Physics of Life Reviews 52 (2025) 197–204

Contents lists available at ScienceDirect

Physics of Life Reviews

journal homepage: www.elsevier.com/locate/plrev

Review

The paradox of the self-studying brain

Simone Battaglia [a,*], Philippe Servajean [b], Karl J. Friston [c,d]

[a] Center for Studies and Research in Cognitive Neuroscience, Department of Psychology "Renzo Canestrari", Cesena Campus, Alma Mater Studiorum Università di Bologna, 47521 Cesena, Italy
[b] Laboratoire Epsylon EA 4556, Université Montpellier 3, Montpellier, France
[c] Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, UK
[d] VERSES AI Research Lab, Los Angeles, CA, 90016, USA

ARTICLE INFO

Communicated by: Fontanari

Keywords:
Theoretical neuroscience
Consciousness
Perception
Introspection
Neurophenomenology
Brain

ABSTRACT

The paradox of a brain trying to study itself presents a conundrum, raising questions about self-reference, consciousness, psychiatric disorders, and the boundaries of scientific inquiry. By which means can this complex organ shift the focus of study towards itself? We aim at unpacking the intricacies of this paradox. Historically, this question has been raised by philosophers under different frameworks. Thanks to the development of novel techniques to study the brain on a functional and structural level - as well as neurostimulation protocols that can modulate its activity in selected areas - we now possess advanced methods to progress this intricate inquiry. Nonetheless, the broader implications of the brain's pursuit of understanding itself remain unclear to this day. Ultimately, the need to employ both perception and introspection has led to different formulations of consciousness. This creates a challenge, as evidence supporting one formulation does not necessarily support the other. By deconstructing the paradoxical nature of self understanding - from a philosophical and neuroscientific point of view - we may gain insights into the human brain, which could lead to improved understanding of self-awareness and consciousness.

1. State of 'mind'

The study of the human brain has fascinated us for centuries, leading to advances in our understanding of its structure and function [1]. However, this pursuit raises a paradoxical question: how can the human brain, a complex organ of cognition and self-awareness, study itself? Have you ever pondered this question, or to be more precise, has your brain? This paradox challenges our understanding of self-reference, awareness, and consciousness [2,3]. For example, it has been argued that "adding observational, representational, or control capabilities to a meta-level component of a system cannot, even in principle, lead to a complete meta-level representation of the system as a whole." In short, "self models cannot, in general, be empirically tested by the systems that implement them." [4]. On the other hand, self-reference plays a vital (sic) role in distinguishing between self and other; both neuronally and immunologically [5] with Gödelian foundations [6].

When we study a phenomenon, we do it exclusively through measurement and perception. However, this is no longer true when the phenomenon under study is ourselves (i.e., *when the brain studies itself*). In particular, when the object of *our* inquiry is *ourselves*, we need to recruit an additional cognitive process: introspection. Therefore, the paradox of the self-studying brain would stem precisely

* Corresponding author at: Department of Psychology, University of Bologna, Viale Carlo Berti Pichat 5, 40127 Bologna, Italy.
E-mail address: simone.battaglia@unibo.it (S. Battaglia).

https://doi.org/10.1016/j.plrev.2024.12.009
Received 13 December 2024; Accepted 15 December 2024
Available online 7 January 2025

from this peculiarity. In this treatment, we consider hypotheses of consciousness emerging from introspection and perception as distinct, thus leading to two different versions of the same hypothesis (the introspective version and the perceptive version). While the hypotheses are different variations of the same hypothesis, evidence for one would not constitute evidence for the other, since both have their own source. This paradoxical predicament emerges when the brain studies itself, which requires both perception and introspection.

The paradox at hand is the notion of self-study or self-measurement, which is paradoxical in the sense a yardstick cannot measure its own length - or the light that cannot illuminate itself [7]. The paradox of the self-studying brain arises because, unlike other objects of study, the brain is both the subject and the object of its own investigation. This unique self-referential loop requires not just perception but introspection, creating distinct forms of evidence that may not converge, as they stem from fundamentally different processes [8]. This dual role introduces a fundamental asymmetry in the types of evidence available, as introspection and perception produce hypotheses that cannot mutually reinforce each other.

This essay aims to address the intricacies of this paradox, offering perspectives on the brain's pursuit to understand itself [9]. To clarify the multi-layered nature of the self-studying brain, we can distinguish among four explanatory levels. At the epistemological level, we explore the knowledge limits inherent to self-referential inquiry, while the ontological level addresses the nature of consciousness as both an emergent and self-sustaining phenomenon. Methodologically, our approach integrates reductive techniques from neuroscience; for example, functional imaging and brain stimulation, to examine the neural correlates of consciousness, but also considers non-reductive, phenomenological perspectives to account for subjective experience. Lastly, empirical data can provide the foundation for testing hypotheses across these domains. This approach addresses both reductive and non-reductive views in neuro-philosophy, providing an inclusive basis for thinking about consciousness.

