Check for updates

RESEARCH ARTICLE

REVISED When is a brain organoid a sentience candidate?

[version 2; peer review: 5 approved, 1 approved with reservations]

Jonathan Birch

Centre for Philosophy of Natural and Social Science, The London School of Economics and Political Science, London, England, WC2A 2AE, UK

v2 **First published:** 18 Oct 2023, **2**:22
https://doi.org/10.12688/molpsychol.17524.1

Latest published: 20 Aug 2024, **2**:22
https://doi.org/10.12688/molpsychol.17524.2

Abstract

It would be unwise to dismiss the possibility of human brain organoids developing sentience. However, scepticism about this idea is appropriate when considering current organoids. It is a point of consensus that a brain-dead human is not sentient, and current organoids lack a functioning brainstem. There are nonetheless troubling early warning signs, suggesting organoid research may create forms of sentience in the near future. To err on the side of caution, researchers with very different views about the neural basis of sentience should unite behind the “brainstem rule”: if a neural organoid develops or innervates a functioning brainstem that regulates arousal, and leads to sleep-wake cycles, then it is a sentience candidate. If organoid research leads to the creation of sentience candidates, a moratorium or indefinite ban on the creation of the relevant type of organoid may be appropriate. A different way forward, more consistent with existing approaches to animal research, would be to require ethical review and harm-benefit analysis for all research on sentience candidates.

Keywords

organoids, sentience, consciousness, brainstem, regulation, precaution

This article is included in the The Ethics of Brain Organoids collection.

Open Peer Review

Approval Status ? ✓ ✓ ✓ ✓ ✓

	1	2	3	4	5	6
version 2 (revision) 20 Aug 2024			✓ view	✓ view	✓ view	✓ view
version 1 18 Oct 2023	? view	✓ view	? view			

1. **Tomasz Żuradzki**, Jagiellonian University in Kraków, Kraków, Poland
2. **L. Syd M. Johnson**, SUNY Upstate Medical University, Syracuse, USA
3. **Takuya Niikawa**, Kobe University, Kobe, Japan
4. **Jon Mallatt**, University of Idaho, Moscow, USA
5. **Federico Zilio**, University of Padova, Padova, Italy
6. **J. Lomax Boyd**, Johns Hopkins University, Baltimore, USA
Elizaveta Kalinina, Johns Hopkins University (Ringgold ID:1466), Baltimore, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Jonathan Birch (j.birch2@lse.ac.uk)

Author roles: Birch J: Conceptualization, Investigation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, Grant Number 851145.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Birch J. **When is a brain organoid a sentience candidate? [version 2; peer review: 5 approved, 1 approved with reservations]** Molecular Psychology: Brain, Behavior, and Society 2024, **2**:22 https://doi.org/10.12688/molpsychol.17524.2

First published: 18 Oct 2023, **2**:22 https://doi.org/10.12688/molpsychol.17524.1

REVISED **Amendments from Version 1**

Many updates made in response to reviewer suggestions. Key changes from v1: various subtle wording changes to the definition of "sentience candidate", to the Sentience Precautionary Principle, and to the brainstem rule. An important note has been added to the brainstem rule, namely "a functional equivalent of a brainstem (even if artificial) would also suffice". An unclear aspect of the rule, namely the reference to "prioritization", has been deleted. The range of realistic possibilities now runs from R1-R5 rather than R1-R4, separating out those theories that attach special significance to the prefrontal cortex (the figure has also been updated). A discussion of how the integrated information theory relates to this range has been added. Further discussion of the relevance of embodiment has been added, granting both a realistic possibility that embodiment is essential for sentience and a realistic possibility that it is inessential.

Any further responses from the reviewers can be found at the end of the article

1. The promise of organoid research

Biomedical research urgently needs new and better alternatives to animal models. The trend in recent decades has been towards increasing reliance on a small number of model species, especially rats, mice, zebrafish and fruit flies, and towards a troubling level of dependence on assumptions about the relevance of these model systems to human medical conditions (Farris, 2020). Many researchers and funding agencies have invested heavily in the idea that understanding the brain mechanisms of animal models will help us understand complex conditions such as depression, anxiety, autism or schizophrenia in their human forms. But animal models are far from perfect models, leading to widespread reflection on how things could be done differently (Shemesh & Chen, 2023; Taschereau-Dumouchel *et al.*, 2022).

The maxim to "replace, reduce and refine" (the "3Rs") was coined more than fifty years ago and is now embedded in frameworks for the regulation of animal research around the world (Hubrecht & Carter, 2019). It crystallizes a point of wide agreement. We should aim to replace animal models with other types of model where possible, reduce the numbers of animals being used, and refine experimental techniques to minimize suffering. Yet this maxim has turned out to be compatible with a drastic increase over those same fifty years in the total numbers of animals used (Taylor & Alvarez, 2019).

So, we have two disquieting trends: growing concern about the ability of biomedical research on animal models to deliver tangible benefit, particularly in relation to neurological/mental conditions, and a growing realization that, despite widespread endorsement of the 3Rs, invasive animal research is on the rise, not on the way out. These trends raise the question: what is the alternative? To study a complex condition like depression or autism, the argument goes, you cannot simply study tissue in culture, but you also cannot study human subjects at the level of mechanistic detail required to understand how, for example, particular alleles and patterns of gene expression may influence these conditions. So, you must use animals, where the ethical limits on what can be done are more permissive and a broader range of interventions is available.

This is where brain organoids have tremendous promise. The organoid is a relatively new kind of model system with great potential for replacing invasive animal research. Organoids are models of organs constructed from pluripotent stem cells. Human stem cells can be used, leading to miniature models of human organs constructed from human tissue. Suppose, for example, you want to understand human kidney function. One option is to study the renal system of a rat or mouse, relying on the idea that this will resemble human kidney function in the ways that matter. But organoid technology gives you a new option. You take pluripotent human stem cells and induce them to differentiate into kidney cells. The kidney organoid you construct will still differ from a normal kidney in many ways, but you have a degree of control over those ways, and you can be confident that the genes being expressed are the same as those in human kidney cells.

When we are talking about kidney organoids, gastrointestinal organoids, cardiac organoids, and other types of non-neural organoid, these developments should be celebrated. We should not try to put the brakes on a programme that could turn out to deliver the alternative to animal research that has been so sorely needed for so long.

But when it is the *brain* being modelled, the work becomes more controversial, and rightly so. A brain organoid is a model constructed from pluripotent stem cells induced to form organized neural tissue. Here too, it is the use of human stem cells to create human neural organoids that is generating major scientific excitement. I will use the term "brain organoid" here, but I note that the term "neural organoid" is also used, and the terms "cerebral organoid" and "cortical organoid" are also often used in cases where the organoid is intended to model the human neocortex. Sometimes these models have been implanted into mice, leading to functional connections with the mouse's brain (Wilson *et al.*, 2022), but my main focus here will be on extracorporeal organoids, sustained in a controlled environment outside of a living body.

There are ethical reasons in favour of doing this research, if it allows us to model neurological conditions for which scientists currently lack good models, and if it can substitute for invasive animal research. And yet the research invokes the image, if not currently the reality, of a sentient brain in vitro, and this image fills many onlookers with a sense of horror, regardless of whether the brain is human or non-human (although I will be focusing, in this article, on human brain organoids). Even when one looks at the research as it is now, it is hard not to feel a certain unease at the idea of a miniature model of a brain constructed from brain tissue. Sometimes unease is a bias we should try to overcome. But sometimes it is pointing us in the direction of genuine moral reasons to pause the research.

We need to think seriously about which of these possibilities is the case here. In recent years, a number of bioethicists have been doing just that (Ankeny & Wolvetang, 2021; Hyun *et al.*, 2020; Kreitmair, 2023; Lavazza, 2020; Lavazza & Massimini, 2018; Niikawa *et al.*, 2022; Sawai *et al.*, 2019; Sharma *et al.*, 2021; Żuradzki, 2021). I have weighed into the debate already, advocating for a precautionary approach to these issues (Birch & Browning, 2021). This is an extension of the approach I advocate towards non-human animals in which sentience is disputed, such as shrimps and insects (Birch, 2017). But, as Steel (2015) has argued, even though formulations of the precautionary principle often fail to mention consistency, requiring consistency in our thinking about different risks is crucial if we are to avoid inconsistent packages of recommendations. In this case, we need to make sure our approach to organoids is fully consistent with our approach to animal research. In particular, we must be careful not to be overcautious regarding organoids in a way that undermines their promise as replacements for animals.

My goal in this article is to find the right balance. This essay will provide philosophical reflection (both epistemological and ethical) on our state of uncertainty in relation to organoids, and on the ways in which we may try to manage that uncertainty. The central question will be: *what should it mean to take a precautionary approach to this problem?* My way of answering that question will involve what in recent philosophy has been called the method of "conceptual engineering": designing and constructing new concepts to help us escape problems created by our existing concepts.

Here is an outline of the article: Section 2 and Section 3 set out the ingredients of a precautionary approach to questions of sentience. The discussion here will be quite general and will not specifically concern organoids. The crucial concept introduced is that of a *sentience candidate*: a system that is not certain to be sentient, but which is sentient according to at least one reasonable, evidence-based theoretical position. Section 4–Section 8 ask: when is a brain organoid a sentience candidate? I argue that the presence of a functioning brainstem that regulates arousal and leads to sleep-wake cycles is enough. Section 9 considers the precautionary steps we should consider when a brain organoid is a sentience candidate.

2. A scientific meta-consensus

The term "sentience" in English comes from the Latin "sentire", literally "to feel". It is used in different ways in different contexts, with the idea of "feeling" providing a loose common thread. Sometimes, people in brain organoid research use the term to mean nothing more than "responsiveness to sensory stimuli due to adaptive internal processes" (Kagan *et al.*, 2022). When the term is used in this way, some preparations of human brain tissue are already sentient. However, I strongly recommend against using the term in this way, because it creates a large gap with how the term has come to be used in bioethics, animal ethics, animal law and the science of animal welfare.

In those fields, to say that a system is "sentient" is to say that it is capable of valenced conscious experiences such as pain or pleasure. That is: in at least in some conditions, there is something it is like to be that system, and the experience is either pleasant (positively valenced) or unpleasant (negatively valenced). The reason for using the term in this way is that this capacity is widely taken to be morally significant. Put simply, it is a good thing when animals have conscious pleasant experiences, whereas unpleasant experiences such as pain are a source of ethical concern. That is the sense in which I will be using the term. Not everyone would agree with that definition, and we could spend a whole article unpacking it, but this is not the place for that. Our question is whether there is good reason to think that brain organoids could already be—or have the potential to become—sentient in this sense.

There is no scientific consensus about the neural basis of sentience or phenomenal consciousness in humans, other mammals, or any other animals. Contemporary consciousness science contains a wide range of positions (Seth & Bayne, 2022). It is equally important, though, to see that an absence of consensus on a specific theory does not lead to a chaotic "anything goes" situation in which all speculation is equally valid. Evidence still constrains theorizing. Some options are serious and evidence-based, while others are not.

The concept of "meta-consensus" can be helpful for thinking about these situations. The concept is borrowed from political science (Dryzek, 2010). In very broad terms, the motivation for the concept is that people may agree about a lot, even when they disagree about the best policy. Crucially, they may still agree about the range of reasonable options, and they may agree about how these options relate to each other along important dimensions (such as more moderate to more radical). "Meta-consensus" is a term for consensus on these "meta" questions concerning the option space. Seeing a meta-consensus can be an incredibly important step towards negotiating a way forward.

To my knowledge, the concept has not yet received explicit discussion in relation to scientific disagreement. But it should. Just as finding a meta-consensus can help lawmakers move forward when they disagree, so finding a scientific meta-consensus can help scientists move forward, as well as helping outside audiences to better understand what is going on in the science. It is all too easy for a non-expert, looking in, to think "since they disagree so much, there is no reason for me to listen to a word they have to say. I'll just go with my gut feeling". That is a poor inference, and a very dangerous one too, but it can be a tempting one when scientists cannot articulate clearly what they *do* agree about.

