

Insights in consciousness research, volume II

Edited by

Xerxes D. Arsiwalla, Antonino Raffone and
Luca Simione

Coordinated by

Monia D'Angiò

Published in

Frontiers in Psychology



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-6379-3
DOI 10.3389/978-2-8325-6379-3

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Insights in consciousness research, volume II

Topic editors

Xerxes D. Arsiwalla — Wolfram Research, Inc., United States

Antonino Raffone — Sapienza University of Rome, Italy

Luca Simione — UNINT - Università degli studi Internazionali di Roma, Italy

Topic coordinator

Monia D'Angiò — Sapienza University of Rome, Italy

Citation

Arsiwalla, X. D., Raffone, A., Simione, L., D'Angiò, M., eds. (2025). *Insights in consciousness research, volume II*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-6379-3

Table of contents

| | |
|-----|--|
| 04 | Editorial: Insights in consciousness research, volume II Monia D'Angiò, Luca Simione, Xerxes D. Arsiwalla and Antonino Raffone |
| 07 | The necessary and sufficient mechanism of consciousness in a layered mind Zenán Ruan |
| 12 | Minimal self-consciousness and the flying man argument Shaun Gallagher |
| 21 | Quantifying empirical support for theories of consciousness: a tentative methodological framework Asger Kirkeby-Hinrup |
| 37 | Entangled brains and the experience of pains Valerie Gray Hardcastle |
| 45 | The maps of meaning consciousness theory Scott Andersen |
| 51 | Artificial intelligence, human cognition, and conscious supremacy Ken Mogi |
| 59 | Cognitive styles and psi: psi researchers are more similar to skeptics than to lay believers Marieta Pehlivanova, Marina Weiler and Bruce Greyson |
| 71 | How-tests for consciousness and direct neurophenomenal structuralism Sascha Benjamin Fink |
| 86 | Interfacing consciousness Robert Prentner and Donald D. Hoffman |
| 91 | Crisis of objectivity: using a personalized network model to understand maladaptive sensemaking in a patient with psychotic, affective, and obsessive-compulsive symptoms Aleš Oblak, Matic Kuclar, Katja Horvat Golob, Alina Holnthaner, Urška Battelino, Borut Škodlar and Jurij Bon |
| 106 | Reflexivity gradient—Consciousness knowing itself Zoran Josipovic |
| 116 | Quantum-like Qualia hypothesis: from quantum cognition to quantum perception Naotsugu Tsuchiya, Peter Bruza, Makiko Yamada, Hayato Saigo and Emmanuel M. Pothos |



OPEN ACCESS

EDITED AND REVIEWED BY
Christopher Gutland,
Zhejiang University, China

*CORRESPONDENCE

Monia D'Angiò
✉ monia.dangio@uniroma1.it

RECEIVED 30 March 2025

ACCEPTED 21 April 2025

PUBLISHED 07 May 2025

CITATION

D'Angiò M, Simone L, Arsiwalla XD and
Raffone A (2025) Editorial: Insights in
consciousness research, volume II.
Front. Psychol. 16:1602845.
doi: 10.3389/fpsyg.2025.1602845

COPYRIGHT

© 2025 D'Angiò, Simone, Arsiwalla and
Raffone. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Editorial: Insights in consciousness research, volume II

Monia D'Angiò^{1*}, Luca Simone^{2,3}, Xerxes D. Arsiwalla⁴ and
Antonino Raffone¹

¹Department of Psychology, Sapienza University of Rome, Rome, Italy, ²Institute of Cognitive Sciences and Technologies, Consiglio Nazionale delle Ricerche, Rome, Italy, ³Faculty of Interpreting and Translation, UNINT, Università degli Studi Internazionali, Rome, Italy, ⁴Wolfram Institute for Computational Foundations of Science, Champaign, IL, United States

KEYWORDS

consciousness, non-dual awareness, phenomenology, minimal self, artificial intelligence, pain

Editorial on the Research Topic

Insights in consciousness research, volume II

Advancing upon the scientific program of the inaugural Research Topics in this series on insights and rising stars in consciousness research (Arsiwalla et al., 2023; Srinivasan et al., 2023), this second edition seeks to explore classic debates in consciousness science, such as distinguishing between the most promising contemporary theories of consciousness, while also offering fresh perspectives and new insights into the progress of this field, including current reflections on its connection to artificial intelligence.

One of the most debated issues in consciousness research concerns its neural correlates (NCCs). Although researchers often aim to distinguish proper NCCs from their prerequisites and consequences (Aru et al., 2012; Seth and Bayne, 2022), new approaches are being developed in the field. Fink proposed a framework based on direct neurophenomenal structuralism, which directly relates neural structures to the structures of phenomenal experience without postulating intermediate levels of explanation. To achieve this, the author introduced a classification of four “sufficiency tests” designed to determine which systems are conscious (Which-test), when they are conscious (When-test), their conscious content (What-test), and how they are phenomenally experienced (How-test). According to the author, the How-test is best approached through direct neurophenomenal structuralism. These methodologies should guide experimental investigations of consciousness and the formulation of hypotheses regarding NCCs.

In the same vein, Josipovic argues that conscious awareness does not require the mediation of mental representations. As such, a dedicated network, distinct from the neural correlates of cognitive processing, should account for the dynamics of consciousness. In his theory of the reflexivity gradient of consciousness, Josipovic highlights that consciousness research predominantly investigates its phenomenal aspects, such as content, arousal level, and cognitive processing, often neglecting consciousness itself. This non-dual awareness, with its inherent, non-representational reflexivity, is characterized by an implicit-explicit gradient of experience that is independent of both the content of experience and the state of experiencing.

Pehlivanova et al. conducted an original study to test whether different cognitive styles [i.e., actively open-minded thinking (AOT) and need for closure (NFC)] influence how psychic researchers, compared to academics from other disciplines and lay believers, evaluate data on such phenomenological experiences. Results showed that psychic academics exhibit a level of AOT similar to that of other academics, with both groups differing from lay believers. This demonstrates that psychic researchers possess strong critical thinking skills and are not biased in their engagement with research on psychic phenomena.

Qualitative phenomenology is a cross-disciplinary methodology that has been applied in various fields of study. An interesting application can be found in clinical psychiatric research. Oblak et al. conducted a single-case study of a patient with psychiatric comorbidities, collecting data over 2 years to construct a personalized network model (PNM) explaining psychiatric disorders within a phenomenology-informed framework. By incorporating various measures, including phenomenological, neuropsychological, and language assessments, the resulting PNM identified a core maladaptive pattern of sensemaking and disorders of self described as “the crisis of objectivity.” These data demonstrate that PNM can be effectively incorporated into qualitative phenomenological methods applied to clinical psychiatric research.

Another aspect related to qualia (phenomenal experience) is the minimal self, a first-person, pre-reflective self-awareness. Gallagher proposed that the minimal self is linked to both the sense of ownership and the sense of agency, which pertains not only to bodily actions but also extends to cognitive processes such as thinking and imagining—implying that we are the agents of our own cognition. Similarly, the sense of ownership is not limited to bodily ownership alone. However, in everyday life, directly perceiving minimal experience can be challenging. Only specific phenomenal practices, such as meditation, sensory deprivation, or experimental conditions, can provide insights into the experience of the minimal self.

A framework to investigate the qualitative aspects of consciousness was established by Tsuchiya et al., utilizing quantum theory to formulate the Quantum-like Qualia (QQ) hypothesis. Traditionally, qualia are treated as fixed points in a dimensional space, assuming they can be measured without alteration. However, empirical evidence suggests that internal attention can modify qualia during measurement. In this model, qualia, encompassing all possible aspects of experience, are referred to as “observables,” while sensory inputs and internal conditions (e.g., attention) are considered “states” that influence “measurement outcomes,” resulting from their interaction. The predictions of the QQ hypothesis align with experimental findings, offering new perspectives on the relationship between consciousness and attention.

According to Andersen, some aspects, such as evolutionary biology, Occam's Razor, and Hume's Dilemma, are often overlooked or inadequately addressed in existing models of consciousness. In an attempt to incorporate these aspects, the author proposed the Maps of Meaning theory of consciousness, which is grounded in a first-principles approach to defining consciousness and integrate psychology, neuroscience, religion,

and philosophy. In this theory, consciousness is conceptualized as the inevitable byproduct of having multiple goals and the continuous process of evaluating and prioritizing these goals to guide action in the world.

Instead of introducing new theories of consciousness, some authors have focused on models for evaluating and distinguishing existing theories (Kirkeby-Hinrup) or integrating them (Ruan).

Kirkeby-Hinrup proposed a methodological framework to better explain and quantify the evidence supporting theories of consciousness. Two approaches are currently used in the literature: (1) collaboration between proponents of different theories to develop paradigms that test their respective predictions (ARC; e.g., Consortium et al., 2023; Melloni et al., 2023); and (2) the establishment of a set of criteria to assess the scope and explanatory power of each theory regarding conscious phenomena, largely independent of empirical data (CRIT; Doerig et al., 2021). Building on these two approaches, the author introduced the “quantification to the best explanation” (QBE) method, based on Bayesian confirmation theory, to complement and address the shortcomings of the existing approaches.

Ruan proposed an integrative approach aimed at unifying existing theories of consciousness. In this process, two key aspects must be considered: first, ensuring that the theories being examined genuinely address consciousness itself by properly defining different global states of consciousness; second, critically evaluating the methods and strategies used to study consciousness. Instead of merely attempting to unify theories of consciousness (ToCs), the author proposed a layered architecture of the mind as a potential way to reconcile even competing theories. In this model, multiple signals are processed simultaneously, involving several brain regions and mechanisms. The formation of multiple, temporary zones of consciousness, which can be arbitrarily bounded, results in experiences with specific and distinct attributes.

Due to advancements in artificial intelligence (AI) technology, another debated issue in consciousness science is whether AI could exhibit conscious properties. Prentner and Hoffman offered insights on the potential inclusion of AI within a framework of consciousness. Their approach is based on the conscious agent theory (CAT; Hoffman and Prakash, 2014), which relies on rigorous mathematical assumptions and emphasizes the fundamental role of agency in selecting a particular experience from a set of possible experiences, making it probabilistically measurable. In this view, experience itself constitutes the first-person aspect of consciousness, while its consequences are what can be observed and measured. Alongside CAT, the interface theory of perception (ITP; Hoffman et al., 2015) conceptualizes perception as a kind of interface with the world, enabling an agent to interact with reality. Within this framework, consciousness is understood as a network of conscious agents that represent themselves through interfaces, forming a self-reflective, non-dual awareness.

Building on reflections about AI, Mogi explores the potential computational role of consciousness as an alternative approach to studying consciousness beyond phenomenology. While several cognitive functions, such as attention regulation, adaptation to new contexts, and embodied cognition, may be uniquely associated with conscious processing, it remains unclear which computations are specifically tied to consciousness.

Moreover, the study introduces the concept of “conscious supremacy”—inspired by quantum supremacy—to distinguish computations that require consciousness from those that can be performed unconsciously.

Like consciousness, pain is a complex state involving a range of qualia and psychological (cognitive) processes. Gray Hardcastle offers an insightful perspective on studying pain, suggesting that pain, rather than being localized to a single brain region, emerges from a widespread activation pattern that partially overlaps with other sensory and cognitive processes. From a connectivity-based perspective, multiple brain areas contribute to various functions rather than operating in isolation. Given the heterogeneity of neuronal responses, also the experience of pain—like consciousness—might be dynamic and adaptive, shaped by shifting patterns of brain activity over time, rather than being reducible to fixed neural mechanisms.

The articles included in this Research Topic provide a perspective on the multifaceted nature of consciousness research, drawing on scientists from various cognitive science disciplines. In addition to existing theories, many new conceptualizations have been proposed in light of recent advancements and empirical evidence in the field (Andersen; Fink; Gallagher; Josipovic; Tsuchiya et al.), some of which incorporate conceptualizations of AI (Mogi; Prentner and Hoffman). Several methodological proposals have been developed to assess existing theories (Kirkeby-Hinrup; Ruan). Importantly, the ongoing debate on consciousness also has significant implications for clinical research and practice (Gray Hardcastle; Oblak et al.; Pehlivanova et al.). Insights from

consciousness research encompass diverse themes and approaches, offering a complex perspective on the fascinating and intriguing phenomenon of consciousness and its many facets.

Author contributions

MD: Writing – original draft. LS: Writing – review & editing. XA: Writing – review & editing. AR: Writing – review & editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Arsiwalla, X. D., Srinivasan, N., Simone, L., Kleiner, J., and Raffone, A. (2023). Editorial: rising stars in: consciousness research 2021. *Front. Psychol.* 14:1205982. doi: 10.3389/fpsyg.2023.1205982
- Aru, J., Bachmann, T., Singer, W., and Melloni, L. (2012). Distilling the neural correlates of consciousness. *Neurosci. Biobehav. Rev.* 36, 737–746. doi: 10.1016/j.neubiorev.2011.12.003
- Consortium, C., Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., et al. (2023). An adversarial collaboration to critically evaluate theories of consciousness. *BioRxiv* 2023–06. doi: 10.1101/2023.06.23.546249
- Doerig, A., Schurger, A., and Herzog, M. H. (2021). Hard criteria for empirical theories of consciousness. *Cogn. Neurosci.* 12, 41–62. doi: 10.1080/17588928.2020.1772214
- Hoffman, D. D., and Prakash, C. (2014). Objects of consciousness. *Front. Psychol.* 5:577. doi: 10.3389/fpsyg.2014.00577
- Hoffman, D. D., Singh, M., and Prakash, C. (2015). The interface theory of perception. *Psychon. Bull. Rev.* 22, 1480–1506. doi: 10.3758/s13423-015-0890-8
- Melloni, L., Mudrik, L., Pitts, M., Bendtz, K., Ferrante, O., Gorska, U., et al. (2023). An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory. *PLoS ONE* 18:e0268577. doi: 10.1371/journal.pone.0268577
- Seth, A. K., and Bayne, T. (2022). Theories of consciousness. *Nat. Rev. Neurosci.* 23, 439–452. doi: 10.1038/s41583-022-00587-4
- Srinivasan, N., Simone, L., Arsiwalla, X. D., Kleiner, J., and Raffone, A. (2023). Insights in consciousness research 2021. *Front. Psychol.* 14:1182690. doi: 10.3389/fpsyg.2023.1182690



OPEN ACCESS

EDITED BY

Luca Simone,
Università degli studi Internazionali di Roma
(UNINT), Italy

REVIEWED BY

Alexander Fingelkurts,
BM-Science, Finland
Stuart Hameroff,
University of Arizona, United States

*CORRESPONDENCE

Zenan Ruan
✉ znuan@zju.edu.cn

RECEIVED 21 August 2023

ACCEPTED 14 September 2023

PUBLISHED 28 September 2023

CITATION

Ruan Z (2023) The necessary and sufficient
mechanism of consciousness in a layered mind.
Front. Psychol. 14:1280959.
doi: 10.3389/fpsyg.2023.1280959

COPYRIGHT

© 2023 Ruan. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other forums is
permitted, provided the original author(s) and
the copyright owner(s) are credited and that
the original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

The necessary and sufficient mechanism of consciousness in a layered mind

Zenan Ruan^{1,2*}

¹Center for the Study of Language and Cognition, School of Philosophy, Zhejiang University, Hangzhou, China, ²Department of Automation, School of Mechanical Engineering and Automation, Zhejiang SCI-TECH University, Hangzhou, China

KEYWORDS

consciousness, Gazzaniga, theories of consciousness, IIT, GNWT

Introduction

The study of consciousness is becoming one of several significant challenges at the frontiers of science, in contrast to its previously being off-limits. With the application of binocular rivalry, split brain, blindsight, and other paradigms by passionate pioneers in the last century (Seth, 2018), empirical theories of consciousness have emerged in neuroscience. Currently, the situation has reached a critical point of both hope and challenge in that a large number of theories of consciousness (ToCs), each with specific empirical support, have claimed their respective plausibilities, and their proposed conjectures have led to diverging predictions (Del Pin et al., 2021; Signorelli et al., 2021; Seth and Bayne, 2022; Yaron et al., 2022). Various theories have been discussed, and it appears that this issue is becoming more prevalent. Currently, the lack of collaboration between different groups and fields hinders the advancement of theories of consciousness. However, a fundamental theory which is not limited by the boundaries of individual theories is expected to emerge in the future (Koch, 2018).

In this process, four major kinds of ToCs have garnered the most attention (Seth and Bayne, 2022): Integrated Information Theory (IIT) (Tononi, 2008; Oizumi et al., 2014; Tononi et al., 2016), Global Neural Workspace Theory (GNWT) (Dehaene, 2014; Mashour et al., 2020), Higher-Order Theory (HOT) (Lau and Rosenthal, 2011; Brown et al., 2019), and Recurrent Processing Theory (RPT) (Lamme, 2018) and Predictive Processing Theory (PP) (Seth and Hohwy, 2021).

Briefly, IIT identifies any conscious experience with the maximally irreducible cause-effect structure of the system in the corresponding state; GNWT proposes that the global workspace, triggered by widespread neural ignition and the sharing of information across several cognitive modules, is the key to conscious access; HOT is based on the higher-order structure of conscious experience in which “I” am aware of “something” (the representation of “something” is first-order). At the same time, RPT and PP emphasize the importance of top-down processing in conscious mental activity.

Rather than attributing consciousness to neural activities, a fifth approach has identified consciousness with underlying physical processes across multiple spatiotemporal scales. As a typical and noted paradigm, Orchestrated Objective Reduction (Orch OR, cf. Hameroff and Penrose, 2014) theory claims that mental aspects like understanding, free will, or insight cannot be Turing machine computable based on Gödel’s incompleteness theorems (Penrose, 1999). It associates consciousness with quantum mechanical processes. The Field Theories of Consciousness, which compare uncertain particle-like and wave-like phenomena as the “neuron-wave duality” (John, 2001), propose that the widespread electromagnetic (EM) fields in brains could be the physical correlates of consciousness (Hunt and Jones, 2023).

Their rivalries are likely to yield a winner through empirical tests (or remove inappropriate theories from the competitive stage to the extent possible) and eventually enable contemporary theories to move toward falsifiable unification (Ellia et al., 2021). Since the preparations begun in 2019, there has been an initial *adversarial collaboration* between IIT and GNWT (Reardon, 2019; Melloni et al., 2021), a project aimed at falsifying various ToCs and breaking down the barriers between them. With the implementation of Chalmers winning the “25-year wager” with Koch on unraveling the mechanism of consciousness at the meeting of the Association for the Scientific Study of Consciousness (ASSC) in 2023, the preliminary result of the adversarial collaboration has been published: neither of them matches their tests perfectly (Lenharo, 2023).

Block (1995) advocated an early distinction between P-consciousness, which focuses on the experiential properties of consciousness (qualia), and A-consciousness, which focuses on the cognitive functions of consciousness (e.g., linguistic activities). Regarding these two aspects of consciousness, GNWT and HOT generally refer to the so-called A-consciousness, whereas IIT and RPT might refer to P-consciousness. This seems to explain why IIT would maintain that the maximum integrated information should be generated in the posterior cortex, whereas the prefrontal cortex, which GNWT emphasizes, would not be necessary for IIT (cf. Koch et al., 2016; Boly et al., 2017; Odegaard et al., 2017). As Doerig et al. (2021a,b) discussed in the hard criteria for testing ToCs, some ToCs associate consciousness singularly with their preferred properties and mechanisms, which are likely to be necessary but insufficient. Similarly, Lamme (2018) comparison of her RPT with other ToCs led to the conclusion that “missing ingredients” exist in all of these necessary theories.

The trend of unifying the theories of consciousness

In the Chinese context, the classic metaphor of “blind men feeling the elephant” is often used to describe how people each grasp only a particular facet of a thing and therefore perceive the same thing differently because of the discrepancies in the facets to which they have been exposed. In a practical investigation, however, following the method the blind men do may not be such a bad start, as it suggests that we have been exposed to at least parts of the fact and that by correlating this knowledge, we will come to a complete understanding.

The recognized trend toward unifying ToCs has become more widely adopted, such as Wiese (2020) advocating a “minimal unifying model” (MUM) that would be compatible with the major theories. In an attempt to integrate multiple ToCs, Safron (2020) combined IIT, GNWT, and PP to construct a comprehensive theory. This was a remarkable effort, and it would be more explanatory if it incorporated more theoretical and experimental evidence, and could further respond to the conflicts between the remaining theories. As for HOT, Brown et al. (2019) argued that realizing a global workspace requires higher-order metacognition. The Attention Schema Theory (AST) (Graziano, 2019a,b), another current theory of consciousness, has also attracted much attention; Graziano et al. (2020) previously attempted to integrate their AST

with GNWT, HOT, and other theories into a standard model of consciousness. In their response, Panagiotaropoulos et al. (2020) agreed with Graziano et al., at least on the orthogonal dimensions of the model of consciousness.

Nevertheless, some cruxes must be considered when comparing and contrasting the various theories. First, we must correctly touch the “elephant” and not something else; otherwise, for example, the integration of a model of finger movements (obviously not consciousness) into a model of consciousness would be troublesome; second, we also need to consider whether the methods or strategies used are appropriate. Regarding ToCs, for the first question, we need to cautiously confirm the diverse global states of consciousness (Bayne et al., 2016; McKilliam, 2020). A transformation in the global state of consciousness would result in a marked shift in the structure of the entire experience, as if going from one inner world to a very different one, rather than a simple change in the intensity or content of the experience. As for the second issue, Lau (2022), in his new book, analyzes in detail the ways in which current experimental methods can lead to biased interpretations of results. The rise of “no-report paradigms” (Tsuchiya et al., 2015), even “no-cognition paradigm” (Block, 2019), recently manifested a practical step forward in this regard.

Being careful of both concerns above, we might effectively have a series of necessary elements if to suppose as A, B, and C... for each indicates a model and corresponding mechanism, such as A referring to IIT. Based on the present approaches to unification, the fundamental theory would be an integration of these elements, i.e.,

the fundamental theory = A and B and C and ...

Ideally, this result would be a *necessary and sufficient* condition for consciousness, also referred to as “minimally sufficient” (Fink, 2016). Is this always true? Is it possible that by integrating more and more candidate theories, our model could become more and more accurate? It is also important to note that such attempts at unification often overlook the fifth physical approach.

The architecture of a layered mind for consciousness

Gazzaniga (2018), “the father of cognitive neuroscience,” suggested his unique view of how consciousness arises based on many instances of abnormal brains he had been exposed to during ward rounds and split-brain research. For machines, confronting a breakdown is a better way for engineers to access and understand how they work. Similarly, neurological diseases indirectly provide an excellent window into the mechanisms of the mind, which Gazzaniga used to explore consciousness in the brain. For his strategy, Gazzaniga considered diverse global states and appropriate methods. He then argued that consciousness is the overall manifestation of the coordination of the diverse basic instincts of the mind, like a symphony without a conductor; in such a distributed system, individuals operate relatively independently, and different combinations of them can exhibit different patterns of performance (see Figure 1).

Unlike other theories, this view does not specify whether cortical activity is sufficient or necessary for consciousness. Gazzaniga found that our brains were resilient. A computer with

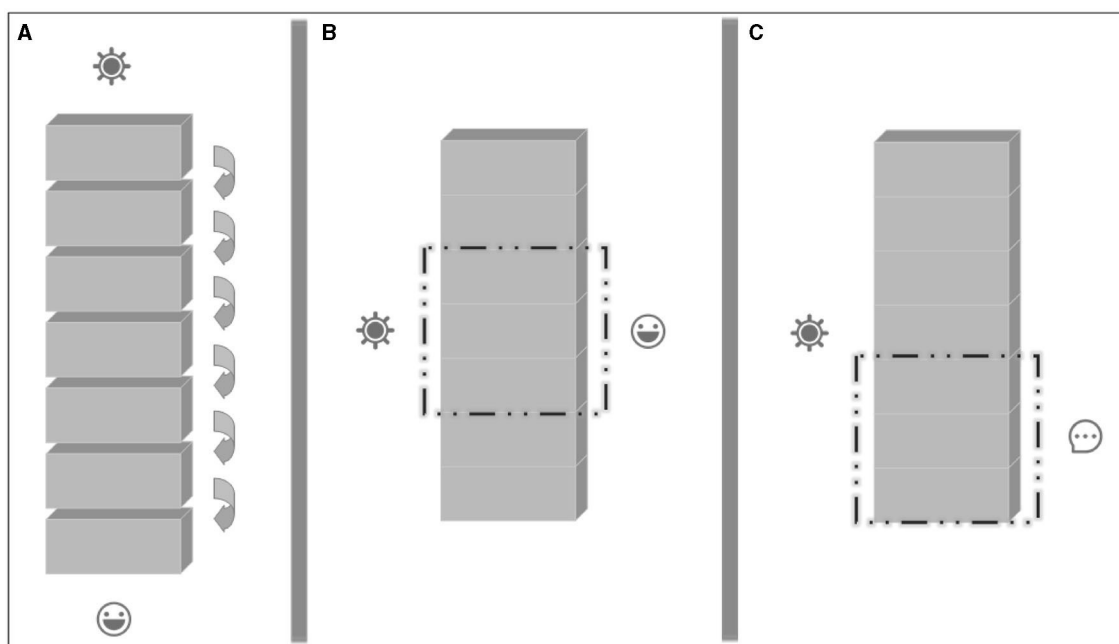


FIGURE 1

The architecture of layered mind for consciousness. **(A)**: In the architecture of traditional information processing, the stimulus signals are processed sequentially in modules, of which each specific form of information is only the product of a specific processing step, and would finally constitute experience in the so-called imaginary “module of consciousness”; **(B)**: However, in the architecture of layered mind, signals are processed simultaneously in various layers, each of which is a candidate for a temporary “zone of consciousness”; **(C)**: The arbitrary bonds of different layers bring specific types of experiences with different structures and attributes, such as the intervention of higher cognitive functional layers to bring the experience of the conceptual component.

many severely damaged components would be rendered wholly paralyzed, but the damaged brains in the wards had still been functioning well in a way. There is no palace in the cortex and no part that acts like the core of a computer. Not only are the frontal cognitive modules and the posterior higher sensory cortex candidates for consciousness, but the entire cortex is also an evolutionary expansion of earlier forms of consciousness. In addition to Damasio (2010), Gazzaniga believes that the subcortical affective system may act as an “engine,” with which any cortical module can collaborate to produce a unique conscious experience accompanied by a sense of self. Additionally, Seth (2021) recently endorsed Damasio’s illumination of the role of emotion in generating experiences in his theory of consciousness based on PP. If we consider a layered architecture for consciousness, the formulation of the above integration should be

the necessary and sufficient model = the “engine” and (A or B or C or ...)

In his recent work, Block (2023) distinguished our perception from cognition, which he used to argue against what he called “cognitive theories of consciousness.” Layered architecture can reconcile this apparent contradiction. From the perspective of the architecture of the layered mind, different global states may result from diverse brain regions and mechanisms. Eventually, both IIT and GNWT, as well as various other important ToCs, will be assessed for their indicative roles within a synthetic model in the meaning of layered architecture.

If we explore this architecture radically into more essential ranges, it may extend to a general version that the physical approach may help out. Our brains, as complex systems, have many components and layers of subsystems, and both Orch OR and EM fields can operate as a hierarchy across multiple levels of the brain. Hameroff (2022) argued the orders of magnitude in frequency in microtubules inside each neuron. The proponents of EM fields describe them from micro to macro scales as “stuff” of phenomenology, patterns of experiences, and phenomenal objects, respectively (Fingelkurts et al., 2013; Hales and Ericson, 2022).

Further work will focus on determining the specific interpretative position of each ToC within the layered model and will help unravel the interaction protocols between the components in the model.

Discussion

In this opinion article, we reviewed the stalemate that various theories of consciousness, each with its specific empirical support and respective plausibility, and attempted to unify these theories. Contrasting a layered architecture with the unification of traditional viewpoints suggests that it may be a more conducive approach to profoundly understand consciousness and may be compatible with competing theories.

Author contributions

ZR: Conceptualization, Formal analysis, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Social Science Fund of China (Grant Nos. 21ZD0b1 and 18ZDA029).

