25 September 2023

When is a brain organoid a sentience candidate?

Jonathan Birch

Centre for Philosophy of Natural and Social Science,
London School of Economics and Political Science,
Houghton Street, London, WC2A 2AE, UK.
j.birch2@lse.ac.uk
http://personal.lse.ac.uk/birchj1
@birchlse

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 brainstem-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 registers and prioritizes its needs, 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.

Key words: organoids, sentience, consciousness, brainstem, regulation, precaution

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 the payoffs have been limited, 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" was coined more than fifty years ago and is now embedded in frameworks for the regulation of animal research around the world. 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). If the total number of scientists and labs soars, as it has done in the last fifty years, then the total number of animals used is likely to soar as well, even if every scientist in every lab is sincerely attempting to replace, reduce and refine.

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, because the cells *are* 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.

As Hank Greely (2021) has observed, brain organoids present an "onrushing ethical dilemma". There are strong 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 human brain in vitro, and this image fills many onlookers with a sense of horror. 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 the human brain constructed from human 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 a crucial part of a precautionary approach to any issue is *consistency* in our thinking about different risks (Steel, 2015), and 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: Sections 2 and 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 constructed 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. Sections 4-8 ask: when is a brain organoid a sentience candidate? I argue that the presence of a functioning brainstem that registers and prioritizes the organoid's needs, regulates arousal, and leads to sleep-wake cycles, is a sufficient condition for sentience candidature. 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, where it captures the idea that people may agree about a lot, even when they disagree about the best policy (Dryzek, 2010). In particular, 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). 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 views about sentience (defined as the capacity for valenced conscious experience) are reasonable, as long one holds them non-dogmatically, remaining open-minded about the other reasonable views:

R1. Sentience depends on the mammalian neocortex and cannot be achieved in any system without a neocortex.
R2. Sentience depends on the neocortex in mammals but can also be achieved by other structures performing relevantly analogous functions (such as the avian pallium).
R3. Sentience does not depend on 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 cannot be achieved in any other way.
R4. Sentience does not depend on the neocortex even in mammals and can be achieved in at least a minimal form by integrative subcortical mechanisms crucially involving the midbrain. It can also be achieved by other structures performing relevantly analogous functions (such as the central complex in insects).

These four positions are ordered from less inclusive to more inclusive. The fourth position is the most inclusive, in the sense that the distribution of sentience in the natural world is likely to be the widest if this position is correct. By contrast, it is not reasonable, given current evidence, to dismiss both the neocortex and the midbrain as irrelevant and defend a view more inclusive than R4. For example, the evidence does not support a view on which the hindbrain or spinal cord can support sentience by themselves or a view on which sentience can persist following

brainstem death. It would also be unreasonable to defend a view less inclusive than R1, such as the view that sentience depends on a capacity for natural language.

There is no consensus about which of the four reasonable options is correct, and each option can be fleshed out in many different ways. For example, Merker (2007), Panksepp (1998) and Solms (2021) all defend midbrain-centric theories that I see as neutral between R3 and R4. Barron and Klein (2016) have defended R4, focusing on the case of bees. Feinberg and Mallatt (2016) and Ginsburg and Jablonka (2019) are also in the R4 camp. Lau (2022) and Dehaene (2014) have put forward cortex-centric functionalist theories most naturally interpreted as versions of R2 (the perceptual reality monitoring and global neuronal workspace theories, respectively). R1 is perhaps the least popular option at present. Beck and Eccles (1992) defended a version of R1, proposing an overtly dualist theory on which the six-layered structure of the neocortex is crucial, so that a functionally analogous but differently organized structure would not suffice.

The consensus lies instead at the meta-level, in the idea that everyone should be able to grant the reasonableness of holding any of the positions R1-R4, given current evidence, provided they are held open-mindedly, in a way that responds to new evidence. They are all realistic possibilities that must be taken seriously. Moreover, everyone should be able to agree on the ordering of these views from less inclusive to more inclusive (**Figure 1**). And everyone should be able to agree on the severe challenges facing any view that sees both the neocortex and the midbrain as unimportant, or any view that regards a fully functional mammalian neocortex as insufficient. That is not to say that such views can be decisively ruled out with 100% certainty, but only to say that current evidence warrants assigning them very low probability.