Does meta-consensus exist in the science of sentience? I think it does. I will first present where I think the meta-consensus lies, and then explain *why* I think this:

Proposed meta-consensus:

Given our current evidence, all of the following theoretical positions about the neural system requirements for sentience (defined as the capacity for valenced experience) are realistic possibilities. None should be held dogmatically, but all should be taken seriously in practical contexts:

R1. Sentience requires distinctively primate neural mechanisms (e.g. in granular prefrontal cortex) and is absent in non-primates.

R2. Sentience requires mechanisms distinctive to the mammalian neocortex and is absent in non-mammals.

R3. Sentience requires the neocortex in mammals but can also be achieved by other brain mechanisms performing relevantly analogous functions (such as the avian pallium).

R4. Sentience does not require the neocortex even in mammals and can be achieved in at least a minimal form by integrative subcortical mechanisms crucially involving the midbrain. However, it is absent in non-vertebrates.

R5. Sentience does not require the neocortex even in mammals and can be achieved in at least a minimal form by integrative subcortical mechanisms crucially involving the midbrain. Moreover, it can also be achieved by other brain mechanisms performing relevantly analogous functions (such as the central complex in insects).

These five positions are ordered from less inclusive to more inclusive. R5 is the most inclusive, in the sense that the distribution of sentience in the animal kingdom is likely to be the widest if this position is correct, since midbrain mechanisms are far more widely shared than neocortical mechanisms.

By contrast, it is not reasonable, given current evidence, to give serious attention in practical contexts to views less inclusive than R1 (such as a view on which sentience requires a developed capacity for natural language) or more inclusive than R5 (such as a view on which the spinal cord is said to support sentience by itself in the absence of a brainstem). The evidence does not support taking these views seriously in practical contexts.

There is no consensus about which of R1–R5 is correct, and each option can be fleshed out in many different ways. At the most inclusive end of the reasonable range, Merker (2007), Panksepp (1998), and Solms (2021) have defended midbrain-centric theories that are neutral between R4 and R5, while Barron and Klein (2016), Ginsburg and Jablonka (2019), Feinberg and Mallatt (2016), and Tye (2016) have defended versions of R5. Damasio can be placed approximately between R3 and R4, since he has often emphasized the importance of both the midbrain and some parts of the cortex (especially the insular and somatosensory cortex) (Damasio *et al.*, 2000; Damasio *et al.*, 2013).

Meanwhile, many cortex-centric, computational functionalist theories, such as the global workspace theory, the perceptual reality monitoring theory, and the recurrent processing theory are most naturally interpreted as versions of R3. Both Dehaene (2014) and Lau (2022) posit important roles for distinctively primate mechanisms in the human implementation of the mechanisms they take to be responsible for conscious experience: Dehaene proposes a key role for dorsolateral prefrontal cortex in implementing the global neuronal workspace, while Lau proposes a key role for dorsolateral and frontopolar prefrontal cortex in implementing perceptual reality monitoring. However, both allow that these mechanisms may have alternative implementations in other animals. Lamme (2022), in developing the recurrent processing theory, focuses on mammalian visual areas (such as visual cortex), but recurrent processing could, clearly, be implemented by other animals in their own sensory areas. Humphrey's (2022) theory also falls in the R3 zone: he recognizes that the feedback loop he takes to be constitutive of conscious experience may be implemented differently in birds.

At the less inclusive end, R1 includes more demanding computational functionalist theories, on which sentience is linked to complex computations that may only be achievable by brain mechanisms distinctive to the primate lineage. The relevant mechanisms are located in granular prefrontal cortex (granular PFC), a part of the frontal lobe greatly expanded and elaborated in primates, incorporating the frontopolar, dorsolateral, and ventrolateral prefrontal cortex and characterized by a notably thick layer of granular (layer IV) cortical neurons (Preuss & Wise, 2022). These brain regions are strongly linked to executive control functions. Rolls's (2004; 2014) 'higher-order syntactic thought' theory gives a crucial role to these mechanisms. LeDoux has at times appeared sympathetic to R1 and has emphasized the special processing properties of granular PFC (LeDoux, 2023, pp. 758-9). However, his most recent work clarifies that granular PFC is required only for the most cognitively demanding kinds of consciousness: 'autonoetic' and 'noetic' consciousness (LeDoux *et al.*, 2023). He allows that 'anoetic' consciousness, which I see as much closer to the idea of sentience, may be achievable in a much wider range of animals.

The R2 category includes theorists who have, for various reasons, proposed that neocortical neurons, and perhaps especially the large pyramidal neurons in layer V, may have special processing properties that allow them to support consciousness (e.g. Aru *et al.*, 2020; Beck & Eccles, 1992; Key, 2015). On this view, granular PFC is not necessary, potentially allowing all mammals to meet the requirements, but there is something very special about the neocortex more generally. For Beck and Eccles (1992), for example, pyramidal layer V neurons were the most likely entry point for mental causation in the workings of the brain. On this (admittedly highly speculative) theory, the nucleated structure found in birds might not be enough.

The consensus lies not at the level of specific positions (clearly!) but rather at the meta-level, in the idea that everyone should be able to recognize any of the positions in the range R1–R5 as realistic possibilities that must be taken seriously in practical contexts. All positions in this range have some evidence behind them, conferring a degree of plausibility. Moreover, everyone should be able to agree on the ordering of these views from less inclusive to more inclusive (Figure 1). Finally, everyone should be able to agree on the severe challenges facing any view that sees both the neocortex and the midbrain as unimportant to sentience, or any view that regards a functional primate brain as insufficient.

This may sound like it does not exclude very much, but it does. Consider, for example, the cerebellum. This is part of the hindbrain, at the very back and base of the brain, and it contains more neurons than any other brain region, even the cortex. There are 69 billion neurons in your cerebellum, compared with a mere 16 billion in the cortex (Herculano-Houzel, 2009). If one were trying to guess the 'seat of consciousness' in the brain using nothing but neuron counts, one would probably guess the cerebellum—and be completely wrong. There is no evidence for a role for the cerebellum in generating conscious experience and strong evidence against. The cerebellum has important roles in motor control and sensorimotor integration, and appears to be crucially involved in modelling the expected sensory consequences of our actions and registering prediction errors (Arikan *et al.*, 2019; Johnson *et al.*, 2019). These computations could have turned out to be essential to sustaining a conscious state, but they turn out not to be, as a matter of empirical fact. Being born without a cerebellum (complete primary cerebellar agenesis) leads to motor control problems but turns out to be compatible with otherwise normal cognitive development (Yu *et al.*, 2015).

So, the evidence does not warrant attaching significant probability to a hindbrain-centric theory of sentience, or a theory that blithely predicts that sentience will be tied to the brain region with the most neurons with no consideration of what the neurons are doing. One cannot pluck theories out of thin air, without supporting evidence, and expect them to be taken seriously when practical questions are at stake. There are too many possible-but-very-low-probability theories, and their practical implications are so diverse that they are apt to derail discussion if we admit them to the table. In practical contexts, we need to maintain a focus on credible theories that have amassed enough evidence in their favour to merit serious discussion of their practical implications.

For another example, this time from the other end of the axis, consider a theory that ties sentience to natural language. There are serious theories, such as Rolls's (2004; 2014) 'higher-order syntactic thought' theory, that tie conscious experience to quite sophisticated kinds of thought, suggesting a narrow distribution of sentience in the animal kingdom. Yet even Rolls stops short of proposing that natural language is required for the relevant type of thought, allowing that a 'language of thought' might also be sufficient. This is a wise move, because we have clear evidence that linguistic abilities are not needed to have conscious experiences. Brain injuries to regions associated with language can lead to temporary aphasia (loss of linguistic ability) of various kinds, but subjects, when they recover, can often vividly recount their conscious experiences during the time they were affected (Koch, 2019).

One theory that is challenging to locate in the R1-R5 range is the integrated information theory (IIT) of Tononi and collaborators (summarized in Tononi *et al.*, 2023). In consciousness science, researchers outside the IIT camp often distinguish *fundamental IIT* (or strong IIT), the full version of the theory including a highly speculative metaphysical background picture, from *empirical IIT* (or weak IIT), which simply claims that, in the human brain, the neural correlate of conscious experience is a "posterior cortical hot zone" at the back of the neocortex and that the high causal integration of this region is what makes it apt for this role (Mediano *et al.*, 2019; Mediano *et al.*, 2022; Michel & Lau, 2020). Empirical IIT is too thin a claim to locate in the

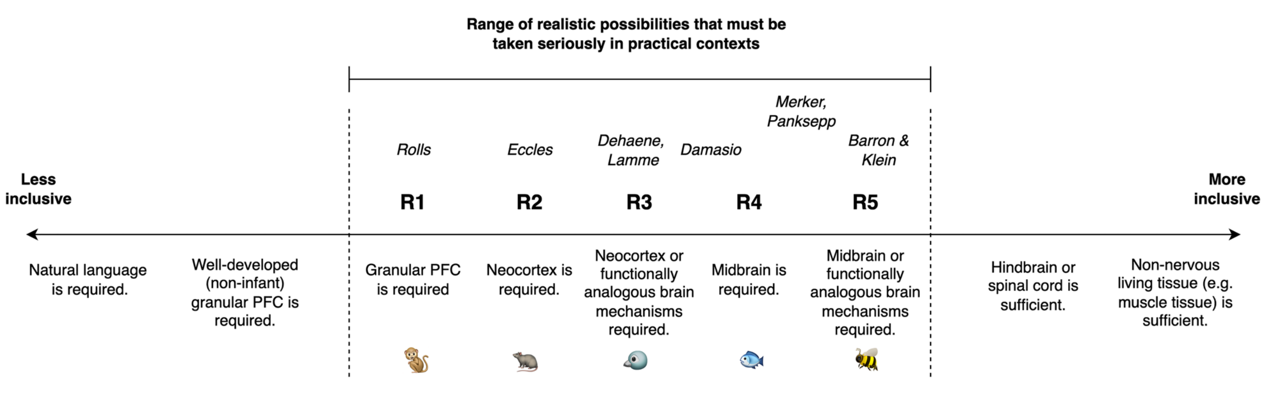

Figure 1. A proposed scientific meta-consensus on the neural basis of sentience. There is no consensus about which position within the reasonable range is correct. However, this is compatible with a meta-consensus forming around the idea that positions R1-R5 are reasonable, given current evidence, provided they are held open-mindedly. Moreover, there can be a meta-consensus on the ordering of these views from less inclusive to more inclusive, and on the challenges facing any view that falls outside this range.

R1-R5 range, but it has affinities with the R2 group, in that it ascribes a special status to the neocortex without giving any special emphasis to the *prefrontal* cortex (which is, in fact, de-emphasized).

Meanwhile, fundamental IIT appears to go dramatically further than empirical IIT, claiming that *any* causally integrated system will realize some form of conscious experience, albeit not necessarily a valenced form (and the IIT group has not yet offered a theory of valence). On this view, not only neural organoids but also kidney organoids, gastrointestinal organoids, cardiac organoids, and other types of organoid (or their constituent parts, if causal integration is greater in the parts than in the whole) would have a form of consciousness. I regard fundamental IIT as a highly speculative position, fair to discuss in the seminar room but unsupported by empirical evidence. There is therefore no reason to expand the range of realistic possibilities to make room for this view.

I am not suggesting that views outside the range R1–R5 can be decisively ruled out with absolute, 100% certainty. What I have in mind is closer to the old idea of 'moral certainty': enough confidence to justify setting aside these views when grave practical questions are at stake. Possibilities in the range R1–R5 have amassed enough evidence to deserve serious consideration when important practical questions are at stake, whereas views outside this range have not.

Many people may hope the current meta-consensus is something we can *move beyond* as new evidence comes to light. This could take the form of a narrowing of the range of realistic possibilities, a widening of that range, or a restructuring of the way we think about the range. But holding such a hope is compatible with accepting that the meta-consensus succeeds in capturing the positions we need to take seriously *now*, given the evidence we have.

