction is disappointing from an epistemological point of view: Why should empirical beliefs about the past be exempt from sceptical doubts, if even observational beliefs are dubitable? The sceptical argument should explain *why*, not just argue *that* some kind of knowledge is

\(^2\) This is, of course, only a very rough sketch of Putnam’s argument, for fuller versions cf. Wright 1992, Brueckner 2012, Button 2013: ch. 12.
impossible. But the restriction to present empirical beliefs is only the result of Putnam-proofing the scenario and does not reveal interesting epistemological differences within our empirical beliefs. This problem is not meant as a knock-down objection. One may hope to solve it by augmenting the sceptical argument with an additional step that somehow extends the result about present perceptual beliefs to all empirical beliefs (e.g. inverse paradigm case argument or piecemeal scepticism, cf. Brueckner & Altschul 2010, Smith 2016). Whatever the merits of such proposals, it is preferable to solve the problem without epicycles.

The second problem is the evidence problem: It is all to easy to underestimate how much evidence we have against recent envatment (for some glimpses cf. Tymoczko 1989: 295, Dennett 1991: 3–7, Thorpe 2017: 13–16): First, there is neurophysiological and technological evidence against recent envatment: Last year human brains could not even be kept alive in vivo long enough, electrodes could not yet be connected to brains on a large scale, computers were not powerful enough to run the simulation etc. Second, there is economic evidence: Even if practically possible, envatting humans is bound to consume a lot of resources and is not a routine procedure. Third, there is folk psychological evidence: Even if practically possible, there is no plausible motivation for envatting me instead of some other person. Evil guys with funds are on different missions. Fourth, there is evidence stemming from the smooth continuity in my life last year. If I was envatted last year, I must have been kidnapped. To cover up the kidnapping, evil scientists must pick the lock silently, shoot the dog before it barks, sedate me without waking me and transport me to their lab without family members or neighbours calling the police – not an impossible feat, but highly improbable. What is worse, the scenario is supposed to work for everybody. Scepticism is not restricted to those who like me are not paranoid and rich enough to sleep in a panic room, but claims that nobody – independent of their sleeping habits – can rule out the sceptical scenario. Fifth, the improbability of the scenario is raised even further if it includes that the earth or even the whole universe – except the
BIV, of course – has been annihilated after envatment. It is difficult to come up with a more outlandish possibility.

To sum up, lifelong envatment is appealing as a sceptical scenario because it robs me of all evidence so that I cannot even tell what the probability of being in such a scenario is. In contrast, recent envatment leaves me with so much evidence that I can reasonably dismiss it based on circumstantial evidence. That circumstantial evidence does not guarantee the scenario’s falsity does not rescue the sceptical argument: That our empirical evidence rarely hands out guarantees reminds us of our fallibility, but is a far cry from scepticism.

4 A better response to Putnam’s argument

The result may seem to pose a dilemma: Either the scenario is suited for a sceptical argument, but not Putnam-proof or it is Putnam-proof, but too easy to dismiss (cf. Thorpe 2017: 2). But that conclusion is premature: So far we have looked only at two versions of the BIV scenario. However, there are other versions in which there are enough causal connections left for the victim to entertain the thought that it is in that scenario, but not sufficient evidence for dismissing the scenario. In fact, going back to Putnam’s original description of the BIV scenario gives us a hint for how to fix recent envatment:

“He [= the evil scientist] can also obliterate the memory of the brain operation, so that the victim will seem to himself to have always been in this environment.”

(Putnam 1981: 6)

This remark addresses a worry mentioned already: By obliterating memories the evil scientist can cover up the kidnapping so that the victim does not suspect that something is amiss. But once memory alteration is allowed, the sceptical toolbox suddenly contains many

---

3 In Nozick’s version the evil scientist is even more powerful: “for any reasoning […] we can imagine the psychologists […] feeding it to their tank-subject, along with the (inaccurate) feeling that the reasoning is cogent” (1981: 167f.).
more scenarios. If the scenario tells a convincing story why my memory is untrustworthy, both the evidence problem – if memory cannot be trusted, I do not have any evidence to dismiss recent envatment – and the distinction without a difference problem – beliefs about the past are no longer treated differently – can be solved.