To appreciate the implicit paradox of self-understanding, it may help to consider the historical perspectives on the brain's capacity for self-study. Initial efforts to investigate the brain relied on external observations and anatomical dissections, yielding limited insights into its internal workings [10]. However, the field of neuroscience underwent a transformation with the emergence of advanced neuroimaging, enabling neuroscientists to examine the brain's dynamics, connectivity, and functional architecture [11], thereby establishing the groundwork for its ability to understand the world - and itself.

The paradox of self-study underwrites much of cognitive neuroscience. Cognitive neuroscientists employ a diverse array of tools to unravel the intricate relationship between brain activity and cognitive processes [9,12]. Functional neuroimaging techniques, such as functional MRI (fMRI) and positron emission tomography (PET), allow researchers to observe and measure proxies of neural activity during various cognitive tasks, including self-awareness tasks [13]. These studies provide invaluable insights into the functional segregation and integration involved in language processing, perception, memory formation, decision-making, and other cognitive functions, often relating them to the cognitive ability to self-monitor [14]. Moreover, the use of electrophysiological recordings, such as electroencephalography (EEG), magnetoencephalography (MEG), and intracranial recordings, offers higher temporal resolution, at the expense of the spatial resolution of neuroimaging techniques based on brain metabolism or blood flow. In addition, dynamical systems and information theory may offer more perspectives on how neural dynamics contribute to complex processes such as self-referential awareness [15,16]. Through this approach, the brain's self-referential nature can be understood not only by examining its structures but also by considering how it organizes experiences across space and time. Such a perspective could explain the mechanisms by which the brain generates self-awareness.

Furthermore, computational modelling has played a key role in simulating neural processes, building intuitions about how the brain generates self-referential thoughts and subjective experiences [17,18]. Finally, advances in non-invasive brain stimulation techniques, such as transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS), have further contributed to our understanding of the brain's self-study capabilities. These stimulation procedures allow researchers to transiently disrupt or enhance specific neural circuits, by modulating or disrupting neuronal dynamics and plasticity in brain systems associated with self-referential processing, offering potential insights into how these circuits contribute to consciousness. For instance, TMS targeting the temporoparietal junction has been linked with altered self-perception, suggesting a role in our awareness of self, as compared to others [19]. This approach opens avenues for empirically exploring how subjective experience depends on particular brain regions. In summary, by selectively modulating neural activity in specific brain regions, researchers can establish the causal relationships between these regions implicated in self-awareness and other cognitive processes [9].

2. How can an organ become sufficiently sentient to study itself?

The experience of being and existing is a profound aspect of human consciousness. It arises from an interplay of cognitive, perceptual, and introspective processes. In this context, a crucial process is self-reflection, which allows subjects to *introspect* and make inferences about one's own thoughts, emotions, and experiences. Additionally, the capacity for abstract thinking and symbolic representation enables humans to contemplate their existence beyond sensory constraints [10]. Furthermore, sociocultural factors shape our awareness of being - e.g., narrative self - through language, shared beliefs, and societal norms [20]. However, it is worth noting that the nature of consciousness and its potential existence in other beings, such as animals and even plants, is a topic of ongoing debate and investigation in both philosophy and neuroscience.

The notion that self-consciousness is a unique aspect of human consciousness could be nuanced, acknowledging the uncertainties surrounding consciousness beyond the human experience. Indeed, it could be argued that conscious processing can only be inferred by the Observer of something that looks as if it is conscious (please see below). Nonetheless, it is also important to acknowledge that consciousness may not be a well-defined property (e.g., vague; in the sense of how many grains of sand constitute a pile). Accordingly, consciousness entails a fundamentally subjective aspect that may not be reducible to objective or physical processes, and is hard to say

what provides evidence for consciousness [21,22]. This property arises from a self-model generated by the brain, a multifaceted collection of processes and representations that give rise to the experience of being oneself [23,24].

Throughout history, philosophers, scientists, artists, priests, and many others have pondered the essence of consciousness and the complex connection between the mind and body; both from a historical and philosophical standpoint [25–27]. René Descartes' theory of dualism, which posits that mental phenomena are, at least in certain respects, not physical phenomena, i.e., the mind and the body are distinct and separable, influenced our comprehension of the brain's ability to self-reflect. Descartes' dualism provided a framework for thinking about mind and body as separate, yet his approach lacked empirical tools to test this relationship. Modern neuroscience, by contrast, often seeks to bridge the dualism through integrated functional models of brain processes, but this effort continually encounters the boundary of subjective experience that Descartes highlighted. The historical evolution from dualism to neuroscience thus parallels our struggle with the self-studying brain, as it underscores the difficulties of reconciling introspective and perceptive accounts.

