Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Theories of consciousness and a life worth living

Liad Mudrik[1,2], Myrto Mylopoulos[3], Niccolo Negro[1,4] and Aaron Schurger[5,6,7,8]

What is it that makes a life valuable? A popular view is that life's moral worth depends in some way on its relationship to consciousness or subjective experience. But a practical application of this view requires the ability to test for consciousness, which is currently lacking. Here, we examine how theories of consciousness (ToCs) can help do so, focusing especially on difficult cases where the answer is not clear (e.g. fetuses, nonhuman animals, unresponsive brain-injured patients, and advanced artificial systems). We consider five major ToCs and what predictions they offer: Integrated information theory, Higher-Order Thought Theory, Recurrent Processing Theory, Global Neuronal Workspace Theory, and Attention Schema Theory. We highlight the important distinction between the capacity and potential for consciousness and use it to explore the limitations in our ability to draw firm conclusions regarding an entity's consciousness on the basis of each theory.

Addresses
[1] School of Psychological Sciences, Tel Aviv University, Israel
[2] Sagol School of Neuroscience, Tel Aviv University, Israel
[3] Department of Philosophy and Department of Cognitive Science, Carleton University, Canada
[4] Monash Centre for Consciousness and Contemplative Studies, Department of Philosophy, Monash University, Australia
[5] Department of Psychology, Crean College of Health and Behavioral Sciences, Chapman University, USA
[6] Institute for Interdisciplinary Brain and Behavioral Sciences, Chapman University, USA
[7] INSERM U992, Cognitive Neuroimaging Unit, NeuroSpin Center, France
[8] Commissariat à l'Energie Atomique, Direction des Sciences du Vivant, I2BM, NeuroSpin Center, France

Corresponding author: Mudrik, Liad (mudrikli@tauex.tau.ac.il)

Current Opinion in Behavioral Sciences 2023, **53**:101299

This review comes from a themed issue on **Consciousness on the Borders of Life and Death**

Edited by **Gerry Leisman, Amedeo D'Angiulli, Calixto Machado, Charlotte Martial** and **Olivia Gosseries**

For complete overview of the section, please refer to the article collection, "Consciousness on the Borders of Life and Death (2023)"

Available online 9 September 2023

Received: 27 May 2023; Revised: 15 July 2023;
Accepted: 2 August 2023

https://doi.org/10.1016/j.cobeha.2023.101299

Introduction

It is an obvious truth that humans value life and lament its loss. Less evident, however, is what it is about life that makes it valuable, or justifies our moral concern for its continued preservation. Here, we consider the view that the moral worth of a system[1] depends in some way on its relationship to consciousness or subjective experience. Accordingly, we argue that modern theories of consciousness (ToCs) should contribute to this debate by offering guidance regarding the question of whether or not consciousness is present, or could be present in the future, in a given organism or entity. This could help navigate bioethical concerns regarding various entities whose relationship with consciousness is not fully understood, including fetuses, nonhuman animals, individuals diagnosed with unresponsive wakefulness syndrome, and even advanced artificial systems.

Before we do so, three notes are in order. First, we acknowledge that a range of positions is available regarding the relationship between moral worth and consciousness (see Refs. [1,2] for discussion of each). First, one might hold that consciousness is necessary, but not sufficient, for a life with moral worth. On such a view, only conscious beings have such worth, but not all of them, as possession of other properties is also required. A second option is to take consciousness to be both necessary and sufficient for moral worth. On this view, all conscious beings have moral worth and the moral worth of an entity's life derives solely from its conscious experiences. A third option is to view consciousness as sufficient but not necessary for moral worth, holding that other properties might also be sufficient for moral worth (e.g. intelligence or cognitive ability). On this view, all conscious entities have moral worth, but some nonconscious entities may also possess such worth, due to other candidate properties. Here, we will remain neutral on which is the correct stance to take. Since all three approaches assign a critical role for consciousness in determining moral worth, we will focus on how ToCs might help us determine *which* entities have moral worth on such approaches.

