THERE ARE ACTUAL BRAINS IN VATS NOW

Adam Michael BRICKER

ABSTRACT: There are brains in vats (BIVs) in the actual world. These "cerebral organoids" are roughly comparable to the brains of three-month-old foetuses, and conscious cerebral organoids seem only a matter of time. Philosophical interest in conscious cerebral organoids has thus far been limited to bioethics, and the purpose of this paper is to discuss cerebral organoids in an epistemological context. In doing so, I will argue that it is now clear that there are close possible worlds in which we are BIVs. Not only does this solidify our intuitive judgement that we cannot know that we are not BIVs, but it poses a fundamental problem for both the neo-Moorean (i.e. safety-based) anti-sceptical strategy, which purports to allow us to know that we aren't BIVs, and the safety condition on knowledge itself. Accordingly, this case is especially instructive in illustrating just how epistemologically relevant empirical developments can be.

KEYWORDS: BIVs, cerebral organoids, scepticism, neo-Mooreanism, safety

0. Introduction

When Hilary Putnam first introduced the "case of the brains in a vat," he categorised BIVs as a "science fiction possibility."[1] A mere three decades later, this science fiction possibility is now a rapidly advancing research programme in developmental and molecular biology. Taking only a slight liberty with the word "vat," but none whatsoever with the word "brain," we might now observe that there are quite literally brains in vats. These BIVs—cerebral organoids—are not quite as Putnam imagined. They are not harvested, but rather grown from stem cells *in vitro*. More significantly, they are not yet comparable to conscious human brains, but closer to that of a three-month-old foetus.[2] Nevertheless, the science of cerebral organoids is advancing at a remarkable rate,[3] and bioethicists are already

[1] Hilary Putnam, *Reason, Truth and History* (Cambridge: Cambridge University Press, 1981), 5.

[2] Iva Kelava and Madeline A. Lancaster, "Dishing out mini-brains: Current progress and future prospects in brain organoid research," *Developmental Biology* 420, 2 (2016): 199-209.

[3] See Kelava and Lancaster, "Dishing out mini-brains," Stefano L. Giandomenico and Madeline A. Lancaster, "Probing human brain evolution and development in organoids," *Current Opinion in Cell Biology* 44 (2017): 36-43.

looking forward to a future containing conscious cerebral organoids.[4] As I will argue below, we are unsettlingly close to the world Putnam imagined.

This paper explores an especially noteworthy epistemological consequence of the modal closeness of BIV-containing worlds: We cannot know that we aren't BIVs. Here we will see this consequence play out not only at the intuitive level, but also with regard to the neo-Moorean anti-sceptical strategy, which purports to allow us to know that we aren't BIVs. In short, not only is the safety condition unable to successfully anchor an anti-sceptical argument, but it now faces a sceptical problem of its own. In arguing along these lines, this paper will proceed in the following way: First, I will discuss the BIV scenario through the lens of epistemology, focusing specifically on its relationship with neo-Mooreanism (§1). Next, I will provide a brief, non-technical introduction to the actual-word BIVs that are cerebral organoids (§2). Finally, I will argue that the present status of cerebral organoids in the actual world indicates that there are close possible worlds in which any number of us are cerebral organoids (§3). It follows directly from this that we cannot know that we aren't cerebral organoids, that neo-Mooreanism is unable to explain how we might have this knowledge, and that the safety condition on knowledge now faces its own sceptical problem.

1. Brains in Vats

In this section I discuss the epistemological significance of brains in vats. After introducing the case of BIVs, I'll discuss neo-Mooreanism, the predominant anti-sceptical strategy that seeks to explain how we might know that we aren't BIVs.

