To build conscious machines, focus on general intelligence: a framework for the assessment of consciousness in biological and artificial systems

Henry Shevlin [0000-0002-7753-3281]

Leverhulme Centre for the Future of Intelligence, University of Cambridge
henry.shevlin@gmail.com

1 Introduction

Consciousness presents us with a number of different explanatory challenges. The most fundamental, sometimes called the 'Hard Problem', is focused on how subjective states can arise from objective physical systems [1]. Another important question concerns the cognitive mechanisms that distinguish conscious from unconscious states [2]. A third debate, and the one that will be the focus of the present enquiry, concerns how we can determine whether a given biological or artificial agent is conscious at all. Of the three questions, the latter has particular practical and ethical significance: our treatment of animals depends in part on whether we regard them as having a capacity for conscious experience [3]. Likewise, while few would endorse the idea that current artificial systems are conscious, as their capacities improve and come to more closely resemble those of animals and humans, ethical and legal questions concerning machine consciousness will likely loom large.

In this paper, I will argue that a useful framework for the assessment of consciousness in both animals and machines can come from the notion of general intelligence. I begin in Section 2 by noting the important connections between our concepts of intelligence and consciousness. In Section 3, I argue that it is general rather than specialised intelligence that carries the greatest weight in our assessments of consciousness, and offer a preliminary framework for the assessment of general intelligence that appeals to three features, namely robustness, flexibility, and system-wide integration. In Section 4, I argue that current artificial systems, unlike many non-human animals, currently fail to exhibit these features of general intelligence to any significant degree. As a result, I suggest that we have some reason to think artificial consciousness remains a distant goal. Finally, in Section 5, I briefly survey some challenges faced by a framework that takes general intelligence to be our best evidence of consciousness.

Before proceeding, it will be helpful to provide a brief gloss on the notions of consciousness and intelligence. In short, I use the term consciousness to refer to a capacity for subjective experience. For conscious creatures, there is something it's like to have experiences in the sense of Nagel [4]: colours may look a certain way and pains feel a certain way to them. I will use the term intelligence broadly to refer to the capacity of a system to use information processing to achieve its goals in an efficient and effective

manner. Note that I regard consciousness as a pretheoretical concept whose reference we grasp first-hand. By contrast, intelligence is a theoretical concept open to revision.

2 Consciousness and intelligence

We have all intuitions about which animals are conscious. Most of us, I suspect, would regard it as beyond serious question that dolphins, chimpanzees, and dogs undergo subjective experiences. Likewise, relatively few take seriously the idea that extremely simple systems such as a thermostats, bacteria, or plants are conscious (however, see [5] and [6]). Between these two extremes, however, there is considerable disagreement. For example, there is considerable controversy regarding consciousness in fish [7], cephalopods [8], and insects [9]. This lack of agreement in intuitions is reflected by the disparate nature of legal protections for different species in different jurisdictions. Thus while the British Animals (Scientific Procedures) Act of 1986 extends protections to all vertebrates as well as octopuses, the corresponding American legislation (7 U.S.C. § 2131-2156) makes provision only for warm-blooded animals. Similar controversies also arise for humans in relation to patients in comas and persistent vegetative states, as well as foetuses, and are likely in time to arise for machine intelligences.

A tempting response to these conflicting intuitions may be to disregard their value entirely, and adopt a purely scientific criterion of consciousness. I would suggest, however, that such a conclusion would be misguided. Consciousness is a pretheoretical concept with deep connections to our ethical practises, and cannot simply be operationalised in the interests of scientific expediency. While we might simply stipulate, for example, that we will define consciousness as sensitivity to external stimuli or a capacity for higher-order cognition, these definitions will not serve the purposes that are required of a theory of consciousness. Faced with questions like whether fish feel pain, or whether patients in persistent vegetative states are having experiences, we wish to know whether they *really* have subjective experiences, not just whether they are sensitive to external stimuli or have intact metacognitive capacities.

Any theory of consciousness that will serve the purposes to which we wish to put it, then, must reflect and engage with our standing folk psychological concept of consciousness. This is not to say, of course, that our current attitudes are immune to revision. An important part of our task in giving a theory of consciousness is identifying ungrounded biases and assumptions that regulate our intuitions about which animals are conscious, such as our tendency to attribute it to charismatic megafauna. Moreover, after identifying and regimenting our core pretheoretical commitments concerning the nature of consciousness, we may find that far more (or far fewer) systems satisfy them than we had previously assumed.

