The Integrated Information Theory of Agency

Hugh Desmond[1,2]
Philippe Huneman[1]

[1] Institute of History and Philosophy of Science and Technology, Université Paris 1 Panthéon-Sorbonne – CNRS.

[2] Department of Philosophy, University of Antwerp

Published in Brain and Behavioral Sciences
doi:10.1017/S0140525X21002004, e0

Abstract

We propose that measures of information integration can be more straightforwardly interpreted as measures of agency rather than of consciousness. This may be useful to the goals of consciousness research, given how agency and consciousness are "duals" in many (though not all) respects.

Once consciousness is analysed as "efficient network activity", it is manifested across a broad range of systems, and its meaningfulness as a concept becomes diluted. That is, as Merker et al. successfully show, a fundamental challenge for the integrated information theory (IIT) of consciousness (Tononi 2008; Tononi et al. 2016).

As a way forward, we propose that measures of information integration can more straightforwardly be interpreted as measures of agency rather than of consciousness. Because agency can be defined in terms of behavioural patterns, it avoids the problems arising from quantifying "first-person perspective" properties by means of "third-person perspective" measures. As the conceptual dual of consciousness, agency may, thus, deserve a more prominent place in consciousness research.

Agency and Integration. Agency is increasingly of interest to biologists, as many developmental patterns and behaviors (including those of plants) are characterized in agential terms. It is part of a trend to assign a greater theoretical role to organisms as

such in our understanding of evolution: organisms do not simply passively undergo evolutionary processes but actively shape their selective environment (Laland, Matthews, and Feldman 2016) and respond in a goal-directed manner to opportunities or "affordances" in their environment (Walsh 2015). At a very general level, "agency" refers to how organisms exhibit goal-directed behaviours in response to environmental change.

Talking about goals in this way activates old worries about teleology and anthropomorphism (i.e., agency as human-like intentionality). However, in practice, the pay-off for explaining behaviors as goal-directed lies in accounting for patterns of behavioral robustness: an organism's "goal" is simply what it attempts to achieve through various means, even when it is perturbed or challenged by an environmental change. In other words, agency refers to how a (1) small number of goals can account for patterns of connectivity between (2) a large number of possible environmental states, and (3) a large number of possible behaviors.

This explanatory structure describes a bow-tie architecture (see Figure 1) where environmental states and behaviors are integrated in virtue of the "goals" present. Note that the figure does not illustrate any fine-grained connections between environmental states and behaviors. What it does illustrate is the general explanatory structure of agency, where the "goals" are used to explain how environmental states and behaviors are informationally integrated (for an information-theoretic treatment of scientific explanation: see e.g. Desmond 2019).

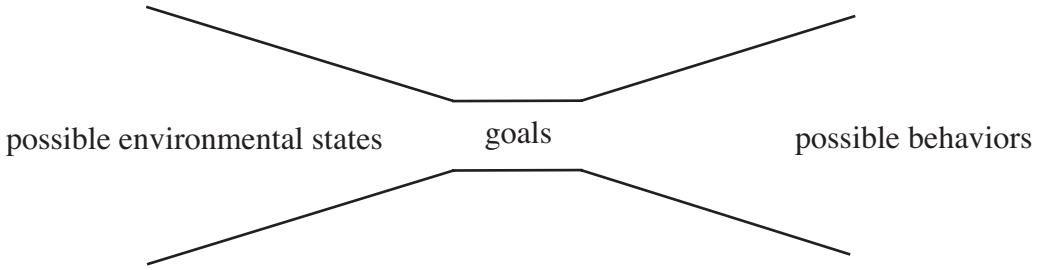

Figure 1: The bow-tie architecture of agency.

When the issue is put in this way, suggestive parallels with theories of consciousness emerge. The IIT posits consciousness as the integration of various experiential properties (Tononi 2008), but also global workspace theory (Dehaene and

Changeux 2011, 11) posits a similar structure, where consciousness is a global broadcast mechanism, integrating input and output systems.

However, unlike theories of consciousness, an integrated information theory of agency would not need further justification of why such bow-tie architectures should be identified with agency. Goal-directedness is a third-person concept and in this way it can be unproblematically fleshed out in terms of input-output patterns. Agency as a concept just *is* a type pattern of connectivity between environmental states and behaviors. There is no need to posit a counterpart to "qualia" or some ineffable subjective quality.

The underlying reason for this is that agency is an explanatory concept rather than one that refers to an empirical state-of-affairs. Informally, agency could be said to be more like "Newton's law of inertia" rather than to "white snow". If one is habituated to thinking of agency in terms of intentionality (or the presence of some form of mentality), this view of agency may require a gestalt-switch. The concept of agency imposes a structure on observed behavior, and if the observed behavioral patterns do not exhibit this general structure, there is simply no need to describe them as "agential" (see discussion in Desmond and Huneman 2020).