Let's begin with Putnam's original BIV scenario:

> Imagine that a human being (you can imagine this to be yourself) has been subjected to an operation by an evil scientist. The person's brain (your brain) has been removed from the body and placed in a vat of nutrients which keeps the brain alive. The nerve endings have been connected to a super-scientific computer which causes the person whose brain it is to have the illusion that everything is perfectly normal. There seem to be people, objects, the sky, etc; but really all the person (you) is experiencing is the result of electronic impulses travelling from the computer to the nerve endings.[5]

[4] Megan Munsie, Insoo Hyun, and Jeremy Sugarman, "Ethical issues in human organoid and gastruloid research," *Development* 144, 6 (2017): 942-945, Andrea Lavazza and Marcello Massimini, "Cerebral organoids: Ethical issues and consciousness assessment," *Journal of Medical Ethics* 44, 9 (2018): 606-610, Joshua Shepherd, "Ethical (and epistemological) issues regarding consciousness in cerebral organoids," *Journal of Medical Ethics* 44, 9 (2018): 611-612.

[5] Putnam, *Reason, Truth and History,* 5-6.

When Putnam first put forth this BIV scenario, it was for the purpose of illustrating a semantic point about reference.[6] Nevertheless, brains in vats proliferated throughout philosophy in the decades following, including not only the philosophy of language, but also the philosophy of mind, metaphysics, and epistemology.[7] Given the breadth of the impact of the BIV scenario, here I will focus only on the epistemological significance of brains in vats. Within an epistemological context, the BIV scenario is something like the contemporary successor to Descartes' Evil Demon. While Descartes imagined a wholesale illusion of the external world as the product of supernatural forces, not the presumably naturalistic processes employed by the evil scientist, the upshot is the same: Were we deceived by an evil demon, just as if we were BIVs, all our experiences would be indistinguishable from what they are now.

This type of scenario forms the basis for the familiar sceptical argument from closure: It seems we intuitively judge that we cannot know that we aren't BIVs. After all, were we BIVs, all our experiences might be identical to how they are in the actual world. Moreover, intuitively, it also seems that knowledge is closed under known entailment. That is, if S knows that p, and S knows that p entails q, then S knows that q. However, if we commit to both these claims, this entails that we cannot have any knowledge of the external world that is inconsistent with being a BIV, which is to say most all knowledge of the external world. Following closure, knowing about the external world would entail knowing that we aren't BIVs.

Defeating this sceptical argument has been one of the primary aims of contemporary epistemology. One widely employed strategy involves denying or restricting closure in some way, which would allow us to have knowledge of the external world despite not knowing that we aren't BIVs.[8] However, here I will focus on one predominant anti-sceptical approach that purports to allow us to know that we aren't BIVs: neo-Mooreanism.

In his infamous argument against scepticism, G.E. Moore[9] offered a simple inversion of the above sceptical argument: If (i) we know some ordinary

[6] Roughly, Putnam maintained that the thoughts of a BIV cannot refer to entities in the external world. Therefore, the thought, "I am a BIV," cannot be true, even when thought by a BIV.

[7] For an overview, see *The Brain in a Vat,* ed. Sanford Goldberg (Cambridge: Cambridge University Press, 2016).

[8] For an introduction, see Steven Luper, "Epistemic Closure," *The Stanford Encyclopedia of Philosophy,* ed. Edward N. Zalta (Spring 2016 Edition), https://plato.stanford.edu/archives/spr2016/entries/closure-epistemic.

[9] George Edward Moore, "Proof of an External World," *Proceedings of the British Academy* 25 (1939): 273–300.

proposition, e.g. that we have hands, and (ii) we know that this ordinary proposition entails the negation of the sceptical hypothesis (i.e. that we aren't BIVs), then (c) we know the negation of the sceptical hypothesis. While this argument finds little love in contemporary epistemology, multiple prominent philosophers (see below) have defended a more robust version of Moore's argument, dubbed "neo-Moorean" by Pritchard.[10] The key difference between Moore and neo-Moorean arguments is that the latter approach seeks a broader theoretical motivation for our ability to know that we aren't BIVs.[11] While Moore appealed, unconvincingly, only to our common sense, pre-theoretic intuitions about knowledge of the external world,[12] neo-Moorean arguments seek to establish this via some general epistemological principle.

