AJOB Open peer commentary (2023, 14, 2, 202–204)

https://doi.org/10.1080/21507740.2023.2188303

Searching for consciousness in unfamiliar entities:

The need for both systematic investigation and imagination

Authors: Sarah Diner (University of Bonn), Maxence Gaillard (University of Oslo, Université catholique de Louvain)

The possibility that human cerebral organoids (HCOs) develop consciousness is one of the main concerns driving current ethical discourse. Evaluating existing evidence, Zilio and Lavazza in their article rely on a selection of theories of consciousness as an extensive overview of the theories of consciousness available in the scientific literature would be out of reach. There are indeed good reasons not to pursue out of hand an exhaustive overview: (i) the number of theories of consciousness "on the market"; (ii) the lack of consensus on what should be considered as a serious candidate or not; (iii) the lack of a common nomenclature (in definitions and concepts) that should guide us to a possible resolution of the theoretical discussion; (iv) the evolving nature of the field of consciousness studies, both in terms of the landscape of dominating theories and in experimental tools (Cova, Gaillard, Kammerer 2021). In Kuhnian (1996) terms, the observer turning to consciousness studies and looking for a dominant paradigm to rely on would instead face a chaos of proposals and experiments that are typical of emerging fields or times of crisis. Selecting some theories, and then applying a sample of those arbitrarily chosen, seems ineluctable.

This being said, any serious attempt to search for consciousness in unfamiliar entities is akin to looking for signs of extraterrestrial life in exobiology: it should be both systematic in the approach and wild in imagination. Yet, because of the lack of consensus and the number of competitors, picking some theories won't either do justice to all the research conducted in the domain nor increase our chances to mobilize the "right" theory, if there is one. We have to be very explicit on the criteria for considering some theories and unfold all their assumptions. However, a portion of wildness is also required. Looking at consciousness emerging in a dish won't be probably just adapting existing theories at the margins. Because we are talking of unfamiliar entities, we have to be open to any form of consciousness and its manifestation. As life forms on exoplanets are probably different from the ones we know today on earth, exobiologists have to mobilize everything we know about life on earth, even in its most extreme forms, and even extending these boundaries by imagination, before looking outside for something that might qualify as a sign of life.

A first step would be the delineation of different levels of theoretical accounts. For instance, a theory can be considered as a general explanatory scheme of the emergence of a phenomenon. The global neural network theory of consciousness (GNWT) or the integrated information theory (IIT) are theories in that they provide a framework that explains the emergence of consciousness in terms of biological function based on properties of the nervous system. Theories can be grouped into families, or paradigms, as they might present similarities or shared patterns of explanation. In that case, the "embodied" approach is such a paradigm, or, as the SEP puts it, the "embodied" approach is "a wide-ranging research program" (Shapiro and

Spaulding, 2021): a family of theories that considers the body to matter, specifically in the emergence of consciousness − contrary to neurocentric theories (Gaillard 2021). Embodiment is thus not a single theory but a basic hypothesis with many variations depending on the field of inquiry. If theories are rather general and stable, models can be considered fine-grained descriptions of events happening at the psychological and biological levels that build on a respective theory. Models can thus be improved and corrected without requiring changes in regard to the theory they presuppose. For instance, a model will propose brain localizations and will be testable in experiments with human participants reporting their conscious states while their brain activity is recorded. If the model does not match the observed outcome of the study, it can be adapted without changing entirely one's "theory of consciousness". At another even lower level, one can find concepts, which are the building blocks of theories. Concepts are used to delineate different aspects of reality that the theory is supposed to describe. Typical concepts in the field of consciousness studies are: conscious state, awareness, sentience, experience, and so on. The concept of consciousness in itself is notoriously vague and might correspond to different understandings of the notion depending on the theoretical context. Different theories might use different concepts (e.g., not all theories accept the access/phenomenal consciousness distinction even if widely used) and some concepts are developed in the framework of a particular theory ("preconscious" in GNWT).

If we want to develop an assessment tool for unfamiliar entities, a general theory will not suffice. At some point, we will have to turn to a specific model that makes predictions on a given system and provides clarification on the concepts we are mobilizing. Zilio and Lavazza are aware that we cannot generally apply everything we know about human consciousness to make assumptions about conscious states in HCOs. This is for instance the case when they criticize the current discourse for drawing inferences across species based on analogies in brain activity. Reasoning that has far-reaching implications for the potential harm we conceive admissible when performing research on neurodevelopment or brain disorders using HCOs (Diner forthcoming).

Yet, when it comes to the development of an assessment tool for these novel systems the authors seem to shy away from the consequences of their own thought. This particularly pertains to the tool favored by Zilio and Lavazza, the Perturbation Complexity Index (PCI) which is a model that builds on the IIT. While the authors themselves note that an internal validation strategy is required for applying the tool to HCOs, in the final step of their argument they again draw from knowledge acquired in humans. At this point, Zilio and Lavazza infer the course of direction of PCI values legit to assume rising levels of consciousness from correlations observed in patients. Being more imaginative about the ways consciousness forms in a dish, however, questions the legitimacy of such a line of thought because other mechanisms than in humans might guide the emergence of consciousness. Turning to a specific model that makes predictions on a given system might thus provide further grounds for predictions on the kinds of conscious states present in HCOs.

Even though adapting the model is only one of the challenges that IIT faces right now, delineating the level of theoretical commitment (from paradigm to model) might be particularly helpful to the conceptualization of assessment tools for HCOs. Because such changes might not affect "theories" or basic assumptions at a more general level.

Without any doubt, the manifold ways in which HCOs can form pose a challenge for the detection of consciousness using these models and the development of an assessment tool for these systems. However, if one starts to look at the wide range of ways these systems form from a different angle, HCOs might actually also spur research on consciousness since they allow

exploiting a wide range of conscious states for research, i.e., variability is also productive (Gaillard and Botbol-Baum 2022). Nevertheless, the insights these systems grant might be difficult to be subsumed under an overarching theory of consciousness because lab researchers might need to adapt their models to whatever they encounter as unfamiliar entities.

References:

Cova, F, M. Gaillard, and F. Kammerer. Is the phenomenological overflow argument really supported by subjective reports? 2021. *Mind & Language*. 36:422–50. doi.org/10.1111/mila.12291

Diner, S. under revision. "Similarity-based views in anticipatory ethics: from consciousness to pain in precautionary discourse on cerebral organoids".

Gaillard, M. 2021. Neuroessentialism, Our Technological Future, and DBS Bubbles."*Neuroethics* 14 (S1):39–45. doi.org/10.1007/s12152-019-09407-6

Gaillard, M, and M. Botbol-Baum. 2022. Pursuit of Perfection? On Brain Organoids as Models. *AJOB Neuroscience* 13 (2):79–80. doi.org/10.1080/21507740.2022.2048735

Kuhn, T. 1996. *The Structure of Scientific Revolutions*. 3rd ed. Chicago, IL: University of Chicago Press.

Shapiro, L., and S. Spaulding. 2021. Embodied Cognition. *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2021/entries/embodied-cognition/

Zilio, F., and A. Lavazza. forthcoming, Consciousness in a Bioreactor? Science and Ethics of Potentially Conscious Human Cerebral Organoids *AJOB Neuroscience*.