# Cognitive theories of consciousness and the trivial realization argument

## Introduction

An important goal for scientific theories of consciousness is to enable us to make more informed judgements about which beings are conscious. In the case of fellow humans, such ascriptions are typically unproblematic. Things become harder, however, when dealing with patients in vegetative states; harder still when we consider non-human animals; and hardest of all when speculating about consciousness in non-biological systems.

In this short paper, I wish to explore an important worry relating to the ascription of consciousness. In short, I claim that many of the current leading theories are vulnerable to *trivial realisations*; that is, they are prima facie committed to consciousness in the kind of simple artificial system that a computer scientist could build or design with relatively little difficulty. This is what I will term the trivial realisation argument (TRA). I begin in Section 1 by briefly summarising the current state of the theories of consciousness debate, and single out a set of theories that I term *cognitive* theories of consciousness that I take to be particularly vulnerable to the TRA. In Section 2, I lay out the trivial realisation argument in more detail. In Section 3, I consider possible responses that cognitive theories of consciousness can make to the TRA. Finally, in Section 4, I suggest an alternative 'whole system' approach for the ascription of consciousness that sidesteps the TRA altogether.

## 1. System consciousness and the theories of consciousness debate

A central goal in the study of consciousness has been to develop an account to explain why some psychological and neural states are conscious and others are not. Frameworks that attempt to offer such explanations are typically known as *theories of consciousness*, and the business of assessing and critiquing such theories as the *theories of consciousness debate*. While the theories of consciousness debate draws on longstanding philosophical and scientific issues, as a specialised area of cognitive science involving active collaboration between philosophers and scientists its modern history began in the 1980s, with the emergence of framework such as Rosenthal's Higher-Order Thought theory (Rosenthal 1986) and Baars' Global Workspace account (1988). Since then, a plethora of other

theories have emerged, with notable examples including Tononi's Integrated Information Theory or IIT (Tononi 2008), biological theories of the kind defended by Ned Block and Victor Lamme (Ned Block 2009; Lamme 2010), and Dehaene and Naccache's Global Neuronal Workspace (Dehaene and Naccache 2001).

While no consensus has emerged from the debate concerning which theory of consciousness is to be preferred, the debate as a whole has been highly productive, inspiring important scientific and clinical work on the assessment of consciousness in vegetative state patients (Owen et al. 2006; Casali et al. 2013) and giving rise to now canonical theoretical concepts such as Block's distinction between phenomenal and access-consciousness (Block 1995).

While the theories of consciousness debate has mainly focused on what distinguishes conscious from unconscious states in individual human subjects, it also has fairly direct ramifications for the ascription of consciousness to non-human systems. Thus if a theory posits that some specific mechanism or process gives rise to consciousness in the human case, we might reasonably expect non-human systems that implement that same mechanism or process to have conscious states. However, most theories of consciousness have held off from making claims about which non-human systems are or could be consciousness.

Whether or not we can reasonably expect a theory to make claims about consciousness in non-human systems depends, of course, on its broader commitments. Theories that take consciousness to have a still poorly understood biological basis (Block 2009; Searle 1980), for example, might consistently claim that we simply lack the requisite information about the fundamental mechanisms of consciousness to make confident ascriptions outside central cases. This particular kind of agnosticism, however, is not available to those theories that identify consciousness with some specific cognitive process such as higher-order thought or global broadcasting. By their lights, we already have a good theory of the specific forms of information processing involved in consciousness, and the only remaining challenge in ascribing consciousness to non-human systems is to establish which of them implement the relevant processes. This may of course allow us to reserve judgment about consciousness in systems whose internal architecture is still poorly understood, but we must be committed at least to specific conditional claims about consciousness in non-human systems: if the

system implements the relevant processes or mechanisms, it will *ipso facto* have conscious states.

I take it that many of the leading theories – including global workspace theory, higher-order thought theory, and integrated information theory – fit this latter description, and I will refer to them collectively as *cognitive* theories of consciousness. It these theories that face a particular challenge from the TRA, as I will now describe.