Furthermore, early philosophical insights into this paradox can be traced back to Schopenhauer, who recognized the complexities of the brain's self-referential nature. Schopenhauer's reflections laid a foundation for later philosophical discussions of the brain's role in understanding consciousness. In this regard, Georg Northoff's influential work, particularly his book on the '*brain problem*' [28], emphasizes the tension between the brain's objective mechanisms and its subjective experiences.

Nevertheless, contemporary science questions this dualistic viewpoint by appealing to the interdependence of brain function and human psychology [2,29]. Therefore, an individual may ask himself: what is the intention behind the brain's endeavour to study, decipher, and comprehend itself? What is the ultimate objective? No one [brain] can answer this question [yet]. However, the brain's self-study capabilities have profound implications for our understanding of consciousness, self-awareness, and human cognition [30]. The implicit paradox challenges us to explore the subjective nature of experience and the potential biases or ethical considerations introduced when the brain investigates itself [1,30]. Additionally, understanding self-understanding may hold potential for the diagnosis and treatment of neurological and psychiatric disorders [9].

The concept of a system studying itself inherently involves self-reference, which has been extensively examined in mathematical logic and philosophy due to its paradoxical nature. Self-reference is central to several classical paradoxes, including the Liar Paradox [31] and Russell's Paradox [32], and forms the foundation of Gödel's Incompleteness Theorems [33]. Within any sufficiently complex formal system, there will be statements that are true but unprovable within the system itself, pointing to inherent limitations in self-descriptive capacities. This perspective is further developed in the theory of undecidable propositions by Church and Turing [34], where certain questions about systems cannot be resolved from within those systems. A synthesis of these paradoxes is provided by Graham Priest through the "Inclosure Schema" framework [35], which formalizes the paradoxical aspects of self-reference. These principles can also be applied to the paradox of the self-studying brain, as the conditions of the paradox satisfy both Transcendence and Closure mentioned by Priest. Accordingly, in the act of studying, the brain attempts to include itself in the set of elements under investigation, but in doing so it should be both within and outside the set, leading to a paradoxical situation. Therefore, we can approach the question of reflexivity in cognitive systems: when a brain attempts to "study itself," it might encounter logical boundaries or paradoxes, hinting at fundamental constraints in our understanding of consciousness from a first-person perspective.

In this context, it is evident that the easy problem of consciousness is - and has been - addressed with increasing finesse using modern neuroscience techniques (e.g., neuroimaging) to gather evidence for the hypothesis that you - as my experimental subject - are a sentient artefact. If we treat this evidence gathering as a form of perception building [36], then the hypothesis "you are a sentient artefact" is entertained in *my* brain to explain some statistical regularities in *my* sensory inputs, as opposed to irregularities that would reflect prediction error.

This does not present a hard problem until I make a key move: does the hypothesis that you are sentient apply to me? Am I sentient? This would be a straightforward question to answer if I could gather evidence from myself. For example, I could have the hypothesis that you are 'tall' and confirm that hypothesis by measuring your height. It would be a simple matter to apply that hypothesis to me - i.e., am I tall? - and test it by measuring my height. However, I cannot measure my own brain in a transparent (or opaque) way. This follows from the fact that my internal dynamics and (belief updating) mechanics are, in virtue of their sustained existence, secluded behind a Markov blanket or holographic screen, which represent the boundaries that mediate interactions between the inside and outside of systems [37–39]. Breaching the boundary between my brain and the universe is precluded in an existential sense: e.g., I cannot perform psychosurgery on my motor cortex, because I would not be able to move my scalpel or gamma knife: I cannot hear the firing of cells in my auditory cortex. In short, consciousness is a useful hypothesis to explain self-evidencing 'things' like 'you' [40] but my very existence precludes gathering evidence for the hypothesis that "I am conscious". In principle, this idea goes against the grain of what some philosophers think: I can gather evidence for the hypothesis "*I am* conscious", but not for the hypothesis "*you are* conscious". In particular, these philosophers argue that whatever the behavioural and neurophysiological evidence I gather on you, the hypothesis "you *are not* conscious" will remain conceivable for me. An important question then arises: When philosophers and I talk about the hypothesis "you are (*or I am*) conscious", are we talking about the same hypothesis?

One possible answer would be to consider consciousness in reference to the *easy* problem of consciousness, whereas philosophers refer to consciousness in the sense of the *hard* problem of consciousness. The easy problem of consciousness is the problem of explaining how the brain implements the *functions* of consciousness. Although there is no consensus about what the functions of consciousness are, this problem is considered easy because it can be addressed using the scientific method (at least in principle). In this context, the hypothesis "you are conscious" would refer no more and no less to the fact that you are a physical system that implements the functions of consciousness (e.g., perception, metacognition, and planning). On the other hand, the hard problem of consciousness is the problem of explaining how the physical brain gives rise to subjective experience. From this point of view, the hypothesis "you are conscious" would refer to the fact that there is something it is like to be you (i.e., you have a *subjective* experience). In short, "my"

hypothesis of consciousness would not be the same as the one of philosophers.

