Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 41 (2017) 93–98

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Skepticism revisited: Chalmers on *The Matrix* and brains-in-vats

Richard Hanley

Department of Philosophy, University of Delaware, Newark, DE, USA

Received 10 May 2016; accepted 12 July 2016
Available online 20 July 2016

Abstract

Thought experiments involving The Matrix, brains-in-vats, or Cartesian demons have traditionally thought to describe skeptical possibilities. Chalmers has denied this, claiming that the simulations involved are real enough to at least sometimes defeat the skeptic. Through an examination of the meaning of kind terms in natural language I argue that, though the Chalmers view may be otherwise attractive, it is not an antidote to skepticism.

Keywords: Putnam; Chalmers; Brain-in-a-vat; Matrix; Simulation; Skepticism; Kind terms

1. Standard skepticism, and standard responses to it

In "*The Matrix* as Metaphysics," Chalmers (2015) argues that we in the philosophical tradition have gravely misunderstood hypotheses such as Descartes' demon, the brain-in-a-vat (BIV), and the Matrix. These are not essentially skeptical hypotheses, Chalmers tells us. Rather, they are interesting metaphysical hypotheses.

Chalmers' basic argument is an extension of some points that Hilary Putnam made in "The Meaning of 'Meaning'," and "Brains in a Vat." Putnam (1975, 1981) uses a theoretical background of the causal theory of reference, which Chalmers claims to avoid relying on; Chalmers tells us he wants to derive the causal theory rather than assume it. But since the order of dependence will not matter for my purposes here, I will employ the causal theory for explication of the arguments.

First let's rehearse the tradition that Putnam and Chalmers rebut. Suppose that I am not a BIV. Then I am not now in Tucson. And a good thing, too, since I believe I am not in Tucson right now, and like Russell's pedant I prefer my beliefs to be true. Suppose further that there is a BIV in Tucson right now, being manipulated by clever scientists to have experiences that seem to justify it in believing it is not in Tucson right now. It believes it is not in Tucson right now, and this belief seems to be false, since the BIV *is*, we just supposed, in Tucson right now. Bad news for the BIV.

Moreover, the tradition continues, bad news for me as well, since although it's true that I am not in Tucson right now, nothing in my present experience conclusively rules out the possibility that I am a BIV in Tucson having an experience of not being in Tucson right now, in which case my belief would be false. Given some plausible assumptions, it follows that I don't know much. At least, and this will be our focus, my empirical beliefs about the external world, no matter how justified, fall short of propositional knowledge.

Let's examine this skeptical argument in more detail. Consider:

E-mail address: hanley@udel.edu

http://dx.doi.org/10.1016/j.cogsys.2016.07.003
1389-0417/

A. I know I have hands.

B. I know that if I have hands, I am not a BIV.

C. I know that I am not a BIV.

Given *epistemic closure*, C follows deductively from A and B. (If the knowledge operator K is closed, then "(K) *p*" and "(K)*p entails q*" entail "(K) *q*".) But the BIV hypothesis gives us reason to deny C. To see this, consider that propositional knowledge seems to require the impossibility of error. (If Marta has one ticket in a billion ticket lottery drawn tomorrow, she has *very* strong reason for thinking she am going to lose. Suppose Marta accordingly believes that she is going to lose, and she right. She will lose. Nevertheless, it seems Marta does not *know* that she will lose. The mere *possibility* of error seems to kill knowledge.) Since I *could* be a BIV, having the very same experience as I am having now, then my present experience fails to rule out the possibility of error, and so I do not know that I am not a BIV, even if I am not a BIV.

If C is false, then either A or B is false. But B seems unassailable. The mere description of a BIV entails that the BIV has no hands. By contraposition, anything with hands is not a BIV. Therefore if I have hands, I am not a BIV. Premise B is not vulnerable to the argument just given, since there is no possibility of a BIV having hands, so no knowledge-killing possibility. Therefore, A is false. I do not know that I have hands.

Moreover, there is nothing special about the knowledge claim in A. It is an ordinary claim of empirical knowledge. So it seems that the argument will generalize to any empirical belief. I do not know that I have feet, or hair, or that it is the 21st century, or that Paris is in France, or that the Earth revolves around the sun, and so on.