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 at all for a role for the cerebellum in generating conscious experience, and strong evidence against. The functions of the cerebellum mainly concern muscle control and coordination. Being born without a cerebellum ("complete primary cerebellar agenesis") leads to motor control problems but turns out to be compatible with normal cognitive development (Yu *et al.*, 2014).

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 important practical questions are at stake. We need to maintain a focus on credible, evidence-based theories.

At the other end of the axis, consider a possible theory that denies sentience to pre-verbal human infants (before the age of 2, let's say) on the grounds that they lack a sufficiently developed neocortex. There are, to be sure, serious limits on our ability to gather evidence about infants. The vast majority of experimental work in consciousness science involves human adults or non-human animals. But there is no evidence to support the view that any of the various mechanisms that have been posited to be linked to consciousness, such as the global neuronal workspace or perceptual reality monitoring, are absent in infants.

Moreover, when researchers have looked for signatures of relevant mechanisms in infants, they have found them. A 2013 study found evidence of global ignition in the neocortex (an electrophysiological phenomenon central to the global neuronal workspace theory) in 5-month old infants (Kouider *et al.*, 2013). It is credible to theorize that infants may experience the world in a different way from adults—on this, see Alison Gopnik's (2009) discussion of "lantern" and "spotlight" consciousness—but not credible to theorize they lack conscious experiences altogether.

The same can be said of the idea that sentience requires natural language. There are serious theories, such as Rolls's (2013) "higher-order syntactic thought" theory, that tie conscious experience to conceptual thought in a way that (if the theory were correct) would suggest a narrow distribution of sentience in the animal kingdom. This idea does have a place in the zone of reasonable disagreement. But even Rolls stops short of proposing that natural language is required for the relevant type of conceptual thought, allowing that a "language of thought" (in Fodor's sense) might also be sufficient. This is 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).

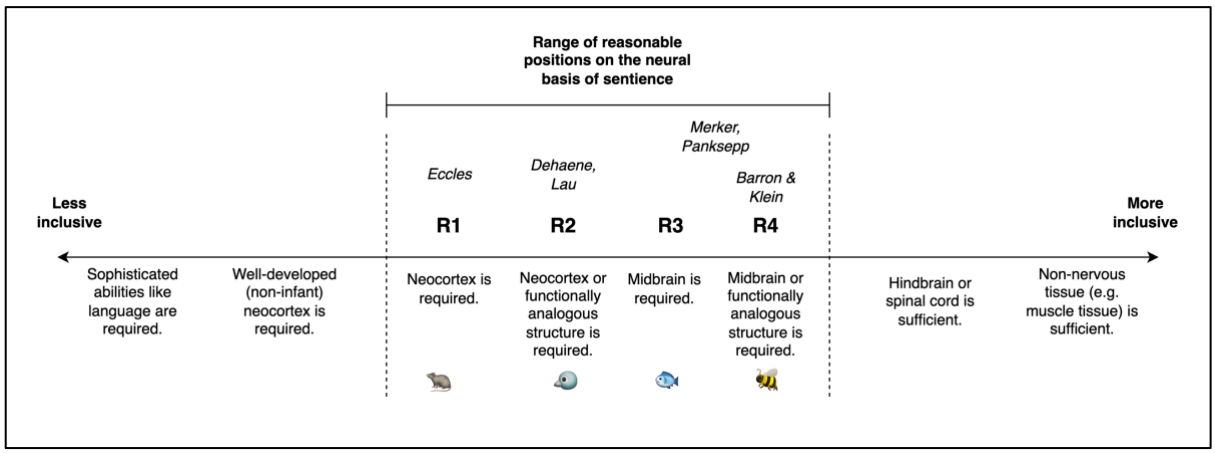

Figure 1: A proposed scientific meta-consensus on the neural basis of sentience. There is no consensus about which position within the reasonable range is correct. However, this is compatible with a meta-consensus forming around the idea that positions R1-R4 are

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.