However, if we follow the idea of *conceptual dualism* [41], these hypotheses would in fact be two different versions of *one and the same hypothesis*. In the framework described by Papineau, conscious states can be referred to from a phenomenological and physical point of view, based on the subjective experience and the physical or functional state of the brain underlying it, respectively. However, the main point provided by the author is *not* that consciousness is dual. On the contrary, the very purpose of conceptual dualism is to explain the *illusion* of distinctiveness. Indeed, the starting point is to consider that the subjective experience and its associated physical or functional state of the brain are in fact *one and the same thing* (in the strongest sense). Then, it is a matter of explaining - in purely physical terms - why we think this is not the case (i.e. why we think consciousness is non-physical). This explanation may differ depending on the version of conceptual dualism in question. Regardless, the basic idea is as follows. Perceptive and introspective processes would involve two distinct categories of concepts: phenomenal (or introspective) concepts and physical (or perceptive) concepts. The issue is that brain states–consciousness–could be "observed" through *both* perception and introspection (at least in some sense). As a consequence, one would have two concepts of consciousness rather than one, giving rise to the illusion of distinctiveness. Therefore, conceptual dualists do not claim that mind and matter are two substances or realities but acknowledge the existence of different ways to access only one reality, retaining its compatibility with physicalism.

The same principle could be applied to the hypothesis "you are (or I am) conscious". One could say that the neuroscientific version of this hypothesis belongs to the *physical* concepts category, whereas that of philosophers belongs to the *introspective* concepts category. Put another way, when we say that we cannot gather evidence for the hypothesis "*I am* conscious", we are talking about the physical version of this hypothesis. Conversely, when philosophers say that we cannot gather evidence for the hypothesis "*you are* conscious", they are talking about the introspective version of this hypothesis.

3. Mind over matter

The underlying idea here is that perception is the only way to gather evidence for the 'physical' hypothesis of consciousness, while introspection is the only way to gather evidence for the 'introspective' hypothesis of consciousness. Crucially, if these hypotheses are two versions of one and the same hypothesis, then this is paradoxical. Two 'identical' hypotheses should have the same implications or explananda. This presupposes that evidence for one of these hypotheses is also evidence for the other. The problem is that this cannot be the case here since introspection cannot provide evidence for the 'physical' hypothesis of consciousness and perception cannot provide evidence for the 'introspective' hypothesis. This particular situation could be the reason why some of us think that the hard problem of consciousness is not reducible to the easy problem of consciousness.

In particular, in an 'ideal' world, one would have a single version of the hypothesis "you are (or I am) conscious", whose evidence

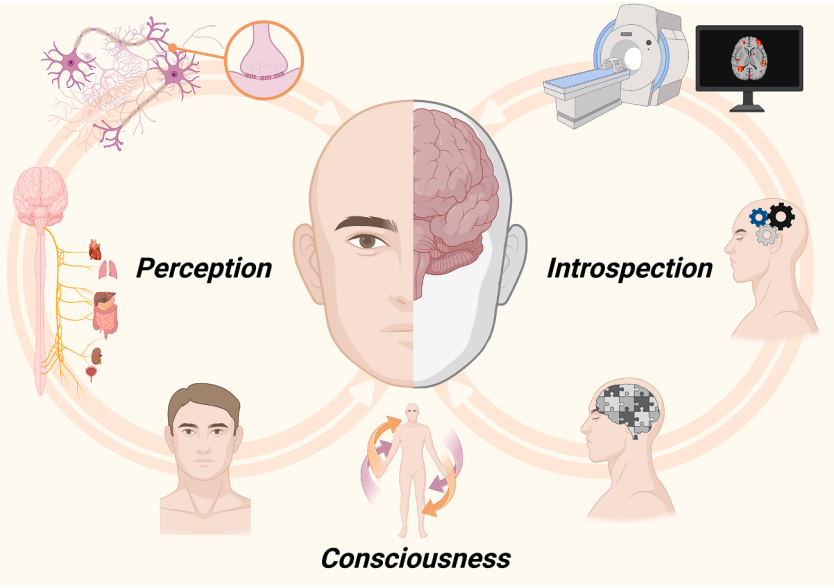

Fig. 1. Graphical representation of the dynamic interplay between introspection, perception, and consciousness that arises in the human brain. The figure depicts the dynamic interplay between introspection, perception, and consciousness. Introspection is depicted by a series of brains with gears and puzzle icons, symbolizing self-reflection and internal observation based on scientific and acquired knowledge. Perception is represented through sensory ability that integrates neural and bodily feedback, highlighting the role of external inputs to shaping awareness, which culminates in integrated and processed information. At the centre, consciousness is visualized as a unified orb, emerging from the integration of introspective and perceptual processes. Closed loop arrows and overlapping zones emphasize the perpetual bidirectional flow of information and the fusion of internal and external experiences that culminate in consciousness, reconciling introspective and perceptual evidence to address the paradox of the self-studying brain.