**Recent memory-altering envatment.** Last year a member of an alien species living on a planet far away from earth was kidnapped and envatted. It underwent a training session devoted to radically altering its memories. This training session affected all its empirical memories, but not its apriori and conceptual knowledge. Otherwise its memory works properly: It can reliably retrieve memories and its working memory is not affected at all. After the training session is completed, the envatted brain is sent to space. A supercomputer stimulates the brain in such a way that it has all the experiences I actually have. This all happens as a means of population control: The alien species prevent overpopulation on their planet by running an envatment lottery. Since they consider it unethical to let the losers know that they have lost, they devised the memory alteration scheme.4

This is a radical sceptical scenario: All the beliefs covered by lifelong envatment are also covered by this scenario.5 The scenario even covers the BIV’s beliefs that brains are bihemispherical, grey and weigh approx. three pounds. In the scenario brains may well be octospherical, blue, weigh approx. twenty pounds with the BIV only seeming to remember having seen brain scans showing two hemispheres etc. Thus, since all neurophysiological, technological, folk-psychological etc. beliefs are false, there is no evidence left that could

---

4 Memory alteration is rarely mentioned in the literature on Putnam’s argument. Brueckner & Altschul 2010: 176 and Gerken 2012: 72 are exceptions, but none of them discusses the permissiveness of memory alteration in a sceptical scenario any further.

5 Since the scenario is designed to be consistent with semantic externalism, the beliefs that water, Churchill etc. exist(ed) are an exception. Surprisingly, McKinsey’s paradox (McKinsey 1991) works for, not against the sceptical argument here: If these beliefs are non-empirical beliefs, as McKinsey’s paradox suggests semantic externalism is committed to, they are exempt from sceptical doubts not because they are true in the scenario, but because they are non-empirical.
be used to dismiss the scenario. Causal connections, however, are not affected in any way so that the BIV can entertain all thoughts it was able to entertain before envatment. Memory alteration is not memory replacement. Causal connections are left intact because memories are not ‘overwritten’ by new ones, but only altered in a way that results in false beliefs.

The fine print of recent memory-altering envatment is worth commenting on: First, all empirical memories are altered. One may wonder why a sceptical scenario with partial memory alteration does not suffice, e.g. restricting memory alteration to those memories that are evidence against recent envatment. A convincing sceptical scenario is one whose victim has no evidence – not even weak evidence – against being in that scenario. Again, it should not be underestimated how many memories have to be altered to achieve this goal. Altering all the neurophysiological, technological, folk-psychological etc. memories that may potentially be adduced as evidence requires altering large swaths of memories. Second, one may wonder whether there is really no evidence left to dismiss this scenario. What about arguing that running this lottery would consume too many resources on a planet already saddled with overpopulation? But even this, rather weak, evidence is ruled out. The aliens are presented as very ethical. They would never kill or neglect a fellow. Moreover, the elaborate memory-alteration is also needed for soothing the lottery’s winners: Those who continue experiencing alien life can reason that they have not lost because only non-envatted aliens experience alien life. Third, in the scenario the victim’s memory is altered in a training phase and the supercomputer no longer interferes with the victim’s memory once the training session is completed. Having memories altered in a training session makes the scenario consistent with memory being distributed over the brain and avoids the need to postulate a ‘memory box’ in the brain to which a supercomputer could regularly feed new memories.

---

6 For some hints at how the training phase could look like in principle cf. Ramirez et al. (2013), Liu et al. (2014). Their experiments rely on optogenetics – the neurones of mice are altered to be light-sensitive –, but we can easily add to the scenario that the aliens’ brains are light-sensitive.
5 A good response to Putnam’s argument?

