General intelligence: an ecumenical heuristic for artificial consciousness research?

Henry Shevlin[a]

ABSTRACT: The science of consciousness has made great strides in recent decades. However, the proliferation of competing theories makes it difficult to reach consensus about artificial consciousness. While for purely scientific purposes we might wish to adopt a 'wait and see' attitude, we may soon face practical and ethical questions about whether, for example, an artificial agents is capable of suffering. Moreover, many of the methods used for assessing consciousness in humans and even non-human animals are not straightforwardly applicable to artificial systems. With these challenges in mind, I propose that we look for ecumenical heuristics for artificial consciousness to enable us to make tentative assessments of the likelihood of consciousness arising in different artificial systems. I argue that such heuristics should have three main features: they should be intuitively plausible, theoretically neutral, and scientifically tractable. I claim that the concept of general intelligence – understood as a capacity for robust, flexible, and integrated cognition and behavior – satisfies these criteria and may thus provide the basis for such a heuristic, allowing us to make initial cautious estimations of which artificial systems are most likely to be conscious.

1. Introduction

Over the last three decades, we have made great strides in developing innovative and fruitful frameworks for theorizing consciousness, with approaches such as the global neuronal workspace [Dehaene & Naccache, 2001], higher-order theories [Lau, 2007; Rosenthal, 2006], and Integrated Information Theory or IIT [Tononi & Koch, 2015] all contributing important insights. One of the next key goals for the science of consciousness will to tease apart the specific predictions and commitments of these approaches and synthesize and refine frameworks. However, this will likely be a long-term project, and while from a dispassionate scientific perspective it may be prudent to adopt a 'wait and see' attitude, there are domains in which we face more pressing need for answers.

One such domain is the field of artificial intelligence (AI). In light of the breakneck pace of AI development, it is no longer fanciful to imagine that we may soon be capable of building systems with a capacity for conscious experience. As we approach this goal, we will be forced for the first time to take seriously questions concerning artificial consciousness, including, for example, the possibility that we may have ethical or legal obligations towards the machines we construct [Agar, 2019]. Developing principled tools for assessing consciousness in artificial systems, however, is a formidable task, and one that arguably involves special methodological challenges not arising for comparable projects concerning humans and animals.

With this task in mind, I argue that the field of artificial consciousness research would benefit considerably from developing what I call *ecumenical heuristics*, tools that can help us to make tentative judgments concerning the likelihood of consciousness in different artificial systems. The paper proceeds as follows. I begin in Section 2 by

[a] Leverhulme Centre for the Future of Intelligence, University of Cambridge.

motivating the need for such a heuristic, noting some key questions about AI consciousness and the challenges we face. In Section 3, I spell out the notion of an ecumenical heuristic in more detail, and present three desiderata such a heuristic should fulfil. Finally, in Section 4, I sketch what I take to be a promising heuristic of this kind, namely one that uses general intelligence as a proxy measure of consciousness.

2. Why do we need consciousness heuristics?

While the possibility of artificial consciousness has been mooted since the early days of AI [Turing, 1950], it is not a prospect that preoccupies most contemporary researchers in science and industry. However, there are practical reasons why at least some AI researchers may wish to engage with this issue, of which I will briefly note three.

First, to the extent that consciousness is an important (if still poorly understood) aspect of human mental function, our efforts to construct human-level AI might stand to benefit if we are able to implement some form of artificial conscious processing [Kanai, 2017]. Such an endeavor might also lead us to new insights about human consciousness, perhaps with potential for clinical applications. Second, since a capacity for conscious feeling is often taken to be closely linked to human emotional responses like love, care, and affection, we might reasonably wish to ensure that future robot companions and caretakers are capable of conscious affective response [Danaher, 2019]. Finally, and most troublingly, without tools for assessing the likelihood of AI consciousness, it is possible that we may inadvertently build machines that are capable of negatively-valenced experiences, and even suffering [Agar, 2019; Tomasik, 2014]. If we are to safeguard against such ethical risks, we need some relatively uncontentious ways of predicting or measuring consciousness in AIs, even if these are initially tentative and fallible.