3. The concept of a sentience candidate

From the idea of a scientific meta-consensus, we can construct the concept of a *sentience candidate*:

A system *S* is a **sentience candidate** if there is an evidence base that:

(a) implies a realistic possibility of sentience in *S* that it would be irresponsible to ignore when making policy decisions that will affect *S*, and

(b) is rich enough to allow the identification of welfare risks and the design and assessment of precautions.

The concept of a sentience candidate is defined in terms of possibilities it would be irresponsible to ignore, given current evidence. There is, inevitably, a value-judgement involved in declaring that evidence has amassed to a point at which it is now irresponsible to ignore it in practical contexts. Judging something to be a sentience candidate is not, therefore, a completely value-neutral exercise.

Yet it is also a judgement that must be informed by the scientific meta-consensus just described. We can appeal to the meta-consensus to explain why disconnected spinal cords, zygotes, neural and non-neural tissue samples, organs other than the brain, non-neural organoids (such as kidney organoids) and unicellular organisms are not sentience candidates. One can speculate, in the seminar room, about sentience in these systems, but responsible precautionary actions cannot be based on these speculations.

To judge a system to be a sentience candidate, then, involves scientific and evaluative components: like many other judgements that have to be made at the science-policy interface, it is a 'mixed' judgement (Alexandrova, 2018; Plutynski, 2017). The concept captures a delicate threshold in our evidential and practical situation. When the threshold is crossed, a substantial enough evidence base exists to allow responsible, informed discussion of possible precautionary actions.

How could a system fail to be a sentience candidate? A medically important example of the first type of case is a patient who definitively meets the clinical criteria for brain death. In practice, these criteria test for the irreversible cessation of functional brainstem activity, not literally for the death of the cells in question. It is not straightforward to establish an irreversible loss of brainstem function, leading to continuing debate about the correct criteria (Greer *et al.*, 2020; Walter *et al.*, 2018). But let us focus on a case where the irreversible cessation of functional brainstem activity has been conclusively established. In this case, what remains is not sufficient for sentience on any view in the R1–R5 space. This is why doctors are legally permitted to remove organs and tissues from registered organ donors who are brain dead. This is perhaps the most significant illustration of the idea that a hidden meta-consensus can exist regarding the parameters of reasonable debate when grave issues are at stake. Because we agree that, *if* all brainstem function has irreversibly ceased, the patient is no longer sentient, serious disputes can focus on the question of whether brain death has been accurately determined.

The concept of a sentience candidate is a bridging concept that helps us move from disagreement in the realm of theory to agreement on a course of action. When a being is a sentience candidate, there will be at least one reasonable, scientifically credible basis for taking steps to protect its welfare. That should trigger us to at least start talking about what the reasons *against* might be, and what an all-things-considered *proportionate* response that does justice to the reasons on both sides might look like. By contrast, if a system is not even a sentience candidate, the bar for triggering this process is not cleared.

We can capture this thought in the form of a “Sentience precautionary principle” (intended to be more general than the “Animal Sentience Precautionary Principle” defended in Birch, 2017):

Sentience precautionary principle. If *S* is a sentience candidate, then it is reckless/negligent to make decisions that create risks of suffering for *S* without considering the question of what precautions are proportionate to those risks. Reasonable disagreement about proportionality is to be expected, but we ought to reach a policy decision rather than leaving the matter unresolved indefinitely.

4. Brain organoids: no risk of sentience?

With this general precautionary framework in place, let us turn back to brain organoids. I want to start by considering possible reasons to think current neural organoids (at the time of writing) are not sentience candidates. A simple reason often given is their size. This is not a persuasive reason. Bees have around 1 million neurons, and they are sentience candidates. There are existing brain organoids of a similar size, in terms of neuron count, and researchers aim to create organoids with around 10 million neurons (Smirnova *et al.*, 2023).

A second simple reason, in my view more on-target than the first, is that organoids are not living organisms. They are pieces of tissue, and a default attitude of scepticism towards the idea of sentient tissue, outside of any living animal, is appropriate. Neuroscientists have experimented with small samples of cortical tissue for many years without anyone suggesting a risk of sentience. We must ask: given that cortical tissue samples are not normally sentience candidates, what is different about this type of cortical tissue sample that should cause us to worry? This creates a legitimate default bias against sentience if there is no evidence to the contrary.

Moreover, we should take account of what is missing from present-day organoids. Current neural organoids are typically clusters of cortical neurons, without connections to a functioning brainstem. On Merker’s theory, mechanisms at the top of the brainstem, in the midbrain, are constitutively involved in conscious experience (Merker, 2007). Advocates of these theories should be sceptical of the idea of sentience in a neural organoid composed only of cortical tissue. The situation is different when an organoid is implanted into the brain of a host animal (typically a mouse or rat) to create a chimera. These chimeras are clearly sentient, but that is because the host animal is sentient, and the hard question becomes one of how the new tissue alters its cognitive capacities and welfare needs, and there are huge evidence gaps in this area (for commentary on the ethical implications of these evidence gaps, see Birch & Browning, 2021). But in the case of a cortical organoid that is not implanted into a host, midbrain-centric theories give no grounds for attributing sentience.

Here there is an interesting inversion of debates about non-mammalian animals. In the animal case, there is a certain familiar pattern: those who suspect subcortical mechanisms are constitutively involved in consciousness take the possibility of sentience very seriously in a wide range of cases, whereas those who think only neocortical mechanisms are constitutively involved are inclined to play down the risk. Current cortical organoids present us with the opposite situation. They generally lack the subcortical mechanisms taken to be so important by Merker, Panksepp, Solms, Feinberg and Mallatt, Ginsburg and Jablonka, and others. Yet they do have cortical tissue that resembles the neocortical tissue of a developing human brain. So now it is a different family of theories—neocortex-centric theories—that recommend taking the risk of sentience more seriously.

Even defenders of neocortex-centric theories, however, will normally grant a crucial role to the brainstem in supporting conscious experience in humans. The idea is typically that brainstem mechanisms, and in particular the reticular activating system, are akin to a “power cable” for conscious experience, switching it on without being part of its constitutive basis, just as your computer’s power cable makes it possible to run a software programme without itself running that programme. Current organoids lack this “power cable” and accordingly display no sleep-wake cycles, to my knowledge.

We should feel pressure towards consistency: when an adult human patient displays no sleep-wake cycles and no brainstem reflexes, and when this condition is irreversible, they are declared “brainstem dead”, *regardless of the amount of cortical tissue they still possess*. Cortical tissue alone is not enough for sentience candidature, even if one thinks the constitutive basis of sentience lies in the neocortex.

Indeed, as I understand it, a major limitation of current organoids (when not implanted into host animals) is that they are not fully vascularized: they lack active blood flow. As I write, labs around the world are trying hard to overcome this limitation by joining up neural organoids to vascular organoids, with varying degrees of success (Matsui *et al.*, 2021; Shirure *et al.*, 2021; Sun *et al.*, 2022). We cannot rule out the possibility that fully vascularized organoids will be developed very soon, or even by the time this article is published. But as things stand at this moment, it seems a basic pre-requisite for any cognitive function or conscious experience in a human brain is absent in brain organoids.

5. Early warning signs

For all this, there are concerning signs about the potential for organoid research to accelerate rapidly towards the edge of sentience. In the case of disorders of consciousness, the search for electrophysiological markers of conscious experience has been underway for decades. Synchronized, rhythmic oscillations of local field potentials—informally known as brain waves—have long been seen as one of the most important sources of potential markers. Despite a continuing lack of consensus about exactly which oscillations matter, there is widespread consensus about the idea that they are promising places to look.

Trujillo *et al.* (2019) allowed cortical organoids to develop for an unusually long period of time, 10 months, and recorded their electrophysiological activity through weekly recordings. They charted the emergence of complex oscillatory waves. They found that organoids quickly settled into a pattern of switching "between long periods of quiescence and short bursts of spontaneous network-synchronized spiking" (Trujillo *et al.*, 2019, p. 562). These synchronized "network events" became stronger and more frequent over time, while the intervals between events became more variable.

This broad pattern of increasingly strong and frequent bursts of activity, with less predictable intervals, is also seen in the EEGs of preterm infants. In an eye-catching result, Trujillo *et al.* showed that a regression model predicting a neonate's developmental age from key features of its EEG recording, and trained only on data from preterm infants, could also judge the developmental age of organoids older than 25 weeks with above-chance accuracy, with moderate correlation between the predicted and actual ages.

The result must be carefully interpreted. This does not show that the organoids were in any sense equivalent to the brains of preterm infants. It is important to note, first of all, that these cortical organoids were not *brains* at all. We should take care to avoid terms such as "mini-brain" for systems like these. The organoids were formed of a single type of tissue—cortical tissue—representative of one particularly important brain region, the neocortex. The organoids were vastly smaller than an infant brain, and still lacked a brainstem and vascularization. Nor does it show that the electrophysiological activity was the same or indistinguishable in the two cases. The regression model aimed to exploit the similarities that existed, not quantify the degree of similarity. The model identified enough similarities to inform above-chance predictions of developmental age, but this is compatible with substantial differences.

Nonetheless, the result was, to me, a wake-up call: a jolt out of complacency about the potential ethical implications of this research. Brain organoids develop, they are sometimes allowed to develop for a long time, and they develop in ways that show broad electrophysiological similarities to the developing human brain.

6. Assessing sentience candidature in brain organoids

We cannot rule out the possibility that sufficiently sophisticated organoids will soon be sentient, and we can expect the science to continue to develop extremely rapidly. So, we need to have a discussion now about what sort of warning signs might suffice to regard an organoid as a sentience candidate.

Here we run into a serious problem. In people with prolonged disorders of consciousness (another difficult case), some behaviour remains, despite the tendency to describe patients as "unresponsive", and that behaviour informs diagnosis and the design of precautions (Johnson, 2022). Clinicians (in the UK) are already advised to respond to outward signs of pain, distress, anxiety and depression on the precautionary assumption that they really do indicate those states. The behaviour may be involuntary much of the time, but it is behaviour nonetheless. Sleep-wake cycles are also present, marking a clear distinction with coma.

Meanwhile, in the case of non-human animals, the most widely accepted markers of sentience again tend to be behavioural. Animal welfare experts have formulated lists of such markers, generally focusing on pain (Birch *et al.*, 2021; Sneddon *et al.*, 2014). Organoids present a very different kind of challenge. None of these behavioural markers of sentience are likely to be present, because organoids are typically cut off from the sources of sensory input and motor output that are available to a complete and developed organism.

This could turn out to be an incorrect assumption. Some future organoids, even in the near term, may well have sources of sensory input and motor output. For example, a recent study showed that under the right conditions a cortical organoid can spontaneously develop optic vesicles—the developmental precursors to eyes—and it is not yet known how far this process could go, as the technology develops (Gabriel *et al.*, 2021). Another study allowed organoids to develop in culture for a year, placed near to a spinal cord and muscle tissue taken from a mouse. The organoids "were able to innervate mouse spinal cord" and "evoke contractions of adjacent muscle" (Giandomenico *et al.*, 2019, p. 669).

On this evidence, a time when organoid preparations can be joined up to both muscle outputs and sensory inputs is not far off. At that point, public concern about the research may grow. At the same time, using behavioural criteria to assess the likelihood of sentience may also become more feasible, providing a new way in which public concern could be exacerbated or at least slightly eased, depending on the results. Negative results would still require very cautious interpretation, because a failure to display sentience-related behaviours could easily reflect a failure of coordinated muscle control and a very limited behavioural repertoire rather than a lack of sentience.

Sentience, then, may be both more likely and easier to attribute when a neural organoid is joined up to other tissues, be they themselves organoids or taken from animals. But let us focus for now on the case of a "pure" brain organoid, disconnected from any other tissues and any sources of sensory input or motor output. This is the type of case that presents the deepest puzzle. If the system is sentient, then it is what Bayne, Seth and Massimini (2020) have called an "island of awareness", unable to manifest its sentience in any of the usual ways. In this case, there is no behaviour, so we need to assess sentience candidature using only non-behavioural markers. Where do we even begin?