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 is a ***sentience candidate*** if there is at least one reasonable view (i.e. one view compatible with the scientific meta-consensus) according to which systems of the type in question are sentient.

There is a space of reasonable views, but it is not unconstrained; scientific evidence supplies significant constraints. Sentience candidates are systems to which sentience will be reasonably attributed by proponents of at least one view in the space.

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. The basic thought is this: 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 a system is a sentience candidate, then precautions aimed at reducing the risk of causing it suffering may be proportionate. Reasonable disagreement about proportionality is to be expected, but we ought to reach a policy decision rather than leaving the matter unresolved indefinitely.

There are two main types of case in which a system can *fail* to be a sentience candidate. One is a case in which there is clear evidence that it lacks a functioning forebrain, functioning midbrain, or anything relevantly functionally analogous to either structure, according to any serious, credible theory of what the relevant functional analogies are. The other is a situation in which there a total or near-total lack of evidence one way or the other, making it impossible to make a credible, evidence-based case either for or against sentience.

A medically important example of the first type of case is a patient who is brainstem dead. There is a strong scientific consensus around the idea that brainstem death implies the

irreversible loss of the capacity for conscious experience. The cortical tissue may still be alive but it can no longer function. The spinal cord may still be functional but it is not able to generate experiences by itself. This is why doctors are legally permitted to remove organs and tissues from registered organ donors who are brainstem dead but still on life-support, saving countless human lives.

This is perhaps the most significant illustration of the idea of a hidden consensus regarding the parameters of reasonable debate, which is what my proposed "meta-consensus" is trying to capture. For all the obvious disagreement in consciousness science, we are *very* far from a situation in which people are seriously debating whether sentience persists in a human after brainstem death, despite the loss of all relevant brain mechanisms. That is a view beyond the pale, outside the zone of reasonable disagreement. When a decision to switch off the life-support of a brainstem-dead patient is challenged in court, the courts can correctly cite a secure scientific consensus around the proposition that sentience does not survive brainstem death.

Examples of the second type of case arise when we think about invertebrates. For a very wide range of animals, including octopuses, crabs, lobsters and insects, a credible, evidence-based case for sentience can be mounted, provided we are willing to entertain the Merker/Panksepp/Solms view on which mechanisms functionally analogous to those in the vertebrate midbrain are sufficient for sentience (Barron & Klein, 2016; Birch *et al.*, 2021). So, these animals are sentience candidates. But in other cases, including snails, spiders and oysters, we are faced with a frustrating lack of evidence of the right kind, one way or the other.

This can lead to a temptation to loosen up the concept of a sentience candidate, allowing species to count as sentience candidates where there is no evidence-based case for sentience, but where at least some people's intuitions lean towards sentience. I advise against this, because of the serious practical role I want the concept of a sentience candidate to play. The role of the concept is to trigger evidence-based discussions of proportionate steps to manage welfare risks. Where the evidence base is simply not rich enough to guide the design of precautions or to allow assessments of their proportionality, the right response is to enrich the evidence base as a matter of priority, not to take a guess at what might or might not help to mitigate welfare risks. Evidence, not intuition, needs to guide our thinking at the edge of sentience.

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 (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 place 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. The possibility of sentience is real.

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" (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 a brainstem-dead human is not a candidate for sentience. 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 such a person can never regain consciousness. As already noted, this common ground has major clinical significance, because it makes organ donation possible. It is because there is a robust consensus around the idea that brainstem death implies the irreversible loss of the capacity for conscious experience that doctors are legally permitted to remove organs and tissues from patients who are brainstem-dead but still on life-support.

