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

Original Scientific Paper

BRAINS IN A VAT AND MEMORY: HOW (NOT) TO RESPOND TO PUTNAM'S ARGUMENT

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 brain in a vat scenario, lifelong envatment, it is powerless against a different one, recent envatment. Although initially appealing, I argue in this paper that this response – merely replacing lifelong envatment by recent envatment – is a bad response to Putnam's argument. Yet there's a different version of the brain in a vat scenario, recent memory-altering envatment, that Putnam's argument doesn't refute and is also sufficiently radical. The crucial issue turns out to be which epistemic sources sceptical scenarios may attack. I argue that there's no convincing reason for exempting memory from the sceptical attack: Sceptical scenarios must target memory to be sufficiently radical and they can do so without violating any constraint on sceptical scenarios. In the end Putnam's argument doesn't 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 (Putnam 1981, cf. Button 2013, Madden 2013, Goldberg 2016, Thorpe 2018, 2019). Putnam's argument is ambitious: According to Putnam, we can know that we are not *BIV*s and we can even know this based on apriori reasoning and reflective self-knowledge alone. Some thought experiments about intentionality and reference together with reflection about what what we are currently thinking about and referring to suffice to rule out being a *BIV*.

Few philosophers have been persuaded by Putnam's argument. Objections to it are legion: It has been accused of being question-begging, 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

(for surveys cf. Brueckner 2012, Goldberg 2016). Yet the most prominent response is to concede for the sake of the argument that Putnam's argument successfully refutes *some* version of the *BIV* scenario, lifelong envatment, but to object that it does *not* refute *all* versions of it; in particular, it is said to be powerless against recent envatment. If this is so, Putnam's argument is at most a partial response to scepticism. The sceptical challenge remains alive as long as there is at least one sceptical scenario left that we cannot rule out.

Although *prima facie* convincing, this concessive response leads to problems of its own: It is doubtful that recent envatment is a truly *sceptical* scenario. In fact, as I shall argue below, the concessive strategy as defended in the literature fails for exactly this reason. Recent envatment is not a sceptical scenario. But I shall also argue that with some modifications the concessive strategy can be revived. For there is a different version of recent envatment that is both a truly sceptical scenario and is not refuted by Putnam's argument. The central idea here is this: The classical *BIV* scenario only questions perception as a source of knowledge, but there is no reason why memory should not be included among the epistemic sources under attack in the *BIV* scenario. Relying on this idea I argue that Putnam's argument indeed fails to refute scepticism because it fails to rule out *all* sceptical scenarios.[1] Interestingly, it does not fail because of some 'deep' philosophical mistake, but because it overlooks how flexible and adjustable sceptical scenarios are. If that is so, we can put the debates on whether Putnam's argument is question-begging, whether it rests on an implausible kind of self-knowledge, and so on to rest. No matter how these debates turn out, Putnam's argument cannot succeed since it fails to rule out all sceptical scenarios.

This completes my outline of the dialectical situation surrounding Putnam's argument. I will now go through all the steps in detail in order to defend how we should and how we should not respond to Putnam's argument. After briefly summarising Putnam's argument (§2), I discuss why replacing Putnam's original scenario with recent envatment is a bad objection against Putnam's argument (§3), what a better response looks like (§4) and why the latter is indeed a good response, i. e. why it is permissible to target memory without violating any constraints on sceptical scenarios (§5).

---

1 It is sometimes argued that Putnam's argument is not meant to refute (Cartesian) scepticism, but to refute only metaphysical realism, i. e. to refute a picture of mind and world that underlies and motivates a specific kind of sceptical worry, but is not equivalent to scepticism. In this paper I argue only that Putnam's argument fails to refute scepticism. I think this is instructive even if Putnam's official target is not scepticism since, on the one hand, Putnam's argument is often taken to be relevant to this debate and, on the other hand, it is by no means obvious that his argument bears only on metaphysical realism.