Notice that necessary truths escape the argument just given (as the treatment of B suggests), and our resultant skepticism need not be global. We may indeed know that $2 + 2 = 4$, or that bachelors are unmarried, or that every villain in Denmark is an arrant knave. But global empirical skepticism is quite bad enough. For instance, it seems to render much of science less valuable than it is.

Here are three important responses to the argument for global empirical skepticism. First, we can be fallibilists about knowledge. We might point out that empirical skepticism seems to follow directly from the lottery analogy above. Yet what could be more reasonable than assuming—as scientists do—that things really are as they appear to be? So perhaps we should just deny that knowledge entails the impossibility of error. Second, we could use similar considerations to instead admit that we don't have empirical knowledge, but hold on to the idea that we have—at least when we are doing science well—what really matters. Third, we could deny that knowledge is closed under entailment, thereby holding on to A and B in the skeptic's argument, but denying that C follows from their conjunction. These responses all have something going for them. But if Chalmers is right, there's at least one class of cases where no such response is required, because premise B is false is such cases.

2. The Putnam-Chalmers semantic partiality response to skepticism

Putnam used considerations from his Twin Earth thought experiment to argue that a BIV could not have the thought that it was possibly a BIV, and so could not falsely believe that it was not a BIV. If you haven't heard of Twin Earth, it's at another location—in our universe but far, far, away—and bears an uncanny resemblance to Earth. In fact, just about everything on Twin Earth is qualitatively the same except for the chemical composition of the colorless, odorless liquid that fills the lakes and rivers. On Earth it is of course H_2O, but on Twin Earth it is some other substance, not occurring on Earth, that we'll dub "XYZ." There is no H_2O on Twin Earth. Every functional role that water plays on Earth is played by XYZ on Twin Earth, and vice versa. Now suppose for simplicity that there is a language of thought, and for me it is English. On Earth, I think water-ish thoughts, and thereby think about H_2O. But it seems my Twin's water-ish thoughts are about XYZ. Does this matter? Yes. If I were to be magically transported to Twin Earth, pointed to some water-ish stimulus and thought, "That's water," I would be wrong. So it's part of the meaning of the English word "water" that it is H_2O and not some other substance.

Putnam's explanation is in terms of a causal theory of meaning for natural kind terms like "water." My water-ish thoughts, and utterances using "water," are about H_2O and not XYZ (or, as it has become known, "twater"), because my water-ish thoughts are historically connected to H_2O and not XYZ. My twin's water-ish thoughts are about XYZ and not H_2O, because his water-ish thoughts are historically connected to XYZ and not H_2O. And so my twin's thoughts are not in English, but rather in Twenglish. The causal theory is not limited to natural kind terms, moreover. Ordinary proper names seem to have their reference determined in a similar way.

So my mental token "I am not in Tucson right now" is true because the indexical "I" picks out Richard Hanley, "right now" picks out a certain time *t* and "Tucson" picks out a certain desert city; and Richard Hanley is not in that city at *t*. Now compare me with a BIV having qualitatively identical experiences, and which has only ever been a BIV. If Putnam is correct, then the BIV's mental token "I am not in Tucson right now" is *not* in English. For the BIV's token "Tucson" does *not* pick out the desert city that my mental token "Tucson" does.

Again, the causal theorist explains this in terms of a causal network of public language tokens of "Tucson" that I am appropriately linked to, but the BIV is not. And the network I am linked to is ultimately grounded in a certain desert city, whereas the BIV has no such link. Hence the

BIV can have no thoughts about Tucson, and so cannot falsely think that it is not in Tucson right now. The same point goes for natural kind terms. So, just as water-ish thoughts on Twin Earth are not thoughts about H_2O, brain-ish thoughts in a BIV are not about *brains*. The BIV is not party to a causal network of public tokens of the English word "brain", and so cannot have thoughts about brains, even though it is one!

Can we just stop there? This can seem an excessively negative strategy. On Twin Earth, water-ish thoughts are about some natural kind, after all—they are about XYZ, the stuff that plays the water-ish role on Twin Earth. And on Twin Earth, Tucson-ish thoughts are about a desert city, after all—they are about Twucson, the place that plays the Tucson-ish role on Twin Earth. If the BIV's empirical thoughts weren't about anything at all, then the BIV seems to altogether lack empirical beliefs, and so would anyway lack the empirical beliefs required for empirical knowledge.