A second note relates to a subtle but crucial distinction between the *capacity* and the *potential* for consciousness. The *capacity* for consciousness corresponds to currently

[1] We use the term 'system' to refer to both biological organisms and artificial systems.

meeting all the necessary conditions to support conscious states, even if, at the present moment, the being happens not to be in a conscious state (e.g. someone in a dreamless sleep). Conversely, the *potential* for consciousness corresponds to the ability to meet both the necessary and sufficient conditions for consciousness in the future, when these are not yet satisfied (e.g. an embryo). It is for ethicists to determine whether the critical criterion for moral worth lies with the potential or the capacity for consciousness; here, we only explain how ToCs can translate these terms into empirical tests for determining if a system has the former, the latter, or neither.

Third, consciousness is often conceptualized as human-consciousness, and ToCs have been developed either based on data mostly from human participants (e.g. [3]. For an overview of how theories are supported by empirical data, see Ref. [4]) or on contemplations about the nature of consciousness [5,6], which tend to be anthropocentric. But the focus here is on nonstandard cases that are not necessarily similar to human-consciousness. A good ToC should account for multiple-realizability [7], the possibility for consciousness to be realized in different, nonhuman systems. Thus, we must abstract away from the human domain upon which these theories are usually built. The problem is how to justify this abstraction, to validate the applicability of ToCs to nonstandard domains [8–10]. Owing to the brevity of this paper, we note this problem without tackling it, to put the suggestions below in the right context.

Theories of consciousness and consciousness in nonstandard cases

Here, we focus on five ToCs, and show that, although they are not sufficiently developed so as to provide a concrete test for consciousness, they still might have some insights to offer. We further assert that any comprehensive theory of consciousness must be such that it can be used to develop reliable means to detect consciousness, as this might be its most critical contribution to society at large.

One theory that tries to address this issue is Integrated Information Theory (IIT [5,11]). IIT claims to be able to determine the level of consciousness of a given system, using a metric called 'phi' derived mathematically from the system's connectivity and activation patterns. Notably however, the theory does not offer a threshold value for phi above which a system is conscious (although the theory does suggest a lower bound on the threshold, such that a complex with markedly lower phi than that of a human brain during deep dreamless sleep, would have a negligible quantity of consciousness). Without such a threshold, it is hard to determine what level of phi (i.e. consciousness) should suffice for moral worth. In addition, it is currently impossible, due to insufficient computational power, to actually compute phi for real-life cases.

Another class of theories, called Higher-Order Theories [6,12], holds that being in a mental state M (e.g. a state of pain) that is conscious is equivalent to having a higher-order representation (HOR) of oneself as being in state M. Accordingly, a system that cannot form such HORs has no potential for consciousness, while a system that is capable, even if not doing so at present, has the capacity for consciousness. The practical challenge, however, is the lack of an agreed-upon *marker* of HORs in nonstandard cases, where subjective report is not available.

Recurrent Processing theory (RPT [13,14]) seems to set a much lower bar for both the potential and the capacity for consciousness. RPT is a first-order ToC, not requiring anything beyond the first-level representation. As such, it claims that consciousness occurs when there are recurrent connections, and hence recurrent (as opposed to feedforward) processing (RP), within the relevant area. Thus, if the system is built in a solely feedforward manner, it lacks the capacity for consciousness. If it has feedback connections, it has the capacity and would be conscious once these are activated. Yet here, there is no clear-cut criterion for such recurrency; is a single recurrent action potential enough? And how would that be quantified and measured?

The Global Neuronal Workspace Theory (GNWT [3,15]) proposes that consciousness is associated with the activation of a global network of specialized neurons that broadcast information throughout the brain. Thus, a system devoid of such a workspace would not be conscious. According to GNWT, the global broadcast should be signaled by an all-or-none ignition response. While this might serve as a marker for consciousness, it might not be exhaustive (i.e. one could envision a system exhibiting an all-or-none response without having a global workspace). And without this marker, it is difficult to determine if a global broadcast exists or not.