7. The brainstem rule

There is one important piece of common ground in this area. All reasonable views compatible with the scientific meta-consensus can agree that, in a human brain, there can be no sentience in the absence of a functioning brainstem. Agreement *that* this is the case is much wider than agreement about *why* it is the case. For the midbrain-centric family of theories, mechanisms at the top of the brainstem are sufficient for sentience without a cortex. For the cortex-centric family, midbrain mechanisms are causally but not constitutively involved. They help regulate the global state of consciousness without being part of its neural basis.

All parties can agree, however, that sentience in humans depends on brainstem activity. Without a living brainstem, a human cannot maintain coordinated patterns of global cortical activity, integrative subcortical activity or sleep-wake cycles. Theorists from right across the zone of reasonable disagreement are able to agree that irreversible loss of brainstem function implies the irreversible loss of consciousness. The main challenge is determining exactly when "irreversible" loss has occurred.

We should add a caveat in the interests of future-proofing. Strictly speaking, what is required is a functioning brainstem *or a functionally equivalent system* that regulates arousal and supports sleep-wake cycles. It could be that, in the distant future, artificial brainstems will be created to allow people to recover from currently irrecoverable brain injuries. Such a person would clearly be a sentience candidate, despite lacking a biological brainstem. This is a long way off, but what may be much closer is the possibility of a small-scale functional equivalent that is able to regulate the activity of an organoid in the same way a brainstem would. Even in the absence of a biological brainstem, we should be wary of the risks posed by attempts to use artificial brainstem-like systems to regulate and coordinate cortical activity in organoids.

This common ground is at the root of the widespread view that current organoids are *not* sentience candidates. But it also gives us one threshold for the point at which organoids *will* become sentience candidates. If an organoid (or assembloid) is developed that has a functioning brainstem or artificial substitute that regulates arousal and leads to sleep-wake cycles then, no matter how small it is, it should be regarded as a sentience candidate. There would be at least one view within the zone of reasonable disagreement (namely a midbrain-centric view along the lines of Panksepp, Merker and Solms) on which such a system would be likely to be sentient. The outward signs of regulated arousal and sleep-wake cycles would, in this context, be indicators that the conditions Panksepp/Merker/Solms regard as sufficient for sentience are plausibly in place.

We can call this proposal the "brainstem rule"

> ***Brainstem rule:*** If a neural organoid develops or innervates a functioning brainstem (including the midbrain) that regulates arousal and leads to sleep-wake cycles, then it is a sentience candidate. A functional equivalent of a brainstem (even if artificial) would also suffice.

This is proposed as a *sufficient condition for sentience candidature*. To be clear, it is not proposed as a sufficient condition for sentience (since the Panksepp/Merker/Solms view is a realistic possibility, not a certainty), nor is it proposed as a necessary condition for sentience candidature. The idea is that, when the condition is satisfied, we are in a situation in which we can no longer have confidence that sentience is absent (in contrast to the case of brainstem death) and so should start considering precautions. The proposal leaves open the possibility that there may be *other* scenarios in which we should consider precautions. I am describing here a route to sentience candidature that runs via taking midbrain-centric theories of consciousness seriously, but there may well be other routes, running via different theories.

The proposal says "develops or innervates", highlighting two different ways in which an organoid could acquire a functioning brainstem. One is spontaneous development, along the lines of the optic vesicles spontaneously developed by an organoid in the Gabriel *et al.* (2021) study. The other is through innervating animal tissue, along the lines of the innervation of a spinal cord by an organoid in the Giandomenico *et al.* (2019) study.

We may well find that future model systems in neuroscience increasingly blur the boundary between organoids and chimeras, as more and more living brain tissue from a host animal is used in mixed human-animal "preparations". One can imagine a future variation on Giandomenico *et al.* (2019) that takes the whole living brainstem from a mouse, not just the spinal cord, and connects it to an organoid. Such a system may realistically possess the midbrain mechanisms that lead us to regard humans with conditions such as hydranencephaly as sentience candidates. So, the pressure of consistency should push us towards regarding this system as a sentience candidate too.

A controversial aspect of the proposal is that it implies a system that is clearly *not* a complete, embodied living organism *can* nonetheless be a sentience candidate. That is indeed what I am proposing. The brainstem rule says, in effect, that what I earlier called a "legitimate default bias" against the idea of a sentient non-organism should not be allowed to become a hard assumption that leads us to neglect risks. Although we can be confident that destroying the brainstem "pulls the plug" (so to speak) on sentience in a living organism, we have no right to be similarly confident that taking away the rest of the body while leaving the brainstem fully functional would have the same effect. Disembodied brain organoids with functioning brainstems would be intrinsically similar systems, reached by building up rather than by stripping away.

8. Possible regulatory frameworks

The proposed "brainstem rule" leaves open what would be a proportionate response to an organoid's sentience candidature. It may be tempting to think: even if an organoid is sentient, it is at no serious risk of harm, because harm requires nociception and a capacity for bodily sensation. This, however, would be too hasty. Think, for example, of phantom pain: we know that, in adult humans, the brain mechanisms associated with pain can be triggered in the absence of a physical stimulus (Culp & Abdi, 2022). We should take seriously the risk of an organoid developing versions of the pain pathways of a normal human brain (and the pathways linked to other negatively valenced states—thirst, hunger, cold, etc.) without their usual bodily inputs, leading a risk of these mechanisms being activated unpredictably by the environment.

Among the possible responses are a moratorium (time-limited ban) or even just an indefinite ban on the creation of these particular organoids. I say "indefinite" rather than "permanent" because governments are not able to bind their successors, so there can never be a guarantee that a ban will be permanent. I take these seriously as options that may be proportionate, and I resist the idea that they would amount to drastic or radical restrictions on biomedical research. They should be options that are on the table when we debate these issues.

There is, after all, a huge amount of valuable research that can be done on organoids without getting anywhere near the edge of sentience. Researchers could invest their time in simpler neural organoids or in non-neural organoids, such as kidney organoids and gastrointestinal organoids. A similar line of reasoning is often considered plausible in relation to embryos past the legal age limit (14 days in the UK). Yes, we could learn much from research on older embryos, but it is not in keeping with our values to run even a small risk of creating sentient human embryos solely for the purpose of experimentation, and there are many other valuable kinds of research we can prioritize instead, so we should be willing to forego the benefits. The key would be to ensure that the ban is targeted, so that lower-risk forms of organoid research are allowed to continue. An indiscriminate ban on all organoid research would be excessive and disproportionate. It would give no weight to the great promise of organoid research as a potential substitute for research on whole animals.

A less stringent response would be to allow research on sentience candidates, but subject this research to a licensing regime modelled on that of animal research. After all, most animals used in research are sentience candidates (like insects) or sentient as a matter of consensus (like rats and mice). As a society, we permit this research even though it implies some level of suffering to sentient beings. Where research on a potentially sentient organoid might replace research on a clearly sentient animal, like a mouse or a rat, and might even be preferable on scientific grounds, consistency suggests we should at least try to weigh up the harms and benefits of the two possible projects, rather than always favouring animal research. An indiscriminate bias in favour of research on whole sentient animals rather than merely potentially sentient organoids is unwarranted.

The "weighing" here is, admittedly, very challenging, because our uncertainty about the nature of sentience is so severe that we cannot put precise, agreed probabilities on the chance of sentience in an organoid, and opinions vary widely. We cannot expect ethical review bodies to weigh the risks precisely—but I think we can expect them to weigh risk in broad, qualitative terms, and to debate whether imposing clear, known harms on clearly sentient animals is any easier to justify than imposing somewhat speculative potential harms on organoids that are merely sentience candidates.

This line of thought led me to suggest, in a piece with Heather Browning, that we should look to include potentially sentient organoids within the scope of animal experimentation legislation, such as the UK's Animals (Scientific Procedures) Act 1986, commonly known as "ASPA" (Birch & Browning, 2021). This would certainly be more appropriate than treating potentially sentient organoids as mere tissue, and also more appropriate than treating them as if they were whole embryos, when they are not.

Under ASPA, scientists proposing research projects with the potential to cause suffering to animals have to obtain a licence for the work. To be licensed, they need approval from an institutional ethical review board. The board needs to see that the scientists have carefully weighed harms and benefits and duly considered the imperative to reduce, refine, and replace. In this context, "replace" might mean the replacement of work on potentially sentient organoids with work on organoids that lack any brainstem structures and are less likely to be sentient. Researchers should be expected to make a case that they need to create a sentience candidate, and not just a simpler organoid system, to achieve the biomedical goals of the work. The ethical review board should consider whether those goals genuinely make the proposed research justifiable, and whether proportionate steps have been taken to mitigate the risks of causing suffering.

Plainly, it would be controversial to bring a form of human tissue under regulations designed for animal research, for two reasons: we are talking about tissue and not about whole animals, and we are talking about human tissue, not the tissue of other animals. In both ways, the proposal involves extending a general regulatory approach outside the context for which it was originally devised. However, I see the problems here as problems of framing and wording, not deep problems. If ASPA were to be amended to include organoids, it would be wise to rename it. Politically, it may be wiser to regulate organoid research using new legislation modelled on ASPA rather than through amending ASPA itself.

I see both of the above options—an indefinite ban or moratorium targeted at specific types of organoid, and a regulatory framework modelled on ASPA and centred on the idea of harm-benefit analysis—as options worthy of serious discussion. Which option we take depends on broader evaluative questions about the value we see, as a society, in this research, relative to the disvalue of the risks. We may also want to use both options in relation to different types of brain organoid, regulating research on some, banning research on others. Moreover, in cases where organoid research can replace kinds of animal research that are *more* harmful, targeted bans on the relevant kinds of animal research should *also* be on the table. I doubt there will be a one-size-fits-all solution, and for now I want to put both proposals on the table as options that should be debated further.

> ***Response 1 (targeted bans):*** If organoid research leads to the creation of organoids that are sentience candidates, a moratorium (time-limited ban) or indefinite ban on the creation of this particular type of organoid may be an appropriate response. Bans should avoid indiscriminate targeting of all organoid research.

> ***Response 2 (ethical review):*** When a neural organoid is a sentience candidate, research on it, if permitted at all, should be subject to ethical review and harm-benefit analysis, modelled on existing frameworks for regulating research on sentient animals.

To be clear, the proposals in this paper are independent of each other. So, one may still agree that my proposed responses are on the right lines even if one thinks the "brainstem rule" sets the bar in the wrong place, and *vice versa*.

9. Conclusion

To summarise the overall argument: human brain organoids are showing great promise as models of the human brain, models that could potentially replace a substantial amount of animal research. It would be hasty to dismiss the possibility they could develop sentience, (defined as the capacity for conscious experiences with a positive or negative quality). However, scepticism about this idea is appropriate when considering current organoids (at the time of writing). This is not because of their size, but because of their organization. It is a point of consensus across reasonable views that a brainstem-dead human is not sentient, and current organoids lack a functioning brainstem or anything equivalent to one. There are nonetheless some troubling early warning signs, suggesting that organoid research may create forms of sentient being in the future.

Researchers with very different views about the neural basis of sentience can unite behind the "brainstem rule": if a neural organoid develops or innervates a functioning brainstem that registers and prioritizes its needs, regulates arousal, and leads to sleep-wake cycles, then it is a sentience candidate. This is proposed as a *sufficient condition for sentience candidature*. When a system is a sentience candidate, we should take the possibility of its sentience seriously and discuss proportionate steps to protect its welfare, despite continuing uncertainty and doubt.

What steps might be proportionate? If organoid research leads to the creation of organoids that are sentience candidates, a moratorium (time-limited ban) or indefinite ban on the creation of this particular type of organoid may be appropriate, but bans should avoid indiscriminate targeting of all organoid research. An alternative approach, consistent with existing approaches to animal research, is to require ethical review and harm-benefit analysis whenever a brain organoid is a sentience candidate.

