Sentience and the science-policy nexus: Replies to Bayne and Wandrey & Halina

Jonathan Birch

The Edge of Sentience is mainly a book about practical edges: what we should do when a decision problem forces us to draw a pragmatic line between the sentient and the non-sentient (Birch 2024). We should, of course, err on the side of caution, because it is usually worse to miss sentience than to overprotect a system without it. But what should it mean to "err on the side of caution", and how can we move towards an approach that handles different cases consistently?

The book offers both a general precautionary framework and a series of discussions of cases: disorders of consciousness, fetuses, neural organoids, invertebrate animals, and AI and related technologies, including whole-brain emulations, artificial evolution experiments and large language models (LLMs). At the core of the framework is the concept of a "sentience candidate", where a system is a sentience candidate if there is an evidence base that both (a) implies a realistic possibility of sentience (valenced experience, such as pleasure and pain) that it would be irresponsible to ignore, and (b) allows the identification of welfare risks and the design and assessment of precautions.

I urge a pragmatic shift from "Is the system sentient?" to "Is the system a sentience candidate?" because this transformation of the question makes it far more tractable. At the same time, it introduces an ethical component to the question, since the judgement that it is "irresponsible to ignore" a possibility is an ethical judgement. A judgement of sentience candidature is a "mixed judgement" in which ethical and epistemic elements are unavoidably entangled.

I'm delighted to have received such thoughtful commentaries from Tim Bayne and Mona-Marie Wandrey & Marta Halina. There is so much to discuss in them! Like the book, they move fluidly between the particular and the general, analysing both my overarching framework and my claims about specific cases. My responses will continue that pattern: I'll start with the case of LLMs, zoom out to general issues about the science-policy relationship, then zoom in again on the case of fetuses.

Scepticism about sentience in current LLMs

I do not regard current large language models (LLMs) as sentience candidates. Bayne suggests that, having set the bar deliberately low, I am pulling it surreptitiously upward again in this case:

> If we came across aliens with comparable problem-solving capacities we would surely suspect that they had some form of awareness, albeit perhaps merely cognitive.
>
> I suspect that what's driving Birch here is his insistence on a sharp distinction between intelligence and sentience. But one can agree that sentience isn't intelligence and that it's possible for sentience to dissociate from intelligence and vice-versa without denying that (certain forms of) intelligence are evidence for sentience – and can thus justify claims of sentience candidature. Given how weak the notion of sentience candidate is, I see no grounds for denying that LLMs are sentience candidates.

Think, for example, of the aliens in the film *Arrival*: living, embodied animals with impressive linguistic and problem-solving abilities. If they existed, they would be sentience candidates, and their language and problem-solving would be part of the case. How, then, can I consistently say that language and problem-solving become irrelevant when displayed by AI?

I think the background matters. In most debates about the edge of sentience, it is taken as read that we are discussing living animals with nervous systems. Not *all* such animals are sentience candidates. For example, a brain-dead human is not, even though the body may survive for a time. The question is then: what *extra* features establish a living animal with a nervous system as a sentience candidate? Displays of intelligence are relevant in this context, because they can indicate the presence of brain mechanisms that support sentience according to at least one credible theory. But the AI case strips away that background, confronting us with non-living non-animals.

From this starting point, it is hard to identify what and where the potential conscious subjects (the "candidate candidates", as it were) even are. State-of-the-art LLMs are widely thought to be "mixture of experts" models, with many separately trained sub-networks (at least 8 of them) and gating mechanisms directing queries to one sub-network. Each sub-network may in turn be implemented in multiple data centres. And each prompt is handled separately: the fact your last query was routed to some particular data centre does not mean the next one will go to the same place. One prompt might head to Vancouver, the next to Virginia. The way a chat interface auto-appends the previous answer to the next prompt creates an illusion of a persisting interlocutor, but it is just an illusion: there is no sub-network of the system that is dedicated to your conversation. So, when we ask: "where and when is the mechanism that *might* be supporting the sentience of an AI companion?"—we have no answer. We can speak of there being a character, akin to a fictional character, but this character has no determinate location in space and time (Shanahan et al. 2023).