On such a reading, Putnam's argument might anyway fail to undercut skepticism about empirical beliefs. The traditional worry is that my present true belief that I am not in Tucson would be false if I were a BIV. Suppose a Putnamian says don't worry. Your present belief wouldn't be false, and it wouldn't even be not true. (And not because it would lack what the logical positivists called "cognitive value.") Rather, you wouldn't have your present belief at all. This answers one skeptical challenge, to the effect that even if true and as justified as they could be in the circumstances, your empirical beliefs can never be justified enough for knowledge. But here's a different skeptical challenge: if you are a BIV the mental states that you think are empirical beliefs are not empirical beliefs at all; they are not even false. And no matter how much justification you have for thinking your present states are at least false, you cannot rule out the possibility that they are not even false. Bad news for you.

Fortunately, Putnam is not committed to the reading just given, since he allows that the BIV might be referring to something else instead. Chalmers adopts this strategy, and argues that a BIV would have (the typically assumed number of) true empirical beliefs. Just as Twin-Earthers have relevant experience of Twucson playing the Tucson-ish role, the BIV has relevant experience of something—Chalmers calls it "Tucson*" that plays the Tucson-ish role for the BIV. So the BIV's Tucson-ish thoughts are about Tucson*. Moreover, just as Twin-Earthers have relevant experience of XYZ playing the water-ish role, the BIV has relevant experience of something—call it "brains*," that play the brain-ish role for the BIV. So the BIV's brain-ish thoughts are about brains*.

How can there be such things as Tucson* and brains*? They are, according to Chalmers, *virtual objects*. Such virtual objects are possible, he claims, because a certain hypothesis is possibly true: the *Computational Hypothesis* that "microphysical processes throughout space-time are constituted by underlying computational processes." Roughly, the idea is that if the Computational Hypothesis is true, then it is possible to simulate microphysical processes and anything—such as Tucson, brains, water, and vats—that supervenes on microphysical processes. So a BIV, or someone in the Matrix, can be in appropriate causal contact with virtual objects, and that are available to be part of the content of their thoughts.

Now I am, in fact, a friend of virtual objects. So I am just going to grant their possibility. The following quote from Chalmers illustrates the idea's application (he is considering a series of objections (Chalmers, 2015):

> **Objection 5**: You just said that virtual hands are not real hands. Does this mean that if we are in the matrix, we don't have real hands?
> Response: No. If we are *not* in the matrix, but someone else is, we should say that their term "hand" refers to virtual hands, but our term does not. So in this case, our hands aren't virtual hands. But if we *are* in the matrix, then our term "hand" refers to something that's made of bits: virtual hands, or at least something that would be regarded as virtual hands by people in the next world up. That is, if we *are* in the matrix, real hands are made of bits. Things look quite different, and our words refer to different things, depending on whether our perspective is inside or outside the matrix. This sort of perspective shift is common in thinking about the matrix scenario. From the first-person perspective, we suppose that *we* are in a matrix. Here, real things in our world are made of bits, though the "next world up" might not be made of bits. From the third-person perspective, we suppose that someone *else* is in a matrix but we are not. Here, real things in our world are not made of bits, but the "next world down" is made of bits. On the first way of doing things, our words refer to computational entities. On the second way of doing things, the envatted beings' words refer to computational entities, but our words do not.

We should not extend this explanation too far, however. Chalmers grants that some terms might be shared between my language and the BIV's:

> **Objection 7**: An envatted being thinks it performs actions, and it thinks it has friends. Are these beliefs correct?
> Response: One might try to say that the being performs actions* and that it has friends*. But for various reasons I think it is not plausible that words like "action" and "friend" can shift their meanings as easily as words like "Tucson" and "hair". Instead, I think one can say truthfully (in our own language) that the envatted being performs actions, and that it has friends. To be sure, it performs actions in *its* environment, and its environment is not our environment but the virtual environment. And its friends likewise inhabit the virtual environment (assuming that we have a multi-vat matrix, or that computation suffices for consciousness). But the envatted being is not incorrect in this respect.