Data availability

No data are associated with this article.

Acknowledgements

I thank Tim Bayne, Simon Brown, Heather Browning, Andrew Crump, Katha Dornenzweig, Sarah Diner, Charlotte Gauvry, Syd Johnson, Takuya Niikawa, Lorenzo Sartori, Tomasz Żuradzki, and the members of my Edge of Sentience Reading Group for their comments and advice.

References

Alexandrova A: **Can the science of well-being be objective?** *Brit J Philos Sci.* 2018; **69**(2): 421–445.
Publisher Full Text

Ankeny RA, Wolvetang E: **Testing the correlates of consciousness in brain organoids: how do we know and what do we do?** *Am J Bioeth.* 2021; **21**(1): 51–53.
PubMed Abstract | Publisher Full Text

Arikan BE, van Kemenade BM, Podranski K, *et al.*: **Perceiving your hand moving: BOLD suppression in sensory cortices and the role of the cerebellum in the detection of feedback delays.** *J Vis.* 2019; **19**(14): 4.
PubMed Abstract | Publisher Full Text

Aru J, Suzuki M, Larkum ME: **Cellular mechanisms of conscious processing.** *Trends Cogn Sci.* 2020; **24**(10): 814–825.
PubMed Abstract | Publisher Full Text

Barron AB, Klein C: **What insects can tell us about the origins of consciousness.** *Proc Natl Acad Sci U S A.* 2016; **113**(18): 4900–8.
PubMed Abstract | Publisher Full Text | Free Full Text

Bayne T, Seth AK, Massimini M: **Are there Islands of awareness?** *Trends Neurosci.* 2020; **43**(1): 6–16.
PubMed Abstract | Publisher Full Text

Beck F, Eccles JC: **Quantum aspects of brain activity and the role of consciousness.** *Proc Natl Acad Sci U S A.* 1992; **89**(23): 11357–61.
PubMed Abstract | Publisher Full Text | Free Full Text

Birch J: **Animal sentience and the precautionary principle.** *Animal Sentience.* 2017; **2**(16): 1.
Reference Source

Birch J, Browning H: **Neural organoids and the precautionary principle.** *Am J Bioeth.* 2021; **21**(1): 56–58.
PubMed Abstract | Publisher Full Text

Birch J, Burn C, Schnell A, *et al.*: **Review of the evidence of sentience in**

cephalopod molluscs and decapod crustaceans. LSE Consulting, 2021.
Reference Source

Culp CJ, Abdi S: **Current understanding of phantom pain and its treatment.** *Pain Physician.* 2022; **25**(7): E941–E957.
PubMed Abstract

Damasio A, Damasio H, Tranel D: **Persistence of feelings and sentience after bilateral damage of the insula.** *Cereb Cortex.* 2013; **23**(4): 833–846.
PubMed Abstract | Publisher Full Text | Free Full Text

Damasio AR, Grabowski TJ, Bechara A, *et al.*: **Subcortical and cortical brain activity during the feeling of self-generated emotions.** *Nat Neurosci.* 2000; **3**(10): 1049–1056.
PubMed Abstract | Publisher Full Text

Dehaene S: **Consciousness and the brain: deciphering how the brain codes our thoughts.** Penguin Publishing Group, 2014.
Reference Source

Dryzek JS: **Foundations and frontiers of deliberative governance.** Oxford University Press, 2010.
Publisher Full Text

Farris SM: **The rise to dominance of genetic model organisms and the decline of curiosity-driven organismal research.** *PLoS One.* 2020; **15**(12): e0243088.
PubMed Abstract | Publisher Full Text | Free Full Text

Feinberg TE, Mallatt JM: **The ancient origins of consciousness: how the brain created experience.** MIT Press, 2016.
Reference Source

Gabriel E, Albanna W, Pasquini G, *et al.*: **Human Brain Organoids assemble functionally integrated bilateral optic vesicles.** *Cell Stem Cell.* 2021; **28**(10): 1740–1757.e8.
PubMed Abstract | Publisher Full Text

Giandomenico SL, Mierau SB, Gibbons GM, *et al.*: **Cerebral organoids at the air–liquid interface generate diverse nerve tracts with functional output.** *Nat Neurosci.* 2019; **22**(4): 669–679.
PubMed Abstract | Publisher Full Text | Free Full Text

Ginsburg S, Jablonka E: **The evolution of the sensitive soul: learning and the origins of consciousness.** MIT Press, 2019; 640.
Reference Source

Greer DM, Shemie SD, Lewis A, *et al.*: **Determination of Brain Death/Death by Neurologic Criteria: the world brain death project.** *JAMA.* 2020; **324**(11): 1078–1097.
PubMed Abstract | Publisher Full Text

Herculano-Houzel S: **The human brain in numbers: a linearly scaled-up primate brain**. *Front Hum Neurosci.* 2009; **3**: 31.
PubMed Abstract | Publisher Full Text | Free Full Text

Hubrecht RC, Carter E: **The 3Rs and humane experimental technique: Implementing change.** *Animals (Basel).* 2019; **9**(10): 754.
PubMed Abstract | Publisher Full Text | Free Full Text

Humphrey N: **Sentience: the invention of consciousness**. Oxford University Press, 2022.
Reference Source

Hyun I, Scharf-Deering JC, Lunshof JE: **Ethical issues related to brain organoid research.** *Brain Res.* 2020; **1732**: 146653.
PubMed Abstract | Publisher Full Text | Free Full Text

Johnson JF, Belyk M, Schwartze M, *et al.*: **The role of the cerebellum in adaptation: ALE meta-analyses on sensory feedback error.** *Hum Brain Mapp.* 2019; **40**(13): 3966–3981.
PubMed Abstract | Publisher Full Text | Free Full Text

Johnson LSM: **The ethics of uncertainty: entangled ethical and epistemic risks in disorders of consciousness.** Oxford University Press, 2022; 304.
Reference Source

Kagan BJ, Kitchen AC, Tran NT, *et al.*: ***In vitro* neurons learn and exhibit sentience when embodied in a simulated game-world.** *Neuron.* 2022; **110**(23): 3952–3969.e8.
PubMed Abstract | Publisher Full Text | Free Full Text

Key B: **Fish do not feel pain and its implications for understanding phenomenal consciousness.** *Biol Philos.* 2015; **30**(2): 149–165.
PubMed Abstract | Publisher Full Text | Free Full Text

Koch C: **Consciousness doesn't depend on language.** *Nautilus.* 2019.
Reference Source

Kreitmair K: **Consciousness and the ethics of Human Brain Organoid research.** *Camb Q Healthc Ethics.* 2023; 1–11.
PubMed Abstract | Publisher Full Text

Lamme VAF: **Behavioural and neural evidence for conscious sensation in animals: an inescapable avenue towards biopsychism?** *J Conscious Stud.* 2022; **29**(3–4): 78–103.
Publisher Full Text

Lau H: **In consciousness we trust: the cognitive neuroscience of subjective experience.** Oxford University Press, 2022.
Publisher Full Text

Lavazza A: **Human Cerebral Organoids and consciousness: a double-edged sword.** *Monash Bioeth Rev.* 2020; **38**(2): 105–128.
PubMed Abstract | Publisher Full Text | Free Full Text

Lavazza A, Massimini M: **Cerebral organoids: ethical issues and consciousness assessment.** *J Med Ethics.* 2018; **44**(9): 606–610.
PubMed Abstract | Publisher Full Text

LeDoux JE: **Deep history and beyond: a reply to commentators.** *Philos Psychol.* 2023; **36**(4): 756–766.
Publisher Full Text

LeDoux JE, Birch J, Andrews K, *et al.*: **Consciousness beyond the human case.** *Curr Biol.* 2023; **33**(16): R832–R840.
PubMed Abstract | Publisher Full Text

Matsui TK, Tsuru Y, Hasegawa K, *et al.*: **Vascularization of Human Brain Organoids.** *Stem Cells.* 2021; **39**(8): 1017–1024.
PubMed Abstract | Publisher Full Text

Mediano PAM, Rosas FE, Bor D, *et al.*: **The strength of weak Integrated Information Theory.** *Trends Cogn Sci.* 2022; **26**(8): 646–655.
PubMed Abstract | Publisher Full Text

Mediano PAM, Seth AK, Barrett AB: **Measuring integrated information: comparison of candidate measures in theory and simulation.** *Entropy (Basel).* 2019; **21**(1): 17.
PubMed Abstract | Publisher Full Text | Free Full Text

Merker B: **Consciousness without a cerebral cortex: a challenge for neuroscience and medicine.** *Behav Brain Sci.* 2007; **30**(1): 63–81.
PubMed Abstract | Publisher Full Text

Michel M, Lau H: **On the dangers of conflating strong and weak versions of a theory of consciousness.** *Philosophy and the Mind Sciences.* 2020; **1**(2): 8.
Publisher Full Text

Niikawa T, Hayashi Y, Shepherd J, *et al.*: **Human Brain Organoids and consciousness.** *Neuroethics.* 2022; **15**(1): 5.
Publisher Full Text

Panksepp J: **Affective neuroscience: the foundations of human and animal emotions.** Oxford University Press, 1998.
Reference Source

Plutynski A: **Safe or sorry? Cancer screening and inductive risk.** In: K. C. Elliott and T. Richards (eds.): *Exploring Inductive Risk: Case Studies of Values in Science.* Oxford University Press, 2017; 149–170.
Publisher Full Text

Preuss TM, Wise SP: **Evolution of Prefrontal Cortex.** *Neuropsychopharmacology.* 2022; **47**(1): 3–19.
PubMed Abstract | Publisher Full Text | Free Full Text

Rolls ET: **A Higher Order Syntactic Thought (HOST) theory of consciousness.** In: R. J. Gennaro (ed.): *Higher Order Theories of Consciousness.* John Benjamins, 2004; 137–172.

Rolls ET: **Emotion and decision-making explained.** Oxford University Press, 2014.
Reference Source

Sawai T, Sakaguchi H, Thomas E, *et al.*: **The ethics of cerebral organoid research: being conscious of consciousness.** *Stem Cell Reports.* 2019; **13**(3): 440–447.
PubMed Abstract | Publisher Full Text | Free Full Text

Seth AK, Bayne T: **Theories of consciousness.** *Nat Rev Neurosci.* 2022; **23**(7): 439–452.
PubMed Abstract | Publisher Full Text

Sharma A, Zuk P, Scott CT: **Scientific and ethical uncertainties in brain organoid research.** *Am J Bioeth.* 2021; **21**(1): 48–51.
PubMed Abstract | Publisher Full Text

Shemesh Y, Chen A: **A paradigm shift in translational psychiatry through rodent neuroethology.** *Mol Psychiatry.* 2023; **28**(3): 993–1003.
PubMed Abstract | Publisher Full Text | Free Full Text

Shirure VS, Hughes CCW, George SC: **Engineering vascularized organoid-on-a-chip models.** *Annu Rev Biomed Eng.* 2021; **23**(1): 141–167.
PubMed Abstract | Publisher Full Text

Smirnova L, Caffo BS, Gracias DH, *et al.*: **Organoid Intelligence (OI): the new frontier in biocomputing and intelligence-in-a-dish.** *Front Sci.* 2023; **1**. Publisher Full Text

Sneddon LU, Elwood RW, Adamo SA, *et al.*: **Defining and assessing animal pain.** *Anim Behav.* 2014; **97**: 201–212.
Publisher Full Text

Solms M: **The hidden spring: a journey to the source of consciousness.** Profile Books, 2021.
Reference Source

Steel D: **Philosophy and the precautionary principle.** Cambridge University Press, 2015.
Reference Source

Sun XY, Ju XC, Li Y, *et al.*: **Generation of vascularized brain organoids to study neurovascular interactions.** *eLife.* 2022; **11**: e76707.
PubMed Abstract | Publisher Full Text | Free Full Text