The scenario from last section is likely to be met with resistance, even resentment: Lifelong envatment is already a far-fetched thought experiment, but aliens running an envatment lottery overstrains the imagination – too much is too much. But recall that the aim of this paper is to argue for the permissiveness of memory alteration in sceptical scenarios, not to tell a thrilling story. The interesting philosophical question is whether the restriction of sceptical scenarios to sense perception is legitimate. To assess whether recent memory-altering envatment can indeed be used for an interesting and powerful sceptical argument, we should look at some objections to this scenario.

The first objection I want to discuss is the possibility objection: A common constraint on sceptical scenarios is that they must present (what at least appear to be) genuine metaphysical possibilities. This is often taken to require that the sceptical scenario must be easily conceivable, that it must be consistent with our best philosophical and scientific theories about how the mind works and that it does not merely stipulate that, but explains how the beliefs of its victim fall short of knowledge (cf. Cross 2010, Kung 2011). An example for a scenario that does not meet this constraint is the genie in a bottle scenario. Since we do not understand how minds can be realised as genies in bottles and how genies can be deceived, we do not even know what it is we are asked to rule out.

Despite appearances to the contrary, memory alteration clears that bar. We should not reject memory alteration just because we do not yet understand all the details of it. For the same is true of feeding sense experiences. Although the rough outline is clear – just plug a cable into the optic nerve etc. –, the details are all just science fiction. If feeding sense experiences is thought to be sufficiently supported by science, memory alteration is so, too. After all, there already is scientific evidence for the possibility of memory alteration (in animal research, cf. fn. 6). Regarding easy conceivability the best criterion is to look at science fiction movies and popular science books. Those are open to memory alteration: There is at least one classic science fiction movie, Total Recall
(Verhoeven 1990), whose plot is based on the premise of memory implants and at least one bestselling popular science book, The Memory Illusion (Shaw 2016), questioning our steadfast belief in the trustworthiness of memory. Hence, memory alteration is not an outlandish possibility discussed only in obscure epistemology and science fiction circles.

The second objection I want to discuss is the personal identity objection: Memory is deeply connected with personal identity and, therefore, memory alteration endangers personal identity. If envatment involves near-total memory alteration, envatment creates a new person.

Both the main claim – envatment creates a new person – and the underlying assumption – if a new person is created, the sceptical scenario fails – are dubious. The claim that a new person is created clashes with some intuitions about the case: When suspecting that I may be the victim of such a scenario, I suspect that something bad happened to me, I want to go back to my old life, I want my memories back etc. Moreover, memory continuity is not broken completely: If the alien had memories of some event, say its fifth birthday, it still has memories of its fifth birthday after envatment. Although the details of the memories are false – it now remembers its fifth birthday as its sixth, and it was not its birthday, but new year’s eve –, it still remembers a particular event of its past, albeit falsely. Memory alteration should not be confused with memory replacement.

But even if a new person is created, the sceptical argument does not fail. The causal constraint on reference does not require that the causal connection may not involve several persons. After all, I can refer to mammoths and other objects from the distant past because of inherited causal connections. As long as the causal connection between the person before envatment and the person after envatment is sufficiently tight, as in recent memory-altering envatment, the latter can inherit reference from the former.

The third objection I want to discuss is the reference shift objection: Can a BIV whose memory has been radically altered really refer to the things it had causal contact with before envatment? As Evans’ “Madagascar” example (Evans 1973) illustrates, errors can
result in reference being rerouted: Although there is a causal chain from an area of mainland Africa to current utterances of the proper name “Madagascar”, the name does not refer to the mainland area, but to the island.

Even if “the idea that there is a moment at which the languages switch just seems faintly ludicrous” (Button 2013: 159), it is generally agreed in this debate that languages do not switch instantaneously. In the case of “Madagascar” a reference shift occurred only after Marco Polo’s error caught on, but the objection is based on the stronger claim that in recent memory-altering envatment the errors are so pervasive that they result in immediate reference shifts. Combining a transfer to a new environment with memory alteration may accelerate an otherwise slower switch, but does not turn it into an instantaneous one. There is nothing magical about referring to something that is completely misremembered. For example, someone can refer to Churchill even if everything she tells about him is based on false memories. This is so even if she recently moved to a country in which “Churchill” is commonly used as a name for a pop star. Recall that memory alteration is not memory replacement and that in recent memory-altering envatment Churchill exists/-ed so that the BIV can misremember facts about him. In the end the causal constraint is a double-edged sword when used against scepticism: It rules out some error-possibilities, but is at the same time consistent with reference despite widespread error (cf. Burge 2003).