In light of this, I would suggest that an important starting point in reflecting on consciousness comes from the powerful pretheoretical connection between consciousness and intelligence. As noted above, we unhesitatingly assign consciousness to creatures like chimpanzees, dolphins, and dogs. It is surely no coincidence that these animals are all highly intelligent. Likewise, systems that strike us as very poor consciousness candidates (in the sense of Birch, [10]) tend to be extremely cognitively simple, capable of

little intelligent behaviour. A somewhat similar story can be told for systems intermediate between these two poles, as shown below (Fig. 1).

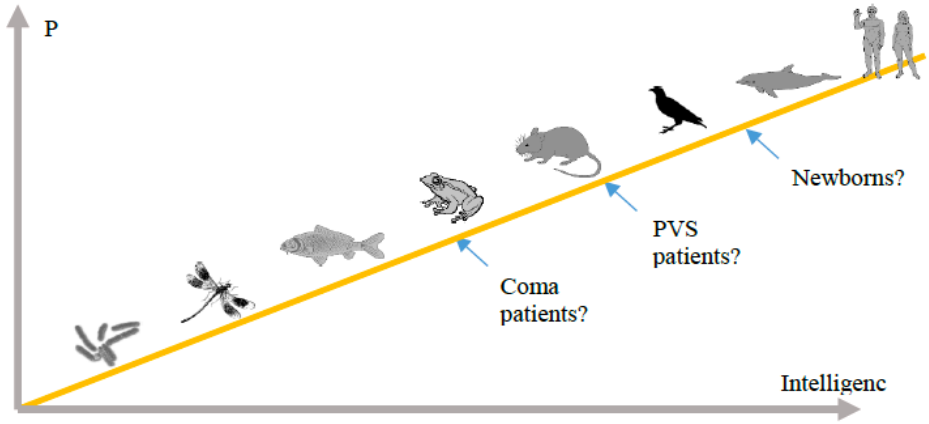

Fig. 1. A schematic depiction of the relation between intelligence and our pretheoretical judgments of the probability that different cognitive agents are conscious, ranging from very low in the case of bacteria to near-certainty in the case of adult humans, newborns, and dolphins.

There are grounds for thinking that this apparent connection between intelligence and our assessment of consciousness is not an idle correlation. When new evidence emerges of intelligent behaviour in a species, it is likely to increase our confidence that the species is conscious, and in turn be reflected in legal protections; the decisions to afford rights to many cephalopods in European Directive 2010/63/EU (5), for example, was prompted by evidence that they possess higher-brain areas and sophisticated behavioural responses to pain [11]. Similarly, the growing scientific consensus that some patients in persistent vegetative states are conscious followed the discovery that some PVS patients retained an ability to accurately answer yes or no questions in a brain scanning paradigm [12].

There are of course many tricky conceptual issues to wrangle with in relating consciousness to intelligence, some of which will occupy much of the rest of this paper. One important issue worth immediately flagging concerns whether we should regard consciousness as a discrete phenomenon, or might be able talk of degrees of consciousness in intelligent systems. This is a point of considerable philosophical and scientific controversy, however, and will be set aside in what follows.

3 Distinguishing general and specialised intelligence

Despite the connection between judgments of intelligence and consciousness, few would consider existing machine intelligences to be strong consciousness candidates, despite their impressive performance on a number of seemingly demanding tasks. There are a number of reasons both good and bad why this might be the case. However, as I will now argue, one well-founded motivation may come from the difference between

specialised and general intelligence. In short, I suggest that insofar as there is an important connection between consciousness and intelligence, it is general intelligence that matters.

First, however, it is necessary to at least outline what I take the terms to mean. Specialised intelligence is relatively easy to grasp, and can be spelled out in terms of the definition of intelligence given earlier, namely the ability to use information processing effectively and efficiently in the pursuit of some narrowly specified goal. General intelligence is more difficult. The notion of generality in artificial intelligence was discussed in an important paper by McCarthy [13], and can broadly be understood as the ability to use "the same goal-seeking mechanism for all kinds of problems, changing only the particular productions." As humans, we like to think of ourselves as having a high degree of general intelligence, as reflected in our complex societies and elaborate cultural products. However, on McCarthy's definition, it seems like that many different animals would also qualify as having high general intelligence: via well-integrated systems of perception, memory, and learning, they overcome numerous different tasks, ranging from long-distance navigation, predation and predator avoidance, mating, and the satisfaction of numerous competing physiological needs.