The preferred epistemological principle of the neo-Moorean response is *safety*: If S knows that p, then there are no close possible worlds in which S falsely believes that p. Assuming, as is standard, that there are no close possible worlds in which we are BIVs, this allows us to have ordinary knowledge of the external world, as there are ordinarily no close possible worlds in which we are mistaken about such things. Moreover, safety even allows us to know that we aren't BIVs, again assuming that there are no close worlds in which we might be BIVs and accordingly believe falsely that we aren't BIVs. In this way, neo-Mooreanism preserves both our knowledge of the external world and closure. Different versions of this of anti-sceptical approach have been advanced by Sosa,[13] Williamson,[14] and Pritchard.[15]

In section 3, I will argue that the core assumption of neo-Mooreanism—i.e. that there are no close possible worlds in which we are BIVs—seems mistaken

[10] Duncan Pritchard, "Recent Work on Radical Skepticism," *American Philosophical Quarterly* 39, 3 (2002): 215-257, 237

[11] See Duncan Pritchard, "Contemporary Neo-Mooreanism," in *Epistemological Disjunctivism* (Oxford: Oxford University Press, 2012).

[12] See Duncan Pritchard, "Mooreanism," in *Epistemological Disjunctivism* (Oxford: Oxford University Press, 2012).

[13] Ernest Sosa, "How to Defeat Opposition to Moore," *Philosophical Perspectives* 13(1999): 141-153.

[14] Timothy Williamson, *Knowledge and Its Limits* (Oxford: Oxford University Press, 2000), chapter 8.

[15] Duncan Pritchard, "Resurrecting the Moorean Response to the Sceptic," *International Journal of Philosophical Studies* 10, 3 (2002): 283-307 and "How to be a Neo-Moorean," in *Internalism and Externalism in Semantics and Epistemology,* ed. Sanford Goldberg (Oxford: Oxford University Press, 2007).

given recent advancements developmental biology. However, first I need to say a bit on what these advancements actually are.

2. Brains *in Vitro*

The purpose of this section is to introduce the basics of cerebral organoids to epistemologists. Accordingly, this discussion will be limited in scope to two epistemologically interesting features of cerebral organoids:[16] (1) Cerebral organoids are quite literally brains in vats;[17] (2) although they are not functionally or structurally analogous to conscious human brains, it seems that this is only a matter of time. I will discuss both points in turn.

Let's begin with a brief overview of these brains *in vitro*. Cerebral organoids begin as stem cells, which, under conditions that partially mimic those that facilitate early embryonic neural development *in utero*, quickly grow into recognisable cortical and sub-cortical structures.[18] Neurons begin to appear in 8-10 days, with recognisable brain structures appearing in under a month.[19] Cerebral organoids famously recapitulate *in utero* development seen between 1 to 4 months of gestation.[20] At the structural level, cerebral organoids display a number of distinct neural regions, including forebrain/hindbrain differentiation, a hippocampus, a choroid plexus, an immature retina, and a cortex with distinct dorsal (including a prefrontal cortex and occipital lobe) and ventral regions.[21] The last of these is most important, as it is the cortical areas that ultimately provide the neural basis of higher-level cognitive function in further developed brains. Beyond these fascinating structural properties, cerebral organoids also display remarkable functional characteristics. Cortical neurons in cerebral organoids form mature neurons with functioning synapses, which display spontaneous neural activity.[22] To be clear, the neurons in the cortex of cerebral organoids—*in vitro* brains that presently develop to the rough equivalent of a three-month-old foetus—actually fire! In short, it is no exaggeration to say that there are presently brains in vats.

---

[16] For a technical introduction, see Madeline A. Lancaster, Magdalena Renner, Carol-Anne Martin, Daniel Wenzel, Louise S. Bicknell, Matthew E. Hurles, Tessa Homfray, Josef M. Penninger, Andrew P. Jackson, and Juergen A. Knoblich, "Cerebral organoids model human brain development and microcephaly," *Nature* 501, 7467 (2013): 373-9.

[17] Again, this might depend on your definition of "vat."

[18] Byoung-il Bae and Christopher Walsh, "What are mini-brains?" *Science*, 342, 6155 (2013): 200-1.

[19] Bae and Walsh, "What are mini-brains?"

[20] See Kelava and Lancaster, "Dishing out mini-brains."

[21] Lancaster et al., "Cerebral organoids model human brain development and microcephaly."

[22] Lancaster et al., "Cerebral organoids model human brain development and microcephaly."

