Criteria for empirical theories of consciousness should focus on the explanatory power of mechanisms, not on functional equivalence

Fahrenfort, J.J.; van Gaal, S.

DOI
10.1080/17588928.2020.1838470

Publication date
2021

Document Version
Final published version

Published in
Cognitive Neuroscience

COGNITIVE NEUROSCIENCE
2021, VOL. 12, NO. 2, 93–94
https://doi.org/10.1080/17588928.2020.1838470

Routledge
Taylor & Francis Group

Criteria for empirical theories of consciousness should focus on the explanatory power of mechanisms, not on functional equivalence

Johannes J. Fahrenfort [a,b,c] and Simon van Gaal [a,b]

[a]Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands; [b]Amsterdam Brain and Cognition (ABC), University of Amsterdam, Amsterdam, The Netherlands; [c]Experimental and Applied Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

ABSTRACT

Doerig and colleagues put forward the notion that we need hard and theory-neutral criteria by which to arbitrate between empirical (mechanistic) theories of consciousness. However, most of the criteria that they propose are not theory neutral because they focus on functional equivalence between systems. Because empirical theories of consciousness are mechanistic rather than functionalist, we think these criteria are not helpful when arbitrating between them.

ARTICLE HISTORY
Received 31 August 2020
Published online 18 November 2020

KEYWORDS
Consciousness; mechanism; functionalism; functionalist explanation; mechanistic explanation

First, we praise the attempt to outline theory-neutral criteria that empirical theories of consciousness should be tested against. We also agree with the first criterion that the authors proposed: paradigm cases in consciousness research warrant explanation if a theory is to be taken seriously. However, the additional criteria put forward were – in our mind – less constructive. These are all functionalist/behaviorist criteria, because they require one to evaluate the presence of consciousness *only* through the behavior of a system (output of a function). This reasoning ignores first-hand knowledge from being such a system (i.e., a human), and the ability to extrapolate from there, as was nicely explained by Tsuchiya and colleagues (Tsuchiya et al., 2020). Here, we focus on another (but somewhat related) problem: the proposed criteria are not theory neutral. Functional explanations provide explanatory power by focusing on functions or goals of phenomena, whereas mechanistic explanations provide explanatory power by appealing to processes, parts and interactions between parts as constituting phenomena.

Importantly, most (if not all) empirical theories of consciousness are mechanistic rather than functionalist. Elsewhere, Doerig and colleagues have tried to argue that some of the important empirical theories of consciousness are in fact functionalist (Doerig et al., 2019), but this is hardly convincing. Typically, empirical theories attempt to establish a *mechanism* in an existing architecture (the brain) through clever experimentation and imaging. This is confirmed implicitly by the authors through repeated references to the word mechanism in the manuscript (it appears 62 times). Indeed, every major empirical theory aims to identify the neural mechanism that constitutes consciousness. Although functionalist explanations are not necessarily incompatible with mechanistic explanations, their goal is very different. Rather than establishing the neural basis of phenomena *in the brain* (which invariably involves mechanistic reasoning), they aim to understand a phenomenon by associating it with a function.

As an example of how functional reasoning in the context of mechanistic theories can go wrong, consider the following example: Imagine that researchers have established that the mechanistic basis of 'memory' is long term potentiation (LTP), the notion that neurons that are repeatedly active together are more prone to fire together in the future. We might say these researchers now understand the most basic mechanism of memory, as it explains how things become associated in a brain. Now let's imagine that some engineers have established that one can also implement seemingly equivalent 'memory' in transistors, using very different mechanisms. Even if this were possible, it would be absurd to claim that the theory about LTP forming the mechanistic basis of biological memory is 'incomplete' or 'wrong' because a functionally equivalent system exists. What counts is whether the mechanism of LTP provides explanatory power, helping us to understand memory in living organisms, not whether a seemingly functionally equivalent phenomenon can also be

CONTACT Johannes J. Fahrenfort fahrenfort.work@gmail.com Department of Psychology, University of Amsterdam, Amsterdam 1001, The Netherlands

implemented by a different mechanism. Theories of consciousness are no different in that respect.

Because functionalist criteria have different goals, they do not resolve much when applied to mechanistic theories. Although the issues the authors put forward are definitely interesting, they are not useful to arbitrate between theories at the current stage of empirical theory formation. Instead, it would be more useful to ask whether current theories of consciousness conform to criteria that identify good mechanistic theories: how are the causality relationships within a mechanism established, and how does the mechanism as a whole provide explanatory power for the phenomenon that requires explanation?

Then, if a mechanism consistently provides explanatory power for consciousness in a system that we have definitive information about (i.e. a human), the correct inferential step would be to assume that different systems (e.g., rats or AI) or smaller systems (10 neurons) with the same mechanism, are also conscious, although possibly degraded or altered. Importantly, this prediction only goes one way: a mechanism explains the phenomenon. A functionally equivalent phenomenon in a different system does not have to have the same mechanistic basis (as in the example of memory above), as that would be a case of reverse inference. Thus, whether one believes that 'true' multiple realizability exists in the case of consciousness, is irrelevant for arbitration between mechanistic theories: it does not disprove them or arbitrate between them. Summarizing, we do not believe the proposed criteria arbitrate between mechanistic theories of consciousness, but instead they might pose a source of confusion for those doing the empirical legwork.

ORCID

Johannes J. Fahrenfort http://orcid.org/0000-0002-9025-3436

References

Doerig, A., Schurger, A., & Herzog, M. H. (2020). Hard criteria for empirical theories of consciousness. *Cognitive Neuroscience*, *25*(1), 1–22. https://doi.org/10.1080/17588928.2020.1772214

Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Special Issue on Introspection*, *72*, 49–59. http://doi.org/10.1016/j.concog.2019.04.002

Tsuchiya, N., Andrillon, T., & Haun, A. (2020). A reply to "the unfolding argument": Beyond functionalism/behaviorism and towards a science of causal structure theories of consciousness. *Consciousness and Cognition*, *79*, 102877. https://doi.org/10.1016/j.concog.2020.102877