A footnote accompanies this response:

> **Note 9**: Why the different response to objection 7, on "action" and "friend"? We noted earlier (note 1) that not all terms function like "water" and "hair". There are numerous *semantically neutral* terms that are not subject to Twin Earth thought-experiments: any two twins using these terms on different environments will use them with the same meaning (at least if they are using the terms without semantic deference). These terms arguably include "and", "friend", "philosopher", "action", "experience", and "envatted". So while an envatted being's term "hand" or "hair" or "Tucson" may mean something different from our corresponding term, an envatted being's term "friend" or "philosopher" or "action" will arguably mean the same as ours.
>
> It follows that if we are concerned with an envatted being's belief "I have friends", or "I perform actions", we cannot use the Twin-Earth response. These beliefs will be true if and only if the envatted being has friends and performs actions. Fortunately, it seems quite reasonable to say that the envatted being *does* have friends (in its environment, not in ours), and that it does perform actions (in its environment, not in ours). The same goes for other semantically neutral terms: it is for precisely this class of expressions that this response is reasonable.

In other words, Chalmers distinguishes between semantically neutral terms like "friend," and what I shall call *semantically partial* terms like "water," "hands" and "hair." According to Chalmers, it's the semantically partial terms that appear in the skeptic's argument, and we can respond that thanks to their partiality, they do not do the work that the skeptic needs of them. To do that, they would have to be semantically neutral.

3. Structural kinds, other kinds, and semantic strangeness

So there can be overlap in the languages of empirical claims between Twin Earth and Earth. But how far does this go? For instance, in both places the water-ish stuff fills the river-ish and lake-ish things. But does Twin Earth have and rivers and lakes? Or does a river or lake have to contain H_2O? Is "river" or "lake" semantically partial?

It seems that the kind terms Chalmers calls semantically neutral are names of broadly *functional* types, whereas those subject to Twin-Earth responses are names of broadly *structural* types. H_2O is structurally different from XYZ, and (if we're not envatted) hair and hands are structurally different from hair* and hands*. Hair* and hands* are "made of bits" and hair and hands are not. A philosopher, though, is a functional type, and Chalmers is a philosopher whether or not he's made of bits.

But now, some trouble lurks. Notice that "envatted" is one of Chalmers' semantically neutral terms. Now suppose that I were envatted, having the experiences I'm having now. If I believed I were not envatted—and I *don't* mean not envatted*, because we don't need the asterisk—would my belief be true? On the one hand, *in the world of the simulation*, which we're granting is virtual but nevertheless real, I would not be envatted. So my belief would, it seems, be true. But on the other hand, as we began by supposing, *I would be envatted*! So my belief would, it seems, be false. It's tempting to add "in the real world," but let's just say, in the *other* world, or O-world. And let's contrast it with the S-world of the simulation.

There are options, of course. If one is an occupant of both the S-world and the O-world, then perhaps semantically neutral terms, although they apply in both worlds, have odd extension conditions. Ordinarily we might think "I am not envatted" is true only if there is no world in which you are envatted, but perhaps instead it's true if there's at least one world in which you are not envatted. That would be odd. Or perhaps one world somehow has priority. (Though I find myself tempted to give priority to the O-world in such a case, and "I am not envatted" then comes out false, anyway.) One problem with such an option is that it works best for terms that are *not* semantically neutral. Moreover, it seems that in any case, I can just go ahead and believe "There is no world in which I am envatted," and thereby believe something false. How could *that* belief be extensionally odd? Is such a belief domain-restricted, willy-nilly?

There are two other ways out. One, Chalmers said "envatted" was *arguably* semantically neutral. So maybe we just discovered that it's arguably *not*. Two, perhaps "envatted" is both semantically neutral and extensionally odd, but as luck would have it only a rare term is in this predicament, and so skepticism does not get much of a foothold.