The fourth objection I want to discuss is the autonomous reasoner objection: This constraint on sceptical scenarios is meant to rule out a variety of uninteresting scenarios such as:

- **Robot.** I am a robot all of whose ‘beliefs’ are regularly externally updated via WiFi, including its ‘beliefs’ that its ‘beliefs’ are based on experience and reasoning.
- **Shortcuts.** There are random shortcuts occurring in my brain all the time.

---

7 Thus, the objection is at odds with the appealing simplicity of Putnam’s argument: Unlike the present objection Putnam relies only on a very minimal, intuitive constraint on reference.
• **Confabulation.** I suffer from a severe confabulation syndrome whose sufferers never realise that they have it.

Of course, I cannot rule out being in such a scenario. Yet this does not mean that the sceptical argument is successful. If I suspect to be in such a scenario, I must suspect that my reasoning about the scenario is affected as well – taking such scenarios seriously is self-undermining. Relatedly, victims of such scenarios lack minimal epistemic autonomy so that the alleged beliefs are no longer the victim’s own beliefs. If the ‘beliefs’ of the victim are directly controlled by something external or are the result of deviant causal processes in the brain, she does not have false beliefs, but the external agent or the deviant process (at most) cause the victim to store a false representation. Analogously, if a book contains a false account of the world (no matter whether it was written intentionally or came about by chance), the paper on which the book is printed does not have false beliefs. Interesting sceptical arguments rely on a scenario in which the victim has beliefs of her own and can reason about them.

Although the autonomous reasoner constraint is central for understanding scepticism, it does not rule out memory alteration. In recent memory-altering envatment the victim’s conceptual knowledge, reasoning skills and working memory are not put into question and questioning empirical memories is not self-undermining. Minimal epistemic autonomy is not threatened by memory alteration either. As long as my present rationality and my present minimal epistemic autonomy is taken for granted, it is my beliefs that I reason about. The training session may alter dispositional beliefs (it does so on at least some conceptions of dispositional belief). For example, even before the newly envatted alien thinks explicitly about it for the first time, it dispositionally believes that it is on earth. However, manipulating dispositional beliefs is consistent with minimal epistemic autonomy. As long as the dispositional beliefs are open to review, the ability to reason critically about one’s own beliefs and sustaining or changing them accordingly is not threatened.
6 Conclusion

If the sceptical scenario presented in this paper is successful, Putnam’s argument fails for reasons that are completely independent of more philosophically loaded objections to it. It fails because it overlooks how flexible and adjustable sceptical scenarios are. Any sceptical scenario must meet some constraints, but, as I have argued, there is none that rules out recent memory-altering envatment. This paves the way for a final observation: If memory alteration is permissible in a sceptical scenario, the sceptical toolbox turns into a Pandora’s box. Once opened, a wide variety of new scenarios emerge in which this or that mental process is manipulated in a way undermining knowledge (cf. Schaffer 2010). Unless we can disallow such manipulations on a principled basis, memory alteration remains in the sceptical toolbox and the prospects for anti-sceptical strategies that, like Putnam’s, are tailored to the specifics of a particular scenario are dim. A convincing anti-sceptical strategy must cope with a plurality of sceptical scenarios, no matter how outlandish and playful they may appear at first sight.

References


Nozick, Robert (1981): Philosophical Explanations. Cambridge/Ms.: HUP.


Ramirez, Steve; Liu, Xu; Lin, Pei-Ann; Suh, Junghyun; Pignatelli, Michele; Redondo, Roger L.; Ryan, Tomás J. & Tonegawa, Susumu (2013): “Creating a false memory in the hippocampus”, in: Science 341 (6144): 387–391.


Verhoeven, Paul (1990): Total Recall [Motion Picture]. USA: Tri Star Pictures.