Acknowledgments

The author thanks Hengwei Li, Shiling Cai, Xuting Cao, and other colleagues for their suggestions on the work presented here

References

- Bayne, T., Hohwy, J., and Owen, A. M. (2016). Are there levels of consciousness? *Trends Cogn. Sci.* 20, 405–413. doi: 10.1016/j.tics.2016.03.009
- Block, N. (1995). On a confusion about a function of consciousness. *Behav. Brain Sci.* 18, 227–247. doi: 10.1017/S0140525X00038188
- Block, N. (2019). What is wrong with the no-report paradigm and how to fix it. *Trends Cogn. Sci.* 23, 1003–1013. doi: 10.1016/j.tics.2019.10.001
- Block, N. (2023). *The Border Between Seeing and Thinking*. New York: Oxford University Press.
- Boly, M., Massimini, M., Tsuchiya, N., Postle, B. R., Koch, C., Tononi, G., et al. (2017). Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? Clinical and neuroimaging evidence. *J. Neurosci.* 37, 9603–9613. doi: 10.1523/JNEUROSCI.3218-16.2017
- Brown, R., Lau, H., and LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends Cogn. Sci.* 23, 754–768. doi: 10.1016/j.tics.2019.06.009
- Damasio, A. (2010). *Self Comes to Mind: Constructing the Conscious Brain*. New York: Pantheon Books.
- Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. New York, NY: Penguin Books.
- Del Pin, S. H., Skóra, Z., Sandberg, K., Overgaard, M., and Wierchoń, M. (2021). Comparing theories of consciousness: why it matters and how to do it. *Neurosci. Conscious.* 2021, 19. doi: 10.1093/nc/niab019
- Doerig, A., Schurger, A., and Herzog, M. H. (2021a). Hard criteria for empirical theories of consciousness. *Cogn. Neurosci.* 12, 41–62. doi: 10.1080/17588928.2020.1772214
- Doerig, A., Schurger, A., and Herzog, M. H. (2021b). Response to commentaries on 'hard criteria for empirical theories of consciousness'. *Cogn. Neurosci.* 12, 99–101. doi: 10.1080/17588928.2020.1853086
- Ellia, F., Hendren, J., Grasso, M., Kozma, C., Mindt, G., Lang, P., et al. J., et al. (2021). Consciousness and the fallacy of misplaced objectivity. *Neurosci. Conscious.* 20, 32. doi: 10.1093/nc/niab032
- Fingelkurts, A. A., Fingelkurts, A. A., and Neves, C. F. H. (2013). Consciousness as a phenomenon in the operational architectonics of brain organization: criticality and self-organization considerations. *Chaos Sol. Fract.* 55, 13–31. doi: 10.1016/j.chaos.2013.02.007
- Fink, F. (2016). A deeper look at the "neural correlate of consciousness". *Front. Psychol.* 7, 44. doi: 10.3389/fpsyg.2016.01044
- Gazzaniga, M. S. (2018). *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*. New York: Farrar, Straus and Giroux.
- Graziano, M. S. A. (2019a). Attributing awareness to others: the attention schema theory and its relationship to behavioral prediction. *J. Conscious. Stud.* 26, 17–37.
- Graziano, M. S. A. (2019b). *Rethinking Consciousness: A Scientific Theory of Subjective Experience*. New York: Norton.
- Graziano, M. S. A., Guterstam, A., Bio, B. J., and Wilterson, A. I. (2020). Toward a standard model of consciousness: reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. *Cog. Neuropsychol.* 37, 155–172. doi: 10.1080/02643294.2019.1670630
- Hales, C. G., and Ericson, M. (2022). Electromagnetism's bridge across the explanatory gap: how a neuroscience/physics collaboration delivers explanation into all theories of consciousness. *Front. Human Neurosci.* 16, 836046. doi: 10.3389/fnhum.2022.836046
- Hameroff, S. (2022). Consciousness, cognition and the neuronal cytoskeleton—A new paradigm needed in neuroscience. *Front. Mol. Neurosci.* 15, 869935. doi: 10.3389/fnmol.2022.869935
- Hameroff, S., and Penrose, R. (2014). Consciousness in the universe a review of the 'orch or' theory. *Physics Life Rev.* 11, 39–78. doi: 10.1016/j.plrev.2013.08.002
- Hunt, T., and Jones, M. (2023). Fields or firings? Comparing the spike code and the electromagnetic field hypothesis. *Front. Psychol.* 14, 1029715. doi: 10.3389/fpsyg.2023.1029715
- John, E. R. (2001). A field theory of consciousness. *Consciousness and Cognition*, 10, 184–213. doi: 10.1006/ccog.2001.0508
- Koch, C. (2018). What is consciousness: scientists are beginning to unravel a mystery that has long vexed philosophers. *Nature* 557, 9–12. doi: 10.1038/d41586-018-05097-x
- Koch, C., Massimini, M., Boly, M., and Tononi, G. (2016). Posterior and anterior cortex—Where is the difference that makes the difference? *Nat. Rev. Neurosci.* 17, 666. doi: 10.1038/nrn.2016.105
- Lamme, V. A. F. (2018). Challenges for theories of consciousness: seeing or knowing, the missing ingredient and how to deal with panpsychism. *Philosoph. Transact. Royal Soc. B.* 373, 344. doi: 10.1098/rstb.2017.0344
- Lau, H. (2022). *In Consciousness We Trust: The Cognitive Neuroscience of Subjective Experience*. New York: Oxford University Press.
- Lau, H., and Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends Cogn. Sci.* 15, 365–373. doi: 10.1016/j.tics.2011.05.009
- Lenharo, M. (2023). Philosopher wins consciousness bet with neuroscientist. *Nature* 619, 14–15. doi: 10.1038/d41586-023-02120-8
- Mashour, G. A., Roelfsema, P., Changeux, J. P., and Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105, 776–798. doi: 10.1016/j.neuron.2020.01.026
- McKilgama, A. (2020). What is a global state of consciousness. *Philoso. Mind Sci.* 1, 58. doi: 10.33735/phimisci.2020.II.58
- Melloni, L., Mudrik, L., Pitts, M., and Koch, C. (2021). Making the hard problem of consciousness easier. *Science* 372, 911–912. doi: 10.1126/science.abj3259

and also the editor and two reviewers for their helpful comments on this manuscript.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Odegaard, B., Knight, R. T., and Lau, H. (2017). Should a few null findings falsify prefrontal theories of conscious perception? *J. Neurosci.* 37, 9593–9602. doi: 10.1523/JNEUROSCI.3217-16.2017
- Oizumi, M., Albantakis, L., Tononi, G., and Sporns, O. (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comput. Biol.* 10, 5. doi: 10.1371/journal.pcbi.1003588
- Panagiotaropoulos, T. I., Wang, L., and Dehaene, S. (2020). Hierarchical architecture of conscious processing and subjective experience. *Cogn. Neuropsychol.* 37, 180–183. doi: 10.1080/02643294.2020.1760811
- Penrose, R. (1999). *The Emperor's New Mind: Concerning Computers, Minds, and The Laws of Physics*. Oxford: Oxford Paperbacks.
- Reardon, S. (2019). Rival theories face off over brain's source of consciousness. *Science* 366, 293. doi: 10.1126/science.366.6463.293
- Safron, A. (2020). An integrated world modeling theory (iwmt) of consciousness: combining integrated information and global neuronal workspace theories with the free energy principle and active inference framework: toward solving the hard problem and characterizing agentic causation. *Front. Artif. Intell.* 3, 30. doi: 10.3389/frai.2020.00030
- Seth, A. (2021). *Being You: A New Science of Consciousness*. London: Faber and Faber Ltd.
- Seth, A. K. (2018). Consciousness: the last 50 years (and the next). *Brain Neurosci. Adv.* 2, 1–6. doi: 10.1177/2398212818816019
- Seth, A. K., and Bayne, T. (2022). Theories of consciousness. *Nat. Rev. Neurosci.* 23, 439–452. doi: 10.1038/s41583-022-00587-4
- Seth, A. K., and Hohwy, J. (2021). Predictive processing as an empirical theory for consciousness science. *Cogn. Neurosci.* 12, 2. doi: 10.1080/17588928.2020.1838467
- Signorelli, C. M., Szczotka, J., and Prentner, R. (2021). Explanatory profiles of models of consciousness—Toward a systematic classification. *Neurosci. Conscious.* 2021, 21. doi: 10.1093/nc/niab021
- Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol. Bullet.* 215, 216–242. doi: 10.2307/25470707
- Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nat. Rev. Neurosci.* 17, 450–461. doi: 10.1038/nrn.2016.44
- Tsuchiya, N., Wilke, M., Frässle, S., and Lamme, V. A. F. (2015). No-report paradigms: extracting the true neural correlates of consciousness. *Trends Cogn. Sci.* 19, 757–770. doi: 10.1016/j.tics.10002
- Wiese, W. (2020). The science of consciousness does not need another theory, it needs a minimal unifying model. *Neurosci. Conscious.* 20, 13. doi: 10.1093/nc/niaa013
- Yaron, I., Melloni, L., Pitts, M., and Mudrik, L. (2022). The contrast database for analysing and comparing empirical studies of consciousness theories. *Nat. Hum. Behav.* 6, 593–604. doi: 10.1038/s41562-021-01284-5



OPEN ACCESS

EDITED BY

Luca Simione,
UNINT - Università degli studi Internazionali di
Roma, Italy

REVIEWED BY

H. Henrik Ehrsson,
Karolinska Institutet (KI), Sweden
Jari Kaukua,
University of Jyväskylä, Finland

*CORRESPONDENCE

Shaun Gallagher
✉ s.gallagher@memphis.edu

RECEIVED 18 September 2023

ACCEPTED 14 November 2023

PUBLISHED 14 December 2023

CITATION

Gallagher S (2023) Minimal self-consciousness
and the flying man argument.
Front. Psychol. 14:1296656.
doi: 10.3389/fpsyg.2023.1296656

COPYRIGHT

© 2023 Gallagher. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

Minimal self-consciousness and the flying man argument

Shaun Gallagher^{1,2*}

¹Department of Philosophy, The University of Memphis, Memphis, TN, United States, ²School of Liberal Arts, University of Wollongong, Wollongong, NSW, Australia

The concept of minimal self-consciousness or “minimal self” is equivalent to a very basic form of first-person, pre-reflective self-awareness, which includes bodily self-awareness, and is related to phenomenal experience (qualia) and sentience. This phenomenological concept plays a role in characterizations of the senses of ownership and agency; in recent debates about Buddhist conceptions of the no-self; in explanations of illusions such as the Rubber Hand Illusion; as well as in characterizations of schizophrenia as a self-disorder. Despite its relevance to these complex investigations, a number of theorists have recently pointed out that the concept is not well defined. In order to provide some clarification about the notion of minimal self and how it relates to bodily and sensory processes this paper reaches back to the ideas expressed in a famous medieval thought experiment proposed in the 11th century: Avicenna’s Flying Man argument. The paper then provides a review of some of the contemporary debates about the minimal self, pointing especially to questions about the role of bodily and social processes.

KEYWORDS

flying man, Avicenna, minimal self-consciousness, sensory input, interoception

Introduction

The phrase “minimal self,” as used in this paper, and as it has been used in both cognitive science and phenomenological philosophy of mind, is equivalent to a very basic form of first-person, pre-reflective self-consciousness, which includes bodily self-awareness. This concept plays a role in phenomenological characterizations of the sense of ownership, and the sense of agency (Gallagher, 2000a; Zahavi, 2017); in recent debates about Buddhist conceptions of the no-self (Albahari, 2011; Siderits et al., 2011); in explanations of illusions such as the Rubber Hand Illusion (Limanowski, 2014; Georgie et al., 2019); as well as in characterizations of schizophrenia as a self-disorder (Nelson et al., 2014). Despite its relevance to these complex investigations, Kim and Effken (2022, 15), have recently pointed out that “there are no clear criteria to define the minimal self except for some vague intuitive feeling of ‘a basic, immediate, or primitive ‘something’ that we are willing to call a self’” (citing Gallagher, 2000a). Likewise, Lang and Viertbauer (2022) outline a plethora of views on pre-reflective self-awareness, and conclude that given this range of interpretations it “is not surprising that there is not only controversy about what is meant by pre-reflective self-consciousness, but moreover whether pre-reflective self-consciousness exists at all . . .”

These controversies relate to other concepts relevant to understanding consciousness, namely, phenomenal experience (qualia) and sentience, especially as the latter is defined by Nicholas Humphrey, who reaches back to the early seventeenth century to find its original meaning: what sensations feel like to the subject who responds to sensory stimuli (2022, 1). In order to provide some clarification about the notion of minimal self and how it relates to bodily and sensory processes I will reach back a bit further in the history of these ideas to a famous medieval thought experiment proposed in the 11th century: Avicenna's Flying Man argument. I'll then review some of the contemporary debates about this concept.

Let me preface the following with a methodological proviso. Like empirical experiments that require the introduction of controls, and like toy models that introduce unrealistic simplifications, thought experiments are also limited in terms of the kinds of results we can attain through their use. In all of these approaches one ends up with some degree of abstraction from the phenomenon that one is attempting to explain. In the following sections I'll often be discussing abstractions. I'll argue, however, that they are insightful abstractions that can provide some direction for further thinking. In that respect the strategy is to point out how they are abstractions and to point to a trajectory that could result in less abstract insights, even if in this paper we don't have the space to pursue these trajectories. In my view, all such trajectories lead toward more embodied and enactive approaches to issues concerning pre-reflective self-awareness. It is specifically the limitations of the more abstract, less embodied views that point us in the right direction.

The flying man

Although philosophers often use thought experiments rather than empirical experiments to further an argument, many philosophers (not only today, but also in the past) engage or have engaged with empirical studies, and Avicenna is no exception to this. As both a physician and a philosopher he conducted empirical medical research and, in the 11th century, published a work entitled *The Canon of Medicine*, which came to be used in western universities until the 16th century. The third volume of this work includes chapters on spinal cord injury (Ghaffari et al., 2022). It's notable that considerations of spinal cord injury have more recently played a strategic role in addressing a question that is roughly similar to the one that Avicenna addresses in his thought experiment on the Flying Man. For example, in behavioral and neuroscientific studies of spinal cord injury Moro et al. (2022), ask whether and to what degree the body's sensory and motor processes, or lack thereof, contribute to or constrain cognition. They develop a positive answer showing how, even in severe cases of body-brain disconnection, deafferentation and de-efferentation, embodied processes continue to play a role in modulating a broad range of cognitive capacities, including spatial perception, motor imagery, the discrimination of biological motion, affordance perception, and so forth. In contrast, Avicenna's seemingly negative answer in the Flying Man argument focuses on just one narrow question about self-awareness. At the end of the first chapter of his treatment of soul in the *Psychology*, Avicenna (Ibn Sina) (1959) argued that a newly created man would be self-aware even if he were

floating in a void with all his senses disabled.¹ The conclusions to be drawn are that the self that one is aware of is not bodily, and self-awareness is not an awareness by means of the senses.

Avicenna presents several versions of the argument. The most extensive one is this:

One of us must suppose that he is created all at once, and created as perfect, but with his sight prevented from seeing anything external [to him]. He is created hovering in the air, or in a void, in such a way that the air does not buffet him so that he would have to feel it. His limbs are separated so that they do not meet or contact one another. He must then reflect as to whether he will affirm the existence of his self [*dhaāt*]. He will not hesitate to affirm himself to exist. He will not, however, affirm things exterior to his members nor the hidden things of his interiors nor his soul nor his brain nor anything else extrinsic. He will affirm himself to exist though he will not affirm the length or the width or the thickness of himself.

If in this situation he were able to imagine a hand or another limb, he would not imagine it as a part of himself, nor as a condition for his self. . . . As to the self whose existence he affirms, it is specific for it that it is identical to him and distinct from his body or his limbs, which he has not affirmed. Thus the alert person has a way to be advised concerning the existence of the soul [or self] as something distinct from the body, or rather distinct from body, and [a way] by which he may understand it and be aware of it. (1959, 15–16; trans. modified from Adamson and Benevise (2018), 148–149).

There is some scholarly dispute about the meaning of the word *dhaāt* (self or essence).² For purposes of this paper, I set aside the ontological-terminological issues in order to focus just on the phenomenology—and for that purpose, I translate *dhaāt* as “self,” following a precedent set by Marmura (1986, 383) who argued, “The primary concern [of the argument] is psychology, not metaphysics.” This allows us to focus on a point that scholars generally agree on, namely that Avicenna designed the flying man to argue that being aware of oneself is independent of any visual, tactile or proprioceptive awareness of one's body or any further content of experience (Avicenna (Ibn Sina), 1959, 225). Thus, he argues that if you are completely unaware of your body and your physical circumstances, you would be unaware of everything except

1 Avicenna didn't use the term “flying man.” Black (2008, 63 n. 3) attributes it to Gilson (1929–1930, 41 n. 1). The term “floating man” is also found in the literature.

2 Hasse (2000, 83) and Adamson and Benevise (2018) argue that in this context it means “essence,” rather than “self,” and contend that Avicenna is attempting to show in opposition to Aristotle that the essence of the soul does not include the body. In contrast, Kaukua (2015, 2020) maintains that the flying man argument was designed to point our attention to our being aware of ourselves independently of any other content of experience. For our purposes, we can follow Avicenna: “We say: what is intended by ‘the soul’ is that which each of us refers to by his saying, ‘I’” (Avicenna (Ibn Sina), 1952, 183; trans. Marmura, 1986, 384). It is notably that G.E.M. Anscombe, without mentioning Avicenna, and rather focused on Augustine and Descartes, dreams up a very similar thought experiment about sensory deprivation to test whether it is the body to which each of us refers by saying, “I.” “Sight is cut off, and I am locally anaesthetized everywhere, perhaps floated in a tank of tepid water; I am unable . . . to touch any part of my body with any other” (Anscombe, 1975, 57).

the “fixedness” of your individual existence (1959, 225). This is a form of self-awareness, Avicenna argues, that is a constituent of the self, and “belongs to it always and in actuality”—a form of natural knowledge that does not depend on contact with another human.

Adamson and Benevise (2018) note that Avicenna elsewhere claims, “we are constantly aware of ourselves, even when asleep,” which they interpret to mean that this is a form of tacit self-awareness.³ Avicenna himself suggests that most of the time we are not “alert” to this awareness, i.e., that we do not have reflective knowledge of it (1959, 226–227). Kaukua (2020, 13–14) interprets this as follows: “most of us have no experience of being aware of nothing but ourselves, given that in the normal circumstances, we are aware of ourselves as embodied agents and subjects of cognition, constantly immersed in our mutual engagement with the world around us. The [flying man] argument is designed to show that self-awareness would remain even if these features normally associated with it were bracketed. It points at something, ourselves, the existence of which we assert without asserting the existence of anybody.” As we’ll see, however, and as frequently noted (see Black, 2008) this doesn’t mean that the self is disembodied. Indeed, self-awareness does not tell us *what* the self is. “Surely, one may be aware of the existence of something, including oneself, without knowing what that thing is, and it is precisely such an awareness of existence that the flying man has.” (Kaukua, 2020, 11). We might argue, however, that self-awareness does tell us one thing about what the self is—it is something that, at a minimum, is capable of self-awareness. And Avicenna says something like this: “[the soul or self’s] awareness of itself is *by nature*, this being a constituent of it and hence belongs to it always and in actuality” (cited in Marmura, 1986, 386; emphasis added).

Self-awareness is a form of what Avicenna calls natural knowledge.⁴ He uses the following example to define natural knowledge. He suggests that if a person is created fully mature and rational, having, however, had no contact with other humans and human institutions, and is confronted with a commonly accepted moral dictum and a self-evident logical truth, he will be able to doubt the first, but not the second (Avicenna (Ibn Sina), 1892, 119; discussed in Marmura, 1986). Although Avicenna recognizes the importance of intersubjective interaction, specifically in the ethical context, he argues that natural knowledge is not something we learn from anyone else. This is the kind of knowledge had by the flying man, i.e., a person born fully mature and rational but having had no human contact. Avicenna thus holds that the self has natural, constant knowledge of itself.

3 Avicenna here is 180° removed from the view of Hume (1739/1978), who equates what we call “self” with the perceptions (sensations) we experience in introspection, and suggests “When my perceptions are remov’d for any time, as by sound sleep, so long am I insensible of myself, and may truly be said not to exist.” Also see Lane (2020) for considerations of whether the minimal self can dissociate from consciousness.

4 “Self-awareness is essential to the soul, it is not acquired from outside. It is as if, when the self comes to be, awareness comes to be along with it. Nor are we aware of [the self] through an instrument, but rather, we are aware of it through itself and from itself. And our awareness is an awareness without qualification, that is, there is no condition for it in any way; and it is always aware, not at one time and not another.... Self-awareness is natural (*gharizah*) to the self, for it is its existence itself, so there is no need of anything external by which we perceive the self. Rather, the self is that by which we perceive the self.” (Avicenna (Ibn Sina), 1973, 160–162; trans. Black, 2008).

The minimal self

Much of the contemporary discussion about the minimal self was motivated by Strawson’s (1997) essay on the self.⁵ There he indicated the methodological primacy of phenomenology over ontology and simply asked what was the most minimal experience of self that we could have. He considers the answer to this to be very basic, and “situated below any level of plausible cultural variation” (§3). His answer is that this basic self is a “mental self.” Although he is a philosophical materialist, and believes that we are wholly material things, the characterization of the self as mental is an answer to the strictly phenomenological question of what we experience.

With respect to this minimal mental self Strawson excludes diachronicity, agency, and personality. For example, he writes:

It seems plain that ... experience of the self does not necessarily involve experience of it as something that has a personality. Most people have at some time, and, however, temporarily, experienced themselves as a kind of bare locus of consciousness—not just as detached, but as void of personality, stripped of particularity of character, a mere (cognitive) point of view. Some have experienced it for long periods of time. It may be the result of exhaustion or solitude, abstract thought or a hot bath. It is also a common feature of severe depression, in which one may experience “depersonalization.” This is a very accurate term, in my experience and in that of others I have talked to. (1997, 420).

Diachronicity is set aside based on Strawson’s own phenomenology (now relatively famous in philosophical circles), that he experiences at best a 3-s-long self, and is not inclined to narrative extensions (also see Strawson, 2004). He states: “I believe the Buddhists have the truth when they deny the existence of a persisting mental self, in the human case, and nearly all of those who want there to be a self-want there to be a persisting self” (1997, 427).

Strawson also excludes the sense of agency, although he does not provide an argument for this exclusion. We can suppose that a sense of agency only comes along as we are engaged in some action, and when we are not, we don’t have a sense of agency, so it can’t be essential. We could add Avicenna’s view on this. He raises the question of whether self-knowledge is mediated through one’s action. This, he argues, is not the case because, “the supposition” of the flying man argument excludes any action. Moreover action is either general or specific. General action does not lead to the knowledge of the particular self. The action would have to be particular; for example, my own individual act. But when I state that I am performing an act, the “I” is prior to my act. My act presupposes the existence of my-self; otherwise I would not refer

5 Strawson does not use the term “minimal self” in his 1997 and 1999 essays, but he does refer to the “minimal case” or form of self-experience. I may be to blame for the term “minimal self” in this phenomenological context (Gallagher, 2000a). I was referring specifically to Strawson’s account, and distinguishing minimal from narrative self. In that article I cite Damasio’s (1999) use of the term “core self” as a related concept. I also use the phrase in Gallagher (2000b), a volume edited by Dan Zahavi.

to it as *my* act (see Black, 2008, and similar points made in regard to object perception in Avicenna (Ibn Sina) (1973), 161).

More positively Strawson defines the “minimal case,” or “the minimal form of self-experience” as a momentary (single) mental subject of experience.⁶ He refers to this as mental or M-experience and asks whether this is clearly a case of *self*-experience, to which he answers “yes,” but notes that we are one step away from the Buddhist idea of non-self.

[I]t is not clear that the minimal case of Self-experience is *ipso facto* the minimal case of M-experience. I suspect that the minimal case of M-experience may be some kind of “pure consciousness” experience, of the kind discussed by Buddhists and others, that no longer involves anything that can usefully be called “Self-experience” at all (1999, 118).

He calls this the “meditative rider” to his positive claims, namely that genuine “M-experience” need not involve an experience of self. If we stay with the concept of minimal self-experience, however, Strawson’s position is close to the standard phenomenological view, namely, that self-experience is not an experience of some object. In this regard he quotes the phenomenologist Louis Sass, who, in turn, references William James: the self “is not, in fact, experienced as an entity in the focus of our awareness, but, rather, as a kind of medium of awareness, source of activity, or general directedness toward the world” (Sass, 1998, 562). Strawson then translates this into the terminology preferred by analytic philosophy: although the self is experienced as a thing of some sort, this “does not require experience of self that is experience (as) of ‘an entity in the focus of awareness’” (1999, 115).⁷ Strawson also quotes the commentary by Zahavi and Parnas (1998), which refers to “the basic self-awareness of an experience,” as “an immediate and intrinsic self-acquaintance which is characterized by being completely irrelational” [Zahavi and Parnas (1998), p. 696]. “Irrelational” here means it does not have a subject-object structure, but rather is solely the subject with the structure of pre-reflective self-awareness.⁸

6 Strawson’s, 1997 paper generated four special issues in the *Journal of Consciousness Studies*, which I edited with Jonathan Shear, and which was then published as a volume, *Models of the Self* (Gallagher and Shear, 1999). Strawson (1999) provided a response to all of the essays. In this response he characterizes the minimal self as:

[1] A subject of experience.

[2] A thing, in some interestingly robust sense.

[3] A mental thing, in some sense.

[4] Single at any given time, and during any hiatus-free or strongly experientially unified period of experience. (1999, 108).

7 This is a point that runs throughout Hutto and Ilundáin-Agurruza’s (2020) essay in which they criticize the phenomenological concept of minimal self—namely the insistence that the minimal self is not something that we experience “as” self or *qua* self. “To acquire a sense of oneself as a self that is distinct from another—a sense of self such that one recognizes oneself *qua* self as featuring in shared experiencing—is a quite sophisticated conceptual achievement” (p. 518). Neither Strawson nor the phenomenologists characterize the minimal self in this way, however. Nor does Avicenna since he contends that we are not “alert” to this awareness, i.e., that we do not have reflective knowledge of it such that we take it as a self (1959, 226–227).

8 This is precisely the view expressed by the classic phenomenologists (Husserl, Sartre, Merleau-Ponty). That is, the notion of the minimal self is tied specifically to a subject’s pre-reflective self-awareness, and this kind of self-awareness is a structural feature of consciousness. The claim is

The phenomenology of minimal self-awareness

This phenomenological conception of the minimal self, however, is not an entirely settled issue, and contemporary debates focus on several questions.

1. Is the minimal self *experiential*, or simply an abstract, formal notion?
2. Is the minimal self *embodied*?
3. Does the minimal self involve *social* existence?

First, a quick answer to the first question is that it is experiential, specifically, as a structural feature of experience it is something that is experienced; it is more than simply a formal principle of the sort defined by Kant. Yet, it is in some regards an abstraction, since for the most part, in our everyday experience, it is never experienced solely in itself without other complications, which may involve embodiment and intersubjectivity. Here we may also start to see the limitations of Avicenna’s flying man in its attempt to abstract away from all sensory experience, including experience of the body.

Second, with regard to the question of embodiment, accounts of the minimal self often include references to proprioception, especially in relation to two features that are typically included in minimal self-awareness: the sense of ownership (or mineness, or “for-me-ness”) and the sense of agency. As we noted, however, the latter is present only when some form of action is involved. This is why Strawson excludes it as essential. To be clear, however, one may have a sense of agency not just for bodily action, in the sense that such action involves proprioception/kinesthesia, as well as efferent processes that may contribute to self-awareness. One may also have a sense of agency for thinking, imagining, remembering, etc. On most interpretations, of course, these cognitive processes involve some embodied aspects (embodied simulation, activations in motor areas, and perhaps even proprioceptive, affective and interoceptive processes), all of which seemingly tell us that we are the agent of such cognitive processes.

The sense of ownership or mineness, however, seems more basic. This was indicated by Avicenna when he suggested that *my* act (or agency) presupposes the existence of *my*-self; otherwise I would not refer to it as *my* act. Again, there are many studies that discuss the sense of bodily ownership—this includes, for example, experiments on the rubber-hand illusion (see Riemer et al., 2019; Ehrsson, 2020, 2023; for a recent review of this literature see Georgie et al., 2019).⁹ Moreover, on some interpretations,

that whenever I am conscious, I am pre-reflectively conscious of being conscious, that is, I am pre-reflectively aware that I am experiencing something. I have a self-awareness of my experience that does not depend on an additional act of consciousness that would reflectively take the first-order consciousness as an object.