This is not a sci-fi future, as it was when Putnam first described the BIV scenario, but an actual feature of the actual world.

At this point, it's important to not get ahead of ourselves. Are there actual brains in vats? Yes. Are these Putnam-esque brains in vats? No. Cerebral organoids aren't yet grown in conditions similar enough to those found *in utero* to allow for later stages of neural development, nor are they vascularised, which is also necessary for later cortical development.[23] In short, today's cerebral organoids are almost certainly not conscious. The vat technology just isn't there yet. Nevertheless, it is not unreasonable to expect that the *conscious* cerebral organoid is only a matter of time. Indeed, there appears to be a growing sense among researchers that this is the case. One indication of this comes from the recent interest in cerebral organoids displayed by medical ethicists, who have already begun laying the groundwork for the ethics of handling conscious cerebral organoids.[24] Moreover, researchers note that cerebral organoid technology is "extremely fast-moving,"[25] and that only a decade ago, today's level of *in vitro* cerebral development was "thought to be unattainable."[26] Further still, as noted by Kelava and Lancaster, current limitations don't appear to be in any way insurmountable:

> It is easy to envisage that in 10–20 years from now (or even less) we will be able to almost fully mimic development of certain tissues *in vitro.* In addition, further improvements in the technique might allow us to model adult brain physiology and disorders of the adult and ageing brain.[27]

As I will argue in the next section, this alone poses a significant problem for both neo-Mooreanism and the safety condition generally. The key here is that while there aren't yet conscious cerebral organoids in the actual world, from the epistemological perspective this doesn't actually matter. They are widely thought to be on the horizon, not a new technology as much as an incremental improvement on existing technology.

---

[23] Kelava and Lancaster, "Dishing out mini-brains."

[24] Lavazza and Massimini, "Cerebral organoids: Ethical issues and consciousness assessment," Shepherd, "Ethical (and epistemological) issues regarding consciousness in cerebral organoids."

[25] Giandomenico and Lancaster, "Probing human brain evolution and development in organoids," 41.

[26] Kelava and Lancaster, "Dishing out mini-brains," 199.

[27] Kelava and Lancaster, "Dishing out mini-brains," 205.

Finally, before moving on, I want to highlight one crucial point that is all too easy to miss if we don't discuss cerebral organoids simultaneously from the epistemological and developmental-neuroscientific perspectives: The very conditions necessary to create neural architecture seen in the postnatal brain, if applied to cerebral organoids, would necessarily transform the cerebral organoid into a sceptical scenario. One of the most basic facts of postnatal neural development is that it is dependent upon perceptual input from external sources.[28] The neural circuits and functional connections found in the mature brain require external signals in order to form. Visual pathways won't develop without visual input; auditory pathways won't develop without auditory input, etc. This means that developing cerebral organoids that are analogous to even the infant brain will require some sort of targeted stimulation of the relevant neural pathways.[29] However, this is all it takes to transform the cerebral organoid into a Putnam-esque BIV. Even crude stimulation of the primary visual cortex, which is easily accomplished non-invasively with current technology (i.e. TMS), will result in consciously perceptible flashes of light or disturbances in the visual field. Moreover, further development of the visual pathway would require more complex inputs, which would in turn produce more complex conscious perceptions. The point here is that simply getting anything close to a mature brain would require giving the cerebral organoid false perceptual input, and the more developed we want it, the more sophisticated this false perceptual content needs to be. That is to say, it requires that we play a role eerily close to Putnam's "evil scientist."

3. Brains in Close Possible Words

The existence of cerebral organoids has immediate consequences for how we think about scepticism and sceptical hypotheses. In this section, I want to discuss these consequences in two different ways. First, at the intuitive level, the existence of cerebral organoids fundamentally changes our judgements about whether we can know that we aren't BIVs. The rapid development of cerebral organoids solidifies

[28] For a textbook introduction, see Dale Purves, *Neuroscience*, (Sunderland, Mass.: Sinauer, 2016), ch. 25.