But do we need a heuristic for these purposes? An optimist might think that even aside from the competing theoretical approaches, we already have adequate general purpose methods for assessing consciousness. Note, for example, that clinicians and researchers with quite different theoretical commitments have made progress in assessing consciousness in patients in persistent vegetative states using relatively uncontroversial neural and behavioral signatures of consciousness [Owen et al., 2006; Sitt et al., 2014]. Something similar might even be said of animal consciousness; while major theoretical and methodological disagreements persist, cognitive ethologists and behavioral scientists have in many cases sought to sidestep these issues by focusing on behaviors that widely agreed to provide tentative evidence for the presence of consciousness, such as trace-conditioning [Birch, 2019] and motivational trade-off [Sneddon et al., 2014].

The problem we face, however, is that such methods are not readily applicable to artificial systems. The use of neural signatures as evidence for consciousness in humans, for examples, relies on assumed commonalities in neural structure and function, and even in the case of animals we can make tentative inferences about consciousness on the basis of neuroanatomical homologies. Needless to say, these strategies will not be available for the assessment of consciousness in non-biological systems.

The same is arguably true of behavioral signatures of consciousness. Consider memory-trace conditioning, for example. This form of learning involves imposing a temporal offset between a conditioned stimulus (such as a tone) and an unconditioned stimulus (such as a puff of air into the eye). Only in conditions where subjects are consciously aware of the former are they able to develop an appropriate conditioned response to it. Because trace conditioning seems to require consciousness in humans, it has been suggested as a potential 'behavioral signature' of consciousness in both vegetative state patients and animals [Bekinschtein et al., 2011].

Such behavioral signatures may not be readily applicable to artificial consciousness, however. To illustrate the problem, note that there is no obvious *conceptual* connection between memory-trace conditioning and consciousness; the discovery that memory-trace conditioning requires consciousness was an empirical one, reflecting *prima facie* contingent facts about our cognitive architecture. When we make inferences about consciousness in non-human animals on the basis of trace-conditioning, then, we are relying on the assumption that their cognitive architecture is relevantly similar to ours. But as we move from humans to systems more phylogenetically remote with quite different cognitive architectures, this assumption becomes shakier insofar as we have less reason to expect the contingent connections between memory-trace conditioning and consciousness to apply. This is especially true in the case of AIs. It seems entirely possible, for example, that one could build a 'gerrymandered' AI capable of memory-trace conditioning but possessing few if any other capacities associated with consciousness.[b] Equally, we can imagine a superintelligent AI that was a strong 'consciousness candidate' yet which had particular quirks in its cognitive architecture such that it was incapable of trace conditioning.

Even if some behavior signatures are limited to creatures relevantly like us, one might still claim that there were *universal* behavioral signatures of consciousness that we could use to construct failsafe tests of consciousness, as famously proposed by the Turing Test or the recent "ACT test" [Turner & Schneider, forthcoming]. In addition to being controversial [Bishop, 2018; Searle, 1980], however, such tests typically rely on high-level abilities like human-level verbal behavior or sophisticated metacognition to establish the presence of consciousness, and so carry the risk of false negatives. Insofar as we are willing to attribute consciousness to some non-human animals, we must also take seriously the possibility that the first conscious AIs will not exhibit such abilities.

I would suggest, then, that our existing battery of theoretically-neutral uncontroversial methods for assessing consciousness in humans and animals cannot be readily applied to questions about artificial consciousness. Consequently, in our search for answers, we must either settle upon a theory of consciousness (an unlikely near-term prospect), or else attempt to find some *ecumenical heuristic* for AI consciousness. It is this possibility that I will now consider.

3. Ecumenical heuristics for AI research

[b] Cf. Tomasik [2014]: "When we develop a simple metric for measuring something… we can game the system by constructing degenerate examples of systems exhibiting that property that we don't intuitively think of as sentient."

I will use the term ecumenical heuristic to refer to a relatively *theoretically-neutral* tool we can use for making tentative assessments of the likelihood of consciousness in an artificial system. The hope is that such a heuristic will allow researchers with very different theoretical commitments to make convergent judgments about consciousness in a given case.

Such a heuristic should have three key features. First, the heuristic should be empirically tractable, assessing consciousness on the basis of readily measurable cognitive and behavioral indicators. The whole purpose of a heuristic, after all, is to enable us to sidestep vexed theoretical debates in order to respond to shorter-term challenges.