I don't recommend either of these alternative ways out, though. First, to put it in my terms, "envatted" is a good candidate for a semantically neutral term precisely because it names a broadly functional type. Second, whatever we say about "envatted," there are many other candidates for functional types lining up to cause the same trouble. Consider "desert" and "city." These seem to me to arguably just as semantically neutral as "friend." If I am not a BIV, then I am not in a desert city right now. Good thing I believe I am not in a desert city right now! But if I were a BIV in Tucson, then I would be a BIV in a desert city right now believing I was not in a desert city right now. Wouldn't I? And would I be wrong, or both right and wrong, or some other weird alternative?

The problem concerns the nature of kind terms and our ability to refer by means of them. A proper name reaches back through appropriate causal links to only one individual, so it is very plausible that I could not think about Tucson if I were a BIV (assuming I'm actually not). But kind terms have this marvelous feature: that their extension is not restricted to those instances that we are appropriately causally linked to.

This does no harm in the cases Chalmers gives us, because beliefs like "I have friends" come out clearly true

and not false. But call the predicament of an occupant of an S-world whose semantically neutral terms reach extensionally into an O-world in a problematic way, *semantic strangeness*. I submit that semantic strangeness offers us a variation on the skeptic's argument, even if everything Chalmers claims about semantically partial terms is correct. First, the possibility that you are in a semantically strange predicament with regard to some proposition undermines knowledge of it. Second, given doubt about whether or not a term is semantically neutral, you cannot rule out the possibility that you are in a semantically strange predicament.

There seems to be but one way to avoid semantic strangeness. Let's go back to Putnam's BIV and its brain-ish thoughts. Is a brain an instance of a structural type or a broadly functional type? Perhaps nothing is a brain unless it's organic, and being organic entails something not realizable in an S-world. Then anything playing the brain-ish role in an S-world is a brain*. Now consider vats, which are members of an artifactual kind. Perhaps nothing is a vat unless it's made of metal or glass or plastic (etc.), and being any of these entails not being realizable in an S-world. Then anything playing the vat-ish role in an S-world is a vat*. And perhaps anything playing the desert-ish role in an S-world is a desert*. Anything playing the city-ish role in an S-world is a city*. And so on. For everything. And for everything*.[1]

Call this suggestion *global structuralism*. According to it, a broadly functional type is still a structural type, at a more fundamental level. It's just that, within a world, the structure required is realized *in everything*, and so it tends to drop out when we're doing the semantics of kind terms and thinking about their extensions.[2]

4. Skepticism revisited

As promising as the global structuralist strategy might seem, it opens us up to a new version of the skeptical challenge. First, imagine the best-case scenario for responding to the skeptic. It's the one Chalmers and Putnam have already described: you are not envatted, and you are imagining what would be true if you were envatted. More precisely, you are in an O-world that is not a world of bits, and you are imagining what would be true if you were in an S-world that is a world of bits, being directly simulated by occupants of the O-world, which is the "next world up," as Chalmers puts it. In such a world, you would think you were not envatted*, and you would be right.[3]

Now consider another possibility. As Chalmers writes, the "next world up" "might not be made of bits," but it also *might* be made of bits. One way for this to be true is if the O-world is an S-world itself, with an OO-world as the "next world up" again. Imagine a scenario in which an occupant of such an O-world of bits is in a BIV-ish role, such that their ("next world down") S-world experience is of not occupying a BIV-ish role. Then they are in a semantically strange predicament, even given global structuralism. How bad is it that such a possibility exists? It depends upon the actual situation. Consider three main variants.

4.1. Variant 1

Suppose that in actuality I am not envatted, and not made of bits. Then there are two reasons that I should not be concerned about a scenario where I am in a semantically strange predicament. The first is that the occupant of the BIV-ish role just described is not a BIV; rather it's a BIV*. And it cannot think about Tucson, or deserts, or cities, or hands or hair. Secondly, if I am an occupant of a world not made of bits, and I am not made of bits, then the occupant of the BIV-ish role (the BIV*) could not *be* me. It's at best Hanley*. So at most I can imagine Hanley* being a BIV*.

4.2. Variant 2

Next, suppose I am envatted, and not made of bits. Then the extension of "BIV" includes all things made of bits and that play the BIV-ish role. So I can imagine something playing the BIV-ish role, and that would be to imagine it being a BIV. But can I imagine *myself* being a BIV? It seems not, since, like Tucson, I am not made of bits. At most, I can imagine Hanley* being a BIV.