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 registers and prioritizes homeostatic needs, 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 registers and prioritizes its needs, 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, prioritization, and sleep-wake cycles would be indicators that the conditions Panksepp/Merker/Solms regard as sufficient for sentience are 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 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*. 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.

9. Possible regulatory frameworks

The proposed "brainstem rule" leaves open what would be a proportionate response to an organoid's sentience candidature. Among the possible responses are a moratorium (time-limited ban) or even just an indefinite ban on the creation of these particular organoids. 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 experimenting on a sentient human being, 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.

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. 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*.

10. 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 and Lorenzo Sartori for their comments and advice.

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.

Competing interests

No competing interests were disclosed.

References

Ankeny, R. A., & Wolvetang, E. (2021, 2021/01/02). Testing the Correlates of Consciousness in Brain Organoids: How Do We Know and What Do We Do? *The American Journal of Bioethics, 21*(1), 51-53. https://doi.org/10.1080/15265161.2020.1845869

Barron, A. B., & Klein, C. (2016). What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences, 113*(18), 4900-4908. https://doi.org/doi:10.1073/pnas.1520084113

Bayne, T., Seth, A. K., & Massimini, M. (2020). Are There Islands of Awareness? *Trends in Neurosciences, 43*(1), 6-16. https://doi.org/10.1016/j.tins.2019.11.003

Beck, F., & Eccles, J. C. (1992). Quantum aspects of brain activity and the role of consciousness. *Proceedings of the National Academy of Sciences, 89*(23), 11357-11361. https://doi.org/doi:10.1073/pnas.89.23.11357

Birch, J. (2017). Animal sentience and the precautionary principle. *Animal sentience, 2*(16), 1.

Birch, J., & Browning, H. (2021, 2021). Neural Organoids and the Precautionary Principle. *The American Journal of Bioethics, 21*(1), 56-58. https://doi.org/10.1080/15265161.2020.1845858

Birch, J., Burn, C., Schnell, A., Browning, H., & Crump, A. (2021). Review of the evidence of sentience in cephalopod molluscs and decapod crustaceans.

Dryzek, J. S. (2010). *Foundations and Frontiers of Deliberative Governance*. Oxford University Press.

Farris, S. M. (2020). The rise to dominance of genetic model organisms and the decline of curiosity-driven organismal research. *PLoS One, 15*(12), e0243088. https://doi.org/10.1371/journal.pone.0243088

Feinberg, T. E., & Mallatt, J. M. (2016). *The Ancient Origins of Consciousness: How the Brain Created Experience*. MIT Press.

Gabriel, E., Albanna, W., Pasquini, G., Ramani, A., Josipovic, N., Mariappan, A., Schinzel, F., Karch, C. M., Bao, G., Gottardo, M., Suren, A. A., Hescheler, J., Nagel-Wolfrum, K., Persico, V., Rizzoli, S. O., Altmüller, J., Riparbelli, M. G., Callaini, G., Goureau, O., Papantonis, A., Busskamp, V., Schneider, T., & Gopalakrishnan, J. (2021, Oct 7). Human brain organoids assemble functionally integrated bilateral optic vesicles. *Cell Stem Cell, 28*(10), 1740-1757.e1748. https://doi.org/10.1016/j.stem.2021.07.010

Giandomenico, S. L., Mierau, S. B., Gibbons, G. M., Wenger, L. M. D., Masullo, L., Sit, T., Sutcliffe, M., Boulanger, J., Tripodi, M., Derivery, E., Paulsen, O., Lakatos, A., &

Lancaster, M. A. (2019, 2019/04/01). Cerebral organoids at the air–liquid interface generate diverse nerve tracts with functional output. *Nature Neuroscience*, *22*(4), 669-679. https://doi.org/10.1038/s41593-019-0350-2

Ginsburg, S., & Jablonka, E. (2019). *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness*. MIT Press.

Gopnik, A. (2009). *The Philosophical Baby: What Children's Minds Tell Us about Truth, Love and the Meaning of Life*. The Bodley Head.