9 There is not universal agreement about the connection between proprioception and the sense of body ownership, although in the rubber hand illusion, where one gains a sense of ownership for the rubber hand (it starts to feel as part of one’s body) the manipulation of proprioception is involved, so that it is subordinated to visual and tactile senses (see Limanowski, 2014). Also, in the absence of proprioception (as in cases of deafferentation) one’s body or body parts can feel alien (unowned) (Gallagher and Cole, 1995). Despite this and other evidence, Humphrey suggests that proprioception “is of little or no importance to establishing

schizophrenic delusions of control and thought insertion represent cases in which the sense of agency is missing, but the sense of ownership remains [“It is my hand that did it, but I did not control it”; “I experience this thought as part of my stream of consciousness, but I did not think it” (Gallagher, 2004; Frith, 2015)]. In the latter respect the sense of ownership does not necessarily involve an explicit sense of body ownership. The more general claim, in regard to the minimal self, is that there is a very basic sense of mineness implicit in experience itself.

This just is the claim that when there is experience, there is a subject of experience, and phenomenologically this can be explained in terms of the temporal structure of consciousness. On Husserl’s account, the retentional structure of consciousness involves retaining in a continuous but fading manner the just-past experience, which allows me to say, for example, that I’ve been listening to a particular piece of music, without engaging in a full-blown act of recollection (Husserl, 1991). This immediate retention includes the sense that it is *my* ongoing experience—that I am the one who has been experiencing the music. The mineness of the experience is built into this structure, and I never have this immediate sense of experience for experience that is not my own.¹⁰

Concerning the question of embodiment, by design of Avicenna’s thought experiment, as in cases of sensory deprivation experiments, we exclude sensory input and bodily movement. With respect to proprioception, when you do not move for some time, your proprioceptive sense of where your limbs are located dissipates. The phenomenology is that in this circumstance, if I don’t move, without vision, I don’t know where my limb is because I can’t feel it. The subjective experience of position sense is not just less vivid or less precise; it has disappeared. To be sure, this is quite temporary. All I have to do is move my limb and proprioceptive awareness returns. Furthermore, however, we should note that in some experimental cases of anesthetic block of the sensory and motor nerves of the arm, the blocking of proprioception does not remove awareness of the limb; rather, a phantom arm is experienced (Melzack and Bromage, 1973), or one has contradictory experiences: an experience of the limb as missing and, at the same time, an illusory experience of the limb as enlarged or swollen or shrunken (Paqueron et al., 2003). One of the reviewers for this paper reports that in the case of a complete experimental ischemic block of one’s arm, which is non-visible behind a screen, proprioceptive awareness of the arm does not dissipate; one still has a sense of it somewhere behind the screen. However, when the screen is removed and the hand is visible, it no longer feels like one’s own arm because (due to proprioceptive drift) the visual input does not match its felt position. Accordingly, eliminating proprioception is not so straight forward, and for this reason in the flying man

experiment we would need to stipulate, in line with Avicenna’s aims, that the person comes into existence in a condition of complete deafferentation (see Gallagher and Cole, 1995; Miall et al., 2021; Gallagher, 2022; for a discussion of empirical cases).

It may be even more difficult to get rid of interoceptive sensation, and in sensory deprivation experiments, these sensations are still operative. Indeed, sensory deprivation experiments suggest that interoception (of beating heart, respiration, hunger, pain, etc.) is enhanced when one removes extrasensory input (Feinstein et al., 2018). This is one important difference between sensory deprivation experiments and Avicenna’s flying man experiment, assuming that interoception is eliminated in the flying man. One might argue that just such interoceptive sensation, what James (1890) calls the “warmth and intimacy” of bodily sensations, or what Fuchs (2013) calls “the feeling of being alive”—a pre-reflective, bodily self-awareness that comprises the background of all intentional feeling—is part of what causally generates the basic sense of mineness for any of my experiences, and constitutively just is what I typically experience as my-self. The sense of body-ownership, then, could be said to depend on the formal temporal structure, the retention of my ongoing experience that, at a minimum, is interoceptive.¹¹ Hence the importance for Avicenna of eliminating interoception, as well as proprioception and exteroception.¹²

11 One reviewer raised an important question about phantom limbs or a phantom body. Would a brain without any somatosensory or other bodily sensory input develop a sense of phantom bodily awareness? Even in cases of congenital absence of limbs individuals experience (aplastic) phantoms (Brugger et al., 2000; Brugger, 2011). One might assume that even the flying man, who, rather than being born, arrives fully mature but without bodily senses, might experience a phantom body. The issue is complicated. A traditional view, which denied apasic phantoms, maintained that having a phantom depended on having had sensory experience with the relevant limb (e.g., Simmel, 1961). The current neuroscientific view is that somatosensory areas of the brain that would be correlated to the missing limb, even if they deteriorate without sensory input, may still generate a phantom. What would Avicenna think? The first mention of phantoms has been attributed to Ambroise Paré in the 16th century. But even if, as Björn Meyerson (in Finger and Hustwit, 2003) suggests, Avicenna in his medical practice must have encountered the phenomenon of phantom pain, it’s not clear how he would go about explaining it. Clearly, we should not attribute an understanding of contemporary neuroscience, plasticity or neural reorganization to him. We could ask what the flying man’s brain would be like. Since Avicenna indicates that he is “created all at once, and created as perfect,” we would expect that he came into existence with a perfectly normal brain but in a complete sensory deprivation condition. In this condition would he experience a phantom body (or body part)? What stimulus would spark this experience of a phantom. If we think that some sensory experience or motor reafference is required, these, as well as bodily pain, phantom or not, are supposedly ruled out by the experiment. And if the phantom was generated by a completely spontaneous activation of the somatosensory cortex, for example, then from the perspective of the flying man’s experience this would be the equivalent of a dream-like phantom or illusion. The question about phantoms is an interesting one, but not one that is easily answered. In this respect Avicenna states: “If in this situation [of the flying man] he were able to imagine a hand or another limb, he would not imagine it as a part of himself, nor as a condition for his self....” The same might be said in regard to the hallucinations that sometimes occur in sensory deprivation experiments (Vosburg et al., 1960; Mason and Brady, 2009).

12 The vestibular sense is another complication and is connected with the idea that the flying or floating man could still have a bodily sense of floating or flying or hovering (as may occur in experiments on out-of-body experiences—Blanke, 2004). Indeed, the vestibular sense may be increased with the loss of other sensory inputs (Horak and Hlavacka, 2001). It’s difficult to discuss vestibular sense on its own since it is tightly connected with other

your sense of self” (2022, 133). It may be that he thinks of proprioception as purely a matter of physiological information and would reject the notion of proprioceptive awareness (see, e.g., Bermúdez et al., 1995 for this distinction)—according to Humphrey there is no phenomenal experience connected with proprioception (2022, 131).

10 For Husserl, we can come to this realization by means of a phenomenological reduction that sets aside any questions about causality. This just is the way that we experience things; and whether such experiences have a causal explanation in terms of neural, proprioceptive, or interoceptive processes is a different question. In this sense, the flying man argument effects something like a phenomenological reduction. If we could put ourselves in the situation of the flying man, we would have this type of pure phenomenological access to our experience.

Finally, some recent critical discussions of the minimal self ask whether the same can't be said about intersubjective or social aspects of experience—that, like the body, they play some implicit role affecting minimal self-awareness. For example, Ratcliffe (2017) argues that the minimal self has to be re-conceptualized in interpersonal terms since the most basic sense of self is developmentally dependent upon other people. Zahavi (2017) responds to this point. He has no problems with Ratcliffe's general claims about the importance of intersubjective dimensions, but he thinks they are irrelevant to the issue concerning the minimal self. He rejects the claim that this basic feature of consciousness “is interpersonally constituted such that young infants who had not yet engaged in sufficient interpersonal relations as well as all non-social organisms would lack phenomenal consciousness and minimal selfhood” (2017, 195). Zahavi's view is consistent with Strawson's view that even non-human (and non-social) animals can have a minimal self. Furthermore, Zahavi points out that there is a shift in Ratcliffe's argument, such that in the end he does not deny a minimal self to infants and non-human animals, but rather would insist that social development brings along a transformation of the minimal self. The idea that social processes may transform the minimal self is an open question for Zahavi, but regardless of how one answers that, given the possibility of social transformation, the issue would no longer be the denial of a non-social minimal self, but a claim about how the minimal self may change in development. “Contrary to the (more) minimal self of an infant, the (less) minimal self of an adult is interpersonally constituted” (Zahavi, 2017, 195). In this case, the proposal by Ratcliffe is not incompatible with Zahavi's notion of minimal self.

We saw that Avicenna defended a similar position, distinguishing natural knowledge from knowledge that we learn from others; minimal self-awareness is a form of natural knowledge that does not depend on others; one can see this in the case of the flying man, since not only is the flying man in a state of sensory deprivation, he is also in a state of intersubjective deprivation. One can think here of the communicative difficulties faced by subjects who are deaf-blind (Gallagher, 2017). Take away all of the other senses, and thereby all social interaction, would there not still remain a self-awareness? At least with respect to the question of the social, seemingly both Zahavi and Ratcliffe would agree with Avicenna's positive answer.

Zahavi also responds in a very similar way to criticisms proposed by Ciaunica and Fotopoulou (2017; see Kyselo, 2016) who contend that the minimal self is intersubjectively constituted. He again accepts the idea that social factors may affect other aspects of the self, and may even transform minimal self-experience. If this were not the case, one would have to consider the minimal self as self-enclosed and not open to the world. Zahavi contends, in contrast, that “qua subject of intentional experience, [the minimal self] is inherently open to the world and others” (2017, 196). More to the point, the phenomenology of the self is not exhausted by the minimal self—there are other aspects of the self (for example, narrative features) that are shaped by intersubjective interactions.

senses (including vision and somatosensory input), and without the other senses it's difficult to know how the vestibular sense would function. It's also the case that one can lose vestibular sense, so, again following Avicenna's aim, we can stipulate the elimination of the vestibular sense.

Ciaunica and Fotopoulou (2017), however, present another argument that centers on interoception. They contend that interoception (the inner feelings of bodily arousal, wakefulness, wellness etc. that accompany physiological changes) is crucial for self-experience, and indeed for the self-other distinction. As indicated above, we can allow that implicit interoception may be an important contributor to minimal self-awareness, and this suggests that the minimal self is embodied, even if I do not, or if the flying man does not experience it as such. Ciaunica and Fotopoulou (2017), however, go beyond this point; they contend, interoceptive modalities depend upon and are changed by embodied interaction with others (see also Crucianelli and Ehrsson, 2023). One can think of physiological and affective regulation by others, not only in infancy, but throughout the life span. In this respect subjective “feeling states” are, at least in part, taken to be the result of such interactions, and do not pre-exist embodied social encounters. One response to this is to accept all but the last point, and rather insist that such interoceptive feeling states do pre-exist encounters with others, arguing, in agreement with Ratcliffe, that they can then undergo transformation in our embodied encounters with others.

This aligns with Zahavi's response, that to read this late transformation into the initial natural phenomenal state would lead to the idea that the self is entirely socially constructed and does not exist outside of social relations—on that view, “human beings, who are deprived of the required social interaction and denied socially mediated attributions of self, would also lack me-ness, be self-less and without consciousness, and therefore remain ‘unconscious zombies’” (2017, 198). This is the view he rejects.

Ciaunica and Fotopoulou (2017), however, do not accept the idea that social interaction is a late achievement. Ciaunica et al. (2021a,b) have argued, for example, that intersubjective interactions already exist between the fetus in the womb and the mother. One might think of this as a kind of primary intercorporeity (Merleau-Ponty, 2012). If one accepts this, then it may be that as consciousness initially emerges, the fetus is already affected by a kind of natural alterity or connection with the other that is somehow intrinsic to pre-reflective experience. There is certainly an argument to be made [and some empirical evidence (see Lymer, 2011)] about proprioception in fetal development providing a self/non-self-distinction. Whether that amounts to a self-other (intersubjective) distinction is an open question.

In this regard, I note that the flying man argument avoids or short-circuits this issue. The flying man “is created all at once, and created as perfect”—apparently not born of a mother, but created by God, where “perfect” seemingly does not depend on having sensory input or encountering others. Perhaps more relevant to the point made by Ciaunica et al. (2021a,b) the flying man is without sensory input (including, supposedly, proprioception, kinesthesia and interoception), especially the kind of sensory input that would provide some kind of access to or awareness of another person. Assuming that Avicenna would want to exclude interoceptive sensation, the flying man would offer resistance to the argument by Ciaunica et al. (2021a,b) since their argument depends on the multisensory basis of pre-reflective experience, specifically touch and interoception (Ciaunica and Crucianelli, 2019). Touch and the other exteroceptive senses are important because they are what allow access to others—touch especially in the case of the fetus. Without sensation of a sort that gives us access to others, would

a minimal form of self-awareness with access only to my own embodied self-experience be possible? This is likely even more minimal than Zahavi would like, since he takes the minimal self to be inherently open to the world and others, which the flying man seemingly is not.

Conclusion: the super flying man

What the flying man argument shows is that the minimal self is something genuinely experiential, but at the same time an abstraction. An abstraction because to arrive at the concept of the minimal self one has to set up a thought experiment where you remove everything that contextualizes human experience, including almost all embodied sensory experience. “Almost,” because, even if one manages to eliminate proprioception and the vestibular sense, it remains a challenge to eliminate interoception. As noted, in sensory deprivation experiments, interoception may even be enhanced when one removes extrasensory input. The elimination of interoception is, of course, an empirical issue. Although the anterior insula has been identified as integrating “all subjective feelings from the body and feelings of emotion” (Craig, 2002, 655), more recent studies demonstrate that it’s much more complicated. Body ownership and multisensory integration involves a complex network that includes frontal and parietal association cortex, such as the premotor cortex and the posterior parietal cortex (Ehrsson et al., 2004; Gentile et al., 2013; Limanowski and Blankenburg, 2016; Guterstam et al., 2019; Chancel et al., 2022; Abdulkarim et al., 2023). Furthermore, there is an additional source of interoceptive sensations – the skin and its somatosensory afferent projections (Khalsa et al., 2009; Rudrauf et al., 2009; Crucianelli and Ehrsson, 2023).

Since this is a thought experiment, we can ideally lesion the projections from skin to somatosensory areas of the brain, as well as knock out any areas responsible for multisensory integration and body ownership and then assume that such operations would entirely eliminate interoception. At this point we would have to leave aside the question of whether we could do something like this and not affect any other of the person’s capacities, so that he would be “perfect” (except for sensation), as stipulated by Avicenna.¹³ This indeed would be a super flying man, with no internal or external sensations.¹⁴ Would he still have a minimal self-awareness—a super-minimal self-awareness?

To answer this question one needs to distinguish between the content and structure of phenomenal consciousness. On Avicenna’s view sensory content is not the determining factor for minimal self-awareness (see Black, 2008, 68–69). Appealing to the flying man argument he argues that self-awareness is completely autonomous and independent of any sensory experience or thought, since one cannot say “I think” or “I experience” without my already having a prior and implicit sense of I. Humphrey (2022) would

have to disagree with Avicenna. On his view, the sense of self depends entirely on having sensory experience, which is equivalent to sentience and phenomenal consciousness. Take away sensory content and no self-awareness is possible.

The disagreement between Humphrey and Avicenna is framed in terms of content. In contrast, Harry Frankfurt suggests an answer that appeals to structure, and abstracts away from content:

What would it be like to be conscious of something without being aware of this consciousness? It would mean having an experience with no awareness whatever of its occurrence. This would be, precisely, a case of unconscious experience (Frankfurt, 1988, 162).

This is consistent with the phenomenological view, which suggests the positive formulation: if the super flying man were still conscious, he would necessarily be minimally self-aware since pre-reflective self-awareness is intrinsic to (or is part of the structure of) consciousness. Strawson and Zahavi, even in presenting an abstract phenomenology of the minimal self-experience, nonetheless hold that the phenomenon is real in the sense that there is in fact some irreducible experience of what it is like *for-me* in the very structure of every experience, whether that experience is complexly rich with sensory input or simple and impoverished in this regard. What it is like is always what it is like *for* someone. For phenomenologists like Husserl and Zahavi, this self-experience would hinge on the intrinsic temporal structure of consciousness. If this intrinsic temporality is a necessary and constituting component of minimal self-awareness, however, would it be sufficient, or would it even work, without sensory input of some sort?¹⁵

A less abstract and more embodied/enactive view is that both structure and content are important. Avicenna had been arguing against this view, especially as it was expressed in Aristotle, who suggests that what we call mind is not any real thing before it thinks or experiences (*De anima* 3.4, 429a23–24). That is, the mind and its structural features are enacted in the process of experiencing. Enactive views reflect this kind of self-production, often conceived as an autopoietic self-organizing process that involves a dynamical coupling of interoceptive, proprioceptive, and exteroceptive factors. Human experience is always complex—embodied and socially contextualized—but it also, arguably, always involves a minimal self-awareness. Avicenna may be right, however, that typically in one’s everyday life one does not know this minimal experience as such. One can gain insight into it only by engaging in certain practices—phenomenology, meditation, philosophical thought experiments, scientific experiments such as sensory deprivation experiments, and so on, all of which involve some degree of abstraction.

¹³ Even if the phenomenology of sensory deprivation came close to the flying man situation (which it doesn’t for reasons stated above), typically the subjects of such experiments are not newly created perfect humans. Also, disruptions of interoception are often associated with experiences of dissociation (e.g., Pick et al., 2020; Kaldewaij et al., 2023).

¹⁴ As one reviewer suggested, the super flying man may just be what Avicenna intended as the flying man.

¹⁵ For the phenomenologists the answer is not clear cut. It depends on how one conceives of the relation between intrinsic temporality, intentional structure, and sensory content (which Husserl calls “hyletic” content), and at least on one embodied interpretation these features of consciousness mutually constrain each other (see e.g., Williford, 2013; Zippel, 2014; Soueltzis, 2023).

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

SG: Writing – original draft, Writing – review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The author thanks the University of Rome-Sapienza for supporting my research on this project as a Visiting Research Prof. of Psychology in 2022. Support was also provided by the Lillian and Morrie Moss Chair of Excellence in Philosophy at the University of Memphis.

References

- Abdulkarim, Z., Guterstam, A., Hayatou, Z., and Ehrsson, H. H. (2023). Neural substrates of body ownership and agency during voluntary movement. *J. Neurosci.* 43, 2362–2380. doi: 10.1523/JNEUROSCI.1492-22.2023
- Adamson, P. and Benevise, F. (2018). The thought experimental method: Avicenna's flying man argument. *J. Am. Philos. Assoc.* 4, 147–164.
- Albahari, M. (2011). "Nirvana and ownerless consciousness," in *Self, no self? Perspectives from analytical, phenomenological, and Indian traditions* eds M. Siderits, E. Thompson, and D. Zahavi (Oxford: Oxford University Press), 79–113.
- Anscombe, G. E. M. (1975). *The first person*. Oxford: Clarendon Press.
- Avicenna (Ibn Sina) (1892). "Pointers and reminders (al-Ishārāt wa-l-tanbihāt)," in *Ibn Sīnā. Le livre des théorèmes et des avertissements*, ed. J. Forget (Leiden: Brill).
- Avicenna (Ibn Sina), (1952). *Ahwāl al-Nafs*, ed. F. Ahwani (Cairo: Dar al-Ma'rif).
- Avicenna (Ibn Sina) (1959). "Psychology ('On the Soul')," in *Avicenna's De Anima: Being the psychological part of Kitaab al-Shifā' ed.* F. Rahman (London: Oxford University Press).
- Avicenna (Ibn Sina), (1973). *Notes: Al-Ta'liqāt* ed. A. R. Badawi (Cairo: Dar al-Ma'rif).
- Bermúdez, J., Eilan, N. and Marcel, A. (eds) (1995). *The body and the self*. Cambridge: MIT/Bradford Press.
- Black, D. (2008). "Avicenna on self-awareness and knowing that one knows," in *The unity of science in the Arabic tradition*, eds S. Rahman, T. Street, and H. Tahiri (Dordrecht: Springer), 63–87. doi: 10.1007/978-1-4020-8405-8_3
- Blanke, O. (2004). Out of body experiences and their neural basis. *BMJ* 329, 1414–1415. doi: 10.1136/bmj.329.7480.1414
- Brugger, P. (2011). "Phantom limb, phantom body, phantom self: A phenomenology of "body hallucinations," in *Hallucinations: Research and practice*, eds J. D. Blom and I. E. Sommer (Berlin: Springer Science and Business Media), 203–218. doi: 10.1007/978-1-4614-0959-5_16
- Brugger, P., Kollias, S. S., Müri, R. M., Crelier, G., Hepp-Reymond, M. C., and Regard, M. (2000). Beyond re-membering: Phantom sensations of congenitally absent limbs. *Proc. Natl. Acad. Sci. U.S.A.* 97, 6167–6172. doi: 10.1073/pnas.100510697
- Chancel, M., Iriye, H., and Ehrsson, H. H. (2022). Causal inference of body ownership in the posterior parietal cortex. *J. Neurosci.* 42, 7131–7143. doi: 10.1523/JNEUROSCI.0656-22.2022
- Ciaunica, A., and Crucianelli, L. (2019). Minimal self-awareness: From within a developmental perspective. *J. Conscious. Stud.* 26, 207–226.
- Ciaunica, A., and Fotopoulou, K. (2017). "The touched self: Psychological and philosophical perspectives on proximal intersubjectivity and the self," in *Embodiment, enaction, and culture: Investigating the constitution of the shared world*, eds C. Durt, T. Fuchs, and C. Tewes (Cambridge, MA: MIT Press).
- Ciaunica, A., Constant, A., Preissl, H., and Fotopoulou, K. (2021a). The first prior: From co-embodiment to co-homeostasis in early life. *Conscious. Cogn.* 91:103117. doi: 10.1016/j.concog.2021.103117
- Ciaunica, A., Safron, A., and Delafield-Butt, J. (2021b). Back to square one: From embodied experiences in utero to theories of consciousness. *PsyArxiv [Preprint]* doi: 10.31234/osf.io/zspm2
- Craig, A. (2002). How do you feel? Interoception: The sense of the physiological condition of the body. *Nat. Rev. Neurosci.* 3, 655–666. doi: 10.1038/nrn894
- Crucianelli, L., and Ehrsson, H. H. (2023). The role of the skin in interoception: A neglected organ? *Perspect. Psychol. Sci.* 18, 224–238. doi: 10.1177/17456916221094509
- Damasio, A. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York, NY: Harcourt Brace.
- Ehrsson, H. H. (2020). "Multisensory processes in body ownership," in *Multisensory perception: From laboratory to clinic*, eds K. Sathian and V. S. Ramachandran (Amsterdam: Elsevier Academic Press), 179–200.
- Ehrsson, H. H. (2023). "Ch 15 Bodily illusions," in *The Routledge handbook of bodily awareness*, eds A. J. Alsmith and M. R. Longo (New York, NY: Taylor and Francis).
- Ehrsson, H. H., Spence, C., and Passingham, R. E. (2004). That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb. *Science* 305, 875–877. doi: 10.1126/science.1097011
- Feinstein, J. S., Khalsa, S. S., Yeh, H., Al Zoubi, O., Arevian, A. C., Wohlrab, C., et al. (2018). The elicitation of relaxation and interoceptive awareness using floatation therapy in individuals with high anxiety sensitivity. *Biol. Psychiatry* 3, 555–562. doi: 10.1016/j.bpsc.2018.02.005
- Finger, S., and Hustwit, M. P. (2003). Five early accounts of phantom limb in context: Paré, Descartes, Lemos, Bell, and Mitchell. *Neurosurgery* 52:686. doi: 10.1227/01.neu.0000048478.42020.97
- Frankfurt, H. (1988). *The importance of what we care about: Philosophical essays*. Cambridge: Cambridge University Press.
- Frith, C. D. (2015). *The cognitive neuropsychology of schizophrenia (Classic Edition)*. Psychology Press.
- Fuchs, T. (2013). "The phenomenology of affectivity," in *The Oxford handbook of philosophy and psychiatry*, eds K. Fulford, K. W. Musgrave, M. Davies, R. Gipps, G. Graham, J. Sadler, et al. (Oxford: Oxford University Press), 612–631.
- Gallagher, S. (2000a). Philosophical conceptions of the self: Implications for cognitive science. *Trends Cogn. Sci.* 4, 14–21. doi: 10.1016/S1364-6613(99)01417-5
- Gallagher, S. (2000b). "Self-reference and schizophrenia: A cognitive model of immunity to error through misidentification," in *Exploring the self*, ed. D. Zahavi (Amsterdam: John Benjamins), 203–239. doi: 10.1075/aicr.23.14gal
- Gallagher, S. (2004). Neurocognitive models of schizophrenia: A neurophenomenological critique. *Psychopathology* 37, 8–19. doi: 10.1159/000077014