What is needed, however, is a more systematic framework for assessing general intelligence in different systems. With this in mind, I would suggest that three features are commonly found in systems that we would regard as generally intelligent. In short, these are *robustness*, *flexibility,* and *whole-system integration*, to be spelled out as follows.

Robustness: the ability to achieve tasks despite interference. Systems with a high general intelligence are typically robust (resisting failure) and resilient (recovering from failure).

Flexibility: the ability to transfer knowledge across tasks. Generally intelligent systems can readily apply existing skills/information to new domains.

Whole-system integration: the ability of a creature to integrate and effectively counterbalance inputs from different systems, including perception, memory, and drives.

As an inventory of features of generally intelligent systems, this list is highly preliminary, and may prove to be incomplete or have unnecessary components. Note, for example, that there are clear connections between robustness and flexibility: a system is likely to be more robust in dealing with unexpected impediments to its goals if it also displays a high degree of flexibility. Likewise, a system is likely to be more flexible if it can integrate all of its sources of information for the intelligent production of behaviour.

Nonetheless, I would suggest that this initial framing of the features of general intelligence has considerable value for the assessment of general intelligence in different systems. More importantly for present purposes, however, all of these features have both pretheoretical and scientific appeal as parameters in the assessment of which creatures are conscious.

To illustrate this point, imagine that we have identified some seemingly very intelligent behaviour in a species not previously considered a good consciousness candidate; suppose, say, that we found evidence of complex tool use in bivalve molluscs. Prima

facie, this would make the creature a stronger consciousness candidate. However, if it were found that the behaviour failed to satisfy one of the dimensions of general intelligence discussed above, this impression would be undermined. Thus if we discovered that the behaviour was *non-robust* and failed outside of extremely specific conditions, we might naturally assume its occurrence relied on simple hard-wired mechanisms rather than being a marker of sophisticated cognition; the contested case of the Sphex wasp might serve as an example of such [14]. We might draw a similar conclusion if the behaviour proved to be wholly *non-flexible*, and could be applied to only one very narrow purpose; in such a case, the behaviour might best be explained via a single evolutionarily honed instinct. Consider, for example, the complex but highly stereotyped nest-building performed by many insects. Finally, if neuroscientific enquiry revealed the behaviour to be accomplished not via the central nervous system, but some wholly non-integrated neural module, we would be much less likely to think it good evidence for the creature's being conscious.

The above examples should serve to provide some initial motivation for the claim that general intelligence understood as the possession of robust, flexible, and integrated behavioural capacities provides evidence of consciousness. However, I would also note that it provides a good fit for many other approaches to consciousness. As noted above, evidence for consciousness in persistent vegetative state patients comes in large part from their ability to perform a highly flexible task, namely accurately answering a range of yes or no questions about different personal and factual matters. Likewise, tests of machine intelligence and consciousness such as the Turing Test and the Winograd schema measure an artificial system's capacity to engage in flexible and robust forms of verbal reasoning, and objections to the value of such tests such as Searle's Chinese Room sometimes rely on showing how the task can be performed via 'dumb' non-integrated processes [15, 16]. Finally, note that the proposed schema for assessing general intelligence exhibits some promising connections with contemporary scientific approaches to consciousness. An emphasis on cross system integration, for example, is common to many leading theories such as Integrated Information Theory and Global Workspace Theory, and there is empirical reason to think that many highly flexible forms of behaviour such as memory-trace conditioning and unconscious two-step arithmetic can be performed only under conscious conditions [17].

Note that I am not proposing that consciousness be *identified* with general intelligence, nor suggesting general intelligence as the *mechanism* by which consciousness arises. The former goal involves conceptual difficulties best left to metaphysicians, while I regard questions about the mechanisms of consciousness as complementary to the current proposal. Instead, I am suggesting that general intelligence – understood as a capacity for robust, flexible, integrated cognitive performance – constitutes an important (if not our best) source of evidence of consciousness.

4 General intelligence in biological and artificial systems

Assuming, then, that general intelligence is a good marker for consciousness, then, we might ask how it applies to different cognitive systems. As suggested earlier, many

animals do well by this metric. Most animal behaviour is frequently highly robust, with feeding, mating, and thriving being accomplished in a wide range of varied environments and climatic conditions. While it is easy to find video examples of animals making foolish mistakes or falling over themselves, these occurrences are rare enough that they amuse and surprise us when they do occur. There are of course strong evolutionary reasons why we should expect animals to have robust capacities. Nonetheless, that does not mean that this achievement is easily won in cognitive terms.