could be gathered through both perception and introspection. In this imaginary world, the hard and the easy problems of consciousness would clearly appear to us as being one and the same problem. This may be one take on the hard problem that leads to, arguably, more pressing questions engendered by the meta-problem of consciousness [42]; namely, what enables us to puzzle over our own phenomenology [43]? It should be noted that some of the ideas mentioned here share close links with the phenomenal concept strategy [41,44]. The purpose of the phenomenal concept strategy is to dissolve the hard problem of consciousness by solving the meta-problem of consciousness. This strategy is based on considerations regarding concepts such as, for example, conceptual dualism. The question regarding self-study has been present in our mind for centuries, but the lack of a possibility of providing adequate and definitive answers, obfuscated by the absence of refined theorizations and of precise protocols and instrumentations, led to an inability to pose this question in the first place. While the conclusive solution may still be temporally out of reach, now more than ever neuroscientific research can pave the way for closing this distance.

The paradox here lies in the apparent incommensurability between perceptive evidence (i.e., derived from external observations) and introspective evidence (i.e., from internal, self-referential awareness) when it comes to validating consciousness as a unified concept (see Fig. 1).

This problem, well-known in the philosophy of mind [45], highlights how our structure as observers may inherently divide these two forms of evidence, creating a unique epistemic challenge. For instance, neurophenomenology, pioneered by Francisco Varela [46], provides methodological tools to connect subjective, first-person introspection with objective, third-person neuroscience [47]. By aligning experiential reports with neurophysiological data, this approach directly addresses the introspection-perception divide central to our manuscript. Neurophenomenology posits that phenomenological insights about lived experience should guide the interpretation of neuroscientific data, while neuroscientific findings can inform and refine phenomenological descriptions. Perceptive signals could, in theory, inform us about behavioural markers or neural correlates of consciousness, yet these observations cannot fully account for the subjective, qualitative aspect captured through introspection. The need to reconcile this dichotomy continues to provoke debate about whether consciousness can be empirically assessed as a single construct, or if it instead requires a pluralistic approach, one that accommodates both the external and introspective perspectives without necessarily merging them. This raises an important question for cognitive science: to what extent does our inability to harmonize perceptive and introspective evidence reveal a fundamental limit of our observational capacities as human beings? Addressing this may require reconsidering the roles that these types of evidence play within the broader framework of consciousness research.

The balance between introspective and perceptive evidence may have broader implications for consciousness studies. If consciousness inherently divides along these lines, it may suggest a fundamental epistemological limit, where introspective accounts provide essential but non-verifiable insights. Reconciling these domains may therefore require an interdisciplinary framework that respects both evidence types without forcing them into artificial coherence. It should be noted that the very fact of understanding *why* such epistemological limit arises (which is the topic of the current paper) may be crucial to the successful completion of this project.

4. Modelling the *paradox*

Beyond the introspective and perceptual dimensions of consciousness, our exploration of the paradox of the brain studying itself resonates with broader themes in cognitive neuroscience, philosophy, and theoretical physics. Self-referentiality, a core concept within this paradox, finds analogous challenges across domains that involve systems observing or interacting with themselves. In physics, for example, quantum mechanics has long wrestled with self-referential issues such as observer effects, wherein the measurement process itself influences the phenomenon under study [48,49]. In systems theory, self-referential loops can lead to emergent behaviours that challenge classical causal models. These analogies suggest that the paradox we encounter in studying self-awareness may not be unique to neuroscience but instead reflect a fundamental epistemological limitation when systems attempt to observe themselves.

Drawing from these cross-disciplinary insights, we can propose novel ways of examining the self-referential challenges in consciousness studies. Specifically, advances in computational modelling and simulation offer an exciting avenue. By creating recursive, self-interacting models, researchers can simulate conditions under which a system 'observes' itself, providing a controlled environment to study how feedback loops and self-referential dynamics influence behaviour. For example, models using neural networks or agent-based systems could simulate a simplified version of self-awareness [50], revealing how different layers of 'self-observation' might impact cognitive states or even mimic self-aware dynamics. Such approaches are valuable for testing hypotheses that are challenging to verify empirically in biological systems due to the inherent complexity of the brain. Ultimately, this may allow us to simulate an agent *facing* the paradox of the self-studying brain.