**2. The Trivial Realisation Argument**

In essence, the TRA holds that, given an appropriately detailed specification of the mechanisms of consciousness as espoused by a particular theory, it is possible to build or at least design a simple realisation of those mechanisms that is not itself a plausible consciousness candidate. We can put the argument a little more formally, as follows.

(1) By the lights of some theory T, it is sufficient for a system to have conscious states that it implements some mechanism M.

(2) It is possible to build or design some simple system S that implements M.

(3) System S is not conscious.

(C) Mechanism M is not sufficient for consciousness and T is false.

I would suggest that versions of the TRA can be applied to most or all of the leading cognitive theories of consciousness, including Global Workspace Theory (Baars 1988; Dehaene and Naccache 2001), Higher-Order Thought Theory (Rosenthal 2005), Integrated Information Theory (Tononi 2008), Attention Schema Theory (Graziano 2013), and the Attended Intermediate Representations Theory (Prinz 2012).

Demonstrating this conclusively for a given case, of course, would require us to spell out the specific details of how we might design trivial implementation of these theories, and this kind of lengthy endeavour lies somewhat beyond the scope of this short paper. Additionally, some of the theories mentioned above lack detailed substrate-neutral specifications, thus making it impossible to show beyond doubt that the general processes they appeal to could be realistically be implemented in a simple system. In light of this, I am content for present purposes to address the TRA as a *conditional*

*challenge* to these theories: if it were discovered that a trivial realisation of a given theory were indeed possible, should we conclude that the theory is false?

Though I will frame the argument in these terms, I believe there is some reason to think that trivial realisations are indeed possible for most or all of the theories described. For one, note that each of the theories above identifies consciousness with some fairly high-level cognitive process, such as selective global information sharing in the case of Global Workspace Theory, or attention to intermediate level sensory representations in the case of Prinz's AIR theory. Insofar as these high-level processes can be implemented independently of the broader capacities of a system (such as its drives, sensorimotor capacities, or general intelligence), it should be possible to realise them in a system that lacks most or all features we associate with conscious systems. Certainly, these theorists give us no reason to think that implementing the mechanism in question – for example, having a global workspace – requires a system to be able to engage in complex and/or intelligent behaviour.

Additionally, it is worth noting that for what is arguably the most fully specified of the above frameworks, namely Integrated Information Theory, such a trivial realisation has already been demonstrated. Mathematician Scott Aaronson has shown that a cognitive system consisting of large number of XOR gates in a simple expander graph would have an arbitrarily high value of 'phi', IIT's measure of consciousness (Aaronson 2014). While of course we cannot generalise from this to a broader claim about the feasibility of such trivial instantiations for other theories, I believe it is suggestive that what is arguably the theory of consciousness spelled out in very fine detail has already proven vulnerable to such an example.

**3. Responding to the TRA**

As noted above, I am content for present purposes to frame the TRA as a conditional challenge to cognitive theories of consciousness. With this in mind, I believe it is worth surveying the range of responses that would be available to the various theories of consciousness if faced with a putative trivial implementation of the mechanism they take to be responsible for consciousness.

I suggest that we can group these responses into three kinds. First, the theorist might simply deny premise (3) above, and accept that this system is conscious, however counterintuitive it may

seem. Second, they might deny premise (2), and insist that, contrary to appearances, the implementation of the mechanism in question does not amount to a genuine realisation of the theory. Finally, the theorist might simply accept the conclusion of the argument as providing a strong case against the use of the theory for ascribing consciousness to non-human systems, while maintaining its viability as a theory of state consciousness in humans.

*3.1 – Biting the bullet*

One option – notably adopted by Tononi in response to Aaronson's Simple Expanders argument – is to grant that even simple systems that implement the relevant mechanisms or information-processing structure are conscious. While it may seem like a radical response to claim that, for example, a robot built by a computer scientist in a weekend using existing technology would be conscious, we should not dismiss this possibility out of hand. After all, it is not clear why we should expect a theory of consciousness to always give intuitive results, as demonstrated by, for example, the persistent popularity of panpsychism among some scientists and philosophers.