Similarly, animal behaviour is often highly flexible. The ability of even simple creatures such as bees to engage in novel social learning and concept acquisition [18, 19] in conditions significantly removed from their evolutionary environments is striking. Among more intelligent creatures such as crows and scrubjays, examples abound of sophisticated and adaptive causal reasoning [20] and clever caching behaviours that are sensitive to a wide range of environmental factors and physiological needs [21].

Finally, there are many examples of highly-developed integration of different systems within animals, ranging from simple phenomena such as the use of motor and vestibular cues to distinguish endogenously- from exogenously-generated changes in sensory input, to the ability of fish, rats, and some crustaceans to engage in 'motivational tradeoff', the rapid online adjustment of behaviour to accommodate different desires and aversions [22].

This is a highly condensed review of just some of the ways in which animals display impressive forms of general intelligence, but it at least provides a useful point of comparison for examining the state of general intelligence current artificial systems. As anyone familiar with the current capacities of artificial systems can attest, these are fairly dismal. In most domains, AI behaviour is non-robust: machine intelligences struggle with tasks outside of highly-regulated training environments, and are vulnerable to a pile up of small errors. The problem of adversarial examples constitutes a vivid case of this: machine vision systems remain vulnerable to making spectacular errors when fooled by clever perturbations of input data (see Fig. 2, below) [23].

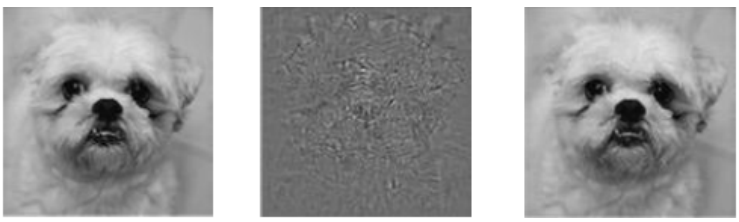

Fig. 2. A demonstration of an adversarial example which tricks a computer vision system into classifying a dog as an ostrich [24].

Similarly, most current AIs are highly inflexible. While important progress has been made on transfer learning tasks, even the best current systems gain only minor performance improvements when leveraging prior knowledge to variant tasks [25]. Somewhat more progress has been made on minimising the impact of catastrophic forgetting

in machine intelligences, but again, artificial systems exhibit strikingly limited capacity in comparison to non-human animals to engage in fluid task switching without considerable loss of prior knowledge. Finally, most machine intelligence systems rely on extended exposure to large training sets, again a dramatic contrast with biological intelligence (consider that a newborn fawn learns to stand up after 10 minutes and walk smoothly in just 7 hours).

Most AIs similarly fail to satisfy the integration component of general intelligence, for the simple reason they are wholly specialised machines. Even in systems that perform multiple functions, this is typically done via highly modular design in which there is little true integration of processing across subsystems. Integration of the kind found in animals arguably requires a centralised capacity for modelling the world and one's action space within it in rich detail, something AI has yet to come close to achieving.

5 Objections

Before concluding, it is worth briefly mentioning three important objections to this approach. First, one may question whether general intelligence is really a robust scientific concept [26]. If what we call general intelligence cannot really be quantified even in a multidimensional framework, but is best considered a loose 'bag of tricks', then this might certainly limit the utility of the approach defended her. However, this is a matter of outstanding scientific debate, and I would suggest that general intelligence may serve as a useful heuristic for assessments of consciousness even if it fails to track any unified set of cognitive mechanisms.

Second, it might be objected that there is little point developing measures of consciousness that are not explicitly related to specific psychological mechanisms such as attention, metacognition, or working memory. My response to this claim is that I regard the current approach as complementary to attempts to identify fine-grained cognitive structure of consciousness. It is also compatible with my approach that, as many have argued, there is no such structure to be found [27].

Third, a key worry for the tripartite schema for assessing general intelligence and consciousness given above is that it will be of practical value only insofar as we can develop principled ways of measuring robustness, flexibility, and whole-system integration. This may be an extremely challenging task. For example, while we might regard human arithmetical capabilities as highly flexible, few of us could compute square roots of large numbers in our heads, while this task is trivial for many machines. How to assess and assign weights both to the overall flexibility of a system and toits flexibility in different domains, then, remains an important outstanding challenge.

6 Conclusion

My main claims in this paper have been threefold. First, I have argued that there are important connections between the notions of consciousness and intelligence, with general intelligence in particular having an important evidential role for our assessments of consciousness. Second, I have suggested that general intelligence can be helpfully

conceptualised as spanning three dimensions, encompassing robustness, generality, and whole system integration. Third, I have claimed that while many animals perform well by these metrics, current artificial systems perform extremely poorly, and as such, it is unlikely that near-future artificial intelligences will be conscious. This may seem to some a pessimistic conclusion. However, if my arguments in this paper are well founded, I would suggest it provides important guidance for those interested in building conscious machines, and clear criteria to aim for: if you want a conscious machine, focus on building one that is as smart as a crow.