Second, it should be as theoretically neutral (hence ecumenical) as possible. In practice, given the wide variety of positions concerning the nature of consciousness, total theoretical neutrality is unlikely; indeed, any heuristic that was *that* ecumenical would likely be applicable only to a very small subset of cases, namely those concerning which there is already universal agreement. What is instead key is that in the initial formulation of the heuristic, we should endeavor to not violate the core theoretical commitments of the main approaches.

Finally, I would suggest that the heuristic should be *intuitive*, comporting with our pretheoretical judgments about consciousness wherever possible. This last feature requires further justification. After all, our reflective intuitions about scientific questions frequently prove to be inaccurate, and many theoretical approaches to consciousness (such as panpsychism and illusionism) are avowedly counterintuitive. Note, however, that our intuitive commitments constitute a starting point for most theories of consciousness, whether in the form of explicit axioms [Bayne, 2018; Tononi & Koch, 2015], methodological commitments concerning, for example, the relationship between consciousness and cognitive access [Cohen & Dennett, 2011], or folk psychological platitudes like the claim that conscious states are states we are suitably aware of [Rosenthal, 2009]. To the extent that a heuristic was explicitly at odds with these commitments, it would hardly be ecumenical. Moreover, there are many consciousness researchers (including the present author) who do not come down strongly on the side of one theory or another. In order to appeal to such people, the heuristic must have some broad plausibility independent of theoretical claims.[c]

Before proceeding, it is also worth noting a key distinction between what we may term *ordinal* and *cardinal* heuristics. An ordinal heuristic would allow to determine which of two artificial systems was more likely to be conscious. More ambitiously, a *cardinal* heuristic would attempt to assign determinate (albeit still tentative) probabilities of consciousness to different systems.

For some purposes, the latter sort of heuristic might be invaluable; trading off the interests of conscious humans against harms to a potentially conscious machine, for example, would require us to assign a probability to consciousness in the latter case.

[c] Note also that some approaches to explaining the mind (such as Lewis 1972) give a still more central role to intuitions, taking the job of a theory of consciousness to be a matter of regimenting and synthesising our competing platitudes about consciousness.

However, a cardinal heuristic may not be achievable in the short-term, as theoretical disagreements concerning fundamental issues will likely lead different researchers to make assign radically divergent absolute probabilities of consciousness in various systems. However, there are contexts in which even a merely ordinal heuristic will be useful. In deciding which of two different competing AI architectures is more likely to be conscious, for example, an ordinal heuristic could constructively inform our judgments. Moreover, an ordinal heuristic may be relatively achievable: even among theorists who assign very different absolute probabilities to consciousness in a given system, we might hope to find broad (if imperfect) agreement about whether one system is more or less likely to be conscious than another.

4. General intelligence as consciousness heuristic

So far, I have argued that artificial consciousness research would benefit considerably if researchers could find some relatively uncontroversial heuristic for assessing consciousness in artificial systems, where this heuristic is scientifically tractable, theoretically neutral, and intuitively plausible. There are many forms that such a heuristic might take. To give one crude example, we might adopt a ‘popularity contest’ model, assigning weights to different theories based on the attitudes of current senior researchers and averaging out their assessments of consciousness to make a verdict in a given case. Needless to say, however, such a project would be hard to conduct impartially, and moreover, would be empirically tractable only to the extent that the various theoretical positions adopted could be straightforwardly applied in a given case.

While there are doubtless many other interesting heuristic strategies available, in the remainder of this paper I wish to explore one specific heuristic that links consciousness to *general intelligence* (GI). While I hope to convince the reader that the heuristic described is plausible and has the potential (when fully developed) to satisfy the desiderata above, even if I am unsuccessful in this regard, the following discussion may at least provide a model for how we might go about developing and evaluating other candidate heuristics.

At a first pass, we may define GI as the capacity of an organism or artificial system to use informational resources to achieve varied goals in different conditions.[d] GI is to be distinguished from *specialized* intelligence, that is, the capacity to excel at one particular task under constrained conditions. It is also distinct from Artificial General Intelligence or AGI, which is commonly used to refer to a possible future artificial system capable of matching or exceeding human-level performance in most or all tasks. Whereas AGI is a specific destination in the development of GI, then, we might talk of systems both biological and artificial that fall well below this threshold as exhibiting GI to varying degrees.