4.3. Variant 3

But suppose I am envatted, and made of bits (since the "next world up" is also made of bits). Then the extension of "BIV" includes all things made of bits and that play the BIV-ish role. Then I *can* imagine myself playing that role, and so can imagine myself being envatted. I can imagine Hanley being a BIV. And that is to imagine Hanley being in a semantically strange predicament.

[1] There seems little point to hanging on to semantically neutral terms that don't cause trouble, like "action" and "friend." (They seem to work mainly when they're relational, and the thinker is one of the relata.) But hold on to them if you like, and ignore them in what follows.

[2] There might be an exception to global structuralism. Chalmers' argument appeals in part to the possibility of the Mind-Body Hypothesis: "My mind is (and has always been) constituted by processes outside physical space-time, and receives its perceptual inputs from and sends its outputs to processes in physical space-time." If the MBH is true, then perhaps you are not ever in an O-world or an S-world. In what follows I will assume that if am not made of bits it is because I am made of some other material entities. But this will not matter to the argument I will give in the following section. A Cartesian dualist who denies that minds are in space still accepts claims about my having spatial properties, but finesses things to avoid claims of exemplification. And if anything, Cartesianism will exacerbate rather than ameliorate the skeptical worry I present.

[3] It would not avail me if I am not a BIV, and yet made of bits, because then I can imagine being a BIV, and that is to imagine a semantically strange predicament.

My problem is that I don't know which of the three variants I am in. My *Cartesian predicament* is of not being able to tell them apart. So nothing in my present experience rules out my being in Variant 3. So nothing in my present experience rules out my possibly being in a semantically strange predicament. Of course, the global structuralist can assure me that if I am not in variant 3, then when I think I am imagining being Hanley being a BIV, I'm not really imagining that, and instead imagining either Hanley* being a BIV or Hanley* being a BIV*. But that at best puts me in a sort of stand-off with the skeptic. Can I play the odds, and say there's a good chance that I can't imagine a skeptical scenario?

There's something else that's a bit odd about the global structuralist strategy, something it shares with the earlier Putnamian response to the skeptic. If I am in Variant 3, then I am not just *imagining* being in a semantically strange predicament; I'm also actually *in* one. So it seems I should hope that I'm not in a semantically strange predicament. Suppose that hope is realized. (The odds aren't bad, after all!) Then I'm not hoping that if I were a BIV, then my empirical beliefs would still be largely true—that would be the Chalmers strategy. Rather, I'm hoping that when I think I imagine that I were a BIV, I'm really mistaken about the content of my imagining. And that's a Cartesian predicament of its own.

We could escape this consequence by appealing to a version of counterpart theory. In its cross-possible-world and cross-time versions, counterpart theory allows occupants of one world or time to satisfy *in absentia* suitably abverbialized open formulae in virtue of having (other-worldly or other-timely) counterparts who satisfy the formulae simpliciter. We could extend this to allow for next-world-up or down counterparts (and next-to-next, and so on), that do the same.

Given global structuralism, this would help only in Variant 2. Only in Variant 2 would I be able, other things being equal, to imagine myself being Hanley being a BIV. Moreover, if Hanley were as imagined a BIV, then Hanley would not be in a semantically strange predicament, and Hanley's empirical beliefs would be largely true. So if I am in Variant 2, then my imagining being Hanley being a BIV would not be a skeptical scenario. So should I hope that I am in Variant 2? (The odds aren't so bad, after all.) But that is to hope I am envatted—a curious response indeed to the traditional skeptical worry!

5. Funding source

This research did not receive any specific grant from funding agencies in the public, commercial, or not -for-profit sectors.

References

Chalmers, D. (2015). The matrix as metaphysics. <http://consc.net/papers/matrix.html>. Accessed 09/30/2015.

Putnam, H. (1975). The meaning of 'meaning'. In K. Gunderson (Ed.). *Minnesota studies in the philosophy of science* (Vol. 7, pp. 131–193). Minneapolis: University of Minnesota Press.

Putnam, H. (1981). Brains in a vat. *Reason, truth, and history* (pp. 1–21). Cambridge: Cambridge University Press.