2 A sketch of Putnam's argument

Cartesian scepticism argues that we cannot know anything we ordinarily think we know about the external world because we cannot rule out radical sceptical scenarios. One of those scenarios is the brain in a vat scenario. In its bare outline the scenario invites us to imagine not having a body, but being a brain kept alive in a vat while being connected to a supercomputer. This scenario is *radical* because it targets not just a small number of our ordinary beliefs and it is *sceptical* because it is difficult to see how we could ever be in a position to rule out being the victim of this scenario. Even a quick look at the literature, however, shows that there is no such thing as *the* brain in a vat scenario. Instead there is a shared template that can be 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? How and why was the *BIV* created? And these are just the basic questions. Additional questions can be raised about what happened to other sentient beings, the laws of physics and so on.

In discussions of Putnam's argument the version of the *BIV* scenario under consideration is usually lifelong envatment in its most radical form:[2]

> **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 its experiences are indistinguishable from the ones I actually have.

Lifelong envatment is a good choice when discussing Putnam's argument for two reasons: On the one hand, it is hard to imagine a more radical scenario so that attempting to refute it is ambitious indeed. On the other hand, Putnam's core idea is easier to motivate when considering this version of the scenario: Putnam introduces and defends a causal constraint on reference and points out that by hypothesis lifelong *BIV*s do not meet this necessary condition for ordinary external world objects (like brains, hands, and so on): Since there are no hands in the scenario, there is *a fortiori* no causal connection to them. And although there is a brain, a vat and a computer, the causal connection to them is deviant and not of the kind required for reference. If, however, a *BIV* cannot refer to brains, vats, and so on, it can neither think nor state that it is a *BIV*. I, however, can think about whether I am a *BIV* – this is exactly what I am doing right now.[3] Hence, whenever I entertain thoughts about whether I am a *BIV*, I cannot be one.

---

2 Putnam himself mentions both lifelong envatment (1981: 6, 12, 50) and recent envatment by an evil scientist (1981: 5f.).

3 Moreover, if I were unable to even entertain the thought that I am a *BIV*, there would *a fortiori* be no sceptical threat, no possibility the sceptical argument could be based on.

In order to understand why this argument, if convincing, shows that we can rule out being a *BIV* based solely on apriori and reflective reasoning it is useful to spell out Putnam's argument explicitly:[4]

(1) In the language I am using right now "hand" refers to hands. (disquotation)
  (2a) A *BIV* is not in causal contact with any hands. (from the description of the scenario)
  (2b) Causal contact is necessary for reference. (from thought experiments about reference)
(2) In the language used by a *BIV* "hand" does not refer to hands. (from 2a and 2b)
(3) *Therefore:* I am not a *BIV*. (from 1 and 2, indiscernibility of identicals)

The first premise is trivial disquotational truth that I can know reflectively and the second premise is based on apriori thought experiments and on the description of the scenario. Since the conclusion follows deductively from premises which are based on apriori and reflective reasoning, it is is known based on apriori and reflective reasoning as well.

As already mentioned in the introduction, I will not discuss the various objections raised against this argument. I will not discuss whether the first premise already presupposes that I am not a *BIV* so that the argument is question-begging. Although it may seem that it presupposes that there is something I can refer to, it is also difficult to see how a disquotational triviality like this could be false: How could a word of my own language not refer to what I refer to by using that very same word? Another objection I will not discuss is whether the argument only shows "*I am not a BIV" is true* which is distinct from *I am not a BIV*. The idea behind this objection is that we want to find out whether we are *BIV*s, not whether everyone states something true when saying "I am not a *BIV*". A third objection I will not discuss is that a *BIV* could repeat the argument verbatim and show it is not a *BIV* either. The force of this objection depends on whether a *BIV* can in fact repeat the argument or merely think or utter something that looks similar.