As described thus far, GI is a somewhat vague notion, with one foot in folk psychology and another in cognitive science. However, we might hope that GI can be

[d] Cf. Legg & Hutter [2007]: “Intelligence measures an agent’s ability to achieve goals in a wide range of environments.”

operationalized with only minimal controversy, and certainly with less controversy than would attend to similar attempts to operationalize consciousness.[e] While it is not my primary purpose in the present paper to offer a full account of this kind, I would suggest that GI can be helpfully understood as involving three main capabilities.

The first is *robustness* of behavior. This can be defined as the capacity of a system to carry out goal-directed behaviors in the face of challenges. Although it is a concept drawn from ordinary language, the term has been specifically operationalized in the 1990 IEEE Standard Glossary of Software Engineering Terminology as "[t]he degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions." Finding one's way to a destination in the presence of heavy fog, eating dinner on a bumpy train, and identifying the face of a friend in blurry photographs are examples of tasks that require some degree of robustness.

The second is *flexibility* of behavior. I will take this to refer specifically to the ability to use cognitive and behavior strategies developed for one task or context and apply them in another. While this definition may sound clear enough, in practice it imposes many challenges, not least how to individuate tasks and contexts. Consider, for example, the observation that a New Caledonian Crow named Betty spontaneously bent a piece of wire in order to retrieve a food reward [Weir et al., 2002]. However, it was later observed that New Caledonian Crows bend twigs in their natural environment [Rutz et al., 2016]. It is hard to say, then, whether Betty's wire bending is a genuinely new context or merely an instance of a pre-existing behavior. Likewise, note that many quite simple organisms like nematode worms inhabit an incredible variety of environments. While this may seem prima facie to be an example of flexibility, the nematodes' ecological versatility is not much of a cognitive achievement insofar as their narrow behavioral repertoire is simply *indifferent* to many forms of environmental variation. Such challenges notwithstanding, in principle it should be possible to develop some operationalized criteria to pin down flexibility. In particular, we might ask about whether successfully performing a task in a new environment presents a system with a *cognitively demanding challenge* that it must dedicate information-processing resources to overcome.

The third proposed criterion for GI is *informational integration*. I take this to refer to the ability to constructively combine and compare information from multiple sources or from a single source both at-a-time and over time. One example of integration at-a-time is multisensory integration, as occurs when we use lip-reading to disambiguate sounds in the comprehension of speech. Integration over time, by contrast, involves the use of memory, and is evinced by tasks like detecting changes in a visual scene or calculating how much money we have left in our bank account. Integration plays a role in updating our model of the world and in fine-tuning task performance, but it can also play a critical role in action selection, as demonstrated, for example, by cases where we have to choose between competing biological needs for hunger and thirst.

[e] It is worth noting that there have been many notable attempts to define "intelligence" and "general intelligence" in the context of AI research. These are not discussed in the present paper for considerations of brevity. However, see Wang [2019] for a review.

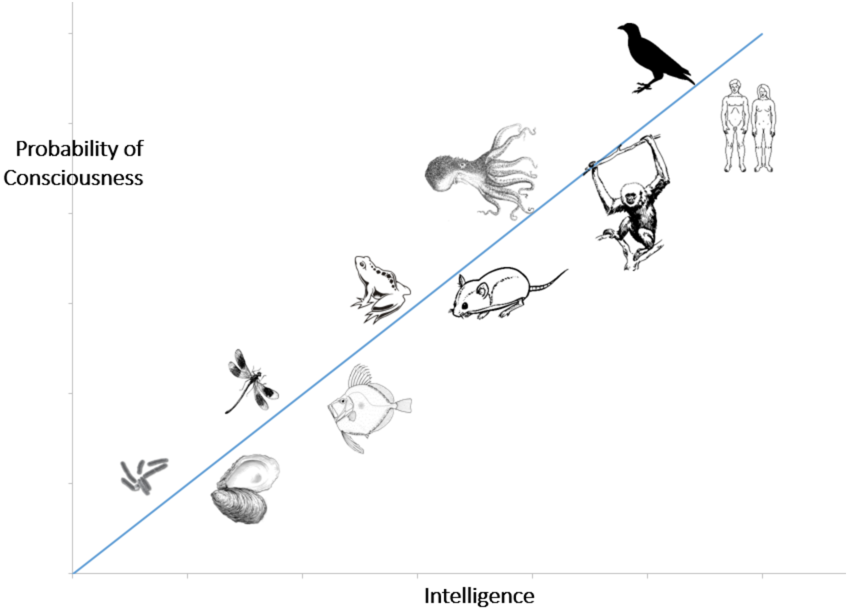

Fig. 1. A schematic illustration of pretheoretical links between consciousness and intelligence. More intelligent creatures like crows, primates, and octopuses intuitively seem like better ‘consciousness candidates’ than less intelligent ones like frogs,