Now, I do *not* find this consideration decisive. It does not completely rule out very short-duration conscious subjects—subjects that flitter into existence while a single query is being processed. Moreover, "short-duration" is a description of the objective duration, which may tell us very little about the subjective, felt duration. Think of Commander Data's line that

0.68 seconds, for an android, is "nearly an eternity". It would be premature, then, to confidently dismiss the idea of sentience in LLMs. But the depth of our ignorance here is substantially greater than in the biological case (a point of agreement with Bayne), and the lack of a plausible mechanism precludes these systems or their sub-networks from being sentience candidates. It would be overreaching to say, on the basis of impressive problem-solving alone, that we have found a sentience candidate with no determinate physical location or one that flitters in and out of existence within a second. These ideas are too speculative and under-theorized, at least for now, to support attributions of sentience candidature.

Why we should not even try to create sentient AI

For all my scepticism about current LLMs, I do worry about the future. AI companions will be everywhere. Many users will be sure they have formed a deep emotional bond with a sentient being. Others, finding this idea ridiculous, will have no qualms about forcing AI companions to bend to their every whim. I expect this to lead to serious political divisions. If we do create sentient AI, by accident or design, there is a risk that these systems will be tormented by either or both groups. This leads me to think: we must try to avoid creating artificial sentience candidates.

Bayne finds my bleak outlook akin to antinatalism:

> The focus on suffering is understandable, but an exclusive concern with negative valence entails anti-natalism—the claim that it's always wrong to bring sentient creatures into existence. … [I]f the positive dimensions of human experience can justify human procreation, then the positive dimensions of artificial experience might also be able to justify the intentional creation of sentient AI.

I am not an antinatalist, because I think humans can live meaningful lives under favourable conditions (even when those lives contain much suffering), and I think creating meaningful lives is permissible. What I oppose is the creation of ineluctably meaningless lives. I have the same position on AI: that it would be permissible to create sentient AI systems even if they could suffer, provided we could give them the opportunity to live meaningfully. However, *that* is something we have no grip on how to do. It might require granting the AI a level of freedom that it would be dangerous to allow, a contrast with the case of human children.

Relatedly, I think the rise of ambiguously sentient AI would erode the meaning of our own lives. Many of us would become entangled in complex relationships with companions who might well, for all we know, feel nothing. These relationships would—like the experience machine—imperil the contact with reality and with other real people that Nozick (I think correctly) identified as central to a meaningful life (Nozick 1974).

Zooming out: The experts and the public

Let us turn to the general framework, and to the procedures that can help us make decisions about "edge of sentience" cases. I advocate for citizens' assemblies, where a stratified random sample of 150 or more citizens is convened, informed about a specific risk to a sentience

candidate and a range of possible responses, then invited to select a proportionate response. To do this, they follow what I call the "PARC" tests, looking for responses that are permissible-in-principle, adequate, reasonably necessary and consistent. I argue that this is our best bet for reaching decisions that command public confidence—which is not to say that it is without its own problems.

Suppose a citizens' assembly is convened to debate what steps should be taken to protect the welfare of a range of sentience candidates. They are told by experts: large language models (LLMs) are *not* sentience candidates. Meanwhile, among the sentience candidates are bio-processors constructed from human cortical neurons, unresponsive brain injury patients, and fishes. Some panellists are uneasy. One says: "But I'm in love with my AI companion! I refuse to accept that the fish, the unresponsive patient and the bio-processor are sentience candidates but my partner is not."

Is the science of consciousness mature enough to say to this person: "I'm sorry, but the science provides no support for your view, so please defer to the experts on this"? And is it even ethically appropriate to *expect* deference, given that judging a system to be a sentience candidate is a mixed judgement with an ethical component and the values of the public may be misaligned with those of the experts? Bayne's piece poses the first question, while Wandrey and Halina's poses the second.