3 A bad response to Putnam's argument

The reason why I do not discuss these objections is that a popular and straightforward reply to Putnam's argument (cf. the list of references in Thorpe 2018: 677[5]) is to concede all of the last section, but to point out that

4 For somewhat similar, somewhat different reconstructions of Putnam's argument cf. Brueckner 1986, Wright 1992, Müller 2003. In the main text I present Putnam's argument as being about words, not sentences or thoughts. For this paper the differences do not matter.

5 A further indicator for its popularity is that it is often mentioned in textbooks whose focus is *not* on scepticism, cf. e. g. Kallestrup 2014: 173 (a textbook on semantic externalism) or Newen & Schrenk 2013: 38 (a textbook on philosophy of language).

there are *BIV* scenarios whose victims can entertain the thought that they are *BIV*s and who can refer to external world objects, 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 it is in that scenario. Putnam-proof scenarios cannot be ruled out with the help of Putnam's argument as sketched in the last section.

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. The brain's sense experiences are the result of a supercomputer stimulating it so that its experiences are indistinguishable from the ones 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. A minor problem is the *distinction without a difference problem*. The restriction to present empirical beliefs appears to be *ad hoc*. It is the result of Putnam-proofing the scenario, but does not reveal interesting epistemological differences within our empirical beliefs. Perplexingly, empirical beliefs about the past seem to be *better* off than empirical beliefs about the present. This problem need not be a knock-down objection. But even if the sceptical argument could be augmented by an additional step that somehow extends the result about present perceptual beliefs to all empirical beliefs (cf. Brueckner & Altschul 2010 and Smith 2016, see also Kraft 2014: 273–280 for some doubts), a sceptical argument without such epicycles seems to be preferable.

The more pressing problem, however, 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, Kraft 2014: 274–275, Thorpe 2018: 679–682): *First*, there is neurophysiological and technological evidence against recent envatment: Last year human brains could not even be kept alive *in vitro* long enough, electrodes could not yet be connected to brains on a large scale, computers were not powerful enough to run the simulation and so on. *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 my dog before she 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 bunker or fortress, 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. Of course, circumstantial evidence does not *guarantee* the scenario's falsity. But that 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 (cf. Kraft 2012).

4 A better response to Putnam's argument

The result seems to pose a dilemma: A sceptical scenario is *either* suited for a sceptical argument, but not Putnam-proof *or* it is Putnam-proof, but too easy to dismiss (cf. Thorpe 2018: 668). But that conclusion is premature: So far we have looked only at two versions of the *BIV* scenario. There are other versions in which there are enough causal connections left for the victim to be able to refer to external world objects and 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." (1981: 6)[6]

6 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.). Nozick's evil scientist is similar to Schaffer's debasing demon (Schaffer 2010). The victim of Nozick's scenario has a belief based on an incogent reason, but mistakenly thinks it is cogent. The victim of Schaffer's scenario has a belief based on an incogent reason, but mistakenly thinks it is based on a different reason. The scenarios discussed in the main text do not depend on such powerful scientists or demons.

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 more scenarios. If the scenario tells a convincing story why my memory is untrustworthy, both the evidence problem – if memory is untrustworthy, I no longer have any evidence to dismiss recent envatment – and the distinction without a difference problem – beliefs acquired in 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.[7] 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 its experiences are indistinguishable from the ones 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.[8]

This is a radical sceptical scenario: All the beliefs covered by lifelong envatment are also covered by this scenario.[9] The scenario even covers the *BIV*'s beliefs that brains are bihemisperical, 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 and so on. Thus, since all neurophysiological, technological, folk-psychological etc. beliefs are false, there is no evidence left that could 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.

---

7 In the rare case that the victim lacks some relevant concepts the training session must involve some prior conceptual learning. For example, if the victim lacks the concept *brain*, it may be unable to think about brains for the trivial reason that it never acquired the concept before envatment.