Nonetheless, I think there are major costs incurred by adopting this approach. For one, we should not underestimate just how dramatic a revision of existing folk attitudes it would constitute. In particular, consciousness is arguably somewhat different from most other psychological notions such as attention, learning, or memory insofar as it is a 'thick concept' (Hare 1952) that carries with it normative considerations. On learning that a system is conscious, many will be inclined to *ipso facto* attribute to it some degree of moral status. While many might be open to saying that a simple robot implements attention, learning, or memory, it will be harder to accept it possesses moral status.

A second perhaps more tendentious worry for this response it misunderstands what a theory of consciousness is for. For many, a guiding constraint in our search for a theory of consciousness is that it systematise as far as possible our pretheoretical platitudes about which states and systems are conscious (Lewis 1972). On this view, we *already have* a good implicit handle on consciousness, and the job of a theory is to regiment that understanding so as, for example, to better handle edge cases. To the extent that a theory does violence to our pretheoretical commitments concerning consciousness (for example, that simple robots are not conscious), it will thus be less attractive than one that can regiment consciousness without doing such violence to these commitments.

*3.2 – Rejecting simple implementations*

A second response open to the cognitive theorist of consciousness is to simply deny that the apparent trivial realisation of their theory truly satisfies its constraints. I should clarify that the case I have in mind is one in which a computer scientist builds a system that is at least *superficially* of the right kind, rather than one where all attempts to implement the relevant mechanism prove hopeless.

To give an illustrative example, imagine that we wish to build a system that implements Rosenthal's HOT theory of consciousness. Somewhat crudely, the theory claims that a first-order cognitive state such as a belief or desire or sensory state such as a representation of colour or pitch (a mental quality, in Rosenthal's terms) is conscious if and only if it is the object of a higher-order thought. We can now speculate about how a computer scientist might attempt to implement such a system in a simple computer. She might begin with a video camera feeding inputs to a processing unit that builds up a model of its local environment. She would then add an internal monitoring unit with a set of symbols for representing different states of the visual processing unit, for example encoding information of the kind "<System S> <is currently in> <Visual Processing State V>".

While this system might on the face of it seem to implement higher-order thought theory, there is a natural response on behalf of the HOT theorist, namely that the states of system are not *psychological* states: the states of visual processing unit are not mental qualities, and the higher-order representations are not thoughts. As a result, we can quite justifiably say that the system is not conscious.

I would note two things about this response. The first is that it shifts the question from what the mechanisms of consciousness are to the question of what constitutes different psychological states (in the strict sense). One risk is that, equipped with this latter, our imaginary computer scientist might still be able to build a system that meets these further constraints while remaining a very simple (and not plausibly conscious) system. The second is that this line of response is arguably *not* available to those cognitive theories of consciousness that are not framed in folk psychological terms. The concept of a 'thought' is a pretheoretical notion that we can reasonably expect to require substantial independent work to pin down. By contrast, theoretical concepts such as the Attention Schema or the Global Workspace are scientific posits. It thus seems harder to justify the claim that we need an

independent account of these terms when they were introduced in large part to help us understand consciousness in the first place.

*3.3 – Accepting the conclusion*

A final important line of response available to the cognitive theorist faced by a trivial realisation would be to accept that the system in question was both an accurate implementation of their theory and was also non-conscious, and to consequently limit their theory of consciousness to the problem of determining which states are conscious in human beings.

While this may be a reasonable move, it has some unattractive features. For one, it dramatically reduces the ambition of the theory of consciousness. Rather than telling us what consciousness *is*, the theory will at best serve as an account of the mechanisms of consciousness just in the human case. Additionally, this account will be *partial*: if mechanism M gives rise to consciousness in humans but not in simple robots, then there must be something *additional* to M present in humans.

**4. Alternatives to cognitive theories of consciousness**

Finally, I wish to make a speculative suggestion concerning how we might attempt to sidestep the TRA altogether. In short, the TRA relies on the fact that we already have strong pretheoretical criteria for the ascription of consciousness (for example, some degree of general intelligence). These criteria are met by humans and many animals, but not by simple robots. The TRA claims we can implement a candidate mechanisms of consciousness in a system that entirely fails to meet such criteria.