- Gallagher, S. (2017). Embodied intersubjective understanding and communication in congenital deafblindness. *J. Deafblind Stud. Commun.* 3, 46–58.
- Gallagher, S. (2022). “Bodily self-awareness and body-schematic processes,” in *Handbook of bodily awareness*, eds A. Alsmith and M. Longo (London: Routledge), 137–149. doi: 10.4324/9780429321542-14
- Gallagher, S., and Cole, J. (1995). Body schema and body image in a deafferented subject. *J. Mind Behav.* 16, 369–390.
- Gallagher, S., and Shear, J. (eds) (1999). *Models of the self*. Exeter: Imprint Academic.
- Gentile, G., Guterstam, A., Brozzoli, C., and Ehrsson, H. H. (2013). Disintegration of multisensory signals from the real hand reduces default limb self-attribution: An fMRI study. *J. Neurosci.* 33, 13350–13366. doi: 10.1523/JNEUROSCI.1363-13.2013
- Georgie, Y. K., Schillaci, G., and Hafner, V. V. (2019). “An interdisciplinary overview of developmental indices and behavioral measures of the minimal self,” in *Proceedings of the 2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, (Oslo: IEEE), 129–136. doi: 10.3389/fnbot.2020.00005
- Ghaffari, F., Taheri, M., Meyari, A., Karimi, Y., and Naseri, M. (2022). Avicenna and clinical experiences in Canon of Medicine. *J. Med. Life* 15, 168–173. doi: 10.25122/jml-2021-0246
- Gilson, E. (1929-1930). Les sources gréco-arabe de l’augustinisme avicennisant. *Arch. d’histoire doctrinale et littéraire du moyen âge* 4, 5–149.
- Guterstam, A., Collins, K. L., Cronin, J. A., Zeberg, H., Darvas, F., Weaver, K. E., et al. (2019). Direct electrophysiological correlates of body ownership in human cerebral cortex. *Cereb. Cortex* 29, 1328–1341.
- Hasse, D. (2000). *Avicenna’s de anima in the Latin West*. London: Warburg Institute.
- Horak, F. B., and Hlavacka, F. (2001). Somatosensory loss increases vestibulospinal sensitivity. *J. Neurophysiol.* 86, 575–585.
- Hume, D. (1739/1978). *A treatise of human nature*, ed. A. Selby-Bigge (Oxford: Oxford University Press).
- Humphrey, N. (2022). *Sentience: The invention of consciousness*. Oxford: Oxford University Press.
- Husserl, E. (1991). *On the phenomenology of the consciousness of internal time (1893–1917), collected works IV*, trans. J. Brough. Dordrecht: Kluwer Academic.
- Hutto, D., and Ilundáin-Agurriza, J. (2020). Selfless activity and experience: Radicalizing minimal self-awareness. *Topoi* 39, 509–520.
- James, W. (1890). *The Principles of Psychology*. New York, NY: Dover, 1950.
- Kaldewaij, R., Salamone, P., Enmalm, A., Östman, L., Pietrzak, M., Karlsson, H., et al. (2023). *Ketamine reduces the neural distinction between self and other-produced affective touch—a double-blind placebo-controlled study*. Available online at: <https://psyarxiv.com/w3ftk/download?format=pdf> (accessed November 26, 2023).
- Kaukua, J. (2015). *Self-awareness in Islamic philosophy: Avicenna and beyond*. Cambridge: Cambridge University Press.
- Kaukua, J. (2020). The flying and the masked man, one more time: Comments on Peter Adamson and Fedor Benevise, ‘The thought experimental method: Avicenna’s flying man argument’. *J. Am. Philos. Assoc.* 6, 285–296. doi: 10.1017/apa.2019.52
- Khalsa, S. S., Rudrauf, D., Feinstein, J. S., and Tranel, D. (2009). The pathways of interoceptive awareness. *Nat. Neurosci.* 12, 1494–1496.
- Kim, N., and Effken, J. A. (2022). Disturbance of ecological self and impairment of affordance perception. *Front. Psychol.* 13:925359. doi: 10.3389/fpsyg.2022.925359
- Kyselo, M. (2016). The minimal self needs a social update. *Philos. Psychol.* 29, 1057–1065.
- Lane, T. (2020). The minimal self hypothesis. *Conscious. Cogn.* 85:103029.
- Lang, S., and Viertbauer, K. (2022). Self-consciousness explained—mapping the field. *Rev. Philos. Psychol.* 13, 257–276.
- Limanowski, J. (2014). What can body ownership illusions tell us about minimal phenomenal selfhood? *Front. Hum. Neurosci.* 8:946. doi: 10.3389/fnhum.2014.00946
- Limanowski, J., and Blankenburg, F. (2016). Integration of visual and proprioceptive limb position information in human posterior parietal, premotor, and extrastriate cortex. *J. Neurosci.* 36, 2582–2589.
- Lymer, J. (2011). Merleau-Ponty and the affective maternal-foetal relation. *Parrhesia J. Crit. Philos.* 13, 126–143.
- Marmura, M. (1986). Avicenna’s “flying man” in context. *Monist* 69, 383–395.
- Mason, O. J., and Brady, F. (2009). The psychotomimetic effects of short-term sensory deprivation. *J. Nerv. Ment. Dis.* 197, 783–785. doi: 10.1097/NMD.0b013e3181b9760b
- Melzack, R., and Bromage, P. R. (1973). Experimental phantom limbs. *Exp. Neurol.* 39, 261–269.
- Merleau-Ponty, M. (2012). *Phenomenology of perception*. London: Routledge.
- Miall, R. C., Afanasyeva, D., Cole, J. D., and Mason, P. (2021). Perception of body shape and size without touch or proprioception: Evidence from individuals with congenital and acquired neuropathy. *Exp. Brain Res.* 239, 1203–1221. doi: 10.1007/s00221-021-06037-4
- Moro, V., Scandola, M., and Aglioti, S. M. (2022). What the study of spinal cord injured patients can tell us about the significance of the body in cognition. *Psychon. Bull. Rev.* 29, 2052–2069.
- Nelson, B., Parnas, J., and Sass, L. (2014). Disturbance of minimal self (ipseity) in schizophrenia: Clarification and current status. *Schizophr. Bull.* 40, 479–482.
- Paqueron, X., Leguen, M., Rosenthal, D., Coriat, P., Willer, J. C., and Danziger, N. (2003). The phenomenology of body image distortions induced by regional anaesthesia. *Brain* 126, 702–712.
- Pick, S., Rojas-Aguiluz, M., Butler, M., Mulrenan, H., Nicholson, T. R., and Goldstein, L. H. (2020). Dissociation and interoception in functional neurological disorder. *Cogn. Neuropsychiatry* 25, 294–311. doi: 10.1080/13546805.2020.1791061
- Ratcliffe, M. (2017). “Selfhood, schizophrenia, and the interpersonal regulation of experience,” in *Embodiment, enaction, and culture: Investigating the constitution of the shared world*, eds C. Durt, T. Fuchs, and C. Tewes (Cambridge, MA: MIT Press).
- Riemer, M., Wolbers, T., and Kuehn, E. (2019). Preserved multisensory body representations in advanced age. *Sci. Rep.* 9:2663. doi: 10.1038/s41598-021-81121-x
- Rudrauf, D., Lachaux, J., Damasio, A., Baillet, S., Hugueville, L., Martinerie, J., et al. (2009). Enter feelings: Somatosensory responses following early stages of visual induction of emotion. *Int. J. Psychophysiol.* 72, 13–23. doi: 10.1016/j.ijpsycho.2008.03.015
- Sass, L. A. (1998). Schizophrenia, self-consciousness and the modern mind. *J. Conscious. Stud.* 5, 543–565.
- Siderits, M., Thompson, E., and Zahavi, D. (eds) (2011). *Self, no self? Perspectives from analytical, phenomenological, and Indian traditions*. Oxford: Oxford University Press.
- Simmel, M. L. (1961). The absence of phantoms for congenitally missing limbs. *Am. J. Psychol.* 74, 467–470. doi: 10.2307/1419756
- Soultz, N. (2023). Kinaesthesia revisited: Kinaesthetic sensation and its temporal asymmetry. *J. Br. Soc. Phenomenol.* 54, 71–90.
- Strawson, G. (1997). The self. *J. Conscious. Stud.* 4, 405–428.
- Strawson, G. (1999). The self and the SESMET. *J. Conscious. Stud.* 6, 99–135.
- Strawson, G. (2004). Against narrativity. *Ratio* 17, 428–452.
- Vosburg, R., Fraser, N., and Guehl, J. (1960). Imagery sequence in sensory deprivation. *AMA Arch. Gen. Psychiatry* 2, 356–357.
- Williford, K. (2013). Husserl’s hyletic data and phenomenal consciousness. *Phenomenol. Cogn. Sci.* 12, 501–519. doi: 10.1007/s11097-013-9297-z
- Zahavi, D. (2017). “Thin, thinner, thinnest: Defining the minimal self,” in *Embodiment, enaction, and culture: Investigating the constitution of the shared world*, eds C. Durt, T. Fuchs, and C. Tewes (Cambridge, MA: MIT Press), 192–199.
- Zahavi, D., and Parnas, J. (1998). Phenomenal consciousness and self awareness: A phenomenological critique of representational theory. *J. Conscious. Stud.* 5, 687–705.
- Zippel, N. (2014). “The hyletic time-consciousness and the embodied subject,” in *Corporeity and affectivity: Dedicated to Maurice Merleau-Ponty*, eds K. Novotny, P. Rodrigo, J. Slatman, and S. Stoller (Leiden: Brill), 35–48.



OPEN ACCESS

EDITED BY

Xerxes D. Arsiwalla,
Wolfram Research, Inc., United States

REVIEWED BY

Jon Mallatt,
Washington State University, United States
Edmundo Lopez Sola,
Neuroelectrics, Spain

*CORRESPONDENCE

Asger Kirkeby-Hinrup
✉ asger.kirkeby-hinrup@fil.lu.se

RECEIVED 20 November 2023

ACCEPTED 27 February 2024

PUBLISHED 14 March 2024

CITATION

Kirkeby-Hinrup A (2024) Quantifying
empirical support for theories of
consciousness: a tentative methodological
framework.

Front. Psychol. 15:1341430.
doi: 10.3389/fpsyg.2024.1341430

COPYRIGHT

© 2024 Kirkeby-Hinrup. This is an open-
access article distributed under the terms of
the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Quantifying empirical support for theories of consciousness: a tentative methodological framework

Asger Kirkeby-Hinrup*

Department of Philosophy, Lund University, Lund, Sweden

Understanding consciousness is central to understanding human nature. We have competing theories of consciousness. In interdisciplinary consciousness studies most believe that consciousness can be naturalized (i.e., consciousness depends in some substantial way on processes in — or states of — the brain). For roughly two decades, proponents of almost every theory have focused on collecting empirical support for their preferred theory, on the tacit assumption that empirical evidence will resolve the debates. Yet, it remains unclear *how* empirical evidence can do this *in practice*. Here I address this issue by offering (a sketch of) a methodology to quantify the divergent sets of empirical support proposed in favor of extant theories of consciousness. This in turn forms the foundation for a process of inference to the best explanation inspired by Bayesian confirmation theory. In interdisciplinary consciousness studies we are blessed with an abundance of theories, but we have reached a point where, going forward, it would be beneficial to focus on the most promising ones. Methods for assessment and comparison are necessary to identify which those are. While future refinement is likely, the methodology for assessment and comparison proposed here is a first step toward a novel way of approaching this through a quantification of empirical support for theories of consciousness.

KEYWORDS

consciousness, inference to the best explanation, Bayesian updating, empirical evidence, theories of consciousness, assessment, comparison

1 Introduction

The field of interdisciplinary consciousness studies (ICS) — i.e., work at the intersection between philosophy of mind, psychology, cognitive science, and neuroscience — has been blossoming over the last decades. Yet, the current state of the field of ICS is precarious, and further development is necessary. In other words, we do not want to remain forever in the current stage of our field, in which we have dozens of theories and no noncontentious way of deciding between them. A positive upshot of this issue has been several proposals of how to assess and compare theories. The (sketch of a) methodology I offer in this paper is a novel proposal for this.

ICS converges (approximately) on the belief that understanding the brain's role in relation to consciousness is central to understanding consciousness *per se*, as well as its associated concepts (e.g., experience, cognition, meta-cognition, emotion, action, and perception). As Weisberg (2014, p. 433) writes: “[...] rooted in empirical data. This is the proper way to approach consciousness.” Weisberg is not alone in this sentiment. In ICS the shared assumption

is that empirical data carries evidential weight in determining the plausibility of a theory of consciousness. But how do we compare the evidential weights of the competing sets of empirical evidence proposed in favor of extant theories of consciousness? Most theories of consciousness on the market are internally consistent conceptual frameworks that propose mechanism (s) underpinning phenomenal consciousness (Doerig et al., 2020 for a useful classification of the different kinds of proposed mechanism; see also Sattin et al., 2021; Signorelli et al., 2021; Schurger and Graziano, 2022). Presently, the field of consciousness studies offers a wide variety of theories [e.g., the Global Workspace Theory of Baars (1996); the first-order theory of Block (1995); the Dispositional Higher-order theory of Carruthers (1998); the Same-Order Metarepresentational Account of Cleeremans et al. (2020); the Global Neuronal Workspace Theory of Dehaene and Naccache (2001); the Predictive Processing Theory of Friston (2013); the Wide Intrinsicity View of Gennaro (1996); the Same-Order Monitoring theory of Kriegel (2007); the Recurrent Processing theory of Lamme (2004); the Attention to Intermediate Representation theory of Prinz (2005); the Higher-Order Thought theory of Rosenthal (1997); the Integrated Information Theory of Tononi et al. (2016); the Higher-Order Global state theory of Van Gulick (2004), to name a few]. While they can be grouped into different ‘families’, they mostly offer mutually exclusive explanations of the structure and function (s) of consciousness (at least they supposedly do. For further discussion see Kirkeby-Hinrup et al., 2023).

Broadly speaking, the questions related to consciousness fall into two distinct domains: the first concerns information processing and behavior (cognitive domain); the second concerns the experience of being — or *what it is like to be* (Nagel, 1974) — conscious (phenomenal domain). Current theories largely agree about the cognitive domain, at least with respect to functional characteristics and behavioral predictions, but they differ with respect to the phenomenal domain. In fact, a major fault line in the debates between theories of consciousness concerns the nature and importance of phenomenality (i.e., what-it-is-like to be conscious). This question roughly divides the field into two camps: proponents of *deflationary* accounts (Rosenthal, 2008, 2012) and those who advance *inflationary* accounts (Block, 2011b). The latter sees phenomenality as widespread in — and central to — consciousness, whereas the former denies this. Yet, both deflationary and inflationary accounts tend to use the same vocabulary, a problem noted by Rosenthal who says: “The phrase ‘what it’s like’ is not reliable common currency” (Rosenthal, 2011, p. 434). When competing theories each are internally consistent, describe the target phenomenon using many of the same concepts — yet disagree about what those concepts actually mean — there is little avenue on conceptual grounds to determine which theory is correct, or even preferable. This has left the conceptual debate largely gridlocked because it is difficult to criticize a theory without begging the question against its underlying conceptual framework. Thus, it is unclear at best if there is an avenue forward in arguing about consciousness solely on conceptual grounds.

However, because most people involved in these debates share the assumption that consciousness can be naturalized (i.e., consciousness depends on physical processes, assumed to occur primarily in the brain), the hope is that empirical evidence may resolve these disagreements by determining which theory is more empirically plausible. Consequently, in recent decades there has been a radical increase in the application of empirical evidence in support of — or to

argue against — theories of consciousness (c.f. Yaron et al., 2022, p. Figure 2b). Proponents of most theories have advanced empirical evidence to illustrate its explanatory power, and/or scaffold its claim to plausibility on a general level. This is reasonable standard scientific practice, and overall a good approach. However, in the last couple of years, attention has turned to how — or whether — empirical evidence actually may do the work for us we hoped it would (determining which theory is most plausible/preferable). This attention has illuminated many issues with respect to how we collect, deploy, assess, and compare empirical evidence in ICS, as often cast in light of well-known considerations from the philosophy of science (Seth, 2009; Del Pin et al., 2021; Kirkeby-Hinrup and Fazekas, 2021; Overgaard and Kirkeby-Hinrup, 2021; Schurger and Graziano, 2022; Kirkeby-Hinrup, 2024). These issues pertain to whether — or how — empirical evidence can help us decide which theory is ultimately most plausible/preferable on a long-term perspective (i.e., which theory is closest to truth(s) about the world with respect to propositions about the phenomenon which we call “consciousness”). Furthermore, even on a short-term perspective do questions about the work empirical evidence can do for us appear. The current abundance of competing theories in ICS can only be a positive thing if there is a way to eliminate theories as part of our scientific process of approximating the truth.

I will, in the next section, examine two existing proposals to gauge the state of the field and set the appropriate context. One — based on *criteria* — proposed by Doerig et al. (2020) consists in assessing and comparing theories according to their explanatory scope and ability to handle principled problems. The other endeavor is of strictly empirical nature and turns on the notion of *adversarial collaboration*, i.e., getting proponents of competing theories to agree on an empirical paradigm on which their theories have differing predictions, and then performing the experiment.¹ In section three, I introduce the general context for my proposal, before presenting the details in section four. Finally, in the fifth section, I offer some concluding remarks.

2 Comparing theories of consciousness

How do we — based on empirical evidence — determine which theory of consciousness is preferable? Currently, there are two prominent approaches to this question (this paper proposes a third). The first approach operates on a principle similar to falsification. The second approach deploys a set of criteria to assess and compare theories of consciousness. Briefly considering each of these is appropriate here because understanding the strengths and/or shortcomings of existing approaches provides anchors for evaluation of the third approach I will present in sections three and four. Consequently, let us consider these in turn.

¹ Accelerating research on consciousness: an adversarial collaboration to test contradictory predictions of Global Neuronal Workspace and Integrated Information Theory; <https://www.templetonworldcharity.org/projects-database/accelerating-research-consciousness-adversarial-collaboration-test-contradictory>. The sizable grant this project received from the Templeton Foundation speaks to the growing recognition that there is a need for new and ambitious approaches to assessing and comparing theories if we are to make progress in interdisciplinary consciousness studies.

The first approach, operating on the principle of falsification, consists of a range of separate projects, and is called “Accelerating Research on Consciousness” (ARC). This enormous and ambitious project rightfully has drawn significant attention and praise in ICS. The methodological approach in ARC is the principle of adversarial collaboration, i.e., testing specific paradigms (agreed upon in advance by proponents of each theory) where competing theories predict different (supposedly concrete and mutually exclusive) empirical measurements. The results of each project are then taken to strengthen the theory whose prediction is confirmed, and (partly) falsify the other(s).

Recently, (Ferrante et al., 2023; Melloni et al., 2023) the results of the first project in ARC have been made public. In this project, predictions of Integrated information theory (IIT) (Tononi et al., 2016; Albantakis, 2020) was compared to those of the Global Neuronal Workspace theory (GNWT) (Mashour et al., 2020). The results were unclear, neither fully supporting either theory, nor fully falsifying either theory. Consequently, in terms of eliminating theories, or assessing which is preferable to the other, the first ARC project did little to move the needle between IIT and GNWT. In subsequent debate, proponents of both theories point to limitations in the data and reach opposing conclusions regarding the involvement of the prefrontal cortex (Ferrante et al., 2023).

However, even if this ARC project had provided — or if the next projects in ARC provide — more conclusive data, another problem remains. The problem is that it is standard scientific practice to revise theories in light of new evidence. So, failing to have your predictions confirmed is likely to be taken as an incentive to further develop a theory, rather than abandon it. That is; proponents of a theory are not immediately inclined to completely abandon a theory if it comes out unsuccessful in an ARC project. To boot, we do not know what the ‘threshold’ for amount — or quality — of evidence is for a theory to be abandoned. Put differently, it is unclear how many — or which kind of — ‘losses’ on ARC projects are sufficient for a theory to be abandoned by its proponents. Problematically however, there is a real risk that this may arbitrarily depend on the individual proponents of a theory. This may raise worries about whether ARC ultimately will be able to deliver results with the requisite ubiquity to falsify a theory to the extent that it is eliminated from further consideration by the field (this worry was echoed by Lucia Melloni when presenting the aforementioned first results of the ARC project at the 2023 ASSC conference in New York with the words: “No one changes their mind” with reference to the Daniel Kahneman, the originator of the adversarial collaboration idea, who had declined to participate in the presentation for that reason). In the long term, whether ARC will be able to change minds remains to be seen, but (assuming an interest in consciousness) it would certainly be in everyone’s interest if it can. Now, in addition to these overall worries (that apply to any way of assessing and comparing theories, including the one proposed below), there is a range of more concrete issues — of either a methodological or practical nature — with ARC. Call the first of these: *targeted theories*. ARC projects inherently treat only a subset of the theories (between two and four theories per project currently).³ This means ARC can never say something about the field as a whole, but only about some specific relation between a few

theories and some specific data. The second issue is that ARC has a *narrow scope*, in the sense that each comparison is based on one or a few paradigms.⁴ Barring some auxiliary framework, this restricts conclusions to the results of the few paradigms, precluding conclusions about overall plausibility.⁵ A third issue is methodological *generalizability*. There are two sides to this issue. The first side is practical, and derives from the fact that, in ARC, paradigms and pipelines deployed to test theory A and B, cannot be applied to test theory C and D. This makes ARC very (time, expertise, money) cost intensive. The second side is methodological; because we are not in a situation where one paradigm ‘fits all’, it is unclear how to compare results from different ARC projects. For instance, if project 1 confirms theory A over theory B, and project 2 confirms theory B over A, which should we prefer?⁶ The fourth issue concerns the *robustness* of ARC results. Because of their specificity, the results from ARC are very sensitive to changes in theories. Therefore, if (aspects of) theory A is revised to account for a failed prediction in an ARC experiment, this will require a whole new ARC project to assess the revised version of the theory. Since revising theories in light of new evidence is standard scientific practice, one would expect such revisions to happen. Finally, the fifth issue is the *cost* of ARC. In line with its ambitious and comprehensive approach (Ferrante et al., 2023; Melloni et al., 2023) the current ARC projects require significant human, financial and institutional resources. On the one hand, this speaks to the scientific rigor, ambitiousness, and effort of ARC. On the other hand, the cost of ARC is prohibitive to the vast majority of researchers in the field, which means it is unlikely to be broadly adopted. The previously discussed issues of *generalizability* and *robustness* further compounds the *cost* issue, since every time a theory is revised (*robustness*, for instance due to results from an ARC project) we need a new tailor made (due to *generalizability*) multi-year multimillion dollar project to assess the new version. This is a steep cost and should raise worries about the long-term feasibility of the ARC approach (especially, if we do not even know what it would take for someone to change their mind).

The second major approach consists in developing and deploying a set of criteria to evaluate and compare theories. The criteria based approach (CRIT) has been advanced by Doerig et al. (2020). They propose two categories of criteria for assessment (e.g., table in Doerig et al., 2020, p. 48). The first category, they dub *criteria*. This category consists of four challenges a theory of consciousness may face depending on the hypothesized mechanisms underpinning consciousness. The second category Doerig et al. call *scope*. Here, they propose to deploy five classical distinctions about consciousness to assess which aspects of the phenomenon are covered by a given theory. CRIT has already been the subject of much debate (Doerig et al., 2021). Here, I highlight four issues that are of particular relevance in the present context. The first of these issues concerns CRIT’s *sensitivity* to empirical evidence. The issue is that CRIT ignores the amount of empirical support of theories outside of satisfying criteria, or the amount of empirical evidence a theory’s meeting of a criterion relies on. While many of the proposed criteria are framed against an empirical background, CRIT only superficially takes into account

2 Where “kind” can be understood either as type of evidence or as strength of evidence.

3 <https://www.templetonworldcharity.org/accelerating-research-consciousness-our-structured-adversarial-collaboration-projects> for further info.

4 Three in the first project (Ferrante et al., 2023).

5 Observe that in a situation where no one ever changes their mind, it is useful to be able to assess overall plausibility because this allows us to still say something about which theories are preferable.

6 This methodological issue is — of course — not unique to ARC.

actual empirical evidence proposed in favor of theories. This means a theory with a lot of empirical support will be scored as equal to a theory with almost no empirical support as long as they satisfy the same criteria. Similarly, CRIT does not take into account the amount of empirical evidence a theory's meeting of a criterion relies on. Theory A is scored as equal to theory B as long as they satisfy the same number of criteria, regardless of their respective sets of evidence. The second issue concerns *arbitration* between theories. Suppose two or more theories satisfy the same number of criteria, how do we decide between them? Given the limited number of criteria and the limited grading system on each criterion (e.g., table in Doerig et al., 2020, p. 48), the possibility of ties is high. *Arbitration* concerns not only how to decide between two or more theories that satisfy identical sets of criteria, but also how we should decide between two or more theories that satisfy the same number of criteria without their sets being identical. In other words, we need to know how to weigh satisfying criterion A against satisfying criterion B. CRIT is certainly useful for an overall classification of theories, but because it is not sensitive to divergent amounts of support, it is insufficient for any fine-grained comparison of theories. The third issue concerns the *flexibility* of the criteria. Now, Doerig and colleagues are explicit that the current set of criteria is not intended to be exhaustive (Doerig et al., 2020, p. 42) and will likely need expansion.⁷ But how many — and which — criteria can we add? One might hope that the answer to this question is that any further criteria will be obvious, and we will come upon all — or most of — these over time (which in turn limits the maximum possible number of criteria as well). Observe, this answer may lead to a debate about what “obvious” entails, to whom it will be obvious, and who gets to decide these questions. This is the fourth issue: *arbitrariness*. For now, I will leave *arbitrariness* to the side since this issue will loom large throughout the text, and instead focus briefly on another upshot of *flexibility*; namely the question of how many criteria we will need to distinguish convincingly between theories (assuming we even can do this in a non-arbitrary way). Presently, any speculation on an exact number of further criteria would be premature. But given that the present set of criteria makes ties likely, it is likely to need expansion in the future. The next thing to note is that the set of criteria that are theory-neutral, obvious, overarching, and important is likely limited (however see, Rosenthal, 2021). This limitation would make any further criteria less central than the nine currently proposed. One reason for thinking this is that, if there were indeed further obvious and important criteria, Doerig and colleagues would have included them in their paper.⁸ Be that as it may, it nevertheless is likely that a future expansion of CRIT will result in increasingly detailed criteria of less and less importance. One positive upshot of adding more criteria is that it seems CRIT may be able to deal with *arbitration* since ties will be less likely⁹ as the number of criteria increases. However, this at the same time would undermine the main appeal of CRIT, i.e., identifying the *overarching principled* criteria a theory of consciousness should satisfy.

In the rest of this paper, I will present a third approach to assessing and comparing theories based on the notion of inference to the best explanation (IBE). Importantly, while I have identified shortcomings of both ARC and CRIT (*targeted theories*, *generalizability*, *robustness*, *cost* and *sensitivity*, *arbitration*, *flexibility*, *arbitrariness*, respectively), and will show that the approach proposed here does not have these shortcomings, I am not advocating that ARC and CRIT have no value, let alone should be given up. The approach here is intended to complement, rather than supplant, ARC and CRIT. There is room for these three approaches, not only to coexist, but to develop also a positive synergy. I will return to this in the concluding remarks.

3 Inference to the best explanation

In the previous section, I discussed two contemporary approaches to assessing and comparing theories. Previously, together with Peter Fazekas (Kirkeby-Hinrup and Fazekas, 2021), I have advocated a third approach based on the notion of inference to the best explanation. Looking at the publications over the last couple of decades, an IBE process seems tacit in much of the work concerned with the relation between empirical evidence and theories of consciousness. One of the most explicit invocations of IBE can be found in the work of Ned Block (2007, p. 486) when he says: “I have in mind [...] the familiar default ‘method’ of inference to the best explanation, that is, the approach of looking for the framework that makes the most sense of all the data [...]” Yet, to my knowledge, outside of my proposal with Fazekas, no one has endeavored to attempt inference to the best explanation in practice in ICS. One reason may be that classical versions of IBE are ill-suited for straightforward application in our situation.¹⁰ This is because we cannot compare theories on their explanatory powers, because there is no consensus on a common explanandum. To elaborate, competing theories do not necessarily have identical explanatory targets (Sattin et al., 2021; Signorelli et al., 2021; Yaron et al., 2021), yet are taken to be mutually exclusive for the reason that they all target the same phenomenon (and they share the assumption that there is only one phenomenon). In the vocabulary of Chalmers (2002), theories differ in their ‘intension’ of the explanandum (the meaning of the word ‘consciousness’), but coincide on its ‘extension’ (the thing in the world picked out by the word ‘consciousness’). In a way, when deploying empirical evidence in assessing and comparing theories of consciousness, we are hoping to resolve disagreements on the *intension* through investigations of the *extension*. The upshot is that we cannot adopt explanatory power as our metric for comparison, since explanatory power depends on the ‘intension’ of the explanandum, which means we would be comparing apples and oranges. Therefore, we must perform IBE on the basis of some other metric than explanatory power. One way to approach this is by collating the respective sets of proposed empirical support of the competing theories, to determine if our observations about the *extension* (empirical evidence) conform to a proposed *intension* (a theory), and how well.

The notion of IBE (sometimes understood as co-extensive with the notion of abduction) is a classic topic in the philosophy of science (Burks, 1946; Harman, 1965; Peirce and Hartshorne, 1974;

⁷ There are already candidates for further criteria (Overgaard and myself have proposed one Kirkeby-Hinrup and Overgaard, 2023).

⁸ This is tenuous of course, given that there may be a range of other reasons elements were not included in a paper.

⁹ Assuming it would not make sense to add a criterion that every theory satisfies.

¹⁰ Another plausible reason is the lack of datasets necessary to carry out IBE, see Section 4.

Minnameier, 2004, 2010; Campos, 2011; Douven, 2021). But, for the reason just given, classical notions are not straightforwardly applicable in our case. Therefore, some clarification is necessary with respect to the way IBE is conceived of here. Firstly, the ‘explanations’ we need to infer to are theories of consciousness (what Block called “frameworks” in the quote above. In Chalmers’ vocabulary, the different proposed *intensions* of consciousness). Secondly, the assessment and comparison are not based on *explanatory* considerations, *per se*. The metric for assessment — and what is being compared — is not a theory’s explanatory power in relation to its targeted explanandum (its *intension*). As just noted, comparing explanatory power in relation to *intension* of the explanandum is problematic because there is no agreement on what a good explanation would entail, because there is no agreement on the exact characteristics of the phenomenon (see, e.g., debate in Rosenthal, 2011; Weisberg, 2011; Block, 2011a,c). To avoid this, the IBE approach could consist in assessing and comparing (i.e., inferring on the bases of) the explanatory power of theories in the empirical domain. In other words, the metric of comparison in this proposal is the ability to explain and predict empirical data.

There are many ways to develop an IBE process. Fazekas and I (Kirkeby-Hinrup and Fazekas, 2021) proposed a four-step process relying on the fact that the first step (assimilation, i.e., data collection) was already far along,¹¹ argued the importance — and demonstrated the feasibility — of the second and third step: compilation and validation (respectively concerned with compiling the proposed evidence for each theory, and validating claims of empirical support on a case-by-case basis). To elaborate, in addition to demonstrating that the second and third steps are feasible by showing what they look like in practice, the main point of the paper was that if we want to decide between theories on the bases of their respective empirical support, we had better know what their respective empirical support is,¹² and whether any given piece of empirical support claimed by a

theory, in fact supports the theory. The step of the IBE process that is developed below is the one we did not treat in that paper, namely the actual comparison of theories,¹³ the fourth and final step. Since the proposal here depends on quantifying the competing sets of proposed evidence, I will call this approach *Quantification to the Best Explanation* (QBE).