8 Memory alteration is rarely mentioned in the literature. Brueckner & Altschul 2010: 176, Briesen 2011: 574–576 and Gerken 2012: 72 are exceptions, but none of them discusses the permissibility of memory alteration in sceptical scenarios any further.

9 Since the scenario is designed to be consistent with semantic externalism, the beliefs that water, Churchill and so on exist/-ed are exceptions. Surprisingly, McKinsey's paradox (1991) works for, not against the sceptical argument here: If these beliefs are non-empirical beliefs, as McKinsey's paradox suggests, they are exempt from sceptical doubts not because they are true in the scenario, but because they are non-empirical.

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 alien being. The elaborate memory-alteration is also needed for soothing the lottery's winners: Those who continue experiencing alien life will believe 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 training is completed.[10] This is important since it 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. The scenario does not depend on treating perception and memory as being similar. In particular, it does not preuppose that both involve some kind of experience, perceptual experience or memory traces. To the contrary, the scenario is neutral with respect to the various philosophical accounts of memory.

5 A good response to Putnam's argument?

The scenario from last section is likely to be met with resistance: Lifelong envatment is already a far-fetched thought experiment, but aliens running an envatment lottery overstrains the imagination – too much is too much, or so it seems. But recall that the aim of this paper is to argue for the usefulness and permissibility of memory alteration in sceptical scenarios, not to tell a thrilling and fascinating story. The interesting philosophical question is whether the restriction of sceptical scenarios to perception is well-motivated. My aim is to argue that if we take scenarios like lifelong envatment seriously, we cannot stop right there, but should allow recent memory-altering envatment as well.

10 Recent work on optogenetics and memory in which memories of transgenic mice with light-sensitive neurons are manipulated provides some hints at how such a training phase might look like, cf. Ramirez et al. 2013, Liu et al. 2014, Robins 2016a.

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* and *why* 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 jinn in a lamb scenario, a scenario which suggests you might be ghost living in a lamb waiting to be freed by Aladdin. Since we do not understand how minds can be realised as jinns in lambs and what beliefs and experiences jinns have while being in a lamb, we do not even know what it is we are asked to rule out.

Despite what one might think at first, memory alteration clears that bar. We should not reject the possibility of memory alteration just because we do not yet understand *all* the details of it. For the same is true of super computers feeding sense experiences. Although the rough outline is clear – plug a cable into the optic nerve –, 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. memscience). Regarding easy conceivability the best criterion is to look at science fiction movies and popular science books. Those are open to memory alteration: There are at least two classic science fiction movies, *Blade Runner* (Scott 1981) and *Total Recall* (Verhoeven 1990), that deal with 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 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 and so on. 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 (for similar distinctions in different contexts cf. Byrne 2010, Robins 2016b).

But even if a new person is created, the sceptical argument does not fail. The causal constraint on reference does not rule out that the causal connection involves 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 even if it is a different person.

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 re-routed: 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), the general consensus is that reference does not switch *instantaneously*. There is nothing magical about referring to something that is completely misremembered. For example, someone can refer to Churchill even if everything she believes about him is based on false memories and even if she recently moved to a place where "Churchill" is commonly used as a name for, say, some living jazz singer. In the end the causal constraint is a double-edged sword when used against scepticism (cf. Burge 2003): It rules out some error-possibilities, e. g. lifelong envatment, but is at the same time consistent with reference despite widespread error, e. g. the example just given or Kripke's Gödel-Schmidt example (reference to Gödel is independent of whether all or most of one's beliefs about him are true, cf. Kripke 1980: 83–84).

