Paradox or illusion?

A comment on “The paradox of the self-studying brain” by Battaglia, Servajean, and Friston

Chris Fields

Allen Discovery Center, Tufts University, Medford, MA 02155, USA
fieldsres@gmail.com; ORCID: 0000-0002-4812-0744

What does it mean to say that the brain studies itself? Battaglia, Servajean and Friston – hereafter “BSF” – begin by assuming the fundamental distinction between introspection and perception, or between the 1st and 3rd person points of view, that has underpinned philosophical discussions of consciousness at least since Descartes. The titular paradox of the self-studying brain stems from the apparent independence of evidence garnered via these two perspectives, or processes, or investigatory methodologies. The fundamental question raised by this paper is, therefore, that of whether introspection and perception really are independent processes that act on fundamentally different kinds of data. They obviously *feel like* independent processes. Descartes argued that they are independent as a matter of logic – he could conceive of circumstances that would challenge any conclusion from perception, but found the idea of a challenge to his introspective conclusion that he existed inconceivable. A similar argument from conceivability underpins the distinction between “easy” and “hard” problems of consciousness (Chalmers, 1996). But is conceivability not, at bottom, a matter of subjective certainty – an epistemic feeling accessed via introspection? To privilege evidence from introspection over evidence from perception to distinguish introspection from perception is to beg the question.

As BSF themselves suggest, adopting the formal approach of the Free Energy Principle (FEP; Friston 2019; Friston et al., 2022) is potentially useful for clarifying the relationship between perception and introspection. Formal models have the advantages of abstraction and generality; the FEP makes no human-specific assumptions and no assumptions about “what it is like” to be any particular organism. It applies to processes implementing perception and action – active inference – in organisms from unicells to humans; hence it spans any suggested boundaries between organisms that do and do not experience introspection or organisms that are and are not sentient or phenomenally conscious. The FEP employs the same assumptions and formal structures at every scale; models of active inference by unicells differ from models of active inference by humans only in their hierarchical complexity. We can, therefore, employ the FEP to develop a generic, scale-free model of perception, and ask what would need to be added or changed to build a FEP-compliant model of introspection. Conceiving of something, moreover, is a mental action, often a covert action. Hence we can use the model to ask: what does it mean to say that something is or is not conceivable?

Models of complex systems compliant with the FEP treat every component of a complex system that has a boundary, i.e. that interacts with its environment only causally, as an active inference agent that uses its available computational resources to minimize its uncertainty about the behavior of its

environment. Multicellular organisms such as humans are collectives of active inference agents at multiple scales; individual neurons, cortical minicolumns, and coherently functional, conditionally independent networks are all active inference agents. Perception and action at the whole-organism scale are outcomes of the collective behavior of these smaller-scale active inference agents, each of which is perceiving and acting on its own local environment. The FEP formalism cleanly and unambiguously differentiates the sensations, actions, and experiences of a composite system from those of its components (Fields et al., 2025). A human being, for example, experiences the familiar external environment, while a cortical minicolumn mainly experiences the behavior of other minicolumns. Neither has unmediated sensory access to the environment of the other. Neuroscientists can measure, using a combination of instruments and theory, some aspects of the environment of a minicolumn, but cannot capture either the richness or the specificity of its sensations or actions.

Within the FEP framework, evidence that an entity X is conscious is input from the whole-system scale environment that results in the evolution of the collective state of a hierarchy of active-inference agents toward an attractor, some dimensions of which associate information about consciousness with information about an entity. The information encoded by the attractor is active, in the sense that being in the relevant state induces, or at least potentiates, particular kinds of actions. In a system capable of speech, these actions could including saying: “X is conscious”; in a system capable of inner speech, they could include consciously thinking “X is conscious” (Fields et al., 2025). One notices that one is thinking “X is conscious” via introspection; one notices that one is saying “X is conscious” via perception. Introspection and perception are distinct ways of noticing the output of the process of associating consciousness with X, not distinct ways of making the association.

What differs between the cases of X being some other person – e.g. Alice – and X being the self? In an FEP model, the only difference is the source, within the environment, of the input. My evidence that Alice is conscious comes from Alice’s body, via my eyes and ears. My evidence that I am conscious comes from my body, via my brainstem (Solms, 2019). The mechanisms of encoding and processing the two kinds of evidence are substantially the same; they employ the same conceptual and language systems. The inferences I make from the data can fail in similar ways. I can infer that Alice is conscious when she is sleepwalking, or infer that she is self-conscious when she is in a flow state. I can infer that I am dead when I am fully conscious – if I have Cotard’s syndrome, as BSF mention – or I can consider myself self-conscious when the self I am conscious of is a delusion.

What then is the difference? We call “perception” both the obtaining of data from the environment and its interpretation via concepts and event models (Hommel, 2019). Yet we call “introspection” just the noticing that we have internally spoken or visualized or otherwise sensed some internally-generated content; the data gathering part is “interoception” or “imagination”. Could this difference in language, reflecting as it does a difference in felt certainty, be the source of the “paradox”? Many other approaches seem to lead to this same conclusion, e.g. the experimental route of Chater (2018) or the evolutionary route of Keijzer (2025). It suggests that the “specialness” of introspection – and of the 1st person perspective – is illusory.

References

Chalmers, D. *The Conscious Mind*. Oxford Uniuversity Press, 1996.

Chater, N. *The Mind is Flat*. Allen lane, 2018.

Fields, C. et al., How do inner screens enable imaginative experience? *Neuroscience of Consciousness*, in press, 2025.

Friston, K. A free energy principle for a particular physics. Preprint https://doi.org/10.48550/arXiv.1906.101842019, 2019

Friston, K. et al. Path integrals, particular kinds, and strange things. *Physics of Life Reviews* 47 (2022) 35–62.

Hommel, B. Theory of Event Coding (TEC) V2.0: Representing and controlling perception and action. Attention, Perception, and Psychophysics 81 (2019) 2139–2154.

Keijzer, F. Full naturalism: The objectivity of subjective points of view. Biological Theory, in press, 2025. https://doi.org/10.1007/s13752-025-00493-9

Solms, M. The hard problem of consciousness and the Free Energy Principle. *Frontiers in Psychology* 9 (2019) 2714.