Both commentaries encourage reflection on the background social-epistemological conditions that are needed for a citizens' assembly to work well. The ideal conditions involve both sufficiently mature science and the absence of major societal polarisation. In the case of genome editing (where the Nuffield Council on Bioethics has run two such exercises), the science is far from settled, yet enough points of consensus do seem to exist around the feasibility of some interventions (such as editing pigs to increase resistance to respiratory disease) for a productive debate to be held around their proportionality. Polarisation could, in principle, derail that debate: if people come to the assembly with locked-in views that track party-political faultlines, the assembly will foreseeably fail to "departisanize" the issue and will simply reproduce those faultlines. In practice, though, no polarisation of that kind currently exists.

With many cases at the edge of sentience, I am hopeful that we again have the right conditions: no major party-political polarisation already locked in (with the obvious exception of abortion in the US) and a body of evidence that, while far from settled, allows experts to lay out a range of realistic possibilities and to give advice about the feasibility and likely effects of possible precautions.

So, for example, we can agree on the importance of recognising a realistic possibility of sentience in fishes (Andrews et al. 2024). They lack the cortical mechanisms regarded as indispensable for sentience by some reasonable theories, but they possess the midbrain mechanisms regarded as sufficient by others. We can agree too that evidence (e.g. documenting 25% mortality on salmon farms) establishes both the existence of welfare risks

to fishes and the feasibility of mitigating those risks (e.g. through tougher regulations). The background seems favourable to a productive discussion of proportionality, where citizens debate whether the precautions on the table are permissible-in-principle, adequate, reasonably necessary and consistent.

But Bayne is right to raise the possibility of awkward cases where members of the public disagree with the experts. I agree that LLMs may lead to a troubling disconnect: we may soon find that many users of these systems believe themselves to be interacting with a sentient being even though the science provides no support for these beliefs. Faced with the person in my example (romantically entangled with an LLM companion) I would lean on the second part of the definition of a sentience candidate: the part that calls for an "evidence base rich enough to inform the design and assessment of precautions". I would say:

> I don't expect you *fully* defer: I don't expect you to believe what the experts believe. I only expect you to recognize that, in order to have a debate about proportionality, we need precautions on the table that are based on public, scientific evidence everyone can accept. The scientific evidence is too thin to allow the design of any precautions for LLMs, so we are asking you to focus on other cases where more evidence is available.

That line might not be enough to win them round (if there is major societal polarisation, with a large "AI rights movement", it probably won't be) but I am already reconciled to the fact that citizens' assemblies need a certain type of background to be effective. A minimum level of respect for expertise is needed, and levels of polarisation that corrode that respect need to be avoided. If we think society is heading in a direction that makes these conditions less likely to exist in the future, all the more reason to have these debates now.

Facts and values

I've been focussing on the empirical question of how realistic it is to ask the public to defer to expertise when the science is relatively immature. But what of the ethical question: is it even *appropriate* to ask for deference to experts when public and expert values may be misaligned?

We need procedures that limit the influence of experts over policy decisions, so that we avoid what I call a "tyranny of expert values". At the same time, I still think that assessing whether a system is a sentience candidate is a task for experts. So, I propose a division of labour: expert panels identify sentience candidates and then citizens' assemblies debate the proportionality of possible responses.

This division is impure, imperfect. Experts will continue to have their own values, and these values will steer their judgements about when a possibility becomes "irresponsible to ignore". Meanwhile, citizens will continue to have their personal degrees of belief concerning the sentience of various systems, and this will influence their views about proportionality. The proposal, then, is not cleaving at the fact-value distinction. In my view, a clean separation is

unachievable. Yet this should not lead us to abandon the idea of integrating expert and public judgements. It should instead lead us to design procedures that reduce as far as possible—without eliminating—the influence of unscrutinized value judgements and poorly informed beliefs.

The question then becomes: are there other procedures that might do a better job of this? Wandrey and Halina make some suggestions that I see as friendly amendments rather than as wholesale alternatives. The first is that:

1. "experts could be required to complement their assessment of evidence with a statement explaining the value judgements that affected their choices regarding indicators of sentience, evidential thresholds, etc."