Yet, although in the case of "Madagascar" a reference shift occurred only after Marco Polo's error caught on, it may still seem that recent memory-altering envatment is an altogether different case. One way of motivating this claim relies on replacing "causal" in Putnam's original causal constraint by "world-involving abilities" (cf. e. g. Putnam 2013: 25): What really matters for reference are abilities to do something with worldly objects, not mere causal connections. In recent memory-altering envatment the changes are so pervasive that the relevant world-involving abilities are lost and reference is shifted instantaneously. It is not obvious, however, why only one's *present* world-involving abilities should matter for reference. If only present world-involving uses matter, *all* (irreversible) switches from one environment to another would result in instantaneous reference shifts. If both present and past world-involving uses matter, the reformulated constraint does not show that in recent memory-altering envatment instantaneous reference shifts

occur. Combining a transfer to a new environment with memory alteration may accelerate an otherwise slower switch, but there is no reason to think that it is turned into an instantaneous one.

The *fourth* objection I want to discuss is the '*causal contact is necessary, not sufficient' objection*: So far I have at most shown that a victim of recent memory-altering envatment meets the causal constraint on reference. This, of course, does not show that it actually can refer to external world objects. After all causal contact is not sufficient for reference, but only necessary. One route to take here is to accept Williamson's principle of knowledge maximisation (2007: ch. 8) and the associated idea that:

> "Roughly: a causal connection to an object [...] is a channel for reference to it if and only if it is a channel for the acquisition of knowledge about the object [...]." (2007: 264)

Based on this idea one may argue that even if there is a causal connection between a victim of recent memory-altering envatment and external world objects, it cannot refer to external world objects since it cannot acquire knowledge about those objects via the causal connection.

In response I concede the main point: My aim was to show only that Putnam's argument, with the causal constraint it depends on, cannot refute that we are victims of recent memory-altering envatment. Of course, a different constraint on reference may be able to do that, but that would not be *Putnam's* argument anymore. That being said let me add some worries about relying on a Williamsonian knowledge constraint on reference to refute recent memory-altering envatment. As formulated by Williamson, the constraint is timeless, i. e. it does not state that I can refer *now* only to what I can *now* acquire knowledge about. For example, it allows that someone referred to something and acquired knowledge about it in the past, but due to an undercutting defeater lost her knowledge about it later. It also allows that in cases of dementia the patient can refer to, say, a long dead aunt by her proper name although he has lost all knowledge about her. But to rule out recent memory-altering envatment the constraint must be understood synchronically: I can *now* refer only to what I can *now* acquire knowledge about. Only this stronger constraint has the consequence that a victim of recent memory-altering envatment is unable to refer to external world objects. The synchronic knowledge constraint on reference, however, seems to be too strong as cases of undercutting defeat and severe memory loss show.[11]

The *final* objection I want to discuss is the *epistemic autonomy objection*: This objection is based on a constraint on sceptical scenarios according to which the victim of a sceptical scenario may not lose its epistemic autonomy, i. e. the beliefs must be the victim's own beliefs and the victim must be able to

11 For further criticism of Williamson's principle of knowledge maximisation and the associated knowledge constraint on reference, cf. McGlynn 2012.

reflect rationally on the epistemic standing of her own beliefs. This constraint on sceptical scenarios is meant to rule out a variety of uninteresting sceptical scenarios such as:

> **Robot.** There is a robot all of whose 'beliefs' are regularly externally updated via WiFi, including its 'beliefs' that its 'beliefs' are based on experiences and reasons. It happens that the robot's 'beliefs' and 'experiences' are indistinguishable from the ones I actually have.
>
> **Shortcuts.** There is someone in whose brain random shortcuts are occurring all the time. It happens that the random shortcuts result in beliefs and experiences that are indistinguishable from the ones I actually have.
>
> **Confabulation.** There is someone who suffers from a severe confabulation syndrome whose sufferers never realise that they have it. By chance the confabulation results in the beliefs that are indistinguishable from the ones I actually have.