Another way of proceeding, then, would be to build these constraints into the theoretical apparatus we employ for ascribing consciousness in the first place. For example, we might attempt to formalise pretheoretical connections between general intelligence and consciousness, and apply these to the system as a whole. Notably, this would not be a cognitive account of consciousness, but a 'whole system' account that attributed consciousness on the basis of broad behavioural capacities.

This approach would not face worries about trivial realisations of consciousness because it would have incorporated into its criteria specifically the factors that distinguish trivial from non-trivial realisations. This need not mean, of course, that it would slavishly track *all* of our intuitions; in the

process of formalising our pretheoretical commitments, we may find contradictions or tensions that need to be resolved. But we might reasonably expect it to exclude extreme cases.

This is a highly incomplete sketch, of course, and leaves many questions unanswered, not least how we might reconcile a 'whole system' theory for the ascription of consciousness with the problem of distinguishing conscious from non-conscious states within a single system. However, it strikes me as offering a potentially valuable and straightforward way around the TRA.

## 5. Conclusion

My goal in this paper has been to flesh out an important challenge related to the ascription of consciousness faced by many theories, namely that they are vulnerable to trivial realisations. I have indicated some ways in which cognitive theories of consciousness might respond to the objection, while also indicating the limitations of such responses. Finally, I provided a brief sketch of another approach to the TRA, namely the whole-system approach. In summary, I would suggest that while challenges relating to trivial realisations are not strictly novel and have been directed to theories in the past, the power and broad applicability of the TRA mark it out as a key obstacle for theories of consciousness to overcome.

## REFERENCES


Aaronson, Scott. 2014. 'Why I Am Not An Integrated Information Theorist (or, the Unconscious Expander).' *Shtetl-Optimized* (blog). 2014. https://www.scottaaronson.com/blog/?p=1799.

Baars, Bernard J. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press.

Block, N. 1995. 'On a Confusion about a Function of Consciousness'. *Behavioral and Brain Sciences* 18: 227–287.

Block, Ned. 2009. 'Comparing the Major Theories of Consciousness'. In *The Cognitive Neurosciences IV*, edited by Michael Gazzaniga, 1111–1123.

Casali, Adenauer G., Olivia Gosseries, Mario Rosanova, Mélanie Boly, Simone Sarasso, Karina R. Casali, Silvia Casarotto, et al. 2013. 'A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior'. *Science Translational Medicine* 5 (198): 198ra105-198ra105. https://doi.org/10.1126/scitranslmed.3006294.

Dehaene, S., and L. Naccache. 2001. 'Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework'. *Cognition* 79: 1–37.

Graziano, Michael S. A. 2013. *Consciousness and the Social Brain*. OUP USA.

Hare, Richard M. 1952. *The Language of Morals*. OUP Oxford.

Lamme, V.A. 2010. 'How Neuroscience Will Change Our View on Consciousness'. *Cognitive Neuroscience* 13: 204–220.

Lewis, David. 1972. 'Psychophysical and Theoretical Identifications'. *Australasian Journal of Philosophy* 50 (3): 249–258.

Owen, Adrian M., Martin R. Coleman, Melanie Boly, Matthew H. Davis, Steven Laureys, and John D. Pickard. 2006. 'Detecting Awareness in the Vegetative State'. *Science* 313 (5792): 1402–1402. https://doi.org/10.1126/science.1130197.

Prinz, J. 2012. *The Conscious Brain: How Attention Engenders Experience*. Oxford: Oxford University Press.

Rosenthal, David. 2005. *Consciousness and Mind*. Oxford University Press UK.

Rosenthal, David M. 1986. 'Two Concepts of Consciousness'. *Philosophical Studies* 49 (May): 329–59.

Searle, John R. 1980. 'Minds, Brains and Programs'. *Behavioral and Brain Sciences* 3 (3): 417–57.

Tononi, G. 2008. 'Consciousness as Integrated Information: A Provisional Manifesto'. *The Biological Bulletin* 215 (3): 216.