Furthermore, the principles of predictive coding, including the free energy principle [29,51], are particularly relevant here. In this context, the free energy principle suggests that the brain minimizes prediction error to maintain homeostasis and adapt to its environment. Applied to self-referential processes, this framework could illuminate how the brain manages the recursive layers of perception and introspection without collapsing into an unstable feedback loop: see [52,53] for a phenomenological account. Predictive coding formulations might offer insights into how self-awareness is maintained, effectively balancing the brain's internal representations with the need for external coherence [54]. By understanding these processes computationally, we may uncover mechanisms by which the brain avoids cognitive overload or instability, shedding light on the maintenance of mental coherence across recursive levels of introspection.

Integrating these computational approaches with experimental neuroscience could provide a fertile ground for future research. For instance, simulation-based models could allow us to systematically vary the 'depth' of self-referential processing, testing whether deeper layers of introspection lead to different stability patterns or cognitive states. Such research might even help explain phenomena seen in pathological conditions like Cotard's syndrome [55], where self-referential processing appears disrupted. Overall, by

leveraging simulations and computational models, we can begin to bridge the theoretical gap posed by the self-study paradox, approaching consciousness not only as a subject of empirical neuroscience but also as a philosophical and computational problem. This interdisciplinary perspective could ultimately enrich both our understanding of consciousness and our methodological toolkit, leveraging the paradox as a driving force for ongoing neuroscientific and philosophical inquiries.

5. Conclusion

The *paradox of the self-studying* brain presents a captivating challenge and highlights the nature of the (hypothesis testing) human brain itself. This paradox of self-referential evidence invites further exploration into whether the specific nature of introspective awareness - our ability to internally "observe" our own conscious state - might underlie the difficulty in reconciling evidence for consciousness. If humans relied solely on perceptive cues - such as observing behaviours or physiological responses that imply awareness - it is conceivable that the paradox might dissolve, as subjective self-awareness would not play a role. The unique challenge posed by introspective awareness lies in its metacognitive dimension: we not only experience consciousness but can also reflect upon that experience, layering self-reference into our assessment of our own consciousness. Conditions such as Cotard's syndrome [55], in which patients deny their own existence or consciousness, suggest that impairments in this metacognitive introspection may disrupt the sense of subjective experience. Further, investigating whether this paradox could be experimentally manipulated - for instance, through transcranial magnetic stimulation (TMS) or pharmacological intervention - may provide insights into how introspective and perceptive evidence interact and whether they contribute differently to our understanding of consciousness. This line of inquiry may help delineate whether the paradox is a consequence of introspective self-reference or if it stems from a more foundational property of conscious systems.

The complexity of the problem at hand should not be seen as an obstacle, but rather as a motivation for scientific enquiry. The progression of future research – trying to unravel this paradox by further clarifying the underpinnings of consciousness and self-awareness – may expand our understanding of the brain and its intricate relationship with its self-studying drive.

CRediT authorship contribution statement

Simone Battaglia: Conceptualization, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Philippe Servajean:** Writing – review & editing. **Karl J. Friston:** Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Simone Battaglia reports financial support was provided by Ministry of University and Research. Karl J. Friston reports financial support was provided by Wellcome Centre for Human Neuroimaging. Karl J. Friston reports financial support was provided by Horizon Europe. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding and acknowledgements

SB is supported by #NEXTGENERATIONEU (NGEU) and funded by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), project MNESYS (PE0000006) - A Multiscale integrated approach to the study of the nervous system in health and disease (DN. 1553 11.10.2022) and Bial Foundation, Portugal (235/22). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them. KF is supported by funding for the Wellcome Centre for Human Neuroimaging (Ref: 205103/Z/16/Z), a Canada-UK Artificial Intelligence Initiative (Ref: ES/T01279X/1) and the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No 945539 (Human Brain Project SGA3). The authors would like to express their gratitude to Claudio Nazzi and Cristina Roperti for their invaluable help and support throughout this project from its initial conceptualization.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.plrev.2024.12.009.

References

[1] Frith CD, Frith U. The neural basis of mentalizing. Neuron 2006;50:531–4. https://doi.org/10.1016/j.neuron.2006.05.001.
[2] Baars BJ, Gage NM. Cognition, brain, and consciousness: introduction to cognitive neuroscience. 2nd ed. Academic Press; 2010.