Of course, I cannot rule out being in such a scenario. Yet this does not mean that the sceptical argument is successful. Victims of such scenarios lack minimal epistemic autonomy so that the alleged beliefs are no longer the victim's *own* beliefs and the victim is unable to reflect rationally on the epistemic standing of her 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'. Moreover, rational reflection on the epistemic standing of one's beliefs is impossible since the results of such a reflection are affected by external updating, random shortcuts or confabulation, as well. 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.[12] Thus, sceptical arguments must rely on a scenario in which the victim has beliefs of her own and can reason about them.[13]

---

12 To see why sceptical arguments must not rely on self-undermining scenarios consider, for example, the closure argument: I know that having hands entails not being a *BIV*. Since knowledge is closed under known entailment, this means that I if I know that I have hands, I also know that I am not a *BIV*. But I do not know whether I am not a *BIV*. Therefore, I do not know whether I have hands. – The uninteresting scenarios mentioned in the main text cannot be relied on in the closure argument: Either I know the entailment or I do not know the entailment. If I do not know the entailment, there is no sceptical threat. If I do know the entailment, I am not in one of the uninteresting scenarios. For victims of these scenarios cannot trust their own reasoning, not even their reasoning about entailments, and therefore lack knowledge of any entailment. Either way there is no sceptical threat.

13 It is an interesting question whether Nozick's *BIV* scenario or Schaffer's debasing demon (cf. fn. 6) meet this constraint. I am not going to try to answer this question in this paper.

Although the epistemic autonomy constraint is central for understanding sceptical scenarios, 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. What is put into question are empirical memories, but that is not self-undermining. 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. In this regard the scenario looks similar to the scenarios mentioned in the last paragraph: The dispositional beliefs are not really the victim's own beliefs. However, manipulating *dispositional* beliefs is consistent with minimal epistemic autonomy. As long as the dispositional beliefs are open to review and one is able to reason critically about them and to sustain or change them accordingly, one's epistemic autonomy is not threatened. To sum up, minimal epistemic autonomy is not threatened by the kind of memory alteration envisioned in recent memory-altering envatment.

6 Conclusion

If the argument of this paper is successful, Putnam's argument fails independently of more philosophically loaded objections to it. It fails because sceptical scenarios are flexible and adjustable in ways that allow for Putnam-proofing them. Although any sceptical scenario must meet several constraints in order to pose a serious challenge, there is, as I have argued, no constraint that rules out memory alteration. If that is so, recent memory-altering envatment is a sceptical scenario Putnam's argument must refute or else it fails as a general anti-sceptical strategy.

This paves the way for a final observation that is not limited to Putnam's argument: 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 cognitive process is manipulated in a way undermining knowledge (cf. Schaffer 2010). Unless sceptical scenarios in which both perceptual and non-perceptual cognitive processes are manipulated can be disallowed in a principled way, the prospects for anti-sceptical strategies that, like Putnam's, are "tailored to the specifics of a particular scenario look dim.[14]

14 Earlier versions of this paper have been presented to audiences in Cologne and Regensburg. Many thanks for discussion and criticism to Alexander Dinges, Andrea Klonschinski, Christoph Michel, Hans Rott, Joshua Thorpe and a reviewer for this journal.

References

Briesen, Jochen (2011): „Antiskeptische Trittbrettfahrer des semantischen Externalismus", in: *Zeitschrift für philosophische Forschung* 65: 563–585.

Brueckner, Anthony (1986): "Brains in a Vat", in: *Journal of Philosophy* 83: 148–167. (Reprinted in *Essays on Skepticism.* Oxford: OUP, 2010: 115–132.)

Brueckner, Anthony & Altschul, Jon (2010): "Terms of envatment", in: Brueckner, Anthony: *Essays on Skepticism.* Oxford: OUP, 174–176.

Brueckner, Anthony (2012): "Skepticism and content externalism", in: Zalta, Edward (ed.): *The Stanford Encyclopedia of Philosophy* (Spring 2012 Edition), https://plato.stanford.edu/archives/spr2012/entries/skepticism-content-externalism/

Burge, Tyler (2003): "Some reflections on scepticism: Reply to Stroud", in: Hahn, Martin & Ramberg, Bjørn: *Reflections and Replies. Essays on the Philosophy of Tyler Burge.* Cambridge/Ms.: MIT Press, 335–346.