[29] It should be noted that "artificial" neural stimulation is well within the means of current medical technology. One excellent example of this comes from cochlear implants, which entirely bypass biological transduction of sound and directly stimulate the auditory nerve with electrical impulses. In principle, there is no reason why a cochlear implant must transmit auditory signals from the external world instead of, say, artificial ones generated to correspond with a simulated reality.

what were previously fluid or ambiguous intuitions: We cannot know that we aren't BIVs. Second, these intuitive attitudes are reflected in the problems cerebral organoids pose for safety and the neo-Moorean anti-sceptical strategy. In short, the existence of cerebral organoids in the actual world means (i) that there are close possible worlds in which we are BIVs and thus (ii) neo-Mooreanism is unable to explain how we might know that we aren't BIVs, or even ordinary information about the external world. Moreover, (iii) the safety condition now appears to *entail* that we cannot have knowledge of the external world, giving us ample reason to reject it on anti-sceptical grounds.

Let's begin with the intuitive judgements engendered by the emergence of cerebral organoids. Put simply, I think we clearly and unambiguously judge that we cannot know that we are not cerebral organoids. This marks a departure from the intuitions elicited by distant-world sceptical hypotheses, which are notably unstable and varied. It sometimes might seem that we can know that we aren't deceived by an evil demon, while in other cases it may not. While advocates of denying closure concede that that we cannot know the negation of sceptical hypotheses, neo-Mooreans have thought we might know the negation of sceptical hypotheses due to a theoretical feature of knowledge, and Moore of course thought it just obvious. The point here is that this sort of ambiguity no longer seems appropriate. Our intuitive response to learning about cerebral organoids is markedly different from that elicited by Putnam's BIV counterfactual. There are already BIVs. There will soon be conscious BIVs, and there is even already talk of connecting "numerous human organoids into working complexes" in the near future.[30] It now seems chillingly obvious that we cannot know that we aren't BIVs, and that any theory of knowledge that entails that we have this knowledge is simply mistaken.

The observation that we intuitively judge we cannot know that we aren't cerebral organoids is reflected in the problems these actual-world BIVs pose for the neo-Moorean response to scepticism. First, the neo-Moorean response, insofar as it is committed to the safety condition on knowledge, only works if there are no close possible worlds in which a sceptical hypothesis maintains. If there are close possible worlds in which we are BIVs, then those are close possible worlds in which we falsely believe that we aren't BIVs, meaning that such beliefs aren't safe. The problem for neo-Mooreanism, and the safety condition on knowledge generally, is that it certainly seems like there are close possible worlds in which we are cerebral organoids. While it is of course difficult to definitely settle questions of modal ordering, I think I might illustrate the modal closeness of these sceptical-

[30] Munsie et al., "Ethical issues in human organoid and gastruloid research," 943.

hypothesis worlds in the following way: First, as discussed above, conscious cerebral organoids seem an inevitability of the near future. With improvements in *in vitro* conditions and complimentary developments in the sophistication of neural stimulation, there will be conscious cerebral organoids in the actual world. The key here is that remarkably little needs to change in the actual world for conscious cerebral organoids, which suggests that there are close possible worlds in which this is the case. Let's call these W1 worlds. Next, not too distant from W1 worlds are worlds in which any one of us is a cerebral organoid. Let's call these W2 worlds. Again, the key here is that not much needs to change from W1 worlds for any given individual to be a conscious cerebral organoid. All that is necessary is that your parents opted to have a lab grow a cerebral organoid you for the purposes of "personalised medicine," as is already being envisioned as a primary application of the technology.[31] As discussed above, developing a mature brain requires complex external stimuli, which your W2 parents opt for in the interest of ensuring your non-organoid siblings won't suffer from neurodegenerative disorders later in life.

The question at this point is whether the changes between the actual world, the closest W1 world, and then the closest W2 world to that are so great as to preclude W2 worlds from being close to the actual world. When viewed in this manner, we can understand just how close W2 worlds are. The change from the actual world to W1 is simply a modest, easily foreseeable advancement in technological capabilities. Moreover, the change from W1 to W2 could be as little as your parents deciding not to grow a cerebral organoid before having you (W1) vs growing a cerebral organoid that is you (W2). Put this way, the modal closeness of W2 worlds to the actual world is striking, and more than a little unsettling. As discussed above, this poses a fundamental challenge for neo-Mooreanism, which requires that there are no such close W2 worlds. Not only is the neo-Moorean argument unable to explain how we might have knowledge that we aren't BIVs, it additionally fails to explain how we might have ordinary knowledge of the external world.