3.1 An intuition about weights of evidence

An initial desideratum is that QBE should avoid the identified shortcomings of ARC and CRIT (*targeted theories, generalizability, robustness, cost, sensitivity, arbitration, flexibility, and arbitrariness*) discussed in the previous section. While QBE avoids many of these easily, two warrant consideration here,¹⁴ namely: *sensitivity* and *arbitration*. These two shortcomings appear to threaten QBE and CRIT equally. To clarify, the sets of empirical evidence proposed for the extant theories of consciousness are *prima facie* incommensurable. One source of the incommensurability is that the sets of empirical support for each of the theories — while in many cases partially overlapping — do not contain exactly the same elements. Thus, the *arbitration* issue reappears on IBE, because we now need a way to weigh the non-overlapping elements against each other. To illustrate, it is unclear whether theory A being supported by the change blindness phenomenon is more “valuable” (as it were) than theory B’s support from the split-brain phenomenon. Yet, many share the intuition that *some* instances of empirical evidence should weigh heavier than others. This raises at least two questions: First: what is the driver of this intuition? And second: which instances? Let us focus here on the first question (the second question will be addressed in subsequent sections). Leaving open whether there are others, here are at least two possible candidate drivers of the intuition that some evidence is more “valuable” (should weigh heavier in IBE) than other (Figure 1).

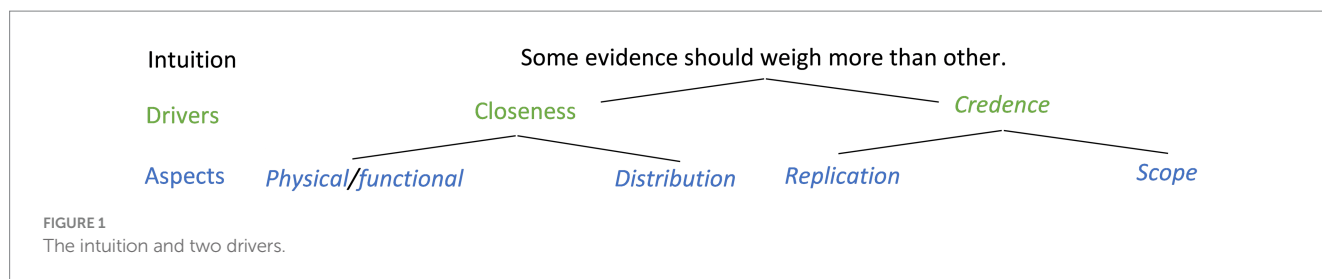
The first candidate as a driver is that the ‘closer’ (applicable) a piece of evidence (a phenomenon) is to the normal human condition (i.e., consciousness *as such*, or consciousness in neurotypical adults) the higher weight it should be ascribed in a comparison process. Call this the *closeness* driver. Closeness could be understood as *physical/functional* closeness, suggesting that studies with human subjects are more “valuable” than animal studies or computational models. Another example of *physical/functional* closeness could be the intuition that studies on neurotypical brains are preferable to studying very rare cases of brain trauma or cognitive dysfunction. Another way to understand *closeness* could be as *distribution*, which can be subdivided into inter-individually and temporally, where the former tracks the number of individuals to which the phenomenon applies and the latter tracks *how often* an individual or group of individuals instantiate the phenomenon. Accordingly, phenomena with high

11 We suggested the first stage — in which proponents of theories collected empirical support — had already been ongoing for a couple of decades, and therefore there was no need to demonstrate its feasibility.

12 Given the complexity of debates between competing theories of consciousness, it is not sufficient to select reports in which empirical evidence is proposed explicitly in support of a theory. In fact, significant (and increasing) parts of the academic exchange between proponents of competing theories take place in so-called proxy debates. Proxy debates, as the name indicates, are not directly about the theories. Instead, they are about specific empirical phenomena or aspects of consciousness, on which the positions taken in the debate are (sometimes tacit) extensions of central views of particular theories. Therefore, when proxy debates deploy empirical evidence, this should be seen as part of the empirical support for a theory. Over the last couple of decades the occurrence of such proxy debates has been relatively steady. Examples include the debates about unconscious perception (Brogaard, 2011; Block, 2016; Peters et al., 2017), non-conceptual content (Brinck, 1999; Jacobson and Putnam, 2016), whether perception is rich or sparse (Kouider et al., 2010; Block, 2011b, 2014b; Knotts et al., 2019), and perceptual precision (Block, 2014a; Prettyman, 2019). A prominent ongoing proxy debate concerns the localization of the neural correlates of consciousness (usually called either the “front vs. back” debate or the “early vs. late” debate). For a small sample of this debate see, e.g., (Lamme, 2003, 2004; Bor and Seth, 2012; Meuwese et al., 2013; Frässle et al., 2014; Kozuch, 2014; Boly et al., 2017; Odegaard et al., 2017; Michel and Morales, 2020).

13 While the methodology proposed here caters to the fourth step and thereby complete the account offered by Fazekas and me, nothing in the below hinges on acceptance of our claims in that paper. Readers uninclined to this view of an IBE process may nevertheless find use for a methodology for quantifying empirical evidence and comparing sets of evidence.

14 The rest will be addressed in the concluding remarks.



distribution (inter-individually, temporally, or both) are ‘closer’ to the explanatory target (‘neurotypical adult consciousness’ because *many* experience the phenomenon *often*) and consequently should be given higher weight in the IBE process.

The second candidate driver of the intuition (that some evidence should weigh heavier than other) concerns our *credence* in the evidence in question. Call this the *credence* driver. According to *credence*, the extent of our knowledge of a given phenomenon seems to matter for the weight it should be ascribed in the IBE process. *Credence*, furthermore, can be subdivided at least into *replication* and *scope*. *Replication* concerns the robustness of our ways of knowing about the phenomenon, i.e., the total amount of studies conducted on it, the amount of replication studies, and the existence of well-established paradigms to investigate it. *Scope* concerns the number of angles we (could) have approached the phenomenon from, i.e., the range of empirical techniques (that can be) applied to it. To elaborate; *replication* considers that phenomena, that have been the subject of thousands of studies and on the basic features of which (independent of any specific theory of consciousness) there is a general consensus, are more valuable than phenomena that have only been recorded very few times and the interpretations of which are widely contentious *outside* of consciousness studies. *Scope*, on the other hand, concerns the number of empirical techniques that have been — or could be — used to investigate the phenomenon. According to *scope*, phenomena that have been measured in many ways (e.g., fMRI, EEG, PET, MEG, ECoG, fNIRS, eye blinks, saccades, eye-tracking, D’, Meta D’, reaction time, introspective report, perceptual awareness scale, to name a few) are more valuable than phenomena that only have been — or only can be — measured using a single or few techniques (e.g., phenomena relying solely on introspective report). Thus — overall on *credence* — the intuition would be that phenomena with either high replication, broad scope, or both should be given higher weight in QBE.¹⁵

¹⁵ Observe, I do not claim that *closeness* and *credence* are the only possible sources for an intuition that some instances of empirical support should be given more weight than others when comparing theories of consciousness. It is entirely possible that there are other sources for this intuition. Nevertheless, since this intuition appears to be rooted at least in a combination of our areas of interest (*closeness*) and scientific principles (*credence*), I will assume we can agree on one or more of the drivers *physical/functional closeness*, *distribution*, *replication*, and *scope*. Importantly, my proposal does not depend on whether the reader subscribes to this intuition (or all of its possible drivers). QBE can be deployed independently of this intuition (for instance as a feasible theory neutral way to assess and compare theories of consciousness, that does not have the shortcomings of CRIT and ARC).

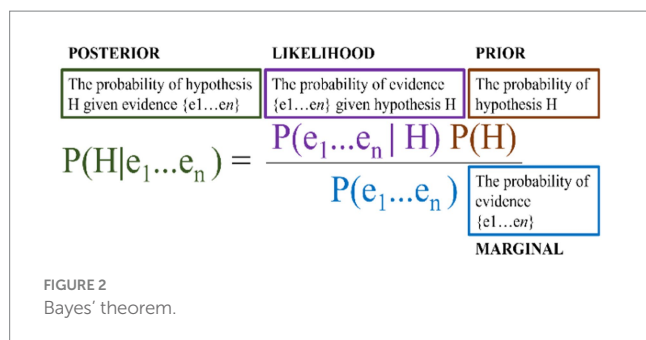
4 Bayesian inference to the best explanation

If we follow the intuition that some evidence should weigh heavier than other, a second shortcoming of the existing approaches that also is a challenge for QBE is *arbitrariness*. In this context, *arbitrariness* concerns *who* gets to assign the weights to the pieces of empirical evidence, and whether this can be done in a theory-neutral and non-contentious way. The “who” matters because, if the assignment of weights in QBE depends arbitrarily on the person performing the comparison, the objectivity of the process is compromised which, for obvious reasons, would be a bad thing. Consequently, a second desideratum for QBE is that it can deliver an objective way (one that does not depend on arbitrary choices of the person performing the comparison) to ascribe weight to proposed empirical support.

As an starting point, Fazekas and I (Kirkeby-Hinrup and Fazekas, 2021) suggested that — in order to stay neutral between theories when evaluating evidence — it is preferable to evaluate each piece of evidence in light of the conceptual framework of the theory it is applied to. This is because using any other conceptual framework (‘intension’ of consciousness) risks begging the question against the theory, i.e., presuming a viewpoint and thereby giving up on objectivity.

From this starting point, since we are aiming to quantify empirical support, we need a way to get numbers, via the sets of empirical evidence proposed in favor of extant theories. The conversion to numbers is made difficult by the way empirical evidence is proposed in ICS, *viz* what we are trying to quantify is really arguments to the effect that some piece of empirical evidence is predicted by — or can be explained in light of — a given theory. Such arguments, in turn, depend on the conceptual framework of the theory, and the mapping of an interpretation of some empirical data to this framework (see Kirkeby-Hinrup and Fazekas, 2021 for details).¹⁶ This kind of

¹⁶ Relevant questions in this regard include whether the same concepts are applied to — or operationalized in (Fink, 2016) — the interpretation in a uniform way. Whether there are equivocations or vagueness in the application of terms from the conceptual framework. It is crucial correctly to identify the theoretical claim defended, since this has implications for the way the argument is evaluated. Plainly, to assess an argument it is imperative to identify what it is an argument *for*. It matters for the assessment whether an argument is about overflow or recurrent processing (even if both pertain to RPT). To understand an argument in this context is to investigate how the proposed interpretations of the empirical evidence map onto the theoretical principle or concept, and illuminating (by extrapolation, if necessary) the argument connecting the



conceptual work does not allow for straightforward quantification (i.e., conversion into numbers).

Before we turn, in the next section, to solving this challenge, let us briefly consider how the numbers will be used once we have them. The inferential process for comparison that forms the core of QBE takes inspiration from Bayesian Confirmation Theory (BCT) (Gelman and Shalizi, 2013; Crupi, 2021) to estimate the strength of evidence in favor of each theory. A positive feature of BCT is that it delivers a posterior probability for each theory *given* all the evidence proposed in its favor and tells us that we should have more credence in higher posteriors than lower ones. I.e., theories with a high posterior are preferable to theories with a lower one.¹⁷ The core of BCT is Bayes' theorem (Figure 2), however, due to the nature of the data QBE quantifies, some accommodation of this is necessary. The rest of this section will be dedicated to clarifying one by one how to reconstruct and understand each of the elements in Bayes' theorem in the context of QBE.

First and foremost, in order to determine the Likelihood and the Marginal in Bayes' theorem we need to know what exactly the evidence ($e_1...e_n$) is in the present context. For QBE an "e" is a claim of empirical support. Such claims, in turn, are generally structured as arguments connecting a given empirical phenomenon with a theory of consciousness, aiming to show that — and/or how — the theory can either predict or explain an observation about the phenomenon. In other words, we are dealing with three components: (1) an empirical *phenomenon*, (2) an *observation* about it, and (3) an

interpretation of the empirical evidence to a theoretical claim. Depending on the principle/concept defended and the empirical evidence, the argument may take various forms. In some cases, a claim of empirical support deploys two or more empirical phenomena interwoven into a complex argument that requires careful analysis before it can be assessed (see, e.g., Brinck and Kirkeby-Hinrup, 2017). In other cases a concept is straightforwardly deployed to explain an empirical phenomenon, but even then, it is important to clarify all the details of the argument. What part of the argument is doing the explaining? Is the explanation reasonable? Are there any errors in the premises or the conclusion? Is there any vagueness or equivocation in the premises or conclusion? Is the interpretation of the empirical phenomenon true to the original empirical reports of the phenomenon? Is there any way we can test empirically whether the explanation offered by the theory is correct? Is the original empirical report framed directly in relation to the theory, or is it given an explanation post-hoc using the theory's framework?

¹⁷ A strength of QBE is that it concretizes the support of theories and simplifies comparison (numbers are easily graspable and easy to compare).

argument.¹⁸ Let us consider these in turn. *Prima facie*, the class of phenomena invoked by extant theories is very heterogeneous allowing many kinds of entries. Examples include pathological conditions such as visual neglect (at varying levels of description, e.g., psychological, behavioral, and physiological), neural processes such as recurrent processes (e.g., as biological, physical, or network-level descriptions), and behavioral phenomena such as visual masking (e.g., as methodology or behavioral descriptions). Consequently, the *phenomenon* concept in IBE must be very inclusive, since limiting the empirical evidence to certain types of phenomena would be arbitrary and risks undesirably biasing QBE against a theory. As a starting point, the *phenomenon* can be conceived of as the definition (and understanding of the network of concepts) through which we pick out the phenomenon in the scientific domains *outside* of consciousness studies. As such, the phenomenon is a (theory-neutral) label we deploy for some state of affairs in the world (purportedly connected to a theory). On this view, the set of proposed empirical evidence of a theory ($e_1...e_n$), is a collection of phenomena purportedly connected to the theory. Next, how does this notion of $e_1...e_n$ impact the likelihood? Traditionally, the likelihood is the probability of the evidence *given* the hypothesis. But in our case, there is no straightforward entailment relation from the evidence to the theory (hypothesis). Currently every theory is (radically) underdetermined by the evidence. Consequently, another connection is needed between the hypothesis and the evidence. One possibility is to conceive of the connection along the lines of a probability that the evidence is *as* the hypothesis explains it. In other words, the likelihood is the extent to which the theory explains or predicts the phenomenon. The next section will unpack this to lay the foundation for the quantification of evidence (that will be discussed in section 4.4).

4.1 The likelihood: from arguments to ordinal rankings

In this section, the objective is to construct the Likelihood variable of Bayes' theorem through the use of an intermediary ordinal categorization of a piece of evidence (each individual e in $e_1...e_n$). As a preliminary consideration, it is imperative to proceed from the assumption that the core principle (e.g., broadcasting in the workspace theories) and core concepts (e.g., overflow, and 'rich' phenomenality in recurrent processing theory) of a theory are valid when assessing evidence proposed of the theory. One reason for this is what is sometimes called *conceptual bleed*. Briefly, in order to make inferences for or against a theory from some empirical datum (the phenomenon) one needs, as a minimum, an interpretation that brings the concepts of the datum and the theory into a shared vocabulary; a kind of conceptual mapping. However, the conceptual mapping impacts (bleeds into) the possible inferences one can make from the observation(s) of the phenomenon. Furthermore, how one prefers to conceptualize and describe phenomena (the *intension* of the explanandum, i.e., consciousness) affects the mapping. This is a natural consequence of the conceptual and theoretical commitments

¹⁸ Part of which involves concepts proprietary to the framework of a theory.

of the researcher, who when interpreting relevant empirical data, will (reasonably) make use of the concepts she thinks best describe and categorize the phenomenon under investigation. Succinctly put, conceptual bleed means that commitments one has in the conceptual domain bleeds into, as it were, considerations and interpretations in the empirical domain. This makes it problematic to evaluate the proposed evidence for a theory “from the outside” (as it were), since the theory has bled into the evidence. One way to safeguard against this is to evaluate each theory on its own terms to avoid begging the question against its conceptual framework (as noted by Fazekas and me, see also above). Importantly, this does not mean that anything goes with respect to claims of empirical evidence. Previous work has shown errors that undermined proposed empirical support even assuming a theory’s conceptual framework. This is possible for instance by mischaracterizing the empirical data (e.g., D’Aloisio’s deployment of aphantasics’ performance on retro cue tasks, see Kirkeby-Hinrup and Fazekas, 2021 section 9) or in cases of unsound deductive arguments (Kirkeby-Hinrup, 2014).

Quantification then is the task of determining the value of a given set of evidence ($e_1 \dots e_n$) which in turn requires determining the value of each piece of evidence (each individual “e” in the set of evidence proposed in favor of a theory),¹⁹ where “value” means “how much” a given *phenomenon* supports a theory (and “how much,” in turn, entails coming up with an actual number). Since there was no immediate way of coming to numbers directly from arguments based on observations, an intermediary element is needed to facilitate the translation. The rest of this section will develop a proposal for this intermediary element. The approach is to categorize arguments on an ordinal scale, which can then serve as an anchor for the actual quantification of evidence (discussed in section 4.4). Categorization of arguments essentially involves assessing them according to some criteria to determine their place on the ordinal scale. For ease of exposition, I will call the result of this assessment the “A-score” of the argument. In addition to facilitating the placement on the ordinal scale, such assessment serves to satisfy a prerequisite for any IBE process, namely determining whether the evidence does in fact support the theory. This is critical since, clearly, we should not count a piece of empirical evidence in favor of a theory unless it in fact supports the theory. So, initially, what is at stake here is whether the proposed connection between a piece of empirical data and a theory of consciousness is sound. Now, if the phenomenon can in fact support the theory (i.e., the argument is coherent), we want some gauge of the amount of support it can lend to the theory, i.e., to assess how *good* the argument is. But what exactly does “good” mean in this context?

Assuming that the argument is sound, and that other pitfalls are avoided (see Table 1) so we can say a phenomenon in fact supports a theory, there are two parameters we can deploy to assess how good a piece of support is. The first is theory-neutral vocabulary, and the second is testability. For instance, it is possible to mount a coherent argument that is nevertheless cached in the conceptual framework of

TABLE 1 The A-score.

| | |
|-------------------------|---|
| Rejected | The <i>phenomenon</i> is incorrectly represented and/or the interpretation of the <i>observation</i> is faulty and/or the <i>argument</i> based on the interpretation is not sound. |
| Coherent but untestable | The concepts deployed in the interpretation of the <i>observation</i> do all the explanatory work. There is no way to test the interpretation — using the exact same empirical <i>phenomenon</i> — that does not rely on presuming the theory and/or concept. |
| Coherent and testable | The interpretation of the <i>phenomenon</i> is testable in principle without presuming the entire theory and/or all concepts deployed in the interpretation. |
| Accepted | The phenomenon has been tested and the argument is sound, and both align with the central principle of the theory, or the defended concept. |

a theory to an extent, where all the explanatory work is done entirely by the concepts of the theory, and there is no way to test the explanation without presuming the conceptual framework. In such a case, we would want to say that the phenomenon does in fact support the theory (because the argument is coherent), but it cannot lend very much support (because the argument is exclusively theory-dependent and untestable). Assuming that there is no smaller amount of support a theory can enjoy than a coherent yet untestable (and otherwise unworkable) argument, let us use this A-score category (“Coherent but untestable”) as the lower bound on our ordinal scale. From this, one can conceive of the next category as merely modifying the testability. The question here is whether the interpretation, or any part of it, is testable (in principle) without presuming the conceptual framework of the theory. Consequently, let us call the second ordinal score “Coherent and testable”. In the final category (A-score), let us collect the evidence that is not only testable in principle (without presuming the conceptual framework of the theory), but has *in fact* been tested. This leaves us with an ordinal ranking of claims of empirical support of four categories (A-score): Accepted, Coherent and testable, Coherent but untestable, and Rejected (Table 1).

4.2 The marginal: from phenomena to ordinal rankings

Traditionally, the marginal in Bayes’ theorem is cashed out as “the probability of the evidence,” but how should this be understood in the present context? Given that we do not have access to any/the objective probability of the evidence (the empirical phenomena claimed in support of a theory), we will again deploy ordinal scores as anchors for quantification. For the A-score we assessed arguments and how much these relied on the conceptual framework of a theory with respect to testability, but neither of these fit well as anchor for a (theory independent) *probability of the evidence* (the *phenomenon*). There are however good candidates for anchors for the marginal inherent in the observations of the phenomenon itself. Here, I will focus on one possible candidate, namely: replication. Initially, three things are worth mentioning with respect to the notion of replication as deployed here.

19 Thus, the process here is to consider the arguments connecting each proposed piece of evidence (each phenomenon) to the theory of consciousness whose empirical support we are evaluating. This entails assessing the proposed evidence for each theory on a case-by-case basis (explication and examples of this process can be found in Kirkeby-Hinrup and Fazekas, 2021).

Firstly, we can note that for every phenomenon there will be a number of replications of a given finding about it. Sometimes, (in case it is a rare or very new study), the number of replications will be zero, and the original finding constitutes the only report of the phenomenon. Now, given that non-existing findings cannot form the bases for claims of *empirical* support, the lowest amount of credence we could have in a phenomenon would then be a single finding that has not been replicated. Secondly, it is worth noting that, plausibly, replication should co-vary with *credence* (discussed in Section 3.1 above), given that we agree that well replicated findings, and well understood phenomena are more credible as evidence. Thirdly, replication allows many values, which in turn allows for multiple ordinals. This makes replication suitable for grouping into different ordinals that can then be used as anchors for quantification.

With these three things in mind, the questions then are: how many ordinals should there be? What should they be? And how do we scale the number of replications of a phenomenon to a category on the ordinal ranking? In the examples below, I will operate with a three-step ordinal ranking categorizing phenomena into “high,” “medium,” and “low” replication (the R-Score). However, given *arbitrariness* discussed above, the number of ordinal rankings should not be up to me, therefore my use of three categories in the examples below is exclusively to keep the example data simple and easy to read, and should not be taken to signal that these are (arbitrarily) set in stone. While my inclination is to think that a relatively small number of ordinals such as this (or 4 in case one prefers a category for single cases with no replication) is most reasonable (and I suspect investigations will find convergence on this, similar to how the PAS scale was developed), nothing in the following hinges on this; there could be arbitrarily many categories (we could create a category for the exact number of replications of each phenomenon) since the methodology deployed in the quantification can accommodate this. For the present purposes, the working assumption merely is that we can meaningfully sort phenomena into three categories reflecting amounts of replication that we call “Low,” “Medium” and “High,” leaving the exact Low-Medium, and Medium-High thresholds unspecified. Nevertheless, because there may be disagreements between researchers (e.g., due to conceptual bleed) pertaining to selection, ordering, and assignment to the ordinal categories, the *arbitrariness* issue in this domain needs to be dealt with. In section 4.6 below, I consider ways to deal with this.

4.3 The prior and scaling

The last element of Bayes’ theorem we need to account for is the prior. Traditionally, the prior is the initial probability of the hypothesis (i.e., the theory). Given that the conceptual debates have come up inconclusive, it seems that assigning a higher initial probability (prior) to one or the other theory would be arbitrary. One way to avoid this is to assign the same initial probability (prior) to every theory. This also reflects the fact that we — as a field — really do not *know* which theory is right.²⁰ But which value should it be set to? Normally, if

we did not know either way, we would set the prior to 0.5 (50%). However, because there are multiple competing theories, the question is not exactly an either-or (fifty-fifty) proposition. An alternative would be to divide full confidence (100%) by the number of theories available. The number of contemporary theories varies between reviews (Northoff and Lamme, 2020; Sattin et al., 2021; Signorelli et al., 2021; Seth and Bayne, 2022). In the examples below, (as a conservative choice) the count is set at 25, and consequently 0.04 priors are used in the example data. Importantly, with respect to the comparison, as long as we stay impartial by assigning the same prior to each theory, the exact number of the prior is inconsequential. However, from a Bayesian perspective the lower and upper bounds are 0 and 1, respectively. So — if one desires to stay within a Bayesian framework — this constrains the scaling, given that no posterior of any theory should end up outside these bounds (<0 or >1) at any point in the quantification process. Similarly, for comparison purposes — as long as we stay neutral and deploy the same values in the quantification of support for each theory — the exact scaling we deploy in the updating function is inconsequential. However, to stay within a Bayesian framework it is desirable that the amount of credence a phenomenon can maximally lend to a theory (the Likelihood) is not such that any individual quantified phenomenon, or the total set of phenomena takes the posterior above one or below zero.²¹

4.4 From ordinals to numbers

In this section, the topic will be how to get numbers from the ordinals (A-scores and R-Scores). To avoid confusion, I will deploy the terms A-value and R-value to signify a given number derived from a specific ordinal score. The central idea in QBE is to deploy the ordinal scores as anchors to provide natural minimum and maximum values (with one or more values of the middle ordinal(s) between). To illustrate: The A-score deploys the categories “Accepted,” “Coherent and testable,” “Coherent but untestable,” and “Rejected.” Not counting rejected evidence, we end up with a three-step ordinal where “Accepted” is better than “Coherent and testable” and both are better than “Coherent but untestable.” The highest ordinal (“Accepted”) is deployed as the natural maximum A-value we would assign in the quantification. Similarly, “Coherent but untestable” is the natural minimum A-value, being the lowest ordinal. As mentioned in the previous section, the exact scaling of the A-value in the updating function is inconsequential, as long as we deploy the same scale for each theory being compared. In the examples here, I will deploy a scale

prior. I am somewhat skeptical of this implementation. On the one hand, I think simplicity is secondary to aspects such as explanatory power, coherence, and empirical support. On the other hand, there are several ways in which theories can be “simple” so this comes with an additional need to eliminate arbitrariness with respect to which aspect(s) of a theory the simplicity should apply to. Nevertheless, at this stage of the development in the model, it is certainly premature to rule anything out, perhaps there is both room and warrant for simplicity considerations somewhere in it.

²¹ A further benefit of this Bayesian constraint on the bounds of posteriors is that it allows us to consider sets of posteriors as probability distributions, which will be relevant for the comparison step discussed in the next section.

²⁰ Perhaps priors could be modified to reflect the simplicity of a given theory, appealing to Occam’s Razor, with simpler theories being assigned a higher

of 1–10 % for the A-values, meaning the lowest increase in posterior a theory can gain from a piece of evidence is 1%, and the highest gain is a 10% increase. Given that the A-Score “Accepted” is the natural max, that means that a piece of evidence with the A-Score “Accepted” should increase the posterior by 10 % (the A-value is 10%). Similarly, a piece of evidence with the lowest ordinal A-Score (“Coherent but untestable”) should increase the posterior by 1 % (the A-value is 1%). In this way the highest and lowest ordinals anchor any scale we decide on. But how do we non-arbitrarily set the A-value of the middle ordinal(s)? Disagreement seems possible on this question. For instance, one might suggest that the A-value of the middle ordinal should be in the middle (or close to) between the min and max, say 5 % on the scale used here. Others might disagree and argue that the testability difference between “Coherent and testable” and “Coherent but untestable” is of such significance that the A-value of the middle ordinal should be closer to 8 or 9 percent, rather than in the middle. So, how do we determine what the A-value of the middle ordinal (e.g., evidence scored as “Coherent and testable”) should be? Critically, we need to deal with *arbitrariness* and not bias the comparison against any theory. This is especially important because the A-value of the middle ordinal influences the posteriors (because it determines the increase a theory gains from a piece of evidence with the A-Score “Coherent and testable.” see Figure 3). In QBE this issue is solved by refusing to fix the middle ordinal to one value. Instead the idea is to calculate the entire dataset for each possible value of the middle ordinal (e.g., using natural numbers 2 through 9 in our example data) and let the collective set of posteriors form the basis for our comparison of theories. The same solution is applied to the R-score. The example data here uses R-values of one to ten,²² and calculates the dataset with each possible R-value for the middle ordinal (“Medium Replication”). Consequently, in our example here, the output of the quantification for a given theory is a set 64 posteriors (8*8) reflecting each combination of the possible middle ordinals of the A-value (2–9%) and the R-value (2–9).