Finally, Attention Schema Theory (AST [16,17]) holds that consciousness is a perceptual attribution or inference that results from having an attention schema. The attention schema is an internal model of our own selective attention, which helps us control it. If so, for a system to be conscious, it should have, at a minimum, selective attention plus a second-order model (schema) of that attention. Yet, this too turns out to be a difficult property to test for. At present, we lack the means to know if a given system indeed has an attention schema, let alone whether or when the attention schema is engaged.

Table 1

A high-level summary of ToCs and their criteria for Lack of consciousness (purple), the Potential for consciousness (blue), the Capacity of consciousness (green), and the Existence of consciousness (yellow). In the lowest row, the main challenge each theory faces for constructing a theory-based test (white). The key requirement of each of the theories is highlighted using italics.

	Attention Schema Theory	Global Neuronal Workspace	Higher Order Thought	Integrated Information Theory	Recurrent Processing
Lack of Consciousness	No *schema of attention*	No *global neuronal workspace* (only encapsulated processing)	No *higher order representations* (only first order)	No *integrated information*	No *recurrent processing* (only feedforward)
Potential for Consciousness	Can the system develop a *schema of attention*?	Can the system develop an architecture allowing for a *global neuronal workspace*?	Can the system develop an architecture allowing for a *higher order representation*?	Can the system develop an architecture yielding *integrated information*?	Can the system develop an architecture allowing for *recurrent processing*?
Capacity for Consciousness	Is the system capable of having a *schema of attention*?	Does the system have a *global neuronal workspace*?	Is the system capable of having a *higher order representation*?	Is the system capable of generating *integrated information*?	Is the system capable of *recurrent processing* in sensory areas?
Existence of Consciousness	Is the *schema of attention* active?	Is the *global neuronal workspace* active?	Does the system have *higher order representations*?	What is the degree of *integrated information* the system has?	Is there *recurrent processing* in sensory areas?
Major challenge to a feasible test	Lack of **means to determine** if a system has an attention schema or if it is active	**Non-exhaustive marker**, undefined **degree of complexity** of the workspace	**No agreed upon marker** for higher order representations	**No threshold** for the level of consciousness	**No criterion** for the required extent of recurrent processing

As the discussion above illustrates, for most theories, the *capacity* for consciousness depends on the way the system is built (i.e. its structure), and its ability to develop the required structure constitutes its *potential* for consciousness. The *presence* of consciousness, instead, is determined by the neural activity within that structure. The main problem here is that the derived tests either beg for a more precise definition, or are only crude proxies of the actual suggested mechanism, with respect both to the structure and the activity patterns (e.g. since we cannot record all the individual components of the GNW, we use the macroscopic nonlinear response as a proxy).

Test cases: how do theories differ?

Although theorists have not reached a consensus as to how to infer the potential for, capacity for, or presence of consciousness in any given system, we might nevertheless gain some insight by looking at where the theories stand, relative to one another, on a given question. For example, on the question "Where does the capacity for consciousness manifest in non-human life (e.g. fish)?" IIT stands out among the abovementioned theories, as being the most liberal in assigning consciousness. According to IIT, there is a continuum in the level of consciousness from very simple central nervous systems, such as in fish, to very complex ones. AST, on the other hand, would probably not attribute consciousness to fish, since they probably do not have the equivalent of an attention schema. GNWT might land somewhere in-between, attributing the capacity for consciousness, depending on the presence of a global workspace architecture. Yet, it is unclear how complex the workspace should be to allow for consciousness [10,18,19], and it is unclear how GNWT would consider a creature that showed signs of being highly intelligent and possibly conscious, without any kind of a global workspace architecture. For example, Octopuses are highly intelligent and thought by many to be conscious [20], but without a parietal or prefrontal cortex (then again, they might have developed a different form of global workspace architecture).