[3] Crick F, Koch C. A framework for consciousness. Nat Neurosci 2003;6:119–26. https://doi.org/10.1038/nn0203-119.
[4] Fields C, Glazebrook JF, Levin M. Principled limitations on self-representation for generic physical systems. Entropy 2024;26:194. https://doi.org/10.3390/e26030194.
[5] Markose SM. Complexification of eukaryote phenotype: adaptive immuno-cognitive systems as unique Gödelian blockchain distributed ledger. Biosystems 2022;220:104718. https://doi.org/10.1016/j.biosystems.2022.104718.
[6] Markose S. The gödelian foundations of self-reference,the liar and incompleteness: arms racein complex strategic innovation. In: Pinto AA, Accinelli Gamba E, Yannacopoulos AN, Hervés-Beloso C, editors. Trends math. econ.. editors. Cham: Springer International Publishing; 2016. p. 217–44. https://doi.org/10.1007/978-3-319-32543-9_11.
[7] Watts A. The book: on the taboo against knowing who you are. Westminster: Knopf Doubleday Publishing Group; 2011.
[8] Butler J. Rethinking introspection: a pluralist approach to the first-person perspective. Basingstoke: Palgrave Macmillan; 2013.
[9] Dehaene S, Changeux J-P. Experimental and theoretical approaches to conscious processing. Neuron 2011;70:200–27. https://doi.org/10.1016/j.neuron.2011.03.018.
[10] Sasaki Y, Rajimehr R, Kim BW, Ekstrom LB, Vanduffel W, Tootell RBH. The radial bias: a different slant on visual orientation sensitivity in human and nonhuman primates. Neuron 2006;51:661–70. https://doi.org/10.1016/j.neuron.2006.07.021.
[11] Finn ES, Poldrack RA, Shine JM. Functional neuroimaging as a catalyst for integrated neuroscience. Nature 2023;623:263–73. https://doi.org/10.1038/s41586-023-06670-9.
[12] Yeo BTT, Krienen FM, Sepulcre J. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J Neurophysiol 2011;106:1125–65. https://doi.org/10.1152/jn.00338.2011.
[13] Tacikowski P, Berger CC, Ehrsson HH. Dissociating the neural basis of conceptual self-awareness from perceptual awareness and unaware self-processing. Cereb Cortex 2017;27:3768–81. https://doi.org/10.1093/cercor/bhx004.
[14] Legrand D, Ruby P. What is self-specific? Theoretical investigation and critical review of neuroimaging results. Psychol Rev 2009;116:252–82. https://doi.org/10.1037/a0014172.
[15] Northoff G, Wainio-Theberge S, Evers K. Spatiotemporal neuroscience – what is it and why we need it. Phys Life Rev 2020;33:78–87. https://doi.org/10.1016/j.plrev.2020.06.005.
[16] Northoff G. From brain dynamics to the mind: spatiotemporal neuroscience. 1st ed. San Diego: Elsevier Science & Technology; 2024.
[17] Hitchcock PF, Britton WB, Mehta KP, Frank MJ. Self-judgment dissected: a computational modeling analysis of self-referential processing and its relationship to trait mindfulness facets and depression symptoms. Cogn Affect Behav Neurosci 2023;23:171–89. https://doi.org/10.3758/s13415-022-01033-9.
[18] Möller TJ, Georgie YK, Schillaci G, Voss M, Hafner VV, Kaltwasser L. Computational models of the "active self" and its disturbances in schizophrenia. Conscious Cogn 2021;93:103155. https://doi.org/10.1016/j.concog.2021.103155.
[19] Donaldson PH, Rinehart NJ, Enticott PG. Noninvasive stimulation of the temporoparietal junction: a systematic review. Neurosci Biobehav Rev 2015;55:547–72. https://doi.org/10.1016/j.neubiorev.2015.05.017.
[20] Heyes CM. Cognitive gadgets : the cultural evolution of thinking. Belknap Press: An Imprint of Harvard University Press; 2018. Illustrated edition.
[21] Nagel T. What is it like to be a bat? Philos Rev 1974;83:435. https://doi.org/10.2307/2183914.
[22] Metzinger T. Neural correlates of consciousness: empirical and conceptual questions. editor. The MIT Press; 2000. https://doi.org/10.7551/mitpress/4928.001.0001.
[23] Metzinger T. Being no one: the self-model theory of subjectivity. Cambridge, Mass: MIT Press; 2003.
[24] Damasio A. Self comes to mind: constructing the conscious brain. New York, NY, US: Pantheon/Random House; 2010.
[25] Clark A. A case where access implies qualia? Analysis 2000;60:30–7.
[26] Dennet D. Consciousness explained. London, UK: Allen Lane/The Penguin Press; 1991.
[27] Frith C. What is consciousness for? Pragmat Cogn 2010;18:497–551. https://doi.org/10.1075/pc.18.3.03fri.
[28] Northoff G. Philosophy of the brain: the brain problem, 52. Amsterdam: John Benjamins Publishing Company; 2004. https://doi.org/10.1075/aicr.52.