Agency and Consciousness as Duals. Whether or not the apparent isomorphism between theories of consciousness and the structure of agency is more than skin-deep is not a question we systematically address here. Instead, we offer a general rationale why the appearance of such isomorphisms should not be surprising. If agency refers to the "activity" of the organism in relation to the environment, consciousness in its broadest sense denotes the "passivity" of the organism. A synonym for consciousness – sentience – makes this passivity clearer: the capacity of "feeling" refers to how an organism "undergoes" its environment (or think of "e-motion": being moved). Agency and consciousness are different sides of the same fundamental coin of organism-environment relationality. One cannot have activity without passivity, and vice versa.

In mathematics, dual concepts are used to integrate two different ways of looking at a same object (Atiyah 2007). Similarly, agency and consciousness can, at a fundamental level, be viewed as "duals." And just as invoking the dual operator in mathematics may help solve otherwise intractable problems, perhaps some of the challenges facing our understanding of consciousness can be addressed by invoking agency. For instance, it is likely that the evolution of consciousness can only be

understood by simultaneously understanding how agency evolved. This is reflected in how greater sensorimotor control has evolved in tandem with various proxies of consciousness such as cognitive systems (van Duijn, Keijzer, and Franken 2006; Godfrey-Smith 2020).

Could an IIT of agency avoid the equivalent of panpsychism, which seems unavoidable once consciousness is naturalized an/or de-anthropomorphized? Panpsychism's dual is "panagentialism". In other contexts, this has been called hyper-agency detection: seeing agency everywhere (cf. Atran 2002). We believe panagentialism can be more easily defused, because of a subtle asymmetry between agency and consciousness (at least as the latter is typically understood). Attributing agency is an explanatory strategy to make sense of behavioral complexity – not a statement about the ontological makeup of the world. Panagentialism is thus simply a (poor) explanatory practise. Note that there may also be "no facts of the matter" regarding consciousness (Carruthers 2020). In that case, panpsychism would also be a poor explanatory practise, and agency and consciousness would be true duals.

REFERENCES

Atiyah, Michael. 2007. "Duality in Mathematics and Physics." Transcription of a talk delivered at the Institut de Matemàtica de la Universitat de Barcelona. https://fme.upc.edu/ca/arxius/butlleti-digital/riemann/071218_conferencia_atiyah-d_article.pdf.

Atran, Scott. 2002. *In Gods We Trust: The Evolutionary Landscape of Religion*. Oxford, UK: Oxford University Press.

Carruthers, Peter. 2020. "Stop Caring about Consciousness." *Philosophical Topics* 48 (1): 1–20.

Dehaene, Stanislas, and Jean-Pierre Changeux. 2011. "Experimental and Theoretical Approaches to Conscious Processing." *Neuron* 70 (2): 200–227. https://doi.org/10.1016/j.neuron.2011.03.018.

Desmond, Hugh. 2019. "Shades of Grey: Granularity, Pragmatics, and Non-Causal Explanation." *Perspectives on Science* 27 (1): 68–87. https://doi.org/10.1162/posc_a_00300.

Desmond, Hugh, and Philippe Huneman. 2020. "The Ontology of Organismic Agency: A Kantian Approach." In *Natural Born Monads: On the Metaphysics of Organisms and Human Individuals.*, edited by Andrea Altobrando and Pierfrancesco Biasetti, 33–64. Berlin: De Gruyter.

Duijn, Marc van, Fred Keijzer, and Daan Franken. 2006. "Principles of Minimal Cognition: Casting Cognition as Sensorimotor Coordination." *Adaptive Behavior* 14 (2): 157–70. https://doi.org/10.1177/105971230601400207.

Godfrey-Smith, Peter. 2020. "Varieties of Subjectivity." *Philosophy of Science* 87 (5): 1150–59. https://doi.org/10.1086/710541.

Laland, Kevin, Blake Matthews, and Marcus W. Feldman. 2016. “An Introduction to Niche Construction Theory.” *Evolutionary Ecology* 30: 191–202. https://doi.org/10.1007/s10682-016-9821-z.
Tononi, Giulio. 2008. “Consciousness as Integrated Information: A Provisional Manifesto.” *The Biological Bulletin* 215 (3): 216–42. https://doi.org/10.2307/25470707.
Tononi, Giulio, Melanie Boly, Marcello Massimini, and Christof Koch. 2016. “Integrated Information Theory: From Consciousness to Its Physical Substrate.” *Nature Reviews Neuroscience* 17 (7): 450–61. https://doi.org/10.1038/nrn.2016.44.
Walsh, Denis. 2015. *Organisms, Agency, and Evolution*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781316402719.