What about human embryos, fetuses, and disorders of consciousness patients? The moral dilemmas about fetuses and embryos are well-known and contrast the two things that humans arguably value the most: freedom and life. The question clearly is not black and white, but consciousness science might be able to shed some light on the matter — assuming consciousness is one of the relevant requirements for protection from being killed [21]. As far as the *potential* for consciousness goes, that is

technically determined at conception. Once conceived, an embryo has the *potential* for consciousness in the future, if allowed to develop normally. An embryo does not, however, have the *capacity* for consciousness, according to any extant theory that we know of, until the budding brain shows some signs of structured activity. Thus, 9–12 [22] weeks seem to be the very earliest threshold for having the capacity for consciousness. Notably, the threshold is probably actually closer to 24 weeks, when the thalamocortical system comes into operation [23], which is a necessary condition for many current theories. As research on brain activity in fetuses, and even infants, is still emerging, also in the context of consciousness [24,25], further studies are needed to determine when the 'right kind' of brain activity (depending on which theory you subscribe to) is present in a fetal brain. As for the *presence* of consciousness, it seems safe to assert that an entity's moral worth does not depend on this criterion. If it did, then it would be acceptable to kill someone in a deep dreamless sleep or under anaesthesia. So the relevant factors for ToCs to address in the context of our main question seem to be the criteria for the *potential* and *capacity* for consciousness, depending on which one is deemed to impart moral worth.

Conclusions

To summarize, none of the theories currently provides a ready-to-use, feasible test for consciousness that can be applied in any arbitrary case. Some have generated successful tests for disorders of consciousness (i.e. IIT [26] and GNW [27]), but it remains unclear whether any such test specifically indexes the presence of consciousness or picks up co-occurring typical features of a healthy working brain, other than consciousness. We argue that greater emphasis should be put on developing such tests that would be (a) more selective (i.e. focused on consciousness rather than related phenomena); and (b) more broadly applicable (i.e. not limited to the healthy adult human brain). This is especially important given the critical ethical implications of these tests, especially now, as Artificial Intelligence systems are rapidly developing and even claimed by some to be sentient. Developing a test for disorders of consciousness is a formidable challenge. Developing a generalized test for consciousness is an even greater challenge. ToCs might one day hold the key to meeting these challenges, with enormous potential ramifications (for a recent attempt to use ToCs for deriving markers of consciousness in AI, see [28]).

CRediT authorship contribution statement

Liad Mudrik: Conceptualization, Writing – original draft, Writing – review & editing, Visualization. **Myrto Mylopoulos:** Conceptualization, Writing – original draft, Writing – review & editing. **Niccolo Negro:** Conceptualization, Writing – original draft, Writing – review & editing. **Aaron Schurger:** Conceptualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

None.

Acknowledgements

This publication was made possible through the support of a joint grant from the John Templeton Foundation and the Fetzer Institute (Consciousness and Free Will: A Joint Neuroscientific-Philosophical Investigation (John Templeton Foundation #61283; Fetzer Institute, Fetzer Memorial Trust #4189). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation or the Fetzer Institute. L.M. is a CIFAR Tanenbaum Fellow in the Brain, Mind, and Consciousness program.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest

1. Lee AY: **Is consciousness intrinsically valuable?** *Philos Stud* 2019, **176**:655-671.

2. • Shepherd J: **Non-human moral status: problems with phenomenal consciousness**. *AJOB Neurosci* 2022, **14**:1-10.
This paper calls into question the intuition that consciousness is necessary for some degree of moral status.

3. Dehaene S, Naccache L: **Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework**. *Cognition* 2001, **79**:1-37.

4. • Yaron I, Melloni L, Pitts M, Mudrik L: **The ConTraSt database for analysing and comparing empirical studies of consciousness theories**. *Nat Hum Behav* 2022, **6**:593-604.
This paper provides an overview of the way four leading ToC have been tested empirically.

5. Tononi G: **An information integration theory of consciousness**. *BMC Neurosci* 2004, **5**:42.

6. Lau H, Rosenthal D: **Empirical support for higher-order theories of conscious awareness**. *Trends Cogn Sci* 2011, **15**:365-373.