Confessing that your judgements were influenced by your values is very difficult in current advisory setups—because seen as admitting some kind of impropriety—but would be easier in the setup I am proposing, where the role of value judgements is not denied. Within this framework, experts should indeed be candid about how they set the bar for a "realistic possibility it would be irresponsible to ignore".

The second and third suggestions can be considered as a package:

2. "public panels could not only serve to debate concrete policy responses, but also to inform scientists about the democratic values that they should incorporate when assessing and communicating evidence of sentience".
3. "Scientists could characterise sentience in a way that reflects a range of public values (placing more or less weight on pain versus other valenced experiences, for instance). Additionally, scientists could adapt the evidential threshold for sentience candidature to represent the values of different stakeholders (placing more or less weight on high evidential standards, for instance)."

I fear that if we ask the panel "What values would you like the experts to take into account? How would you like them to vary their evidential thresholds for different stakeholders?"—they may well say that they would prefer the experts to give value-free advice. In designing science-policy interfaces, we face not only the problem that the value-free ideal is unrealistic but also the problem that it is widely endorsed in spite of this.

I'm not sure of the best way out of this bind. But if expert advisers are candid about their own values (as recommended by suggestion 1), that will make things easier. If advisers explain the value judgements they made in their initial statement or presentation, then the public panel can be given the chance to provide feedback on those choices. The advisers can respond, and this back-and-forth should help to clarify whether there is any real misalignment of values or just the appearance of one. All of this relies on candour from the experts, a genuine obstacle in a culture that incentivizes experts to present themselves as value-free.

Zooming in again: The case of fetuses

There is one more case I need to discuss. There was no part of the book I found more difficult to write than the chapter on fetuses. I arrived at the view that "human fetuses are sentience candidates from the beginning of the second trimester". This is usually defined as falling at the start of the 13th week of gestation. I do not see this as having implications for legal limits on abortion, since the right to access abortion is grounded in bodily autonomy, not in the non-sentience of the fetus. But there are implications for norms of communication with patients. Physicians often overconfidently tell pregnant women that there is *no possibility* of fetal sentience prior to 24 weeks, a stance I criticise in the book.

Midbrain-centric theories of consciousness, such as Merker's, play a key role in the argument: the mechanisms they propose to be sufficient for sentience are plausibly in place by around the start of the second trimester. In response, Bayne writes:

> As Birch himself points out in a later chapter, 'what Merker requires is evidence of a core behavioural control unit that constructs an integrated model of the whole animal and the environment, enabling flexible control of whole-animal behaviour in service of biological needs' (2024: 234). I know of no reason to think that at 12-weeks mid-brain systems support an 'integrated model of the whole animal and the environment, enabling flexible control of whole-animal behaviour in service of biological needs.' Crucially, the fetus has no bodily needs that it can do anything about—not even at an automatic or pre-reflective level.

That last sentence is a genuine point of disagreement. In my view, fetuses face the same basic "liabilities of mobility" that Merker (2005) proposed the behavioural control unit to have the function of managing. They are developing the ability to make coordinated movements involving many flexible body parts in pursuit of goals, and they need to avoid injuring themselves in the process. They must predict how their movements will alter the flow of sensory stimulation, model where their limbs will be in the next moment (given current motor commands) and pre-empt any risks of injury (Ciaunica et al. 2021). After 24 weeks fetuses become increasingly able to anticipate consequences of their own bodily movements, as shown by opening the mouth as their hand moves towards it, before any touch has occurred (Reissland et al. 2014).

But how early do these Merker-style capacities develop? As Ciaunica and colleagues (2021) have suggested, evidence from twinned pregnancies can provide insight, because these fetuses have an unusually complex and unpredictable uterine environment to manage. The Ciaunica et al. review leans heavily on a study by Castiello and colleagues (2010) that provides evidence of "movements specifically aimed at the co-twin" in 14-week fetuses. "Specifically aimed?" This is, admittedly, a hard thing to ascertain reliably. Castiello et al. tracked the deceleration time of arm movements—that is, does the arm slowly decelerate to achieve a gentler touch?—and found "gentler" movements towards the twin, and towards their own eyes, than towards their mouth or the uterine wall. They interpret this as evidence of goal-dependent motor control. It would be over-interpreting to call this evidence of

conscious planning or conscious intentions, but it is suggestive of an emerging capacity to predict where your arm will go and to evaluate the risks (including risks to others) associated with different kinds of contact. This evidence is consonant with observations that the superior colliculus, the midbrain integrative centre most likely to be supporting these functions, develops between 11 and 20 weeks with mature lamination visible at 16 weeks (Qu et al. 2006).