But what about a case where someone wants to deploy more than three ordinals? In these cases there will be two (or more) middle ordinals rather than one. While this increases the combinations in terms of the number of posteriors that need to be calculated, there is nothing inherently problematic with this. Naturally, because two or more middle ordinals are ranked *qua* ordinals, they constrain each other in terms of the values each can have. To illustrate, using a 1–10 scale, if the lower of two middle ordinals has a value of 4, this constrains the possible values of the higher middle ordinal to the numbers [5,6,7,8,9]. In sum, the methodology can easily accommodate cases where more than one middle ordinal is deployed. While the number of posteriors that will be calculated for a theory will increase with the number of middle ordinals, this increase is trivial, and not such that it poses a problem for the methodology.

4.5 Quantification and comparison

The objective in this section is to demonstrate the proposed methodology. Initially, as just discussed the exact scaling of the

parameters is inconsequential for the comparison as long as it is applied to every theory being compared.²³ Unfortunately, there is an immediate obstacle to the demonstration in that there are no datasets on which to demonstrate the methodology. We simply do not have a full view of all — and which — phenomena are claimed in favor of any theory (let alone all theories). While some work has been done on this (Kirkeby-Hinrup and Fazekas, 2021), there is considerable way to go before we have complete compilations for every theory, and have separately assessed each proposed piece of empirical support to derive an A-score and R-score (and any other score we may think of, see below). To address this issue, I wrote a small piece of software²⁴ that generates random datasets for hypothetical theories. The datasets were set to contain between 20 and 40 phenomena, that each had an A-score and an R-score. Generation of A-and R-scores was weighted to output comparatively fewer of the highest ordinal (see footnote 20) to account for the fact that there currently is not much knock down evidence. I generated four hypothetical theories [A,B,C,D], whose resulting datasets sorted in 3 (A-score) by 3 (R-score) matrices are shown in Table 2. I then ran the updating mechanism (using a scaling that allowed the posteriors to stay within Bayesian 0 and 1 bounds) on the datasets of the four hypothetical theories.

For determining the likelihood, a scale for the A-value of one to 10 % was used, meaning the highest A-score ordinal (“Accepted”) implied a 10 % increase, and the lowest ordinal (“Coherent but untestable”) a 1 % increase, with the middle ordinal (“Coherent and testable”) occupying each of the intermediary steps (2–9 percent). The marginal also used a one to ten scale for the R-value. However, the one to ten scale of the R-value was not percent, but rather hundreds. Again, the highest R-score ordinal (“High”) being 10/100, the lowest (“Low”) being 1/100, and the middle ordinal (“Medium”) occupying the intermediate steps (2–9/100). To avoid the counterintuitive result that well replicated studies might decrease a posterior, the marginal was calculated as one minus the R-value (Figure 4). The initial prior was set at 0.04 (based on the assumption that there are at least 25 viable competing theories). The updating itself consisted in iterating through the list of proposed evidence (phenomena) deriving a new posterior after the inclusion of each phenomenon, and that posterior becoming the new prior when updating with the next phenomenon. The end result, with the entire set of proposed evidence processed, is a posterior *given all the evidence*.

Now, because of the way R-values and A-values of the middle ordinals are modeled in QBE the updating has to be carried out with each of possible combination of R-value and A-value. In this example, the result after updating is a dataset for each theory consisting in sixty-four posteriors; *viz* one for each tested combination of values of the middle ordinal of the A-score (2–9%) and R-score (2-9/100). When ordering the datasets according to size of the posterior and plotting them on a graph, (Figure 3) the impact of the value of the middle ordinals of the A-score and R-score is evident. Similarly, by calculating the mean and standard deviation of all the posteriors (Table 3) we can represent the probability that a theory has a given posterior (Figure 5).

²³ In the previous sections, I have mentioned at the parameters that will be deployed here, but I will reiterate them the first time they appear.

²⁴ Not on github, but open source in the sense that I will give you the code if you send me an email.

²² Again: arbitrarily chosen since the exact scaling is inconsequential for the comparison. The important thing is to keep it the same for all theories.

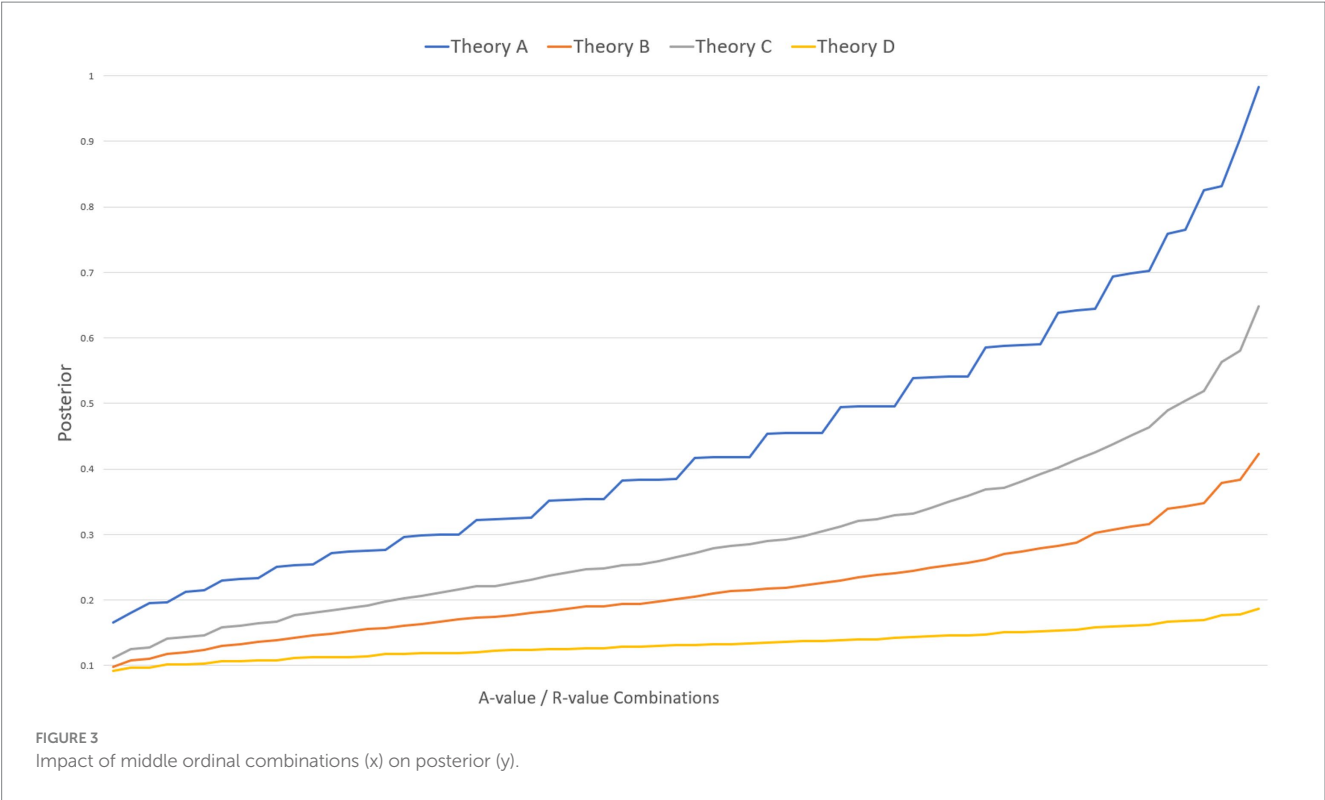


TABLE 2 Hypothetical datasets of theories A, B, C, and D.

| | | Accepted | Coherent and testable | Coherent but unstable |
|---------|-----------------|----------------|-----------------------|-----------------------|
| | Theory B | A-score | | |
| R-score | High | 1 | 0 | 4 |
| | Medium | 1 | 7 | 8 |
| | Low | 0 | 2 | 3 |
| | Theory B | A-score | | |
| R-score | High | 0 | 2 | 1 |
| | Medium | 0 | 6 | 3 |
| | Low | 0 | 4 | 4 |
| | Theory C | A-score | | |
| R-score | High | 0 | 2 | 0 |
| | Medium | 0 | 5 | 8 |
| | Low | 1 | 5 | 4 |
| | Theory D | A-score | | |
| R-score | High | 1 | 1 | 1 |
| | Medium | 1 | 1 | 3 |
| | Low | 0 | 3 | 3 |

While helpful — at a glance — to get an impression of where each theory stands, these are merely ways of depicting the data, and do not suffice as an actual *comparison*.

There are several different ways of going about comparing the numbers. I will next consider a few options. The most straightforward

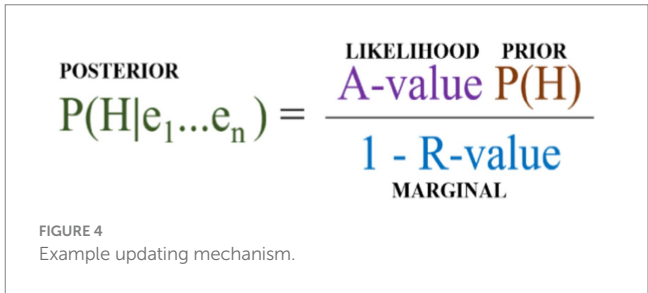


TABLE 3 Means and standard deviation.

| | A | B | C | D |
|---------------------|--------|--------|--------|--------|
| Mean posterior* | 0.4409 | 0.2157 | 0.292 | 0.1329 |
| Standard deviation* | 0.1940 | 0.0744 | 0.1218 | 0.0219 |

*Rounded to 4 decimals.

approach would be to compare directly the mean posteriors of the theories, i.e., collapse the set of posteriors from each theory into a mean posterior for each theory and compare them. The mean posteriors — in themselves — allow for straightforward comparison of the theories on the bases of their respective posteriors (Table 3).

A similar second option could be to take the graphs in Figure 3 and compare the areas under the curve (this could be refined using smaller increments for the calculated A-and R-values and by deploying integrals). However, a more interesting third option may be Z-score comparison (Figure 6). The idea behind Z-scores is to use the mean and SD of posteriors of all theories to create an anchor for how much support a given theory has as compared to

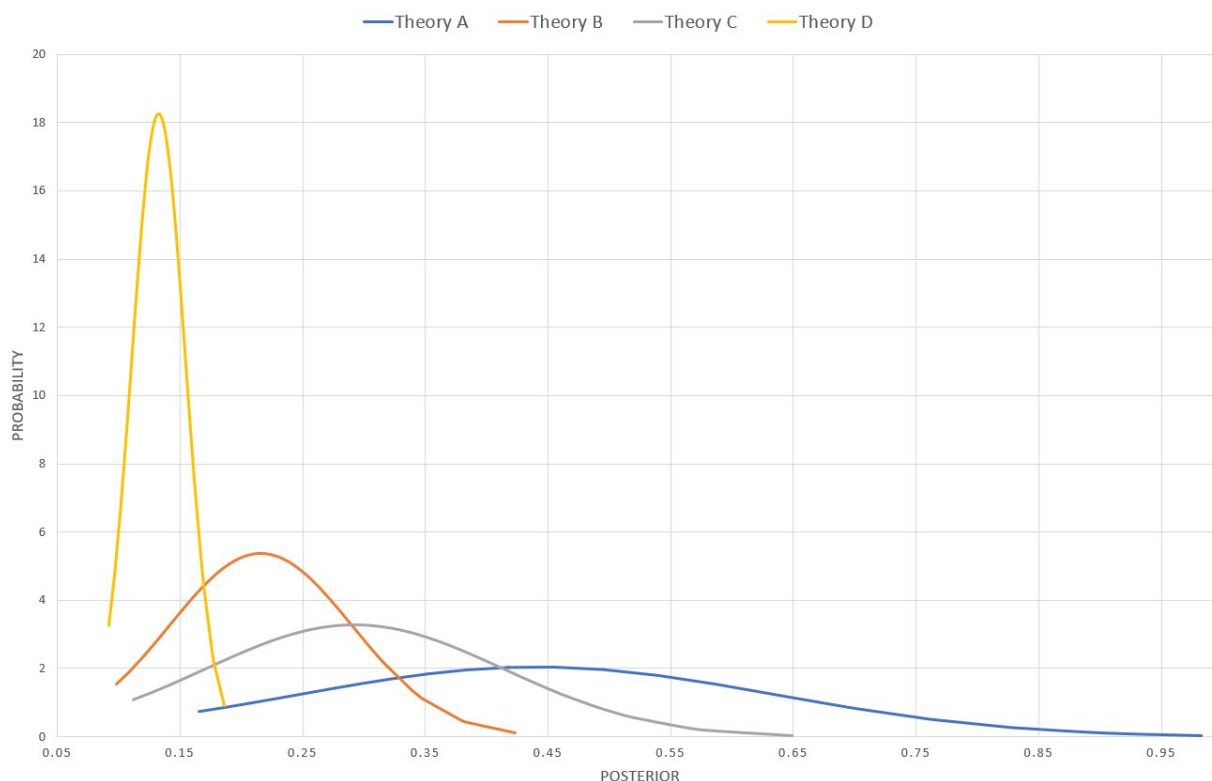


FIGURE 5
Normal distribution of probability (y-axis) that a theory has a given posterior (x-axis).

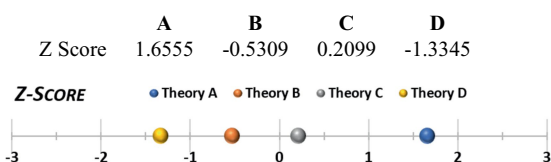


FIGURE 6
Z-score comparison.

the average support theories have. The Z-score shows how many standard deviations a theory's support is from the average. Ranging from -3 to +3, positive Z-scores indicate good support relative to the norm. For comparisons of two concrete theories, a fourth option could be deploying t-tests, either one-tailed (pairwise comparison), or two-tailed, to assess whether the empirical support for theories across the field is truly different. Finally, one could compare theories directly against each other using pairwise ratios of the mean posteriors (Table 4).

One important feature of each of these comparison options is that they are all independent of our scaling choices in the sense that — while it is nice that our data fits within the Bayesian bounds — the comparisons themselves do not depend on this (i.e., we could still do mean posteriors, Z-scores, ratio comparisons etc. if the posteriors were higher than 1). This gives significant flexibility to our choices with respect to scaling, and counters potential issues with arbitrariness in this regard.

4.6 Arbitrariness in scoring, ordinals, and the updating mechanism

In the above, *arbitrariness* has been prevented at every turn, yet three issues remain in this regard that need to be addressed. The first (and most critical) of these is *who* gets to determine the A-score of a piece of proposed empirical support. Given that the A-score directly impacts the amount of support gained from a piece of evidence, if this is left at the whim of the comparer, the whole process is undermined. The solution is straightforward: in the scoring it is necessary to engage with the original authors of a given piece of proposed empirical support.²⁵ Such engagement serves to make certain that the A-score assigned to each piece of empirical evidence is corroborated by the views of the original authors.²⁶ Furthermore, the engagement with the proponents of a given piece of empirical support affords them opportunity to clear up misunderstandings, make corrections, or further specify their argument in light of problems exposed (that result in an A-score they disagree with), or questions that arose in the case-by-case analysis. It is also important to recognize that novel experimental paradigms may impact A-scores. One recent example of this pertains to the pneumatic drill example given in favor of the

²⁵ As well as other relevant proponents of the theory.

²⁶ Remember, their conceptual framework has to be taken for granted, which makes them the authorities on interpretation of their proposed evidence, as noted in section three.

TABLE 4 Pairwise ratio comparison.

| | A | B | C | D |
|-----------------|--------|--------|--------|--------|
| A (mean 0.4409) | X | 2.0442 | 1.5099 | 3.3175 |
| B (mean 0.2157) | 0.4892 | X | 0.7387 | 1.6229 |
| C (mean 0.2919) | 0.6623 | 1.3538 | X | 2.1971 |
| D (mean 0.1290) | 0.3014 | 0.6162 | 0.4552 | X |

distinction between access consciousness and phenomenal consciousness (Block, 1998). The gist of the ‘pneumatic drill effect’ is that upon the disappearance of an (previously un-accessed) auditory source, subjects have a strong intuition that they had been experiencing it all along. By virtue of being a ‘dishwasher example’ (see Kirkeby-Hinrup and Fazekas, 2021, p. 6 for details), previously this would have been classified as ‘Coherent but Untestable’. However, a novel study (Amir et al., 2023) operationalized the same effect with results seeming to corroborate the intuition. Consequently, the pneumatic drill effect more properly belongs in the ‘Coherent and Testable’ category.

Being sensitive to novel findings and engaging with the proponents of theories in this way means that the conclusions reached will properly reflect the views and data in the field, and the datasets deployed in the comparisons accurately reflect the evidence out there and are broadly endorsed. Accusations of arbitrariness in the scoring of evidence are catered to by allowing proponents of theories to spell out the reasoning behind a given piece of proposed evidence, spell out potential testability, or provide updates to arguments. Following any changes to A-scores as a result of such interaction, re-scoring the piece of evidence and re-calculating posteriors is trivial.

The second issue pertains to *arbitrariness* in deciding the ordinal categories and the criteria for each category. To illustrate, whether one deploys a three-step ordinal or a five-step ordinal for the R-score impacts the posteriors of theories because if there is a larger number of ordinal steps this means that two pieces of evidence that are scored as equal on a three-step ordinal (e.g., both in the “Medium” R-score), may not end up in the same ordinal category on a scale with more ordinals (e.g., one may end up in “Upper Medium” and the other in “Lower Medium”). Consequently, since the R-score is the foundation of the R-value, which in turn impacts the posterior, the number of ordinals and their criteria impact the comparison and may bias the comparison against theories whose evidence ends up being ‘worth’ comparatively less if a higher number of ordinals is deployed. A similar problem pertains to the criteria for being scored in a given ordinal. To illustrate, whether 50 or 55 replications is the criterion for “High” replication (the highest ordinal) matters for phenomena with a number of replications between 50 and 54. So, how do we best settle on the ordinal categories, the number of ordinals, and their respective criteria? Again — for by now familiar reasons — no single individual should get to decide these questions. Therefore, it is useful to consider some possible ways of solving this issue.

The first way consists in letting the scientific community decide the categories and criteria. This could either be done as a straight-up crowdsourcing endeavor with questionnaires disseminated through appropriate channels (specialist mailing lists, conferences, websites, or journals), or in a more structured way. One example of such a process is the development of the Perceptual Awareness Scale (“PAS,” see Ramsøy and Overgaard, 2004; Overgaard et al., 2006) which

found that when asked to compose their own scale for visual awareness, subjects’ responses converged on a four step ordinal. PAS is widely recognized as useful (it is probably by far the most deployed scale to assess perceptual awareness in contemporary consciousness science) and has (to my knowledge) never faced significant accusations of arbitrariness. Now, there is of course a sense in which a crowdsourced scale would be arbitrary to the sampled population (i.e., the ‘crowd’). However, this is not the arbitrariness of relevance here given that the sampled crowd is the scientific community, and these are exactly the people whose views we would want the scale to reflect.

The other possible way one could attempt to settle this is by deriving it through data mining the relevant academic body of work. This would consist in surveying the replication numbers of the phenomena to identify the ranges where clustering occurs, and then using the number of clusters to determine the number of ordinals, and the ranges of the clusters to determine the thresholds for a given ordinal. With respect to this kind of data mining, there are a wide range of established algorithms to determine not only the number of clusters in a dataset, but also the values of those clusters (e.g., k-means clustering and x-means clustering to give just two examples). So if we have a dataset containing the number of replications for all the proposed phenomena, we could derive the number of categories (ordinals) and the cut-offs between them.

The third issue concerns decisions about the updating mechanism. In the example above, I used a scale of 1 through 10 percent for the A-score, meaning the prior got multiplied by a number (the A-value) in the range between 1.01 and 1.1. However, there are obviously other ways one could structure such an updating mechanism. To give just one simple alternative: instead of using a multiplication function, one might simply use addition (i.e., by just adding the A-value to the prior). It is trivial that the choice between multiplication and addition matters,²⁷ given that the cumulative effect of several multiplications favors theories with a higher number of proposed pieces of empirical support. This means an updating mechanism deploying multiplication is biased against theories with a low number of proposed empirical support.²⁸ Similarly, the R-value in the example above was modelled as a number between 0.9 and 0.99 (with the highest ordinal being 0.9) to achieve the effect that higher replication scores increased the amount of support a theory received from a phenomenon (because dividing by 0.9 yields a higher posterior than dividing by 0.99). However, there are a multitude of alternative ways in which one could model the R-score in the updating mechanism. One possibility is to use percentages like in the A-value and simply factor the R-score in with the likelihood along with the A-value (this matters because dividing by 0.9 is not equal to multiplying by 1.1). Furthermore, as I will touch on below, one might want to include more elements in the updating mechanism than arguments and replication. In sum, there are many ways to structure the updating

27 Or other possible ways of conceiving of the updating mechanism.

28 One might argue that intuitively this makes sense given that large amounts of empirical support *should* result in higher credence in a theory, when compared to theories with very few pieces of empirical support. This, however, is a separate discussion, and I will leave it to the side for now.

mechanism, it is unclear which is preferable, and choosing between them runs the risk of arbitrariness. Certainly, involving mathematicians (especially statisticians) and philosophers of science would be beneficial to map out different possible updating mechanisms and clarifying their respective implications. In any case, I do not purport that the version presented above is anything more than an early sketch. In fact, I think it is incomplete in the sense that my (arbitrary) preference would be to expand it to account for more features of the evidence in the marginal (I will briefly return to this in the concluding remarks below). Now, given the centrality of the updating mechanism to QBE, it may seem as if *arbitrariness* in this place effectively subverts the whole idea. For instance, it is not an implausible scenario that disagreements about how to structure the updating mechanism may result in multiple competing versions, each with its own group of proponents and no non-arbitrary way to decide which version is preferable. In this case, it may appear that we are back where we started, and QBE has not managed to move forward the debate in any meaningful way. Fortunately, this appearance is misleading. The important progress to notice in this respect is that disagreeing about the updating mechanism is significantly different from disagreeing about the nature of consciousness. One way in which it is different is that discussions on the updating mechanism can be done objectively, in the sense that the subject matter is mathematics (statistics). This means QBE manages to cauterize the conceptual bleed from the theoretical predilections of researchers with respect to consciousness. Put differently, disagreement about the updating mechanism is an *entirely different debate* that can be had without utilizing any of the concepts the disagreements on which were at the root of our problems in ICS.

5 Concluding remarks

Proponents of competing theories of consciousness have spent the better part of almost three decades amassing empirical support for their preferred theory on the assumption that this would somehow resolve the debates. In recent years proposals specifically on *exactly how* empirical support can achieve this have garnered attention.

The approach I have advanced here offers a novel methodological approach to this issue. Throughout I have endeavored to be transparent about the fact that this is not a finished or unproblematic methodology, and that there are several avenues open for future development and refinement. QBE is merely a first approximation of the methodology. Its purpose here is more of a proof of concept that it is possible to quantify empirical support for theories of consciousness in a way that avoids *arbitrariness*, than a fully baked cake. In other words, at this stage QBE is not purported to be either perfect, or entirely noncontentious. Firstly, there may be additional ways of scoring evidence that could either complement or supersede the way proposed here. Secondly, there likely are unexplored ways to quantify the scores. Thirdly, there are many possible ways to construct the updating function in Bayes' theorem. Each of these three avenues of development comes with separate requirements for justification of why it is preferable to other ways of doing the same thing. Or, if full justification is not possible, then the requirements can be for motivation, argument, or rationale, depending on one's position on a range of philosophy of

science issues, and one's epistemological commitments. The proposal offered here is open to exactly that; i.e., that there may be better²⁹ ways to model scoring, conversion, or updating, and the future development of the methodology should be open to change. The modest aim here has been to show that there *is* a model we can develop.

One way to develop the model could be to construct additional ordinals. For instance, with respect to the Marginal in Bayes' theorem, the three remaining aspects³⁰ of the two drivers of the intuition discussed in section 3.1 provide avenues of development.³¹ For instance, it might be relevant to introduce ordinals to score phenomena in accordance with the two aspects of the *closeness* driver. This would mean phenomena would also be scored according to physical/functional closeness (e.g., with categories such as computer models, animal studies, human studies) or Distribution (e.g., with categories such as: Single case, Rare, Common, Prevalent). Similarly, one may want to introduce an ordinal to reflect the other aspect of the *credence* driver (scope). Naturally, each new ordinal one introduces brings a demand for considerations about how this ordinal is then best implemented in the updating mechanism. One strength of QBE is that revisions of both the datasets, scoring, quantification, and updating mechanism are easily handled, which serves to underscore the objectivity, and flexibility of the methodology.

Finally, more should be said on how my QBE avoids the shortcomings of the two current approaches for comparing theories of consciousness, namely the adversarial collaboration (ARC) and criterion-based (CRIT) approaches (As discussed in Section 2). For each of ARC and CRIT, I identified four issues and in Section 3.1, I argued that it was desirable if QBE could avoid these issues. Therefore, a brief summary of how QBE manages to do this is warranted. Firstly, there is no upper limit for the number of theories to which QBE can be simultaneously applied. This means that the issue of *targeted theories* does not pertain to QBE. By considering all available evidence QBE has the broadest possible scope, thereby avoiding the *narrow scope* issue. Similarly, in QBE, the same methodology is applied to all theories thereby avoiding the issue of *generalizability*. The methodology in QBE allows for easy addition, removal, or updating of theories or evidence. This means that the *robustness* issue is also avoided. While the process of collecting and scoring all empirical evidence proposed in favor of every theory constitutes a significant amount of work, it is a one-time effort, in the sense that once the datasets are collected, updating them with further proposed evidence is trivial. This means that while the initial *cost* of QBE is somewhat high, it is nevertheless significantly less than that of ARC (both in the short and long term). Because of the Bayesian updating process, QBE is sensitive to every piece of evidence proposed in favor of a theory. Consequently, by accounting for the total amount of evidence, QBE avoids the *sensitivity* issue. With respect to the *arbitration* issue, the scoring and updating process in QBE make ties highly unlikely. Furthermore, because QBE allows for easy updating,

²⁹ Where "better" can be understood in various ways. To give just a few examples, better could be understood in terms of justification, motivation, simplicity, appeal, alignment with common intuitions, more fine-grained, or more refined mathematically.

³⁰ The R-score already reflects the "Replication" aspect of the *credence* driver.

³¹ Specifically, to my mind, these three remaining aspects each seem relevant to the Marginal in the updating mechanism.

ties will be broken upon the addition — or revision — of a single piece of evidence. QBE is maximally flexible by allowing for any piece of proposed evidence to be scored and added. Thus the *flexibility* issue is also catered to. Finally, at every point it has occurred, I have addressed the *arbitrariness* issue with respect to the topic at hand. Most importantly, by generating sets of posteriors for each theory based on every possible scoring of the evidence, no decisions about evidential weight depend on the judgment of any individual.