Further still, if we accept that there are (1) close possible worlds in which we are cerebral organoids and (2) that safety is a necessary condition on knowledge, this entails that we have little, if any, knowledge of the external world. Given (1), for (at least) most any belief p by S about the actual external world, there will be a close possible world in which cerebral organoid S falsely believes that p. However, the very definition of the safety condition on knowledge is that such beliefs cannot be known. In a fascinating interplay between epistemology and

[31]Kelava and Lancaster, "Dishing out mini-brains," 206.

developmental biology, we can understand that the emergence of cerebral organoids has transformed safety from a plausible inoculation against scepticism into a premise in its own sceptical argument. Accordingly, it seems we now have good reason to reject safety on anti-sceptical grounds.

There is a worry here that all this might be premature. Perhaps the worlds in which we are BIVs won't be close enough to pose a problem for another 20 or so years, when there are conscious cerebral organoids in the actual world. While I understand this sentiment, anti-sceptical strategies and necessary conditions on knowledge shouldn't come with expiration dates. If we think that safety might not be a necessary condition on knowledge 20 years from now, then there is something profoundly mistaken about maintaining that it is now. Beyond this, there is an additional worry that perhaps the emergence of cerebral organoids simply means that we cannot have knowledge about the external world. Unfortunately, I cannot rule this out. However, as such a concession would of course represent a new paradigm in epistemology, I will not address it here.

Additionally, one could object that there isn't actually anything new here, and that any epistemologically interesting work done by modally close BIVs was already possible via modally close dreamers (i.e. close possible worlds in which we could have been dreaming whatever it is we perceive in the actual world).While an understandable impulse, this objection I think exaggerates the extent to which the phenomenal content of dreams might recapitulate that of conscious perception. Yes, there is a sense in which, when we dream, we might form something like false beliefs. However, there is a major concern here that these beliefs will never share the same content as those formed during waking consciousness. When I dream that I'm writing a philosophy paper, I don't form beliefs with comparable content to those I am forming now. Any belief content in dreams, if we might call even it that, is lo-fi, confused, and incoherent. The key to BIVs is that, unlike dreams, their mental content might exactly mimic that of the (waking) actual world. Perceptual input might even follow the same visual and auditory pathways, via the retina and auditory nerve, respectively. Given some perceptual belief that p, with content derived from actual experience, this is what is needed to underwrite a false belief that p. Accordingly, it is clear BIVs can do far more epistemological work here than dreams.

Finally, I want to say a bit more about the significance of this paper. Admittedly, support for safety appears to be waning, not the least among those cited above as initial proponents, and neo-Mooreanism is rather niche. However, it's important not to miss that there is a broader lesson to be learned here about the relationship between empirical science and technological advancement on the one

hand, and epistemological theory on the other. Despite the increasing incorporation of empirical findings and methods into epistemology, this tends to be rather shallow and one-dimensional, focusing largely on the scientific description of patterns of knowledge attribution.[32] Even you aren't especially interested in safety or neo-Mooreanism, it is crucial to understand that even fields like developmental and molecular biology can have non-trivial epistemological implications. While in this case it happens to be for accounts that were already a bit past their prime, the takeaway is still clear. We as epistemologists need to be sensitive to the fact that empirical developments from far outside our orbit might be directly relevant to the accounts we develop. The onus is on us to seek out such findings.

In conclusion, recent advances in developmental biology have precipitated a shift in how we might think about sceptical hypotheses: There are now close possible worlds that contain sceptical hypotheses in which any number of us are BIVs. This poses a foundational problem for the neo-Moorean anti-sceptical strategy, as well as the safety condition on knowledge, and it seems quite clear at this point that we cannot know that we aren't cerebral organoids. Moreover, it serves to underscore the unexpected ways in which empirical developments might be epistemologically relevant.

[32] For example, see Jennifer Nado, *Advances in Experimental Philosophy and Philosophical Methodology* (New York: Bloomsbury, 2016).