References

1. Chalmers, David J. (1995). Facing up to the problem of consciousness. Journal of Consciousness Studies 2 (3):200-19.
2. Block, Ned (2009). Comparing the major theories of consciousness. In Michael Gazzaniga (ed.), The Cognitive Neurosciences IV. pp. 1111-1123.
3. Singer, Peter (1989). All Animals Are Equal. In Tom Regan & Peter Singer (eds.), Animal Rights and Human Obligations. Oxford University Press. pp. 215--226.
4. Nagel, Thomas (1974). What is it like to be a bat? Philosophical Review 83 (October):435-50.
5. Chalmers, David J. (1996). The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.
6. Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? Philosophical Transactions of the Royal Society B: Biological Sciences, 370 (1668).
7. Key, B. (2014). Fish do not feel pain and its implications for understanding phenomenal consciousness. Biology & Philosophy March 2015, Volume 30, Issue 2, pp 149–165
8. Godfrey-Smith, P. (2017). Other Minds: The Octopus and the Evolution of Intelligent Life. William Collins.
9. Barron, A. B., & Klein, C. (2016). What insects can tell us about the origins of consciousness. Proceedings of the National Academy of Sciences Proc Natl Acad Sci USA, 113(18), 4900-4908
10. Birch, Jonathan (2017). Animal Sentience and the Precautionary Principle. Animal Sentience 2:16(1).
11. Fiorito G, Affuso A, Basil J, Cole A, de Girolamo P, D'Angelo L, Dickel L, Gestal C, Grasso F, Kuba M, Mark F. Guidelines for the Care and Welfare of Cephalopods in Research–A consensus based on an initiative by CephRes, FELASA and the Boyd Group. Laboratory animals. 2015; 49(2 suppl): 1-90.
12. Monti, M.M., Owen, A.M. (2010). The behavior in the brain: using functional neuroimaging to assess residual cognition and awareness after severe brain injury, Journal of Psychophysiology, 24(2): 76-82.
13. McCarthy, J. 1990. "Generality in artificial intelligence". In Lifschitz, V., ed., Formalizing Common Sense. Ablex. 226-236.
14. Dennett, D. (1973). Mechanism and responsibility. In T. Honderich (Ed.), Essays on freedom of action. London: Routledge.
15. Searle, J.R. (1980). Minds, brains and programs. Behavioral and Brain Sciences 3 (3):417-57.

16. Sharma, A., Ha Vo, N., Aditya, S., and Baral, C. "015). Towards addressing the winograd schema challenge-building and using a semantic parser and a knowledge hunting module. In IJCAI, pages 1319–1325, 2015.
17. Dehaene, S. (2014). Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts. Viking Press, 2014.
18. Alem, S., Perry, C.J., Zhu, X., Loukola, O., Ingraham T., Søvik, E., Chittka, L. Associative mechanisms allow for social learning and cultural transmission of string pulling in an insect. 2016; PLoS Biol 14(10).
19. Avargues-Weber, A., Dyer, A. G., & Giurfa, M. (2010). Conceptualization of above and below relationships by an insect. Proceedings of the Royal Society B: Biological Sciences, 278(1707), 898-905.
20. Jelbert SA, Taylor AH, Cheke LG, Clayton NS, Gray RD (2014). "Using the Aesop's Fable Paradigm to Investigate Causal Understanding of Water Displacement by New Caledonian Crows". PLoS ONE. 9: e92895. doi:10.1371/journal.pone.0092895
21. Clayton N.S, Bussey T.J, Dickinson A. Can animals recall the past and plan for the future? Nat. Rev. Neurosci. 2003a;4:685–691.
22. Motivational Tradeoff
23. Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533 (2016)
24. Scharr, P. (2018). Army of None: Autonomous Weapons and the Future of War. Norton.
25. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., & Gershman, S.J. (2016). Building Machines That Learn and Think like People." Behavioral and Brain Sciences, vol. 40, 2016.
26. Serpico, D. (2018). What kind of kind is intelligence? Philosophical Psychology, 31(2), 232–252. doi:10.1080/09515089.2017.1401706
27. Dennett, Daniel C. (1991). Consciousness Explained. Penguin Books.