7. Doerig A, Schurger A, Herzog MH: **Hard criteria for empirical theories of consciousness**. *Cogn Neurosci* 2020, **12**:1-22.

8. Block N: **The harder problem of consciousness**. *J Philos* 2002, **99**:391-425.

9. Bayne T, Shea N: **Consciousness, concepts and natural kinds**. *Philos Top* 2020, **48**:65-83.

10. • Shevlin H: **Non-human consciousness and the specificity problem: a modest theoretical proposal**. *Mind Lang* 2021, **36**:297-314.
This paper introduces the 'specificity problem', the problem of applying ToC to systems with a cognitive architecture that differs from that of humans.

11. Tononi G, Boly M, Massimini M, Koch C: **Integrated information theory: from consciousness to its physical substrate**. *Nat Rev Neurosci* 2016, **17**:450-461.

12. Brown R, Lau H, LeDoux JE: **Understanding the higher-order approach to consciousness**. *Trends Cogn Sci* 2019, **23**:754-768.

13. Super H, Spekreijse H, Lamme VA: **Two distinct modes of sensory processing observed in monkey primary visual cortex (V1)**. *Nat Neurosci* 2001, **4**:304-310.

14. Lamme VAF: **Visual functions generating conscious seeing**. *Front Psychol* 2020, **11**:83.

15. Mashour GA, Roelfsema PR, Changeux JP, Dehaene S: **Conscious processing and the Global Neuronal Workspace hypothesis**. *Neuron* 2020, **105**:776-798.

16. Wilterson AI, Kemper CM, Kim N, Webb TW, Reblando AMW, Graziano MSA: **Attention control and the attention schema theory of consciousness**. *Prog Neurobiol* 2020, **195**:101844.

17. Graziano MSA, Webb TW: **The attention schema theory: a mechanistic account of subjective awareness**. *Front Psychol* 2015, **6**:500.

18. Birch J: **Global workspace theory and animal consciousness**. *Philos Top* 2020, **48**:21-38.

19. • Birch J: **The search for invertebrate consciousness**. *Noûs* 2022, **56**:133-153.
This paper defends the idea that the best way to search for consciousness in systems considerably different than humans is to adopt a 'theory-light' approach.

20. Godfrey-Smith P: **Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness**. Farrar, Straus and Giroux; 2016.

21. Warren MA: **The moral significance of birth**. *Hypatia* 1989, **4**:46-65.

22. Berk L: **Child Development**. Pearson Higher Education AU; 2015.

23. Lagercrantz H, Changeux JP: **Basic consciousness of the newborn**. Seminars in Perinatology. Elsevier; 2010.

24. • Moser J, Schleger F, Weiss M, Sippel K, Semeia L, Preissl H: **Magnetoencephalographic signatures of conscious processing before birth**. *Dev Cogn Neurosci* 2021, **49**:100964.
In this paper, a theory-based (GNW) test for consciousness is applied on fetuses.

25. • Passos-Ferreira C: **Are infants conscious?** *Philos Perspect* (Forthcoming).
This paper makes a case for the view that infants are conscious on the basis of neurophysiological and behavioral markers of pain, as well as theoretical considerations.

26. Casarotto S, Comanducci A, Rosanova M, Sarasso S, Fecchio M, Napolitani M, *et al.*: **Stratification of unresponsive patients by an independently validated index of brain complexity**. *Ann Neurol* 2016, **80**:718-729.

27. Bekinschtein TA, Dehaene S, Rohaut B, Tadel F, Cohen L, Naccache L: **Neural signature of the conscious processing of auditory regularities**. *Proc Natl Acad Sci* 2009, **106**:1672-1677.

28. Butlin, P., Long, R., Elmoznino, E.c., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S.M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M.A.K., Schwitzgebel, E., Simon, J., and VanRullen, R. (2023) Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv preprint arXiv:2308.08708*.