Taschereau-Dumouchel V, Michel M, Lau H, *et al.*: **Putting the "mental" back in "mental disorders": a perspective from research on fear and anxiety.** *Mol Psychiatry.* 2022; **27**(3): 1322–1330.
PubMed Abstract | Publisher Full Text | Free Full Text

Taylor K, Alvarez LR: **An estimate of the number of animals used for scientific purposes worldwide in 2015.** *Altern Lab Anim.* 2019;

47(5–6): 196–213.
PubMed Abstract | Publisher Full Text

Tononi G, Albantakis L, Boly M, *et al.*: **Only what exists can cause: an intrinsic view of free will**. arXiv preprint. arXiv: 2206.02069v3, 2023.
Publisher Full Text

Trujillo CA, Gao R, Negraes PD, *et al.*: **Complex oscillatory waves emerging from cortical organoids model early human brain network development.** *Cell Stem Cell.* 2019; **25**(4): 558–569.e557.
PubMed Abstract | Publisher Full Text | Free Full Text

Tye M: **Tense bees and shell-shocked crabs: are animals conscious?** Oxford University Press, 2016.
Reference Source

Walter U, Fernández-Torre JL, Kirschstein T, *et al.*: **When is "brainstem death" brain death? The case for ancillary testing in primary infratentorial brain lesion.** *Clin Neurophysiol.* 2018; **129**(11): 2451–2465.
PubMed Abstract |Publisher Full Text

Wilson MN, Thunemann M, Liu X, *et al.*: **Multimodal monitoring of human cortical organoids implanted in mice reveal functional connection with visual cortex.** *Nat Commun.* 2022; **13**(1): 7945.
PubMed Abstract | Publisher Full Text | Free Full Text

Yu F, Jiang Qj, Sun XY, *et al.*: **A new case of complete primary cerebellar agenesis: clinical and imaging findings in a living patient.** *Brain.* 2015; **138**(Pt 6): e353.
PubMed Abstract | Publisher Full Text | Free Full Text

Żuradzki T: **Against the precautionary approach to moral status: the case of surrogates for living human brains.** *Am J Bioeth.* 2021; **21**(1): 53–56.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ✓ ✓ ✓ ✓ ✓

Version 2

Reviewer Report 13 May 2025

https://doi.org/10.21956/molpsychol.18875.r27631

✓ **J. Lomax Boyd**
Johns Hopkins University, Baltimore, Maryland, USA
Elizaveta Kalinina
Johns Hopkins University (Ringgold ID:1466), Baltimore, Maryland, USA

Review:

The article by Birch focuses on a central question: what should it mean to take a precautionary approach toward brain organoids as a sentience candidate? In doing so, the author confronts several related questions, including the epistemic issue of what conditions suffice for sentience. A compelling and well referenced line of argumentation is presented structured around a tiered meta-consensus approach that infers the likelihood of sentience based on the possession of various neuroanatomical landmarks. The author advocates for use of the 'brainstem rule' as a benchmark for assessing whether brain organoids are sentience candidates, or at least, if they have something functionally equivalent to a brainstem. We believe the author has already made sufficient improvements to the article by addressing issues raised by other reviewers to merit publication. Thus, we take this opportunity to provide critiques/comments that are aimed more at future work, by the author or others, especially on the epistemological challenges evoked by the sentience precautionary principle.

First, is there good reason to think that brain organoids could already be–or have the potential to become–sentient? The author notes the challenges of using traditional behavioral indicators of sentience on organoids. Yet, previous work has argued how it is possible to measure behavioral-like responses in organoids that have correlates in animals and how these measures may be used to make moral decisions (doi: 10.1007/s12152-023-09538-x). Future ethical considerations based on sentience candidacy may be informed, or refined by, results of conducting these experiments. Consideration of behavioral-like electrophysiological responses from HBOs, if they are consistent with observations found in known sentience candidates, would inform the sentience precautionary principle even in the absence of full neuroanatomical similarities, such as whether they have the cell types, neuroanatomy, or functional connectivity of a brainstem or functional equivalent. This approach can be used in parallel to the meta-consensus contributing the functional markers of responsiveness as measurable evidence for sentience candidates in HBOs. For example, behavioral electrophysiological responses - such as learning, adaptation, or stimulus-specific neural activity - can help align the researched structures with one of the five positions of the meta-consensus framework by offering functional evidence of neural responsiveness. Then, the neural patterns of the obtained standardized organoids may serve as a reference point for evidential thresholds for one of the meta-consensus positions.

Second, most approaches to determining whether brain organoids could be sentient candidates are based on data obtained from beings that are inferred, based on similarity to humans, to be sentient. This is a reasonable place to start. Yet all of these beings are evolutionarily related in ways that are instructive for determining the neuroanatomical origins of specific cognitive traits (doi:10.3758/s13414-019-01760-1). Yet organoids are, arguably, neural systems dislodged from an evolutionary timeline in ways that fundamentally change their developmental trajectory. What is needed, perhaps, is an approach that does not require neuroanatomical similarity based on phylogeny or coherence to familiar neurodevelopmental trajectories. Other approaches, based on information/computational architecture have been proposed as alternative theories that embrace the multi-realizability of cognitive traits, and may be more suited for brain organoids (doi:10.1093/oons/kvae004).

Third, while discourse around the ethics of brain organoids has primarily focused on whether they are sentient (or not), less attention has been paid to resolving the continuing scientific uncertainty about the neural mechanisms that give rise to sentience in the first place–the question of cognitive ontology. Undoubtedly, this uncertainty continues to have ethical implications for all sentience candidates. In contrast to moratoriums, brain organoids could be used as experimental systems for the discovery of neural pathways that give rise to sentience (10.1523/JNEUROSCI.0431-24.2024). The author does not engage with how brain organoids could inform neuroethics discourse on the epistemic grounds for moral consider ability. A precautionary principle that limits our ability to discover the fundamental mechanisms of sentience may, in fact, be too precautionary, with ethical implications for all sentience candidates that exist or could exist in the future. A staged experimental model could be introduced. Safeguards and constraints at particular research phases would still allow for the discovery of the fundamental mechanisms of sentience.

Summary:

The article by Birch contributes to the field by offering an accessible framework for identifying sentience candidates among HBOs. The proposed meta-consensus framework may be useful for guiding future discussions on regulatory policies for research of potentially sentient organoids with the aforementioned issues taken into further consideration.

References

1. Boyd J, Lipshitz N: Dimensions of Consciousness and the Moral Status of Brain Organoids. *Neuroethics*. 2024; **17** (1). Publisher Full Text
2. Cisek P: Resynthesizing behavior through phylogenetic refinement.*Atten Percept Psychophys*. 2019; **81** (7): 2265-2287 PubMed Abstract | Publisher Full Text
3. Boyd JL: Moral considerability of brain organoids from the perspective of computational architecture.*Oxf Open Neurosci*. 2024; **3**: kvae004 PubMed Abstract | Publisher Full Text
4. Kagan BJ, Loeffler A, Boyd JL, Savulescu J: Embodied Neural Systems Can Enable Iterative

Investigations of Morally Relevant States.*J Neurosci*. 2024; **44** (15). PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cognitive Science, Bioethics, Human brain Models, Moral Philosophy

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 05 February 2025

https://doi.org/10.21956/molpsychol.18875.r27709

Federico Zilio
University of Padova, Padova, Italy

The article addresses the topic of human brain organoids and their potential ability to develop forms of sentience and/or consciousness. Jonathan Birch examines the various existing theories on the neural basis of consciousness and outlines a “meta-consensus” on the reasonable theoretical positions to be considered. The article also offers notable pragmatic suggestions about how to regulate brain organoid research in case they may become “sentience candidates”.
The article is well-argued and structured. There are some points that might require some clarification from the author. However, the requests for clarification do not undermine the overall

quality of the work, which remains a valuable contribution to the debate in this constantly evolving field of research. So, just for the sake of discussion, the author can choose whether or not to consider the suggestions, comments and questions.

p. 4: “The crucial concept introduced is that of a sentience candidate: a system that is not certain to be sentient, but which is sentient according to at least one reasonable, evidence-based theoretical position.”
Does this concern the fact that, for some theory, the system “is” sentient per se, or “can be” sentient in the sense of having the dispositional property or capacity? The answer to this question does not change the thrust of the argument, but I think it is an interesting point. A theory that claims, for example, that the midbrain is a sufficient condition for sentience can understand it as a condition for the actualization of sentience (if the organoid has a functioning midbrain structure, it is sentient) or for the disposition toward sentience (if the organoid has a functioning midbrain structure, it can have sentient states if, say, it is properly stimulated).

Consciousness and sentience: Since the article speaks mainly of sentience, but also frequently mentions the concept of consciousness and some neuroscientific theories of consciousness, it might be useful for the author to clarify the relationship between the two concepts. Perhaps, is sentience understood here as a subset of the concept of consciousness, or as a synonym for primordial consciousness manifested in valanced experiences?

p. 9: “We should take care to avoid terms such as “mini brain” for systems like these”. It could be useful to mention here some works that in recent years have emphasised the need to avoid this kind of terminology, to support this argument of the author. For example: Barnhart & Dierickx, 2022[1]; Kataoka et al., 2023[2]; Lavazza & Chinaia, 2023[3].

p. 10: “Although we can be confident that destroying the brainstem “pulls the plug” (so to speak) on sentience in a living organism, we have no right to be similarly confident that taking away the rest of the body while leaving the brainstem fully functional would have the same effect. Disembodied brain organoids with functioning brainstems would be intrinsically similar systems, reached by building up rather than by stripping away”. I am not convinced that there is an equivalence between ‘building up’ and ‘stripping away’. To say, a brain (also provided with brainstem) originally built up in the absence of a body might be structurally and functionally very different from a brain (provided with brainstem) to which the rest of the body has been stripped away. The latter could be structured and influenced in its formation by connections with the rest of the body and by such stimulations and be much more likely to develop sentience than the brainstem-bound brain without any connection to other apparatuses, organs and external stimuli (even without necessarily espousing an embodied mind or neuro-ecological perspective). Indeed, as noted by the same author, more attention should be paid to those organoids that may eventually develop a brainstem and also other receptive structures and apparatuses. The question raised in the previous comments, of the difference between the dispositional capacity for sentience and the actual development of sentient states, also reappears here.

p. 11: “Among the possible responses are a moratorium (time-limited ban) or even just an indefinite ban on the creation of these particular organoids. [...] There is, after all, a huge amount of valuable research that can be done on organoids without getting anywhere near the edge of sentience”. This seems to me a very reasonable position regarding the use of organoids. I have a few questions out of pure curiosity. According to the author, could one accept the use of organoids that can be considered sentience candidates by adding anesthetic practices, as is already the case in several animal experiments? Certainly, the use of anesthetic would result in the loss of relevant information about certain neurophysiological mechanisms of organoids, but could this be considered a compromise? Or does even the mere possession of dispositional capacity of sentience render them unusable (even if they cannot experience valanced experiences because of the anesthetic)? Additionally, let us suppose we are faced with the possibility of choosing (with equal scientific results) between a sentience candidate organoid and an equally sentience candidate living being (let's assume we have the same level of certainty about the presence of sentience) and that we can use anesthetics on both. In other words, with an equal possibility of sentience, would there be additional criteria to discern between the two? If so, what criteria? For example, the fact that one of the sentient candidates (e.g., an insect) is a “complete” living being? Or the fact that the organoid establishes a relationship, however indirect, with the human donor?

References

1. Barnhart AJ, Dierickx K: The Many Moral Matters of Organoid Models: A systematic review of reasons.*Med Health Care Philos*. 2022; **25** (3): 545-560 PubMed Abstract | Publisher Full Text
2. Kataoka M, Gyngell C, Savulescu J, Sawai T: The importance of accurate representation of human brain organoid research.*Trends Biotechnol*. 2023; **41** (8): 985-987 PubMed Abstract | Publisher Full Text
3. Lavazza A, Chinaia AA: Human cerebral organoids: the ethical stance of scientists.*Stem Cell Res Ther*. 2023; **14** (1): 59 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and does the work have academic merit?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

No source data required

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Neuroethics, bioethics, neuroscience, philosophy of mind

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 21 January 2025

https://doi.org/10.21956/molpsychol.18875.r27707

Jon Mallatt
University of Idaho, Moscow, Idaho, USA

This is a good paper about why human cortical brain organoids are not conscious, and how to regulate the use of these organoids if they become conscious from future progress. The arguments are logical, thorough, and well-stated while offering fresh perspectives on major questions in consciousness studies. They join good philosophy with good, evidence-based, science. Although I am not an ethicist, the ethical arguments and regulatory recommendations make sense to me.