Greely, H. T. (2021, 2021/01/02). Human Brain Surrogates Research: The Onrushing Ethical Dilemma. *The American Journal of Bioethics*, *21*(1), 34-45. https://doi.org/10.1080/15265161.2020.1845853

Herculano-Houzel, S. (2009, 2009-November-09). The human brain in numbers: a linearly scaled-up primate brain [Review]. *Frontiers in Human Neuroscience*, *3*. https://doi.org/10.3389/neuro.09.031.2009

Hyun, I., Scharf-Deering, J. C., & Lunshof, J. E. (2020, Apr 1). Ethical issues related to brain organoid research. *Brain Res*, *1732*, 146653. https://doi.org/10.1016/j.brainres.2020.146653

Johnson, L. S. M. (2022). *The Ethics of Uncertainty: Entangled Ethical and Epistemic Risks in Disorders of Consciousness*. Oxford University Press.

Kagan, B. J., Kitchen, A. C., Tran, N. T., Habibollahi, F., Khajehnejad, M., Parker, B. J., Bhat, A., Rollo, B., Razi, A., & Friston, K. J. (2022, 2022/12/07/). In vitro neurons learn and exhibit sentience when embodied in a simulated game-world. *Neuron*, *110*(23), 3952-3969.e3958. https://doi.org/https://doi.org/10.1016/j.neuron.2022.09.001

Koch, C. (2019, 24 September, 2019.). Consciousness Doesn't Depend on Language. *Nautilus*.

Kouider, S., Stahlhut, C., Gelskov, S. V., Barbosa, L. S., Dutat, M., de Gardelle, V., Christophe, A., Dehaene, S., & Dehaene-Lambertz, G. (2013). A Neural Marker of Perceptual Consciousness in Infants. *Science*, *340*(6130), 376-380. https://doi.org/doi:10.1126/science.1232509

Kreitmair, K. (2023). Consciousness and the Ethics of Human Brain Organoid Research. *Cambridge Quarterly of Healthcare Ethics*, 1-11. https://doi.org/10.1017/S0963180123000063

Lavazza, A. (2020, 2020). Human cerebral organoids and consciousness: a double-edged sword. *Monash Bioethics Review*, *38*(2), 105-128. https://doi.org/10.1007/s40592-020-00116-y

Lavazza, A., & Massimini, M. (2018, Sep). Cerebral organoids: ethical issues and consciousness assessment. *J Med Ethics*, *44*(9), 606-610. https://doi.org/10.1136/medethics-2017-104555

Matsui, T. K., Tsuru, Y., Hasegawa, K., & Kuwako, K.-i. (2021). Vascularization of human brain organoids. *STEM CELLS, 39*(8), 1017-1024. https://doi.org/https://doi.org/10.1002/stem.3368

Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences, 30*(1), 63-81. https://doi.org/10.1017/S0140525X07000891

Niikawa, T., Hayashi, Y., Shepherd, J., & Sawai, T. (2022, 2022). Human Brain Organoids and Consciousness. *Neuroethics, 15*(1), 5. https://doi.org/10.1007/s12152-022-09483-1

Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press.

Rolls, E. T. (2013). *Emotion and Decision-Making Explained*. Oxford University Press.

Sawai, T., Sakaguchi, H., Thomas, E., Takahashi, J., & Fujita, M. (2019, Sep 10). The Ethics of Cerebral Organoid Research: Being Conscious of Consciousness. *Stem Cell Reports, 13*(3), 440-447. https://doi.org/10.1016/j.stemcr.2019.08.003

Seth, A. K., & Bayne, T. (2022, 2022/07/01). Theories of consciousness. *Nature Reviews Neuroscience, 23*(7), 439-452. https://doi.org/10.1038/s41583-022-00587-4

Sharma, A., Zuk, P., & Scott, C. T. (2021, 2021/01/02). Scientific and Ethical Uncertainties in Brain Organoid Research. *The American Journal of Bioethics, 21*(1), 48-51. https://doi.org/10.1080/15265161.2020.1845866