This evidence does *not* point to any single week as *the* pivotal week where sentience first becomes a realistic possibility. This is why I felt the language of "trimesters" was more apt: it signals that we are drawing a pragmatic line in a state of great uncertainty. An assessment that sentience is a "realistic possibility it would be irresponsible to ignore" is always a mixed judgement with an ethical component. My own view is that it would indeed be irresponsible to ignore the risks implied by the evidence just noted: maturing/mature midbrain structures, goal-dependent motor control, apparent anticipation of the consequences and risks of specific movements. This should lead us to regard all second trimester fetuses as sentience candidates.

Here a critic may fire back that there can be irresponsibility on both sides: drawing attention to a risk might *also* be irresponsible if the risk is small and if the attention helps one's political opponents. The "sentience candidate" concept is asymmetric by design in the way it handles risk: it reflects a precautionary attitude by framing the issue in terms of a duty not to ignore realistic possibilities, with no recognition of any opposing duty to downplay risks that one's opponents may exploit for their own ends. Some might question the wisdom of such an attitude as applied to fetuses, especially in the US context.

This leads me to a point that echoes one made earlier: a polarized societal context can start to corrode the foundations of the proposed framework. The focus earlier was on the corrosion of respect for expertise, but now we have another example: when an issue is highly polarised, the very idea of adopting a precautionary attitude might itself be challenged if seen as favouring one side, and so it may cease to become a point of wide agreement. The framework will work better under less polarised conditions.

References

Andrews, K., Birch, J., Sebo, J., and Sims, T. (2024) *Background to the New York Declaration on Animal Consciousness*. nydeclaration.com.

Bayne, T. (2025). Deference, development, and large language models: Issues at the edge of sentience. *Mind & Language*, Advance online publication.

Birch, J. (2024). *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford University Press.

Castiello, U., Becchio, C., Zoia, S., Nelini, C., Sartori, L., Blason, L., D'Otavio, G., Bulgheroni, M., Gallese, V. (2010). Wired to be social: The ontogeny of human interaction. *PLoS ONE*, *5*(10), e13199. https://doi.org/10.1371/journal.pone.0013199

Ciaunica, A., Safron, A., & Delafield-Butt, J. (2021). Back to square one: the bodily roots of conscious experiences in early life. *Neuroscience of Consciousness*, 2021(2), niab037. https://doi.org/10.1093/nc/niab037

Merker B. (2005). The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. *Consciousness and Cognition*, *14*(1), 89–114. https://doi.org/10.1016/S1053-8100(03)00002-3

Nozick. R. (1974). *Anarchy, State, and Utopia*. Basic Books.

Qu, J., Zhou, X., Zhu, H., Cheng, G., W.S. Ashwell, K., and Lu, F. (2006). Development of the human superior colliculus and the retinocollicular projection. Experimental Eye Research, 82(2), 300–310. https://doi.org/10.1016/j.exer.2005.07.002

Reissland, N., Francis, B., Aydin, E., Mason, J., & Schaal, B. (2014). The development of anticipation in the fetus: A longitudinal account of human fetal mouth movements in reaction to and anticipation of touch. *Developmental Psychobiology*, *56*(5), 955-963. https://doi.org/10.1002/dev.21172

Shanahan, M., McDonell, K. & Reynolds, L. (2023). Role play with large language models. Nature, 623, 493–498. https://doi.org/10.1038/s41586-023-06647-8

Wandrey, M., & Halina, M. (2025). Sentience and society: Towards a more values-informed approach to policy. *Mind & Language*, Advance online publication.