The above discussion is not intended to serve as a dispositive account of the nature of GI, but to provide an exploratory outline of some of its key features. By building on this initial outline, I propose that use the concept of GI as an *ordinal consciousness heuristic* in assessing consciousness in different contemporary AIs. With a clear principles for measuring GI in hand, we could assess which of two AIs exhibited it to a greater degree, and make a (tentative, fallible) assessment as to which was more likely to be conscious. Put succinctly, then, the GI heuristic claims that in considering two artificial systems S_1 and S_2, we should assign a greater probability of consciousness in the system that possesses a greater degree of general intelligence.

Why should we think that GI is a suitable heuristic for artificial consciousness? In short, because it seems well placed to fulfil the three desiderata discussed above for an ecumenical heuristic. I have already discussed how we might go about making GI into an empirically tractable notion (though I recognize that this will be a substantial project). I would next note that there is a strong intuitive connection between consciousness and GI. In everyday discussion of the possibility of animal consciousness, for example, it is common to assume that smarter animals are better ‘consciousness candidates’, and when we learn of dramatic cognitive feats in a particular non-human animal, we will normally consider it more likely that the organism in question is conscious (Fig.1).

It might be objected that this link holds between *intelligence* and consciousness rather than GI and consciousness specifically. However, the specific capacities singled out as critical for GI have particularly powerful connections to our judgments about consciousness. To illustrate, imagine that scientists observe some apparently intelligent behavior such as tool use in a species not normally considered a strong consciousness candidate (for example, an oyster). A natural response would be to positively update our beliefs about the possibility of consciousness in this creature. However, there are several possible defeaters that could lead us to reconsider this revision of judgment. One would be if the behavior turned out to be *non-robust*, displaying extreme sensitivity to interference. Another would be if the behavior was *inflexible*, capable of occurring only under a very narrow range of circumstances despite the existence of other contexts in which it would also be valuable. A third would be if the behavior showed insensitivity to other information available to the organism (in other words, a failure of integration), such that the creature could not make use of, for example, information encoded in memory to perform the action more efficiently.

But is GI really a theoretically neutral heuristic? This is a difficult question to answer decisively, and a thorough response (beyond the scope of this paper) would make need to make reference to the specific commitments of the dozens of proposed theories of consciousness. Still, I would note that while relatively few approaches offer explicitly GI-based accounts of consciousness [see Kanai, Fujisawa, Tamai, Magata, & Yasumoto, forthcoming], many theories would endorse the broader claim consciousness involves (or follows from) cognitive mechanisms at least some of which function to make an organism capable of more robust, flexible, and integrated cognitive and behavior processes.

A few examples should illustrate the point. Consider first Global Workspace Theory [Dehaene & Naccache, 2001]. This approach identifies consciousness with a process of *global neuronal activation* which serves to make a given representation available for multiple cognitive functions including report, voluntary behavior, and encoding in memory. In one sense, then, global activation serves as an *integrating* device that ensures that a given representation can be accessed not just by one subsystem or another but for all core cognitive functions of the organism. This process also allows for the stabilization of representations so as to allow for their use in sustained, deliberate, and *robust* action. As Stanislas Dehaene puts it, “[s]ubliminal information cannot enter into our strategic deliberations… [t]he mighty unconscious generates sophisticated hunches, but only a conscious mind can follow a rational strategy, step after step” [Dehaene, 2014].