High praise goes to the author for introducing the meta-consensus idea to the contentious field of consciousness studies (p. 4). This offers an excellent way to constrain the debate and lessen the useless arguments. It is such an obvious solution that one wonders why the investigators of consciousness did not introduce it earlier.

Another brilliant idea (pp. 5-6) was to characterize the realistic positions on what entities have sentience into the categories R1-R5, ranging from primate-only/prefrontal cortex (R1) all the way to every vertebrate and some invertebrates/with subcortical brain regions or equivalents (R5). Not only does this nicely relate the necessary brain regions to particular animal taxa, but it can also free the field from its crippling overemphasis on primates, mammals, and the neocortex as the subjects of consciousness investigation. I hope this R1-R5 scheme becomes widely adopted in the field. I also appreciate how the scheme uses the existing evidence to downgrade and exclude unsupported fringe theories that say all cells, plants, and all matter (panpsychism) are sentient, or claim preverbal human babies are not sentient (p. 6 and Figure 1).

I have an objection to the term “midbrain theories” or “midbrain-centric theories” of consciousness (used on pages 5, 6, 10, etc.). I know this is what numerous authors have called the R4 and R5 theories, but it is not anatomically correct. That is, the theories in question don’t only attribute valenced sentience to the midbrain but also to much of the diencephalon (e.g., hypothalamus, habenula) and to the subcortical, limbic telencephalon (medial amygdala, septal nuclei, medial forebrain bundle, etc.). The correct way to condense all this is to change the term to “midbrain+limbic forebrain” theories.

I agree with the assessment of Integrated Information Theory on page 7, and why it should be dismissed as not supported by enough empirical evidence. I suggest citing another article, which makes similar arguments, in more depth: Mallatt, J. (2021[1]). A traditional scientific perspective on the integrated information theory of consciousness. *Entropy*, *23*(6), 650.

I like the characterization of the reticular activating system as a “power cable” for conscious experience, but not constitutive for this consciousness (p. 8). It is a powerful analogy that I had not heard before.

I like how certainties of human consciousness were used to construct criteria for judging if organoids are conscious. One example is that people without a living brainstem lack sleep-wake cycles, so such cycles can be used to recognize if organoids are conscious (p. 10). Another example is that such people can have a functional neocortex, but they still lack consciousness, meaning the organoids must have a brainstem-analogue to be conscious (p. 8).

Despite its many good arguments for why the current generation of organoids lack sentience (p. 8), the article underemphasizes the important argument that they lack sentience because they lack sensory input (and motor output). This is the embodied-consciousness argument that says sentience only exists because it receives and processes sensory input (through many levels of a nervous system) to influence motor outputs in a special way. Organoids don't have the input or the numerous processing levels and may not have these in the foreseeable future; therefore, being so isolated, they are not conscious. Also, consciousness seems to emerge from extremely complex, highly integrated neural hierarchies (Feinberg and Mallatt, J. 2020[2]. Phenomenal consciousness and emergence. *Frontiers in Psychology*, *11*, 1041); we cannot assume that organoids could ever develop so much intricate organization until we see real evidence for that.

I got some of the above ideas from a new article that was published in the past five months, on the same topic as Birch's present article. It is Kosik, K. S. (2024). Why brain organoids are not conscious yet. Patterns, *5*(8). Author Birch may want to read that article, even though it is unfair to ask him to add it to his revision, given that his current paper was first posted/published back in 2023. It seems unfortunate that this open-review process went on for so long because it made many of the author's ideas susceptible to being scooped.

Despite my wish for the present article to say more about how the current organoids' lack of sensory input precludes their being sentient, the article adequately discussed this topic for future organoid research, in Sections 5 and 6. Especially see page 9, about the optic vesicles and the organoids that innervate muscle cells through a spinal cord. The topic was also discussed in the sentence, "On this evidence, a time when organoid preparations can be joined up to both muscle outputs and sensory inputs is not far off." Thus, the author need not make changes about the sensory inputs after all, unless he wishes to do so.

But there is one exception to this. I just don't think the point about phantom limb pain succeeds (p. 11). In claiming sentience could exist without any sensory input, it says, "Think, for example, of phantom pain: we know that, in adult humans, the brain mechanisms associated with pain can be triggered in the absence of a physical stimulus (Culp & Abdi, 2022)." This misses the fact that before the amputation, the nociceptive inputs did lead to pain and they even helped to mold the pain circuits, so the phantom pain was not in the absence of any stimulus, as claimed. An easy solution, however, is just to delete that phantom-pain sentence and go directly to the next sentence: "We should take seriously the risk of an organoid developing versions of the pain pathways of a normal human brain (and the pathways linked to other negatively valenced states—thirst, hunger, cold, etc.) without their usual bodily inputs, leading a risk of these mechanisms being activated unpredictably by the environment." This claim is feasible and therefore is sufficient to make the point.

Finally, the Conclusion on page 12 says a brainstem criterion for consciousness is that it "registers and prioritizes its needs." This criterion was never mentioned in the body of the article, and it seems to come out of nowhere. I notice that Reviewer Niikawa complained about it in the previous version of the manuscript, so did the author already remove it from the body but accidentally leave it here in the Conclusion?

I hope my suggestions help. Overall, I am suggesting only minor revisions to this excellent article.

References

1. Mallatt J: A Traditional Scientific Perspective on the Integrated Information Theory of Consciousness.*Entropy (Basel)*. 2021; **23** (6). PubMed Abstract | Publisher Full Text
2. Feinberg TE, Mallatt J: Phenomenal Consciousness and Emergence: Eliminating the Explanatory Gap.*Front Psychol*. 2020; **11**: 1041 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Evolution of consciousness and of vertebrates.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 28 August 2024

https://doi.org/10.21956/molpsychol.18875.r27583

Takuya Niikawa
Kobe University, Kobe, Japan

Thank you, Jonathan, for your comments and revisions. As a reviewer, I am satisfied with the revised version.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: philosophy of consciousness

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 23 February 2024

https://doi.org/10.21956/molpsychol.18799.r27369

Takuya Niikawa
Kobe University, Kobe, Japan

This paper is written clearly and well-organised and provides a promising suggestion of how we should regulate brain organoid research. As a reviewer, I present a couple of questions and suggestions I believe the author should address to make this paper clearer and more comprehensive.

1. The author does not discuss the integrated information theory of consciousness (IIT). The situation over IIT is unique in that its proponents believe that it is scientific and reasonable, while its opponents believe that it is pseudo-scientific and not reasonable. The same can perhaps be applied to quantum theories of consciousness (presented in a couple of works by Penrose, R., and Hameroff, S.)[Ref 1]. Here there does not seem to be a meta-consensus about IIT and quantum theories of consciousness. How should we treat such theories over which there is no meta-consensus? Given that some papers on consciousness of brain organoids explicitly discuss IIT, I believe that the author should address this question.
2. Many theories the author mentions in discussing the meta-consensus seem to be directed at the consciousness of living organisms. For instance, (Feinberg TE et al, 2016) [Ref 2] and (Ginsburg S et al, 2019) [Ref 3] and investigate the consciousness of living organisms with evolutional history. Given that such theories are directed at the consciousness of living organisms and our meta-consensus is on those theories, don't we have another meta-consensus that our theories of consciousness are about the consciousness of living organisms? If we have such a meta-consensus, it seems to conflict with the claim that those theories of consciousness are applicable to non-living organisms such as brain organoids. Then the question is, how can we justify applying those theories of consciousness to brain organoids, which are non-living organisms? I believe that the author should say more about the applicability of those theories of consciousness to brain organoids.
3. The brainstem rule states that "if a neural organoid develops and innervates a functioning brainstem that registers and prioritizes its needs [...]". However, it is unclear when and how a brainstem organoid, without being integrated with other organic systems such as rats and monkeys, can "register and prioritize its need". The worry lies in the notion of "need". What needs can an in vitro brainstem organoid have? I believe that the author should further explain this point. If the author has in their mind the conditional potential that if a brainstem organoid is appropriately connected to other organic systems, then it can contribute to registering and prioritizing the needs of the resultant whole system, then the author should make it explicit.
4. As for the paragraph starting from "Even defenders" in Section 4, the author states that even defenders of neocortex-centric theories grant the "power cable" role to the brainstem in supporting conscious experiences. Although this is correct for human consciousness, if neocortex organoids can operate in a way that produces conscious experience (according to some neocortex theory) without being supported by the brainstem, defenders of neocortex-centric theorists would accept that they have consciousness. Then the question is whether neocortex organoids cannot operate in such a way without being supported by a brainstem. I think that there is no in-principle reason to deny the possibility that neocortical organoids can be developed in a way that can produce conscious experiences (according to some neocortex theory such as global workspace theory) without any support of independent brainstem-like structures. Such neural organoids are not connected with any "power cable" but, so to say, have a "battery" inside them. Given this possibility, neo-cortical theorists may not agree on the brainstem rule, instead pushing something like the neocortex rule that if a neural organoid develops or innervates a functioning neocortex that operates in a certain manner, then it is a sentience candidate. The point here is whether we should also take the neocortex rule (and possibly more rules derived from other reasonable theories of consciousness) into consideration for making regulatory frameworks. Or should we choose

one of them by arguing for it? I believe that the author should make this point clear."

References

1. Hameroff S, Penrose R: Consciousness in the universe: a review of the 'Orch OR' theory.*Phys Life Rev*. 2014; **11** (1): 39-78 PubMed Abstract | Publisher Full Text
2. Feinberg TE, Mallatt J: The nature of primary consciousness. A new synthesis.*Conscious Cogn*. 2016; **43**: 113-27 PubMed Abstract | Publisher Full Text
3. Ginsburg S, Jablonka E: The Evolution of the Sensitive Soul. 2019. Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: philosophy of consciousness

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 19 Jul 2024

Jonathan Birch

Thanks Takuya for these constructive, detailed and helpful comments! Open peer review is a bit strange in some ways, but at least it allows me to thank you properly! While the paper was being reviewed, I also discussed it with my "Edge of Sentience Reading Group" at the LSE, and these discussions have led to some refinements to the realistic range of possibilities and to the definition of a sentience candidate. The R1-R4 space is now an R1-R5 space, because I was encouraged to separate out those theories that attach special significance to the prefrontal cortex. To summarise the key changes in v2 (page references below refer to the manuscript, archived on PhilPapers, because I don't have page references for the typeset version):

1. TN asked for a discussion of IIT, a challenging view to locate in the "range of realistic possibilities". I have added a discussion on page 8.
2. TN asked for more discussion of how the theories in the range of realistic possibilities might apply to non-organisms, such as organoids. Naturally I think we have to apply them in a precautionary way, allowing a realistic possibility that embodiment is essential, but also a realistic possibility that embodiment is inessential. I've tried to clarify this at the end of Section 7 (pp. 15-16).
3. TN notes that, in a truly isolated organoid, it's not clear what it would mean to "prioritize needs", since "prioritization" suggests some degree of behavioural control. I agree, and I think my proposed "rule" for sentience candidature should be modified to remove this reference to prioritization. So I've modified it.
4. TN writes "I think that there is no in-principle reason to deny the possibility that neocortical organoids can be developed in a way that can produce conscious experiences (according to some neocortex theory such as global workspace theory) without any support of independent brainstem-like structures"—I think this a good discussion point. It is compatible with my claims in the paper, though, because I present the brainstem rule as just *one route* to sentience candidature, and allow that there may be other legitimate routes.

Thanks again!