Shemesh, Y., & Chen, A. (2023, 2023/03/01). A paradigm shift in translational psychiatry through rodent neuroethology. *Molecular Psychiatry, 28*(3), 993-1003. https://doi.org/10.1038/s41380-022-01913-z

Shirure, V. S., Hughes, C. C. W., & George, S. C. (2021). Engineering Vascularized Organoid-on-a-Chip Models. *Annual Review of Biomedical Engineering, 23*(1), 141-167. https://doi.org/10.1146/annurev-bioeng-090120-094330

Smirnova, L., Caffo, B. S., Gracias, D. H., Huang, Q., Morales Pantoja, I. E., Tang, B., Zack, D. J., Berlinicke, C. A., Boyd, J. L., Harris, T. D., Johnson, E. C., Kagan, B. J., Kahn, J., Muotri, A. R., Paulhamus, B. L., Schwamborn, J. C., Plotkin, J., Szalay, A. S., Vogelstein, J. T., Worley, P. F., & Hartung, T. (2023, 2023-February-28). Organoid intelligence (OI): the new frontier in biocomputing and intelligence-in-a-dish [Frontiers in Science Lead Article]. *Frontiers in Science, 1*. https://doi.org/10.3389/fsci.2023.1017235

Sneddon, L. U., Elwood, R. W., Adamo, S. A., & Leach, M. C. (2014, 2014/11/01/). Defining and assessing animal pain. *Animal Behaviour, 97*, 201-212. https://doi.org/https://doi.org/10.1016/j.anbehav.2014.09.007

Solms, M. (2021). *The Hidden Spring: A Journey to the Source of Consciousness*. Profile Books.

Steel, D. (2015). *Philosophy and the Precautionary Principle*. Cambridge University Press.

Sun, X.-Y., Ju, X.-C., Li, Y., Zeng, P.-M., Wu, J., Zhou, Y.-Y., Shen, L.-B., Dong, J., Chen, Y.-J., & Luo, Z.-G. (2022, 2022/05/04). Generation of vascularized brain organoids to study neurovascular interactions. *eLife, 11*, e76707. https://doi.org/10.7554/eLife.76707

Taschereau-Dumouchel, V., Michel, M., Lau, H., Hofmann, S. G., & LeDoux, J. E. (2022, 2022/03/01). Putting the "mental" back in "mental disorders": a perspective from research on fear and anxiety. *Molecular Psychiatry, 27*(3), 1322-1330. https://doi.org/10.1038/s41380-021-01395-5

Taylor, K., & Alvarez, L. R. (2019). An Estimate of the Number of Animals Used for Scientific Purposes Worldwide in 2015. *Alternatives to Laboratory Animals, 47*(5-6), 196-213. https://doi.org/10.1177/0261192919899853

Trujillo, C. A., Gao, R., Negraes, P. D., Gu, J., Buchanan, J., Preissl, S., Wang, A., Wu, W., Haddad, G. G., Chaim, I. A., Domissy, A., Vandenberghe, M., Devor, A., Yeo, G. W., Voytek, B., & Muotri, A. R. (2019). Complex Oscillatory Waves Emerging from Cortical Organoids Model Early Human Brain Network Development. *Cell Stem Cell, 25*(4), 558-569.e557. https://doi.org/10.1016/j.stem.2019.08.002

Yu, F., Jiang, Q.-j., Sun, X.-y., & Zhang, R.-w. (2014). A new case of complete primary cerebellar agenesis: clinical and imaging findings in a living patient. *Brain, 138*(6), e353-e353. https://doi.org/10.1093/brain/awu239

Żuradzki, T. (2021, 2021/01/02). Against the Precautionary Approach to Moral Status: The Case of Surrogates for Living Human Brains. *The American Journal of Bioethics, 21*(1), 53-56. https://doi.org/10.1080/15265161.2020.1845868