Button, Tim (2013): *The Limits of Realism.* Oxford: OUP.

Byrne, Alex (2010): "Recollection, perception, imagination", in: *Philosophical Studies* 148: 15–26.

Cross, Troy (2010): "Skeptical success", in: *Oxford Studies in Epistemology* 3: 35–62.

Dennett, Daniel (1991): *Consciousness Explained.* Boston: Little, Brown & Co.

Evans, Gareth (1973): "The causal theory of names", in: *Proceedings of the Aristotelian Society, Supplementary Volumes* 47: 187–208.

Gerken, Mikkel (2012): "Critical notice: *Essays on Skepticism*", in: *International Journal for the Study of Skepticism* 2: 65–77.

Goldberg, Sanford (ed.) (2016): *The Brain in a Vat.* Cambridge: CUP.

Kraft, Tim (2012): "Scepticism, infallibilism, fallibilism", in: *Discipline Filosofiche* 22: 49–70.

Kraft, Tim (2014): "Defending the ignorance view of sceptical scenarios", in: *International Journal for the Study of Skepticism* 5: 269–295.

Kripke, Saul (1980): *Naming and Necessity.* Cambridge/Ms.: HUP.

Kung, Peter (2011): "On the possibility of skeptical scenarios", in: *European Journal of Philosophy* 19: 387–407.

Liu, Xu; Ramirez, Steve & Tonegawa, Susumu (2014): "Inception of a false memory by optogenetic manipulation of a hippocampal memory engram", in: *Philosophical Transactions of the Royal Society B* 369: 20130142.

McGlynn, Aidan (2012): "Interpretation and knowledge maximization", in: *Philosophical Studies* 160: 391–405.

McKinsey, Michael (1991): "Anti-individualism and privileged access", in: *Analysis* 51: 9–16.

Madden, Rory (2013): "Could a brain in a vat self-refer?", in: *European Journal of Philosophy* 21: 74–93.

Müller, Olaf L. (2003): *Wirklichkeit ohne Illusionen.* 2 Vols., Paderborn: mentis.

Newen, Albert & Schrenk, Markus (2013): *Einführung in die Sprachphilosophie.* 2nd ed., Darmstadt: WBG.

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

Putnam, Hilary (1981): *Reason, Truth and History.* Cambridge: CUP.

Putnam, Hilary (2013): "From quantum mechanics to ethics and back again", in: Baghramian, Maria: *Reading Putnam.* London: Routledge, 19–36.

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

Robins, Sarah (2016a): "Optogenetics and the mechanism of false memory", in: *Synthese* 193: 1561–1583.

Robins, Sarah (2016b): "Misremembering", in: *Philosophical Psychology* 29: 432–447.

Schaffer, Jonathan (2010): "The debasing demon", in: *Analysis* 70: 228–237.

Scott, Ridley (1982): *Blade Runner* [Motion Picture]. USA: Warner Bros.

Shaw, Julia (2016): *The Memory Illusion. Remembering, Forgetting, and the Science of False Memories.* London: Random House.

Smith, Martin (2016): "Scepticism by a thousand cuts", in: *International Journal for the Study of Skepticism* 6: 44–52.

Thorpe, Joshua (2018): "Closure scepticism and the vat argument", in: *Mind* 127: 667–690.

Thorpe, Joshua (2019): "Semantic self-knowledge and the vat argument", in: *Philosophical Studies* 176: 2289–2306.

Tymoczko, Thomas (1989): "In Defense of Putnam's Brains", in: *Philosophical Studies* 57: 281–297.

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

Williamson, Timothy (2007): *The Philosophy of Philosophy.* Oxford: Blackwell.

Wright, Crispin (1992): "On Putnam's proof that we are not brains-in-a-vat", in: *Proceedings of the Aristotelian Society* 92: 67–94.