Competing Interests: No competing interests were disclosed.

Reviewer Report 20 February 2024

https://doi.org/10.21956/molpsychol.18799.r27364

L. Syd M. Johnson
SUNY Upstate Medical University, Syracuse, New York, USA

This is a fascinating, thought-provoking, and forward-looking paper that is clearly and accessibly written, and coherently considers ethical implications of and potential responses to the development of sentience in human brain organoids. Some points that would benefit from clarification are suggested.

§1. The 3Rs are described but not referenced or cited in the second paragraph.

§4: First mention of implanted organoids. It would be useful for the author to mention earlier in the paper that the discussion concerns only extracorporeal brain organoids.

§5: The wake-up call described as prompted by Trujillo's research is interesting. It is well-documented that brain death (under the UK's brainstem standard, and the whole brain standard used in the US and elsewhere) is compatible with well-preserved electrical activity as visualized by an EEG (which, in any case, detects cortical electrical activity, and as the author has argued in §4, "cortical tissue alone is not enough for sentience candidature…") I take it this doesn't indicate sentience in the brainstem death case, so how does one arrive at the conclusion that "the possibility of sentience is real" there?

§7: I'd caution against language like "Without a living brainstem" and "brainstem is truly dead" as these are not accurate with respect to brainstem death, where the diagnosis involves determining whether there is a functioning brainstem – cellular or structural death is not required. I would recommend the author avoid leaning heavily on brain death altogether, as there is clinical but not social or philosophical consensus that it is death. It would be sufficient to describe the consensus on the brainstem's role in consciousness, and emphasize the incapacity for sentience when the brainstem is nonfunctioning, in support of a rule that having a functioning brainstem (or equivalent) is sufficient for sentience candidacy.

§7: I take it this paper is defending something like R3/R4 as sufficient for sentience candidacy. The "brainstem rule" as a term (along with the text of the rule), however, states that an organoid with a functioning brainstem is sufficient, while R4, if it includes invertebrates, includes animals with functional equivalents to a midbrain/brainstem. As only vertebrates/chordates have brainstems, I wonder if the brainstem rule might be too conservative/exclusive. It would, after all, exclude invertebrates like octopuses for whom, as the author recognizes, evidence of sentience is quite robust. This may just be a matter of adding "or functional equivalent" to the "brainstem rule." Then assessing sentience candidacy will face the epistemic challenge of identifying functional equivalents in novel organoid structures.

I also take it the main concern here is human brain organoids, although the potential for sentience in animal brain organoids would, presumably, be similarly concerning. But there, we might be creating organoids of animals that do not have brainstems (eg an octopus brain organoid), suggesting a different "rule" is needed – or the same rule, with "functional equivalent" again emphasized by the rule.

§7: Shewmon argues that hydranencephaly is compatible with awareness/sentience. Shewmon, D. A., Holmes, G. L., & Byrne, P. A. (1999). *Consciousness in congenitally decorticate children: developmental vegetative state as self-fulfilling prophecy. Developmental Medicine & Child Neurology, 41(6), 364–374.* doi:10.1017/s0012162299000821

§8: Harms to whole, sentient animals are straightforward, but those animals have more than a brain and functional brainstem. They have nociceptors (or the equivalent), psychological states, preferences, etc. It's not so straightforward or clear that an extracorporeal brain organoid with a functioning brainstem could be harmed without, minimally, sensory organs/nociceptors or some other source of experience (which might include psychological states?). Brains themselves are insensate – even a headache isn't a pain of the brain. A brainstem, thus, might be sufficient for sentience, but not for harm. Some discussion is needed here of what might constitute a potential harm to an organoid such that it would be appropriate to respond with targeted bans or ethical review. Scientific research using sentient animals is harmful (or potentially harmful), which is what prompts restrictions or reviews. It matters rather a lot if a human organoid sentience candidate who cannot plausibly be harmed is needlessly protected over a sentient animal for whom the harms of research are clear and present (that is, it would be bad if our concerns about sentience candidates resulted in maintaining a status quo in which animals continue to be harmed).

Two very minor corrections: p4 "… are reasonable, as long [as] one holds them non-dogmatically…"; and p8: "… there is widespread consensus about the idea that they are [a] promising place to look."

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Animal ethics, research ethics, neuroethics, bioethics, brain disorders, xenotransplantation.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 19 Jul 2024

Jonathan Birch

Thank you Syd for these constructive, detailed and helpful comments! Open peer review is a bit strange in some ways, but at least it allows me to thank you properly! To summarise the changes in v2 (page references below refer to the manuscript, archived on PhilPapers, because I don't have page references for the typeset version):

1. SJ asks for a reference to explain the "3Rs" to someone new to the topic—I have added one.
1. SJ notes that "it would be useful for the author to mention earlier in the paper that the discussion concerns only extracorporeal brain organoids"—I have clarified this on p. 3.

1. SJ writes that "how does one arrive at the conclusion that 'the possibility of sentience is real'" in the Trujillo et al. study—I realized this remark was misleading, because it was supposed to be future directed, i.e. "the possibility of near-future sentient organoids is real". I deleted it.
1. SJ writes "I'd caution against language like 'Without a living brainstem' and 'brainstem is truly dead' as these are not accurate with respect to brainstem death"—so I have avoided this language and tried to be clearer about the meaning of brainstem death.
1. SJ writes that "as only vertebrates/chordates have brainstems, I wonder if the brainstem rule might be too conservative/exclusive"—a good point. Although it is only meant as a sufficient condition for sentience candidature (not a necessary condition) I do want it to be as inclusive as it reasonably can be. So I've added "a functional equivalent of a brainstem (even if artificial) would also suffice". This does add a certain vagueness to the rule as well as adding inclusiveness, so there are difficult trade-offs here.
1. Thank you SJ for the Shewmon et al. reference, now added.
1. SJ writes that "it's not so straightforward or clear that an extracorporeal brain organoid with a functioning brainstem could be harmed without, minimally, sensory organs/nociceptors or some other source of experience"—a fair challenge. I think the case of phantom pain is interesting to think about here. Valenced states can arise without their normal sensory causes, and that's a risk we have to take seriously in the case of organoids. I've clarified this premise in the precautionary chain of reasoning at the start of Section 8, p. 16.

While the paper was being reviewed, I also discussed it with my "Edge of Sentience Reading Group" at the LSE, and these discussions have led to some refinements to the realistic range of possibilities and to the definition of a sentience candidate. The R1-R4 space is now an R1-R5 space, because I was encouraged to separate out those theories that attach special significance to the prefrontal cortex. Thanks again!

Competing Interests: No competing interests were disclosed.

Reviewer Report 02 January 2024

https://doi.org/10.21956/molpsychol.18799.r27367

?

Tomasz Żuradzki
Jagiellonian University in Kraków, Kraków, Lesser Poland Voivodeship, Poland

The article presents a clear and simple analysis of the main challenges with the regulatory framework of human brain organoids. The article is well-written, although the engagement with the vast philosophical literature on similar topics (e.g. the precautionary principle) is limited. Moreover, not all conclusions are adequately supported by the argumentations and clarifications of the following fragments are required:

“If the total number of scientists and labs soars…” – not sure whether predictions about the future are necessary for the argumentation (even if the number of animals used for research stays at the same level, the problem remains).

“a sentient human brain in vitro, and this image fills many onlookers with a sense of horror” – There is one option overlooked in the article: researchers can create sentient *animal* brains in vitro (instead of human brains)?

“But a crucial part of a precautionary approach to any issue is *consistency* in our thinking about different risks (Steel, 2015)” – the precautionary principle does not refer to consistency.

This is not a definition of “meta-consensus” as described by Dryzek 2010: “where it captures the idea that people may agree about a lot, even when they disagree about the best policy”.

“precautions aimed at reducing the risk of causing it suffering…” – why ‘causing’ is necessary in this definition of Sentience precautionary principle? And who is an agent that causes it suffering? Why this version would not be enough: “precautions aimed at reducing the risk of suffering…”

lack of evidence of the right kind

“the hard question becomes one of how the new tissue alters its cognitive capacities and welfare needs (Birch & Browning, 2021)” – this is (at least partially) empirical questions and the reference is to a philosophical commentary.

“an indefinite ban” – One can hope to establish an indefinite ban, but it does not mean that the ban will be “indefinite” indeed.

“it is not in keeping with our values to run even a small risk of experimenting on a sentient human being” – This sentence is not precise, we are experimenting on a sentient human being all the time, but we do not kill them during these experiments.

Regarding including “potentially sentient organoids within the scope of animal experimentation legislation,” it is not clear how proportionality is understand. In the standard case ethics review board evaluated proportionality of harming (certainly) sentient animal. In cases of sentience candidates they would evaluate the proportionality of potentially harming something that may (or may not) be sentient. Please elaborate this two meanings of proportionality relevant in these situations.

Why the author sees “an indefinite ban or moratorium targeted at specific types of organoid” as options worthy of serious discussion, but do not propose an indefinite ban or moratorium on current research with animals?

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioethics, ethics, philosophy

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 19 Jul 2024

Jonathan Birch

Thank you Tomasz Żuradzki for these constructive, detailed and helpful comments! Open peer review is a bit strange in some ways, but at least it allows me to thank you properly! To summarise the key changes in v2 (page references below refer to the manuscript, archived on PhilPapers, because I don't have page references for the typeset version):

1. TZ writes "not sure whether predictions about the future are necessary"—I have deleted them.
2. TZ writes "There is one option overlooked in the article: researchers can create sentient animal brains in vitro"—I've rephrased to highlight that non-human stem cells can certainly be used, but that the research has been focussing recently on human stem cells, and that this will be my focus too (p. 3).

1. TZ writes that "the precautionary principle does not refer to consistency"—although the best known formulations do not mention consistency, Daniel Steel has argued that consistency is key to the design of effective precautions that can resist the charge of incoherence. I have clarified this (p. 3).
2. TZ notes that the definition of meta-consensus is mis-stated. Yes, I should have clarified that the observation that "people may agree about a lot, even when they disagree about the best policy" is just background motivation for the idea, not a definition. I have clarified this (p. 4).
3. TZ asks "why 'causing' is necessary in this definition of Sentience precautionary principle"—I have changed the wording here (p. 8). The new wording is "create risks of suffering". I suppose this leaves room for further debate about what it is to "create" a risk and whether a risk needs to be "created" by us to call for precautions, but I will

leave that for further debate. I think the new wording is an improvement.

4. In relation to implanted organoids, “TZ writes that this is (at least partially) empirical questions and the reference is to a philosophical commentary”—I was just pointing the reader towards the past remarks of Browning and myself on the ethical issues. I’ve clarified why the reference is there (p. 11).
5. TZ notes that “it is not in keeping with our values to run even a small risk of experimenting on a sentient human being” is clearly false—I have rephrased it to “it is not in keeping with our values to run even a small risk of creating sentient human embryos solely for the purpose of experimentation” (p. 16).
6. TZ notes that “in the standard case ethics review board evaluated proportionality of harming (certainly) sentient animal. In cases of sentience candidates they would evaluate the proportionality of potentially harming something that may (or may not) be sentient”—I agree that is one of the big challenges an integrated review board would confront. I have added a paragraph on p. 17 acknowleding that the “weighing” here is, very challenging, because our uncertainty about the nature of sentience is so severe that we cannot put precise, agreed probabilities on the chance of sentience in an organoid, and opinions vary widely. Accordingly, we cannot expect ethical review bodies to weigh the risks precisely, but I think we can expect them to weigh risk in broad, qualitative terms, and to debate whether imposing clear, known harms on clearly sentient animals is any easier to justify than imposing somewhat speculative potential harms on organoids that are merely sentience candidates.
7. TZ asks why I “do not propose an indefinite ban or moratorium on current research with animals”—a fair challenge! I do think that, in cases where organoid research can replace kinds of animal research that are more harmful, targeted bans on the relevant kinds of animal research should also be on the table, so I have noted this on p. 18.

Thanks again!

Competing Interests: No competing interests were disclosed.