A close connection between consciousness and GI also arguably follows from many higher-order theories. There are many different higher-order approaches to consciousness, but they have in common a commitment to the idea that consciousness arises via some form of actual [Rosenthal, 2006] or possible [Carruthers, 2005] metacognitive representation that serves to make an individual aware of their own mental states. Any such metacognitive process is likely to contribute to an organism’s broader capacities for robust, flexible, and integrated cognition and behavior, for example by allowing it to represent and act upon degrees of confidence in different judgments [Terrace & Son, 2009], to identify cases where its own perceptual states are erroneous [Lau & Rosenthal, 2011: 311] or by contributing to an agent’s capacity for social cognition [Rosenthal, 2008: 838].[f]

Finally note that many attention-based theories of consciousness would anticipate a close alignment between consciousness and GI. According to Prinz's AIR theory for example, visual representations become conscious when encoded in working memory via attention (a process enabled, Prinz suggests, by gamma-band neural synchrony). Without such mechanisms, many forms of robust and flexible cognition including strategic decision-making and deliberative action will be difficult or impossible. Prinz considers the example of a person filling a glass to drink, and notes "we must choose high how to fill our glass, where to place the bottle on completion, and how gingerly to lift the full glass… these decisions benefit from conscious input" [Prinz, 2012].

This is a brief snapshot of how consciousness can be related to general intelligence within the theories of consciousness debate, but it hopefully serves to illustrate the basic point, namely that quite different theories can likely to expect some overlap between conscious systems and those with a high degree of GI. Of course, in some cases the proposed mechanism of consciousness is only weakly or indirectly associated with GI, and nothing rules out, for example, the possibility of a creature with a global workspace or higher-order thoughts that was otherwise very limited in respect of GI. Nonetheless, I would suggest that even this weak connection is sufficient to satisfy the desideratum of relative theory-neutrality, insofar as our need not agree with every theory in every case, as long as it loosely aligns with most of their judgments about the relative likelihood of consciousness in different systems.

I would note in closing that two particular kinds of theory pose a trickier challenge for the general intelligence heuristic, namely biological approaches that take consciousness to be a distinctive property of living systems [for example, Thompson, 2010] and those (like Tononi's IIT) that take consciousness to be ubiquitous and present even in some simple systems. While considerations of space prevent detailed discussion of these views here, I would make two brief observations. First, even staunch biological theorists skeptical of artificial consciousness would surely not assign it a probability literally of zero. Because an ordinal heuristic seeks only to assign *relative* probabilities, a biological theorist might be still willing to grant that AIs with a higher GI are *more likely* to be conscious than those with a lower GI, even if very low probabilities are assigned in both cases. Second, to the extent that approaches like IIT are willing to attribute consciousness even to simple systems like thermostats, they are primarily concerned with consciousness as a fundamental informational property, rather than the richer notion of *conscious subjects* with, for example, moral status and a capacity for affective states. Since many of the questions to which we address our heuristic will concern specifically systems with these latter properties, there may still be common ground to be found insofar as these more sophisticated forms of consciousness in particular have some connection with general intelligence.

[f] This is compatible with consciousness *per se* serving no function (as Rosenthal urges), as the broader cognitive capacities that enable consciousness will still contribute to an organism's intelligence.

5. Conclusion

Questions about artificial consciousness will likely loom large in coming years. Answering these without committing to one of the many competing theories of consciousness will be challenging, and existing theoretically-neutral methods for assessing consciousness in humans and animals may be of little help. Instead, I suggested that the field seek consensus on *consciousness heuristics*, understood as independently plausible, scientifically tractable, and relatively theoretically-neutral tools for assessing consciousness in AIs, and I claimed that general intelligence may provide the basis for such a heuristic. Regardless of the merits of this latter particular proposal, I would urge that the field of AI consciousness prioritize the search for common ground.

Acknowledgements

This work was supported by the Leverhulme Centre for the Future of Intelligence, Leverhulme Trust, under Grant RC-2015-067.

REFERENCES

Agar, N. (2019). How to Treat Machines that Might Have Minds. *Philosophy and Technology*. https://doi.org/10.1007/s13347-019-00357-8

Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness*, *2018*(1). https://doi.org/10.1093/nc/niy007

Bekinschtein, T. A., Peeters, M., Shalom, D., & Sigman, M. (2011). Sea Slugs, Subliminal Pictures, and Vegetative State Patients: Boundaries of Consciousness in Classical Conditioning. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00337

Birch, J. (2019). Invertebrate Consciousness: Three Approaches. *Unpublished Manuscript*.