Having argued that QBE avoids the issues identified with ARC and CRIT, it is necessary to reiterate that the objective of QBE is not to supplant these two approaches, but to offer a third approach, to be deployed either independently of — or jointly with — ARC and CRIT. In other words, the different approaches need not be mutually exclusive, but rather may in fact positively interact. For instance, prognosis output from QBE may inform ARC work by indicating relevant theories to test against each other. Reciprocally, findings from ARC projects may be scored and added as evidence in QBE. In a similar vein, CRIT contains meta-theoretic considerations (e.g., regarding what we want theories to explain) that have merit on a general level. It seems there is not only room for co-existence, but also for synergy between the different approaches to assessing and comparing theories of consciousness.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

References

- Albantakis, L. (2020). "Integrated information theory" in *Beyond Neural Correlates of Consciousness*, (Eds.) Overgaard, M., Mogensen, J., and Kirkeby-Hinrup, A. (New York: Routledge), 87–103.
- Amir, Y. Z., Assaf, Y., Yovel, Y., and Mudrik, L. (2023). Experiencing without knowing? Empirical evidence for phenomenal consciousness without access. *Cognition* 238:105529. doi: 10.1016/j.cognition.2023.105529
- Baars, B. J. (1996). Understanding subjectivity: global workspace theory and the resurrection of the observing self. *J. Conscious. Stud.* 3, 211–216.
- Block, N. (1995). On a confusion about a function of consciousness. *Behav. Brain Sci.* 18, 227–247. doi: 10.1017/S0140525X00038188
- Block, N. (1998). How to find the neural correlate of consciousness. *R. Inst. Philos. Suppl.* 43, 23–34. doi: 10.1017/S1358246100004288
- Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behav. Brain Sci.* 30, 481–499. doi: 10.1017/S0140525X07002786
- Block, N. (2011a). The higher order approach to consciousness is defunct. *Analysis* 71, 419–431. doi: 10.1093/analysis/anr037
- Block, N. (2011b). Perceptual consciousness overflows cognitive access. *Trends Cogn. Sci.* 15, 567–575. doi: 10.1016/j.tics.2011.11.001
- Block, N. (2011c). Response to Rosenthal and Weisberg. *Analysis* 71, 443–448. doi: 10.1093/analysis/anr036
- Block, N. (2014a). *The Puzzle of Perceptual Precision: Open MIND*. Frankfurt am Main: MIND Group.
- Block, N. (2014b). Rich conscious perception outside focal attention. *Trends Cogn. Sci.* 18, 445–447. doi: 10.1016/j.tics.2014.05.007
- Block, N. (2016). The Anna Karenina principle and skepticism about unconscious perception. *Philos. Phenomenol. Res.* 93, 452–459. doi: 10.1111/phpr.12258
- Boly, M., Massimini, M., Tsuchiya, N., Postle, B. R., Koch, C., and Tononi, G. (2017). Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? Clinical and neuroimaging evidence. *J. Neurosci.* 37, 9603–9613. doi: 10.1523/JNEUROSCI.3218-16.2017
- Bor, D., and Seth, A. K. (2012). Consciousness and the prefrontal parietal network: insights from attention, working memory, and chunking. *Front. Psychol.* 3:63. doi: 10.3389/fpsyg.2012.00063
- Brinck, I. (1999). Nonconceptual content and the distinction between implicit and explicit knowledge. *Behav. Brain Sci.* 22, 760–761. doi: 10.1017/S0140525X99282180
- Brinck, I., and Kirkeby-Hinrup, A. (2017). Change blindness in higher-order thought: misrepresentation or good enough? *J. Conscious. Stud.* 24, 50–73.
- Brogaard, B. (2011). Are there unconscious perceptual processes? *Conscious. Cogn.* 20, 449–463. doi: 10.1016/j.concog.2010.10.002
- Burks, A. W. (1946). Peirce's theory of abduction. *Philos. Sci.* 13, 301–306. doi: 10.1086/286904
- Campos, D. G. (2011). On the distinction between Peirce's abduction and Lipton's inference to the best explanation. *Synthese* 180, 419–442. doi: 10.1007/s11229-009-9709-3
- Carruthers, P. (1998). Natural theories of consciousness. *Eur. J. Philos.* 6, 203–222. doi: 10.1111/1468-0378.00058
- Chalmers, D. (2002). "The components of content" in *Philosophy of Mind: Classical and Contemporary Readings*, ed. D. Chalmers (Oxford: Oxford University Press), 608–633.
- Cleeremans, A., Achoui, D., Beauny, A., Keuninckx, L., Martin, J.-R., Muñoz-Moldes, S., et al. (2020). Learning to be conscious. *Trends Cogn. Sci.* 24, 112–123. doi: 10.1016/j.tics.2019.11.011
- Crupi, V. (2021). "Confirmation" in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Spring). Available at: <https://plato.stanford.edu/archives/spr2021/entries/confirmation/>
- Dehaene, S., and Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* 79, 1–37. doi: 10.1016/S0010-0277(00)00123-2
- Del Pin, S. H., Skóra, Z., Sandberg, K., Overgaard, M., and Wierzcchoń, M. (2021). Comparing theories of consciousness: why it matters and how to do it. *Neurosci. Conscious.* 2021:niab019. doi: 10.1093/nc/niab019
- Doerig, A., Schurger, A., and Herzog, M. H. (2020). Hard criteria for empirical theories of consciousness. *Cogn. Neurosci.* 12, 41–62. doi: 10.1080/17588928.2020.1772214
- Doerig, A., Schurger, A., and Herzog, M. H. (2021). Response to commentaries on 'hard criteria for empirical theories of consciousness'. *Cogn. Neurosci.* 12, 99–101. doi: 10.1080/17588928.2020.1853086
- Douven, I. (2021). "Abduction" in *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta

Author contributions

AK-H: Conceptualization, Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., Lepauvre, A., et al. (2023). An adversarial collaboration to critically evaluate theories of consciousness. Cogitate consortium. *bioRxiv* 2023:546249. doi: 10.1101/2023.06.23.546249
- Fink, S. B. (2016). A deeper look at the “neural correlate of consciousness”. *Front. Psychol.* 7. doi: 10.3389/fpsyg.2016.01044
- Frässle, S., Sommer, J., Jansen, A., Naber, M., and Einhäuser, W. (2014). Binocular rivalry: frontal activity relates to introspection and action but not to perception. *J. Neurosci.* 34, 1738–1747. doi: 10.1523/JNEUROSCI.4403-13.2014
- Friston, K. (2013). Consciousness and hierarchical inference. *Neuropsychanalysis* 15, 38–42. doi: 10.1080/15294145.2013.10773716
- Gelman, A., and Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *Br. J. Math. Stat. Psychol.* 66, 8–38. doi: 10.1111/j.2044-8317.2011.02037.x
- Gennaro, R. J. (1996). *Consciousness and Self-Consciousness: A Defense of the Higher-Order Thought Theory of Consciousness*. Amsterdam: John Benjamins Pub.
- Harman, G. H. (1965). The inference to the best explanation. *Philos. Rev.* 74, 88–95. doi: 10.2307/2183532
- Jacobson, H., and Putnam, H. (2016). Against perceptual conceptualism. *Int. J. Philos. Stud.* 24, 1–25. doi: 10.1080/09672559.2015.1047164
- Kirkeby-Hinrup, A. (2014). Why the rare Charles bonnet cases are not evidence of misrepresentation. *J. Philos. Res.* 39, 301–308. doi: 10.5840/jpr20148420
- Kirkeby-Hinrup, A. (2024). Interdisciplinary consciousness studies needs philosophers of science. *Filosofiska Notiser* 11, 3–18.
- Kirkeby-Hinrup, A., and Fazekas, P. (2021). Consciousness and inference to the best explanation: compiling empirical evidence supporting the access-phenomenal distinction and the overflow hypothesis. *Conscious. Cogn.* 94:103173. doi: 10.1016/j.concog.2021.103173
- Kirkeby-Hinrup, A., Fink, S., and Overgaard, M. (2023). *The Multiple Generator Model*. ed. Kirkeby-Hinrup, A. New York: New York.
- Kirkeby-Hinrup, A., and Overgaard, M. (2023). Ontogenetic emergence as a criterion for theories of consciousness: comparing GNW, SOMA, and REFCO. *Philos. Mind Sci* 4:9902. doi: 10.33735/phimisci.2023.9902
- Knotts, J., Odegaard, B., Lau, H., and Rosenthal, D. M. (2019). Subjective inflation: phenomenology’s get-rich-quick scheme. *Curr. Opin. Psychol.* 29, 49–55. doi: 10.1016/j.copsyc.2018.11.006
- Kouider, S., De Gardelle, V., Sackur, J., and Dupoux, E. (2010). How rich is consciousness? The partial awareness hypothesis. *Trends Cogn. Sci.* 14, 301–307. doi: 10.1016/j.tics.2010.04.006
- Kozuch, B. (2014). Prefrontal lesion evidence against higher-order theories of consciousness. *Philos. Stud.* 167, 721–746. doi: 10.1007/s11098-013-0123-9
- Kriegel, U. (2007). The same-order monitoring theory of consciousness. *Synth. Philos.* 22, 361–384.
- Lamme, V. A. F. (2003). Why visual attention and awareness are different. *Trends Cogn. Sci.* 7, 12–18. doi: 10.1016/S1364-6613(02)00013-X
- Lamme, V. A. F. (2004). Separate neural definitions of visual consciousness and visual attention: a case for phenomenal awareness. *Neural Netw.* 17, 861–872. doi: 10.1016/j.neunet.2004.02.005
- Mashour, G. A., Roelfsema, P., Changeux, J.-P., and Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105, 776–798. doi: 10.1016/j.neuron.2020.01.026
- Melloni, L., Mudrik, L., Pitts, M., Bendtz, K., Ferrante, O., Gorska-Klimowska, U., et al. (2023). An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory. *PLoS One* 18:e0268577. doi: 10.1371/journal.pone.0268577
- Meuwese, J. D., Post, R. A., Scholte, H. S., and Lamme, V. A. (2013). Does perceptual learning require consciousness or attention? *J. Cogn. Neurosci.* 25, 1579–1596. doi: 10.1162/jocn_a_00424
- Michel, M., and Morales, J. (2020). Minority reports: consciousness and the prefrontal cortex. *Mind Lang.* 35, 493–513. doi: 10.1111/mila.12264
- Minnameier, G. (2004). Peirce-suit of truth—why inference to the best explanation and abduction ought not to be confused. *Erkenntnis* 60, 75–105. doi: 10.1023/B:ERKE.0000005162.52052.7f
- Minnameier, G. (2010). “Abduction, induction, and analogy” in *Model-Based Reasoning in Science and Technology: Abduction, Logic, and Computational Discovery*. eds. L. Magnani, W. Carnielli and C. Pizzi (Berlin, Heidelberg: Springer Berlin Heidelberg), 107–119.
- Nagel, T. (1974). What is it like to be a bat. *Philos. Rev.* 83, 435–450. doi: 10.2307/2183914
- Northoff, G., and Lamme, V. A. F. (2020). Neural signs and mechanisms of consciousness: is there a potential convergence of theories of consciousness in sight? *Neurosci. Biobehav. Rev.* 118, 568–587. doi: 10.1016/j.neubiorev.2020.07.019
- Odegaard, B., Knight, R. T., and Lau, H. (2017). Should a few null findings falsify prefrontal theories of conscious perception? *J. Neurosci.* 37, 9593–9602. doi: 10.1523/JNEUROSCI.3217-16.2017
- Overgaard, M., and Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philos. Mind Sci* 2:37. doi: 10.33735/phimisci.2021.37
- Overgaard, M., Rote, J., Mouridsen, K., and Ramsøy, T. Z. (2006). Is conscious perception gradual or dichotomous? A comparison of report methodologies during a visual task. *Conscious. Cogn.* 15, 700–708. doi: 10.1016/j.concog.2006.04.002
- Peirce, C. S., and Hartshorne, C. (1974). *Collected Papers of Charles Sanders Peirce*. Cambridge, Massachusetts: Belknap Press of Harvard University Press.
- Peters, M. A., Kentridge, R. W., Phillips, I., and Block, N. (2017). Does unconscious perception really exist? Continuing the ASSC20 debate. *Neurosci. Conscious.* 2017:nix015. doi: 10.1093/nc/nix015
- Prettyman, A. (2019). Perceptual precision. *Philos. Psychol.* 32, 923–944. doi: 10.1080/09515089.2019.1598765
- Prinz, J. J. (2005). “A neurofunctional theory of consciousness” in *Cognition and the Brain: Philosophy and Neuroscience Movement*. eds. A. Brook and K. Akins (New York, NY: Cambridge University Press), 381–396.
- Ramsøy, T. Z., and Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenol. Cogn. Sci.* 3, 1–23. doi: 10.1023/B:PHEN.0000041900.30172.e8
- Rosenthal, D. M. (1997). “A theory of consciousness” in *The nature of consciousness: Philosophical debates*. eds. N. Block, O. Flanagan and G. Güzeldere (MIT Press), 729–753.
- Rosenthal, D. M. (2008). Consciousness and its function. *Neuropsychologia* 46, 829–840. doi: 10.1016/j.neuropsychologia.2007.11.012
- Rosenthal, D. M. (2011). Exaggerated reports: reply to Block. *Analysis* 71, 431–437. doi: 10.1093/analysis/anr039
- Rosenthal, D. M. (2012). Higher-order awareness, misrepresentation and function. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 1424–1438. doi: 10.1098/rstb.2011.0353
- Rosenthal, D. M. (2021). Assessing criteria for theories. *Cogn. Neurosci.* 12, 84–85. doi: 10.1080/17588928.2020.1838471
- Sattin, D., Magnani, F. G., Bartsaghi, L., Caputo, M., Fittipaldo, A. V., Cacciatore, M., et al. (2021). Theoretical models of consciousness: a scoping review. *Brain Sci.* 11:535. doi: 10.3390/brainsci11050535
- Schurger, A., and Graziano, M. (2022). Consciousness explained or described? *Neurosci. Conscious.* 2022:niac001. doi: 10.1093/nc/niac001
- Seth, A. (2009). Explanatory correlates of consciousness: theoretical and computational challenges. *Cogn. Comput.* 1, 50–63. doi: 10.1007/s12559-009-9007-x
- Seth, A., and Bayne, T. (2022). Theories of consciousness. *Nat. Rev. Neurosci.* 23, 439–452. doi: 10.1038/s41583-022-00587-4
- Signorelli, C. M., Szczotka, J., and Prentner, R. (2021). Explanatory profiles of models of consciousness - towards a systematic classification. *Neurosci. Conscious.* 2021:niab021. doi: 10.1093/nc/niab021
- Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nat. Rev. Neurosci.* 17, 450–461. doi: 10.1038/nrn.2016.44
- Van Gulick, R. (2004). “Higher-order global states (HOGS) An alternative higher-order model” in *Higher-Order Theories of Consciousness: An Anthology*. ed. R. J. Gennaro, 67–93.
- Weisberg, J. (2011). Abusing the notion of what-it’s-like-ness: a response to Block. *Analysis* 71, 438–443. doi: 10.1093/analysis/anr040
- Weisberg, J. (2014). “Sweet dreams are made of this? A HOT response to Sebastián” in *Consciousness Inside and Out: Phenomenology, Neuroscience, and the Nature of Experience*. ed. R. Brown, vol. 6 (Netherlands: Springer), 433–443.
- Yaron, I., Melloni, L., Pitts, M., and Mudrik, L. (2021). How are theories of consciousness empirically tested? The consciousness theories studies (ConTraSt) database. *J. Vis.* 21:2195. doi: 10.1167/jov.21.9.2195
- Yaron, I., Melloni, L., Pitts, M., and Mudrik, L. (2022). The ConTraSt database for analysing and comparing empirical studies of consciousness theories. *Nat. Hum. Behav.* 6, 593–604. doi: 10.1038/s41562-021-01284-5



OPEN ACCESS

EDITED BY

Xerxes D. Arsiwalla,
Wolfram Research, Inc., United States

REVIEWED BY

Krzysztof Basiński,
Medical University of Gdansk, Poland
Vanda Faria,
Uppsala University, Sweden

*CORRESPONDENCE

Valerie Gray Hardcastle
✉ hardcastle@nku.edu

RECEIVED 21 December 2023

ACCEPTED 15 February 2024

PUBLISHED 15 March 2024

CITATION

Hardcastle VG (2024) Entangled brains and
the experience of pains.
Front. Psychol. 15:1359687.
doi: 10.3389/fpsyg.2024.1359687

COPYRIGHT

© 2024 Hardcastle. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Entangled brains and the experience of pains

Valerie Gray Hardcastle*

Institute of Health Innovation, Northern Kentucky University, Highland Heights, KY, United States

The International Association for the Study of Pain (IASP) revised its definition of pain to “an unpleasant sensory and emotional experience.” Three recent recommendations for understanding pain if there are no clear brain correlates include eliminativism, multiple realizability, and affordance-based approaches. I adumbrate a different path forward. Underlying each of the proposed approaches and the new IASP definition is the suspicion that there are no specific correlates for pain. I suggest that this basic assumption is misguided. As we learn more about brain function, it is becoming clear that many areas process many different types of information at the same time. In this study, I analogize how animal brains navigate in three-dimensional space with how the brain creates pain. Underlying both cases is a large-scale combinatorial system that feeds back on itself through a diversity of convergent and divergent bi-directional connections. Brains are not like combustion engines, with energy driving outputs via the structure of the machine, but are instead more like whirlpools, which are essentially dynamic patterns in some substrates. We should understand pain experiences as context-dependent, spatiotemporal trajectories that reflect heterogeneous, multiplex, and dynamically adaptive brain cells.

KEYWORDS

pain, brain, neural correlate, reduction, navigation, adaptive, multiplex

1 Introduction: defining pain

“All we get are a few specks of time where any of this actually makes any sense.”

Joy Wang

Everything Everywhere All at Once

Intuitively, we think of pains as our bodies’ response to some sort of damage. But in 2020, the International Association for the Study of Pain (IASP) revised its definition of pain such that pain is (only) “an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage” (Raja et al., 2020, p. 1977, *italics mine*).¹ “Pain” now is just the conscious sensation of pain. Our best scientific account of pain has been divorced from what we think of as its typical cause.

This perspective makes the scientific study of pain challenging, to say the least. Three recent recommendations for understanding pain if there are no clear brain correlates include (1) promoting some version of eliminativism (Corns, 2020; Liu, 2023), (2) reviving multiple

¹ Borg et al. (2021) and Coninx et al. (2023a,b) provide good arguments (and some empirical data) for why this position is fundamentally incoherent. I shall not dwell on this possibility here.

realizability and family resemblance models (Borg et al., 2021; Coninx, 2023; Coninx et al., 2023b; Serrahima and Martínez, 2023), and (3) suggesting an intersubjective or affordance-based approach (Oliver, 2022; Coninx et al., 2023a; Fulkerson, 2023).

Here, I adumbrate a different path forward. Underlying each of the above-proposed approaches is the suspicion that there are no specific neural correlates for pain. We do not have an area in the brain devoted to pain processing in the same way that we have a visual pathway, for example. In this article, I will suggest that this basic assumption is incorrect, or, better, misguided. Indeed, as we learn more about brain function, it is becoming clearer that many areas process many different types of information at the same time. The fact that there might be no specific correlates for pain that do not overlap with other sensations (*cf.*, Coninx, 2023) is not indicative of anything unusual about how brains function.

Historically, we have taken a reductive approach to understanding the brain. Consider Nagel's description of how we should explain headaches: once "the detailed physical, chemical, and physiological conditions for the occurrence of headaches are ascertained ... [then] an explanation will have been found for the occurrence of headaches" (Nagel, 1961, p. 366). Taking this sort of reductive explanatory approach means that we learn about brain function through decomposing brain areas into sets of individual cells and then to their individual reactions. We slice the brain into smaller and smaller pieces and then articulate how all these pieces connect to one another and interact as a larger whole. Then, voila! We have explained a brain phenomenon.

But recent work suggests that brain structures, both big and small, are fundamentally interwoven. I shall describe this alternative conceptualization of brain organization and function using a brief history of understanding how brains navigate in space as an example (Mallory et al., 2021; Maïsson et al., 2022). My primary point will be that a variety of brain areas support multiple adaptive behaviors and internal representational schemes. In other words, many brain areas that were once thought to do just one thing turn out to support a wide range of functions, and they do so simultaneously.

I shall use this approach to articulate an alternative way of understanding pain. My wider conclusion, however, will be that philosophical intuitions regarding conscious pains and pain processing (or any sort of brain-related functions) are probably best to be avoided. Instead, how our brains do what they do is rooted deep in our evolutionary history, and their functions do not reflect our linguistic divisions or human conventions. Carving nature at its actual joints will require letting go of many contemporary philosophical categories (*cf.*, Westlin et al., 2023). None of the three putative ways to understand pain are likely to be correct. We do have specific neural correlates for pain and pain experiences, but they are not what one might intuitively imagine them to be.

2 Philosophical agreement with the IASP

The three approaches mentioned above—eliminativism, multiple realizability, and affordances—all essentially accept the IASP's perspective without question. They all agree that pain as a rich and complex experience is not reducible (or fully reducible) to underlying brain activity. For example, in her new book promoting eliminativism, Corns (2020) argues that pain is not a natural kind because it cannot

be scientifically "projected". The cellular interactions that determine instances of pain differ among individuals; therefore, they "undermine ... explanations of pain types or pain as such" (p. 141). Pain cannot be a scientific object of study because its instantiation in brains is not constant across individuals or within individuals over time.

Similarly, Coninx agrees that pain experiences across individuals or even within an individual across time are disunified. Nevertheless, as a proponent of a family resemblance approach, she suggests that "[glossing] over differences between pain cases can prove useful under certain conditions for certain scientific purposes" (Coninx, 2023, p. 186). Even though pain may not be a natural kind, we could use the "resemblance relations" among the neural patterns for pain to create broad but serviceable generalizations that could be used in science or medicine to achieve particular ends, like developing effective treatments. She suggests that in this way, pains could loosely form a sort of "phenomenal kind" (p. 180).

Finally, affordance-based approaches to pain also agree that pain is not (just) a type of brain activity that refers to a perceived bodily condition. For example, Oliver (2022) explains that pain states are experienced from a first-person point of view that is embedded in a rich sociocultural environment and that we ascribe meaning to pain experiences in virtue of our respective communities. Pain is "about the interdependent way multidimensional biopsychosocial factors are of concern to a subject" (p. 18). That is, "pain" refers to these specific integrated experiences of sensation, emotion, and interpretation/evaluation in a particular body as it exists in a specific environment through which the person perceives that they can do/see/experience/think certain things (see also Coninx et al., 2023a). Pain is much more than mere brain activity; we would need to appeal to the relevant aspects of the body and environment to give a proper explanation of a pain experience. Neural correlates alone could never underpin a complete theory (see also Hutto and Myin, 2013).

However, while these affordance-based approaches agree with the other two that pain is complex, they disagree that "pain" refers to dissociable cognitive, affective, and physical aspects (this dissociation then either prevents scientific reduction, as Corns claims, or supports loose generalizations, as Coninx claims.) Instead, the multidimensional biopsychosocial factors exist as a complex whole in an embodied mind. A proper nonreductive science of embodied pain might, thus, be possible (see also Coleman, 2020; Cormack et al., 2022).

Even though the three philosophical approaches all differ on what pain being irreducible to brains implies, they, along with the new IASP definition, all agree that there is no easy one-to-one correspondence between any set of pains and identifiable and consistent brain activity. Many recent neuroscientific investigations into pain also support this perspective: there appears to be a range of different neural structures in different locations across the brain that are involved in pain processing (Apkarian et al., 2005; see also Kucyi and Davis, 2014; Bastuji et al., 2016), and yet none of them seem either necessary or sufficient for the experience of pain (Apkarian, 2017). Additionally, none of these areas are identified with pain exclusively; they are also associated with itch, touch, heat, and difficulty breathing (Evans et al., 2002; Iannetti and Mouraux, 2010; Legrain et al., 2011; Liberati et al., 2016; Dong and Dong, 2018). In the scientific literature, there has been at least the suggestion that there are no underlying mechanisms specific to the experience of pain, nor any clear pattern of activity for it across the brain. Even from science's point of view, it appears more likely than

not that pains are not a natural kind (although Bateu, 2020 and Djordjevic, 2023 suggest a different perspective).

It is easy to see how the IASP reached its revised definition and how many philosophers are coalescing around the idea that pain experience is not a proper object of brain study. Nevertheless, I believe a different (and better) approach is possible. This approach starts by embracing the complexity of pain and the brain in all its glory.

3 A different approach

Perhaps more important than the challenge of the apparent irreducibility of pain is that Corns's, Coninx's, and Oliver's approaches to understanding pain ignore or overlook the question of why the quality of pain is the explanatory target in the first place. Regardless of approach, there is agreement that being in pain is a complex state, one that involves a variety of qualia – negative affect, motivational states, sensation, judgments – along with a variety of psychological processes: memory, attention, mood, alertness (see also Borg et al., 2021; Liu, 2023). Why set all this aside as irrelevant and focus on what appears to be only one aspect of pain?² The IASP's definitional revision to remove pain from its physical substrate means that their new perspective on “pain” misses much of what pains actually are. We must recognize and address the full complexity of pain, including the experience of pain, if we are going to advance the science of pain.

If we take seriously the idea that we need to include all the facets of pain in any scientific theory of pain, the first thing to note is that bodily injury drops out as fundamental to pain. Even though acute injury-based pain is the model for most animal-based pain research and pain theorizing, there are simply too many types of non-injury-related pains to have acute injury be the paradigmatic cause of a pain. Indeed, there is a range of well-defined pain-related disorders. Aside from the challenge of chronic pain, there are also allodynia, arthritis, complex regional pain syndrome type 1, causalgia, chronic fatigue syndrome, fibromyalgia, headaches, irritable bowel syndrome, neuropathic pain, orchialgia, phantom limb pain, radicular pain, temporomandibular disorder, and trigeminal neuralgia, among others. There is also a range of headaches, referred pains, neuromas, and cancer pains, as well as things like menstrual pain (cf., Serrahima and Martínez, 2023), that have no obvious “injury” cause and often no obvious cause at all.

IASP has recognized this issue and has divided pain into three broad categories: nociceptive, neuropathic, and nociplastic. Nociceptive pain refers to what we normally think of as injury-based pain; it includes all pains arising from tissue damage. Neuropathic pain arises from damage to the nerves themselves – things like sciatica. Nociplastic pain means something like “altered nociception;” a pain for which there is no (obvious) disease, lesion, or tissue damage (IASP, n.d.; see also Buldys et al., 2023). Identified only in 2016, we currently have no clear idea what nociplastic pain is, other than a painful condition that has no identifiable cause.

I am mentioning the wide range of pains to underscore that pain is indeed complex and multifarious and may be only roughly unified

in terms of its sensation. Explanations of pain could very well be complex, multifarious, and only weakly unified as well. Recent event-related potential (ERP) research provides a nifty example of how one might (start to) build a theory of such complex phenomena.

In ERP studies, multiple very sensitive electrodes that can measure the electrical impulses that are primarily driven by neural interactions (the EEG waves) are placed on the scalp. If research subjects experience painful stimuli, such as thermal heat on their skin, their brains notice, interpret, and respond to the stimuli. Averaging time-locked brain signals across the skull over multiple similar stimuli produces signature activity patterns, which reflect the brain's response to that sort of stimulus. Sophisticated analytical techniques, combined with known brain structures, allow for some internal localization of the origin of the brain responses. In comparison to fMRI scans, ERP studies provide for better temporal resolution but poorer spatial resolution of stimulus-evoked brain responses. We now know that brains can respond across a variety of frequencies to external stimuli, even when responding to the same stimulus over time. Combining stimuli duration and intensity with brain response duration and frequency as well as location estimates can paint a compelling picture of what the brain is doing with information it is receiving from the external world.

A group of scientists working together across several laboratories recently reported that they have identified brain responses that appear keyed to the transition of a painful stimulus to a pain percept. By varying the intensity and the duration of the painful thermal stimulus and then comparing the various localized neural responses as recorded across the scalp with each other and to subjective pain reports, the scientists could demonstrate that both responses reflected subjective pain ratings for duration and intensity. In particular, the sensation of incidental but extended thermal pain (or rather, the report of such a sensation) co-occurred with a low-frequency waveform (< 1 Hz) originating in the insula and the anterior singular cortex (the medial pain system) and an alpha-band (8–13 Hz) desynchronization in the sensorimotor cortex (the lateral pain system).³ The two waveforms were coupled with each other, with the alpha oscillations fluctuating with the low-frequency waveform (Wang H. et al., 2023). This sort of coupling suggests that the underlying brain structures are responding simultaneously to the same inputs and that whatever is going on inside the brain is distributed and complicated. Further, because the duration of the coupling was correlated with the duration of the pain perception, the waveforms also index the experience of pain. All these data suggest that multiple brain regions are involved in converting stimuli to perceptual awareness of the stimuli.

Additionally, the size of the recorded waves over the insula and the anterior singular cortex varied by reported stimulus intensity. This result aligns with previous EEG and fMRI studies (e.g., Atlas et al., 2014; Woo et al., 2015; Tiemann et al., 2018), which supports the idea that these brain waves are correlated with the brain translating stimulus intensity into concomitant sensations of pain intensity. These responses were in the areas that process the salience

² See also Klein (2015) for another example of this approach to pain or Hall (2008) for an example of this approach to itch.

³ Alpha-band event-related desynchronizations have been associated with cognitive and sensorimotor activity in cortex since at least the 1950s (e.g., Gastaut, 1952).

of stimuli, especially where pain is concerned (*cf.*, Guo et al., 2020). These data also dovetail with data from implanted EEG electrodes in patients being monitored for epileptic foci who experienced a range of durations and intensities of thermal pain under controlled conditions (Caston et al., 2023). We would expect a tight correspondence between pain intensity and pain salience. Thus, the waves in the insula and the anterior singular cortex could reflect the impact the brain's estimation of salience has on the perceived intensity of a stimulus (see also Liberati et al., 2018).