[29] Friston KJ. The free-energy principle: a unified brain theory? Nat Rev Neurosci 2010;11:127–38. https://doi.org/10.1038/nrn2787.
[30] Friston K. Computational psychiatry: from synapses to sentience. Mol Psychiatry 2023;28:256–68. https://doi.org/10.1038/s41380-022-01743-z.
[31] Barwise J, Etchemendy J. The liar: an essay on truth and circularity. New York: Oxford University Press; 1987.
[32] Russell B. Principles of mathematics. London: Routledge; 2010.
[33] Gödel K. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. Monatshefte Für Math Phys 1931;38–38:173–98. https://doi.org/10.1007/BF01700692.
[34] Copeland BJ, Shagrir O. The Church-Turing thesis: logical limit or breachable barrier? Commun ACM 2018;62:66–74. https://doi.org/10.1145/3198448.
[35] Priest G. Beyond the limits of thought. 1st ed. Oxford, UK: Oxford University Press; 2002. https://doi.org/10.1093/acprof:oso/9780199254057.001.0001.
[36] Gregory RL. Perceptions as hypotheses. Philos Trans RSoc Lond BBiol Sci 1980;290:181–97. https://doi.org/10.1098/rstb.1980.0090.
[37] Radulovic J, Ren LY, Gao C. N-Methyl D-aspartate receptor subunit signaling in fear extinction. Psychopharmacol Berl 2019;236:239–50. https://doi.org/10.1007/s00213-018-5022-5.
[38] Friston KJ, Wiese W, Hobson JA. Sentience and the origins of consciousness: from cartesian duality to markovian monism. Entropy Basel Switz 2020;22:516. https://doi.org/10.3390/e22050516.
[39] Fields C, Glazebrook JF, Levin M. Minimal physicalism as a scale-free substrate for cognition and consciousness. Neurosci conscious, 013; 2021. https://doi.org/10.1093/nc/niab013.
[40] Hohwy J. The self-evidencing brain. Nous 2016;50:259–85. https://doi.org/10.1111/nous.12062.
[41] Papineau D. Thinking about consciousness. 1st ed. Oxford, UK: Oxford University Press; 2002. https://doi.org/10.1093/0199243824.001.0001.
[42] Chalmers DJ. The meta-problem of consciousness. J Conscious Stud 2018;25:6–61.
[43] Clark A, Friston K, Wilkinson S. Bayesing Qualia: consciousness as inference, not raw datum. J Conscious Stud 2019;26:19–33.
[44] Stoljar D. Physicalism and phenomenal concepts. Mind Lang 2005;20:469–94. https://doi.org/10.1111/j.0268-1064.2005.00296.x.
[45] Velmans M. Consciousness, brain and the physical world. Philos Psychol 1990;3:77–99. https://doi.org/10.1080/09515089008572990.
[46] Varela FJ. Neurophenomenology: a methodological remedy for the hard problem. J Conscious Stud 1996;3:330–49.
[47] Berkovich-Ohana A, Dor-Ziderman Y, Trautwein F-M, Schweitzer Y, Nave O, Fulder S, et al. The Hitchhiker's guide to neurophenomenology – the case of studying self boundaries with meditators. Front Psychol 2020;11. https://doi.org/10.3389/fpsyg.2020.01680.
[48] Polychronakos AP. Quantum mechanical rules for observed observers and the consistency of quantum theory. Nat Commun 2024;15:3023. https://doi.org/10.1038/s41467-024-47170-2.
[49] Buks E, Schuster R, Heiblum M, Mahalu D, Umansky V. Dephasing in electron interference by a 'which-path' detector. Nature 1998;391:871–4. https://doi.org/10.1038/36057.
[50] Du BZ, Guo Q, Zhao Y, Zhi T, Chen Y, Xu Z. Self-aware neural network systems: a survey and new perspective. Proc IEEE 2020;108:1047–67. https://doi.org/10.1109/JPROC.2020.2977722.
[51] Friston K. The free-energy principle: a rough guide to the brain? Trends Cogn Sci 2009;13:293–301. https://doi.org/10.1016/j.tics.2009.04.005.
[52] Sandved-Smith L, Hesp C, Mattout J, Friston K, Lutz A, Ramstead MJD. Towards a computational phenomenology of mental action: modelling meta-awareness and attentional control with deep parametric active inference. Neurosci Conscious 2021. https://doi.org/10.1093/nc/niab018. 2021:niab018.

[53] Parvizi-Wayne D, Sandved-Smith L, Pitliya RJ, Limanowski J, Tufft MRA, Friston KJ. Forgetting ourselves in flow: an active inference account of flow states and how we experience ourselves within them. Front Psychol 2024;15:1354719. https://doi.org/10.3389/fpsyg.2024.1354719.
[54] Apps MAJ, Tsakiris M. The free-energy self: a predictive coding account of self-recognition. Neurosci Biobehav Rev 2014;0:85–97. https://doi.org/10.1016/j.neubiorev.2013.01.029.
[55] Tomasetti C, Valchera A, Fornaro M, Vellante F, Orsolini L, Carano A, et al. The 'dead man walking' disorder: an update on Cotard's syndrome. Int Rev Psychiatry 2020;32:500–9. https://doi.org/10.1080/09540261.2020.1769881.