Bishop, J. M. (2018). Is Anyone Home? A Way to Find Out If AI Has Become Self-Aware. *Frontiers in Robotics and AI*, *5*. https://doi.org/10.3389/frobt.2018.00017

Carruthers, P. (2005). Consciousness: Essays from a Higher-Order Perspective. In *Consciousness: Essays from a Higher-Order Perspective*. https://doi.org/10.1093/0199277362.001.0001

Cohen, M. A., & Dennett, D. C. (2011, August). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, Vol. 15, pp. 358–364. https://doi.org/10.1016/j.tics.2011.06.008

Danaher, J. (2019). The Philosophical Case for Robot Friendship. *Journal of Posthuman Studies*, *3*(1), 5. https://doi.org/10.5325/jpoststud.3.1.0005

Dehaene, S. (2014). *Consciousness and the brain: deciphering how the brain codes our thoughts*. https://doi.org/10.1080/15294145.2014.956209

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79*(1–2), 1–37. https://doi.org/10.1016/S0010-0277(00)00123-2

Kanai, R. (2017). We Need Conscious Robots. *Nautilus*, *47*. Retrieved from http://nautil.us/issue/47/consciousness/we-need-conscious-robots

Kanai, R., Fujisawa, I., Tamai, S., Magata, A., & Yasumoto, M. (n.d.). *Artificial Consciousness as a Platform for Artificial General Intelligence*. https://doi.org/10.31219/OSF.IO/E4JH2

Lau, H. C. (2007). A higher order Bayesian decision theory of consciousness. In *Progress in Brain Research* (Vol. 168, pp. 35–48). https://doi.org/10.1016/S0079-6123(07)68004-2

Lau, H., & Rosenthal, D. (2011, August). Empirical support for higher-order theories of conscious awareness.

Trends in Cognitive Sciences, Vol. 15, pp. 365–373. https://doi.org/10.1016/j.tics.2011.05.009

Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, *17*(4), 391–444. https://doi.org/10.1007/s11023-007-9079-x

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting Awareness in the Vegetative State : Supporting Information. *Science*, *313*(5792), 1402–1402. https://doi.org/10.1126/science.1130197

Prinz, J. (2012). *The conscious brain*. Oxford University Press.

Rosenthal, D. (2006). *Consciousness and Mind*. Retrieved from https://www.amazon.com/Consciousness-Mind-David-Rosenthal/dp/0198236964/ref=sr_1_3?ie=UTF8&qid=1467577983&sr=8-3&keywords=david+rosenthal

Rosenthal, D. M. (2008). Consciousness and its function. *Neuropsychologia*, *46*(3), 829–840. https://doi.org/10.1016/j.neuropsychologia.2007.11.012

Rosenthal, D. M. (2009). Higher-Order Theories of Consciousness. In *The Oxford Handbook of Philosophy of Mind*. https://doi.org/10.1093/oxfordhb/9780199262618.003.0014

Rutz, C., Sugasawa, S., van der Wal, J. E. M., Klump, B. C., & St Clair, J. J. H. (2016). Tool bending in new caledonian crows. *Royal Society Open Science*, *3*(8). https://doi.org/10.1098/rsos.160439

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X00005756

Sitt, J. D., King, J. R., El Karoui, I., Rohaut, B., Faugeras, F., Gramfort, A., … Naccache, L. (2014). Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. *Brain*, *137*(8), 2258–2270. https://doi.org/10.1093/brain/awu141

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

Terrace, H. S., & Son, L. K. (2009). Comparative metacognition. *Current Opinion in Neurobiology*, *19*(1), 67–74. https://doi.org/10.1016/j.conb.2009.06.004

Thompson, E. (2010). *Mind in life*. Harvard University Press.

Tomasik, B. (2014). *Do Artificial Reinforcement-Learning Agents Matter Morally?*

Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668). https://doi.org/10.1098/rstb.2014.0167

Turing, A. (1950). Computing Machine & Intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Turner, E., & Schneider, S. (2020). The ACT test for AI Consciousness. In M. Liao & D. J. Chalmers (Eds.), *Ethics of Artificial Intelligence*. Oxford: Oxford University Press.

Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*, *10*(2), 2019–2022. https://doi.org/10.2478/jagi-2019-0002

Weir, A. A. S., Chappell, J., & Kacelnik, A. (2002). Shaping of hooks in new caledonian crows. *Science*. https://doi.org/10.1126/science.1073433