I do not intend to lean too heavily on these studies to support any particular conclusion about the substrates of pain experiences, no matter how elegant, but I do intend to use them to suggest a different approach to understanding the brain bases of pain, one that embraces the complexity of brain responses as well as the complexity of pain – and one that looks at more than brain regions and their gross responses to stimuli. I suggest that what makes up a sensation is much more complicated and subtle than philosophers have previously assumed. As explained below, brains are not like combustion engines, with energy driving outputs via the structure of the machine, but are instead more like whirlpools, which are essentially dynamic patterns in some substrates. To reach this conclusion, I shall analogize studies of how the brain navigates in three-dimensional space with how the brain creates pain.

4 Animal navigation

Just like pain processing, navigation in a three-dimensional environment is a complex process. To be useful for the organism, it must combine internal goals and desires with external data and motor planning in real time. What the brain's navigational codes are and how they are implemented have been continuously investigated by neuroscientists for over 50 years, going back to when O'Keefe and Dostrovsky (1971) first identified spatially tuned cells, dubbed “place cells,” in the hippocampus. These cells increased their average firing rates as the animal approached the places to which they were “tuned,” thereby creating a “grid map” that represented the navigable environment around an animal. Over the next half a century, additional spatial cells that were tuned to other animal-environment relationships were also discovered, e.g., allocentric head direction cells (Taube et al., 1990), allocentric border cells (Savelli et al., 2008; Solstad et al., 2008), egocentric boundary cells (Wang et al., 2018; Hinman et al., 2019), etc. At first, these types of cells were only found in and around the hippocampus, which led some to conclude that the hippocampus contained each animal's cognitive map of its world, which in turn supported the animal's movement in its environment. Perhaps, the hippocampal formation could be the navigation center of the brain (see, e.g., O'Keefe and Nade, 1978).

However, not surprisingly, that supposition was too facile, and, over time, scientists have identified many navigational tuning cells throughout the brain, including in the brainstem, cerebellum, and cortex. Indeed, navigational processing seems to be widely distributed throughout the brain. Of course, this makes sense, given that animals must tap their sensory systems, their memory systems, and their motor system to be able to move freely and successfully in complex three-dimensional spaces. At the same time, researchers also learned that the codes that brains used to navigate with were extremely dynamic; they did much more than just passively encode

3-D spatial relationships (see Maisson et al., 2022 for a review). Instead, navigational processes seem to be fundamentally integrated into all the other decision-making that animals must undertake to move about in the world in real time. For instance, the medial temporal cortex in mice integrates sensory inputs, the movements of their eyes and head, and a myriad of other cues to generate a map of landmarks in space (Mallory et al., 2021). Hardcastle et al. (2017) determined that these sorts of neural codes are “highly multiplexed,” “heterogeneous,” and “dynamically adaptive” (*italics mine*). Importantly, this complex structure can support a degree of computational flexibility that allows animals to respond to their ever-changing bodily needs in real time as they navigate across complex landscapes (see also Pessoa et al., 2021).

This sort of theoretical advance reinforces the idea that our brains do not comprise a cortex, doing one set of tasks, riding on top of more primitive subcortical regions, doing a different set of tasks. Instead, the brain consists of widely “distributed and entangled” networks (see Pessoa, 2022; Westlin et al., 2023). That is, the brain is not an assemblage of neural circuits but a large-scale combinatorial system that feeds back on itself through a diversity of convergent and divergent bi-directional connections. The moral of this story is that we should understand what brains are in the same way that meteorologists understand whirlpools (or hurricanes) – as dynamic, context-dependent, spatiotemporal trajectories (*ibid.*, pp. 227–228).

Fortunately, as scientists were beginning to realize that animal navigation was even more complex and distributed than originally envisioned, they were also devising new and better ways to analyze brain activity. They moved from the tuning curves of yore, which were simple peristimulus time histograms, to representational similarity analyses, or “RSA.” In brief, RSA makes pairwise comparisons between conditions of an experimental intervention, using distance matrices to capture the similarity of a given measure for neural activity, behavior, or model output. One can then use these sets of comparisons to analyze whether and how the so-called representational distance matrices, or “RDMs,” vary across contexts, species, regions, models, and so on (*cf.*, Nili et al., 2014). Additionally, new holographic optogenetic techniques, operating on a millisecond timescale (Adesnik and Abdeladim, 2021), permit more realistic representations of individual neuronal interactions. With this technique, researchers can also analyze more than the outputs of a small set of single cells in a brain area. For just one example, Allen et al. (2019) recorded neuronal activity from approximately 24,000 cells simultaneously across 34 cortical and subcortical regions. These recordings demonstrated that it takes only approximately 300 milliseconds for salient sensory stimuli to propagate across the entirety of a rat's brain.

In neuroscience's early days, scientists believed that individual neurons had just one primary task, which determined their coding properties. And these properties changed little over time. For example, an individual head direction cell would encode the direction of the head when it was pointing this way but no other (*cf.*, Taube et al., 1990). Neuroscience's job was to functionally identify all the different types of neurons involved in each deconstructed brain process. We can see this approach in our early understanding of vision: simple cells fed into complex cells, which then built up into more hierarchies and more complex hierarchies (*cf.*, Hubel and Wiesel, 1962).

But today, we have more complex statistical tools that we can use to analyze what cells are responding to, which gives us a different perspective on how brains do their work. Instead of encoding just a single property, we now know that most “navigation” cells simultaneously encode head direction, motion, and location (Sargolini et al., 2006); that is, they are *multiplex*. But even though each of these cells is sensitive to virtually all the features important to the animal moving across its environment, what they are sensitive to differs across cells. It would not be surprising if each cell were to have its own unique combination of informational sensitivities. In other words, neuronal responses are also *heterogeneous* (Hardcastle et al., 2017).

Additionally, we have learned that what cells respond to is not the same across time or conditions. The old view was that if a navigational cell encoded position in one way, it will always encode for position and for position in exactly that way. Researchers have described how animal brains navigate in terms of an internal two-dimensional latitude-longitude coordinate map coupled with an internal compass (cf., Moser and Moser, 2016). However, we now know that if an animal is actively navigating, cell responses become increasingly precise. If it is navigating toward a reward, the neurons record where the reward is more accurately. On the other hand, if an animal is moving slowly and randomly, it responds to location and space less precisely (Hardcastle et al., 2017). In other words, brain cells are adaptive: they become more specific when responding to the more important parts of their environment. The resolution of neurons can change depending on how precise the animal needs it to be in that moment. That is, neural responses are *dynamically adaptive*. And furthermore, multiplex, heterogeneous, and dynamically adaptive brain cells open up new ways of envisioning brain function, as random combinations of variables create a broader space in which brains can learn.

In addition to highlighting the complexity of brain response supporting animal navigation, it should also be indicated that a multiplex, heterogeneous, and dynamically adaptive way of understanding brain function and organization does not belie reduction. No one is claiming that because neurons respond dynamically in a complex manner depending upon animal needs, or because no single area or type of neuron appears to respond to only navigational tasks and nothing else, we cannot reduce animal navigation to brain activity. Instead, researchers are spending their careers trying to map out exactly how brains respond to complex navigational challenges in real time, how neurons work together across regions to move animals to food and shelter, away from foes, and toward mates, depending on their specific hierarchy of needs at that moment.

What if pain is expressed in the same way in brains? What if pain is a highly multiplexed, heterogeneous, and dynamically adaptive process of response to adverse stimuli? Just as we do not understand whirlpools by virtue of the location and directional movement of individual droplets of water, perhaps we should not understand pain in terms of brain areas and static neuronal responses. Instead, we want to know how different brain structures and responses “unfold temporally” to support pain experiences and pain behaviors (Pessoa, 2022, p. 227; see also Westlin et al., 2023). In this case, neuroscientists would strive to understand pain by describing “the *joint state* of brain regions and how it changes,” that is, by describing the brain’s “spatiotemporal trajectories” associated

with pain processes and responses (*ibid.*, p. 228, italics in the original). This approach could keep all the fantastic individuality and complexity of pain but also allow for its reduction to the brain.

5 Pain as heterogeneous, dynamically adaptive, and highly multiplexed neural responses

The idea of neural correlates of pain being at least heterogeneous is not new. Melzack (1999) conceived a “pain neuromatrix” over two decades ago. This matrix references an interconnected network of neural areas that support pain processing, as opposed to a single pain region or pathway in the brain. This network generates distinctive patterns of activation that correspond to different pain experiences (Melzack, 2001). It is divided anatomically and functionally into medial and lateral ascending pathways. The medial pathway processes the affective dimensions of pain via a circuit traveling from the parabrachial nucleus to the amygdala and then to the prefrontal cortex and anterior cingulate cortex. The lateral ascending pathway supports the sensory/discriminative dimensions of pain and is composed of the thalamus, somatosensory cortex, and insular cortex. It is worth noting that this division is here for ease of discussion. It is quite clear that the divisions between affective and sensory information are rather artificial and that there is quite a bit of crosstalk between the two ascending pathways (cf., Giesler et al., 1981; Apkarian, 2012.) There is also a descending pathway that starts in the prefrontal cortex and travels back through the anterior cingulate cortex, amygdala, hypothalamus, and periaqueductal gray, which modulates pain signals in brainstem nuclei that project to the spinal cord (Yao et al., 2023).

Importantly, this network can be influenced by attention and stress, among other states (Tracey et al., 2002; Tracey and Mantyh, 2007; Ploner et al., 2011). How they influence pain perceptions varies across individuals and within the same individual over time. Differential reactions to the same painful stimulus appear to be keyed to differential activation in the dorsolateral prefrontal cortex (Crawford et al., 2023). The results suggest that our attentional processes and other salience networks are also tied into the pain neuromatrix. Emotions too will affect pain experiences (Caston et al., 2023). Seeing someone else react negatively to one’s injury will increase one’s own pain response, as will seeing someone else in pain (Wiech and Tracey, 2009; Budell et al., 2010; Bayet et al., 2014; Jauniaux et al., 2019). These effects appear throughout the spine and seem to reflect a separate modulatory system that evaluates environmental threats, which then facilitates or primes pain responses (Khatibi et al., 2023). To make matters even more complicated, vicarious pain and fear modulate self-pain responses differentially (*ibid.*), perhaps reflecting yet more different but overlapping networks connected to pain responses. As Caston et al. (2023) note, “brain dynamics can shift by changing just one aspect of the stimulus-perception-behavior relationship” (p. 14). These results hint at pain responses being dynamically adaptive.

Is pain processing highly multiplex? We are starting to find clues that it is. We know that pain processing is widely distributed across the brain and it interacts with and is impacted by other

processing networks. As a result, pain experiences are context-dependent and highly individualized (Kucyi and Davis, 2014). The neuromatrix hypothesis implies that pain experiences are tied to synchronized activity across multiple distinct brain regions. However, it could also be the case that pain-sensitive neurons are locally intermingled with neurons that are less sensitive to pain-related information within areas. Given this, is the neural encoding of pain information carried in the brain at a coarse-grained regional level or at a fine-grained local level?

Put another way, we already know that virtually no brain areas are exclusively devoted to processing pain and nothing else (Mouraux et al., 2011; Liberati et al., 2016; Salomons et al., 2016; Su et al., 2019). Multivariate pattern analysis of fMRI data demonstrates that activation patterns in these areas differ between painful and non-painful stimuli, even when stimuli intensity is held constant (Liang et al., 2016).⁴ The question is surrounding the relative level of the pattern. Because fMRI data have limited spatial resolution, is it not clear whether these activity patterns are regionally based or locally based. Are there only pain-specific patterns across areas, or are there pain-specific neuronal responses within areas?

Recent studies suggest that the answer is both. Comparing global and regional multivariate pattern analyses of fMRI data for intensity-matched touch vs. pain stimuli, along with functional connectivity analyses between spatial scales, revealed pain-specific patterns at every level of analysis. Furthermore, the spatial distribution of pain-related processing was unique to individuals, which would explain individual variations in pain perception, pain vigilance, and pain expression (Wang S. et al., 2023). These data, of course, do not tell us that individual neurons are sensitive to pain as well as other stimuli. However, they do show that pain processing is strongly intermingled with other sorts of processing and that this intermingling occurs in all levels of organization thus far examined. Pain neurons might be multiplex, or the analyzed voxels could be multiplex, or both. The point is that the processing that underlies pain is not easily localized, nor does it appear to be devoted exclusively to pain and nothing else.

6 Conclusion: Everything, Everywhere, All at Once

If this way of understanding brain function is correct, then the concerns of Corns, Coninx, and others fall away, for their views on what theories of brain function might look like are mistaken. I agree with Corns that pain processes are not “mechanistic,” but I agree not because there is something different or special about pain but because no complex cognitive/emotional brain processes are mechanistic. Therefore, instead of concluding that differences in pain responses across individuals or over time belie scientific explanations of pain, we can see that such dynamic, heterogeneous, and multiplex responses likely represent normal brain functioning. With a different perspective on understanding brain functioning, it is no longer surprising that different neural structures in different locations across the brain can all somehow be involved in pain

processing, but, at the same time, be individually neither necessary nor sufficient for the experience of pain. Eliminativism is not the only path forward.

It is also not scientifically damning that, as Coninx points out, multivariate pattern analyses of neuroimaging data for pain experiences are unique to individual subjects, pain type, and the larger psychosocial and emotional context. Understanding the brain in terms of dynamic, context-dependent, spatiotemporal trajectories would lead directly to this conclusion. At the same time, we perhaps need not abstract away individual differences among pain cases because we have (and are developing more) neuroscientific tools that allow for complex analyses of multiple variables interacting along multiple dimensions. Patience with scientific advancement might be a better strategy than using family resemblances to support only gross generalizations about pain experiences.

Finally, just as navigational challenges for animals are embedded in complex physical and sociocultural environments, so too is pain processing. Both require individualized brain responses. And just as animal navigation is fundamentally understood in terms of complex biopsychosocial trajectories of brain activity through a theoretical multidimensional space, so too are pain states. If a pain experience is the way that the brain dynamically responds to a particular combination of multidimensional biopsychosocial factors, as Oliver and others intimate, we could still have a very robust neuroscience of pain. This sort of complexity does not prevent a neural theory of pain. Affordance-based approaches could be encompassed in these new approaches.

The approach described herein would not reduce pain experiences, or even pain responses, in the way philosophers have traditionally assumed, but it would reflect the most theoretically grounded and analytically advanced perspectives of how brains work. In summary, pain is more than an unpleasant emotional and sensory experience, despite the IASP assertion to the contrary. While it may only be loosely associated with noxious stimuli, it is still a brain-based response to an animal's internal or external environment. As such, it is something that neuroscientists can study in humans and in animal models. Furthermore, as our experimental and analytic techniques improve and grow ever more sophisticated, so too will our theories of pain processing. What comprises pain experiences is much more complicated and subtle than what philosophers have at least previously assumed. It is far, far too early to begin to throw in the towel and proclaim that a detailed understanding of pain as a brain function is beyond the pale. Our work here is only barely beginning.

The approach adumbrated herein is important not only conceptually but also practically, for it will shape how we treat and care for pain patients. Pain being more than just a sensory response, and bodily injury no longer being required for pain, opens the possibility of greater acceptance and more avenues of treatment for patients with historically dubious sorts of chronic pain, like fibromyalgia and chronic fatigue syndrome, as well as for things like menstrual pain, cancer pain, and other pains whose etiology we do not understand. We should also be able to better understand what nociplastic pain is and, therefore, how to treat it. Pain being essentially a whole-brain response that is integrated with other incoming and self-generated signals allows for nuance and differences across individual pains. The hope, my hope, is that this

⁴ See Iannetti et al. (2013) for a description of these techniques.

perspective will ultimately present ways to re-conceptualize the treatment of pain at a fundamental level.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

References

- Adesnik, H. and Abdeladim, L. (2021). Probing neural codes with two-photon holographic optogenetics. *Nat. Neurosci.* 24:1356–1366. doi: 10.1038/s41593-021-00902-9
- Allen, W. E., Chen, M. Z., Pichamoorthy, R. H., Tien, R. H., Pachitariu, M., Luo, L., et al. (2019). Thirst regulates motivated behavior through modulation of brainwide neural population dynamics. *Science* 364:253. doi: 10.1126/science.aav3932
- Apkarian, A. V. (2012). Chronic pain and addiction pathways. Society of Neuroscience Annual Meeting, New Orleans.
- Apkarian, A. V. (2017). “Advances in the neuroscience of pain” in *Routledge handbook of philosophy of pain*. ed. J. Corns (New York, NY: Routledge), 73–86.
- Apkarian, A. V., Bushnell, M. C., Treede, R.-D., and Zubieta, J.-K. (2005). Human brain mechanisms of pain perception and regulation in health and disease. *Eur. J. Pain* 9, 463–484. doi: 10.1016/j.ejpain.2004.11.001
- Atlas, L. Y., Lindquist, M. A., Bolger, N., and Wager, T. D. (2014). Brain mediators of the effects of noxious heat on pain. *Pain* 155, 1632–1648. doi: 10.1016/j.pain.2014.05.015
- Bastuji, H., Frot, M., Perchet, C., Magnin, M., and Garcia-Larrea, L. (2016). Pain networks from the inside: spatiotemporal analysis of brain responses leading from nociception to conscious perception. *Hum. Brain Mapp.* 37, 4301–4315. doi: 10.1002/hbm.23310
- Bateu, T. M. (2020). Pain in psychology, biology, and medicine: some implications for pain eliminativism. *Stud. Hist. Philos. Biol. Biomed. Sci.* 82:101292. doi: 10.1016/j.shpsc.2020.101292
- Bayet, S., Bushnell, M. C., and Schweinhardt, P. (2014). Emotional faces alter pain perception. *Eur. J. Pain* 18, 712–720. doi: 10.1002/j.1532-2149.2013.00408.x
- Borg, E., Fisher, S. A., Hansen, N., Harrison, R., Ravindran, D., Salomons, T. V., et al. (2021). Pain priors, polyeidism, and predictive power: a preliminary investigation into individual differences in ordinary thought about pain. *Theor. Med. Bioeth.* 42, 113–135. doi: 10.1007/s11017-021-09552-1
- Budell, L., Jackson, P., and Rainville, P. (2010). Brain responses to facial expressions of pain: emotional or motor mirroring? *NeuroImage* 53, 355–363. doi: 10.1016/j.neuroimage.2010.05.037
- Buldys, K., Górnicki, T., Kalka, D., Szuster, E., Biernikiewicz, M., Markuszewski, L., et al. (2023). What do we know about nociceptive pain? *Healthcare (Basel)* 11, 1794–1818. doi: 10.3390/healthcare11121794
- Caston, R. M., Smith, E. H., Davis, T. S., Singh, H., Rahimpour, S., and Rolston, J. D. (2023). Characterization of spatiotemporal dynamics of binary and graded tonic pain in humans using intracranial recordings. *PLoS One* 18:e0292808. doi: 10.1371/journal.pone.0292808
- Coleman, S. (2020). “Painfulness, suffering, and consciousness” in *Philosophy of suffering: metaphysics, value, and normativity*. eds. D. Bain and M. Brady (New York, NY: Routledge), 55–74.
- Coninx, S. (2023). The notorious neurophilosophy of pain: a family resemblance approach to idiosyncrasy and generalizability. *Mind Lang.* 38, 178–197. doi: 10.1111/mila.12378
- Coninx, S., Ray, B. M., and Stilwell, P. (2023a). Unpacking an affordance-based model of chronic pain: a video game analogy. *Phenomenol. Cogn. Sci.*, 1–24. doi: 10.1007/s11097-023-09896-0
- Coninx, S., Willemsen, P., and Reuter, K. (2023b). Pain linguistics: a case for pluralism. *Philos. Q.* 74, 145–168. doi: 10.1093/pq/pqad048
- Cormack, B., Stilwell, P., Coninx, S., and Gibson, J. (2022). The biopsychosocial model is lost in translation: from misrepresentation to an enactive modernization. *Physiother. Theory Pract.* 39, 2273–2288. doi: 10.1080/09593985.2022.2080130
- Corns, J. (2020). *The complex reality of pain*. New York, NY: Routledge.
- Crawford, L., Mills, E., Meylakh, N., Macey, P. M., Macefield, V. G., and Henderson, L. A. (2023). Brain activity changes associated with pain perception variability. *Cereb. Cortex* 33, 4245–4155.
- Djordjevic, C. (2023). Pain cannot (just) be whatever the person says: a critique of a dogma. *Nurs. Philos.* 24, e12446–e12454. doi: 10.1111/nup.12446
- Dong, X., and Dong, X. (2018). Peripheral and central mechanisms of itch. *Neuron* 98, 482–494. doi: 10.1016/j.neuron.2018.03.023
- Evans, K. C., Banzett, R. B., Adams, L., McKay, L., Frackowiak, R. S., and Corfield, D. R. (2002). BOLD fMRI identifies limbic, paralimbic, and cerebellar activation during air hunger. *J. Neurophysiol.* 88, 1500–1511. doi: 10.1152/jn.2002.88.3.1500
- Fulkerson, M. (2023). How thirst compels: an aggregation model of sensory motivation. *Mind Lang.* 38, 141–155. doi: 10.1111/mila.12369
- Gastaut, H. (1952). Etude électrocorticographique de la réactivité des rythmes rolandiques Electrocoricographic study of the reactivity of rolandic rhythm. *Rev. Neurol. (Paris)* 87, 176–182.
- Giesler, G. J. Jr., Yezierski, R. P., Gerhart, K. D., and Willis, W. D. (1981). Spinothalamic tract neurons that project to medial and/or lateral thalamic nuclei: evidence for a physiologically novel population of spinal cord neurons. *J. Neurophysiol.* 46, 1285–1308. doi: 10.1152/jn.1981.46.6.1285
- Guo, Y., Bufacchi, R. J., Novembre, G., Kilintari, M., Moayedi, M., Hu, L., et al. (2020). Ultralow-frequency neural entrainment to pain. *PLoS Biol.* 18:e3000491. doi: 10.1371/journal.pbio.3000491
- Hall, R. J. (2008). If it itches, scratch! *AJP* 86, 525–535.
- Hardcastle, K., Maheswaranathan, N., Ganguli, S., and Giocomo, L. M. (2017). A multiplexed, heterogeneous, and adaptive code for navigation in medial entorhinal cortex. *Neuron* 94, 374–387.
- Hinman, J. R., Chapman, G. W., and Hasselmo, M. E. (2019). Neuronal representation of environmental boundaries in egocentric coordinates. *Nat. Commun.* 10:2772. doi: 10.1038/s41467-019-10722-y
- Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154. doi: 10.1113/jphysiol.1962.sp006837
- Hutto, D. D., and Myin, E. (2013). *Radicalized enactivism: basic minds without content*. Cambridge, MA: The MIT Press.
- Iannetti, G. D., and Mouraux, A. (2010). From the neuromatrix to the pain matrix (and back). *Exp. Brain Res.* 205, 1–12. doi: 10.1007/s00221-010-2340-1
- Iannetti, G. D., Salomons, T. V., Moayedi, M., Mouraux, A., and Davis, K. D. (2013). Beyond metaphor: contrasting mechanisms of social and physical pain. *Trends Cogn. Sci.* 17, 371–378. doi: 10.1016/j.tics.2013.06.002
- International Association for the Study of Pain (IASP). (n.d.) Terminology. Available at: <https://www.iasp-pain.org/resources/terminology/>
- Jauniaux, J., Khatibi, A., Rainville, P., and Jackson, P. L. (2019). A meta-analysis of neuroimaging studies on pain empathy: investigating the role of visual information and observers' perspective. *Soc. Cogn. Affect. Neurosci.* 14, 789–813. doi: 10.1093/scan/nsz055
- Khatibi, A., Roy, M., Chen, J.-I., Gill, L.-N., Piche, M., and Rainville, P. (2023). Brain responses to the vicarious facilitation of pain by facial expressions of pain and fear. *Soc. Cogn. Affect. Neurosci.* 18, 1–11. doi: 10.1093/scan/nsac056
- Klein, C. (2015). *What the body commands: the imperative theory of pain*. Cambridge, MA: MIT Press.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Kucyi, A., and Davis, K. D. (2014). The dynamic pain connectome. *Trends Neurosci.* 38, 86–95. doi: 10.1016/j.tins.2014.11.006
- Legrain, V., Iannetti, G. D., Plaghki, L., and Mouraux, A. (2011). The pain matrix reloaded: a salience detection system for the body. *Prog. Neurobiol.* 93, 111–124. doi: 10.1016/j.pneurobio.2010.10.005
- Liang, X., Zou, Q., He, Y., and Yang, Y. (2016). Topologically reorganized connectivity architecture of default-mode, executive-control, and salience networks across working memory task loads. *Cereb. Cortex* 26, 1501–1511. doi: 10.1093/cercor/bhu316
- Liberati, G., Algoet, M., Klöcker, A., Santos, S. F., Ribeiro-Vaz, J. G., Raftopoulos, C., et al. (2018). Habituation of phase-locked local field potentials and gamma-band oscillations recorded from the human insula. *Sci. Rep.* 8:8265. doi: 10.1038/s41598-018-26604-0
- Liberati, G., Klöcker, A., Safronova, M. M., Ferrão Santos, S., Vaz, J.-G. R., Raftopoulos, C., et al. (2016). Nociceptive local field potentials recorded from the human insula are not specific for nociception. *PLoS Biol.* 14:e1002345. doi: 10.1371/journal.pbio.1002345
- Liu, M. (2023). The polysemy view of pain. *Mind Lang.* 38, 198–217. doi: 10.1111/mila.12389
- Maisson, D. J.-N., Wikenheiser, A., Noel, J.-P., and Keinath, A. T. (2022). Making sense of the multiplicity and dynamics of navigational codes in the brain. *J. Neurosci.* 42, 8450–8459. doi: 10.1523/JNEUROSCI.1124-22.2022
- Mallory, C. S., Hardcastle, K., Campbell, M. G., Attinger, A., Low, I. I. C., Raymond, J. L., et al. (2021). Mouse entorhinal cortex encodes a diverse repertoire of self-motion signals. *Nat. Commun.* 12:671. doi: 10.1038/s41467-021-20936-8
- Melzack, R. (1999). From the gate to the neuromatrix. *Pain* 82, S121–S126. doi: 10.1016/S0304-3959(99)00145-1
- Melzack, R. (2001). Pain and the neuromatrix in the brain. *J. Dent. Educ.* 65, 1378–1382. doi: 10.1002/j.0022-0337.2001.65.12.tb03497.x
- Moser, M.-B., and Moser, E. I. (2016). Where am I? Where am I going? *Sci. Am.* 314, 26–33. doi: 10.1038/scientificamerican0116-26
- Mouraux, A., Diukova, A., Lee, M. C., Wise, R. G., and Iannetti, G. D. (2011). A multisensory investigation of the functional significance of the “pain matrix”. *NeuroImage* 54, 2237–2249. doi: 10.1016/j.neuroimage.2010.09.084
- Nagel, E. (1961). *The structure of science. Problems in the logic of explanation*, New York: Harcourt, Brace and World, Inc.
- Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Comput. Biol.* 10:e1003553. doi: 10.1371/journal.pcbi.1003553
- O’Keefe, J., and Dostrovsky, J. (1971). The hippocampus as a spatial map: preliminary evidence from unit activity in the freely moving rat. *Brain Res.* 34, 171–175. doi: 10.1016/0006-8993(71)90358-1
- O’Keefe, J., and Nade, L. L. (1978). *The Hippocampus as a cognitive map*. Oxford: Oxford University Press.
- Oliver, A. (2022). The social dimension of pain. *Phenomenol. Cogn. Sci.*, 1–34. doi: 10.1007/s11097-022-09879-7
- Pessoa, L. (2022). *The entangled brain: how perception, cognition, and emotion are woven together*. Cambridge, MA: The MIT Press
- Pessoa, L., Medina, L., and Desfilis, E. (2021). Refocusing neuroscience: moving away from mental categories and towards complex behaviors. *Philos. Trans. R. Soc. B* 377:20200534. doi: 10.1098/rstb.2020.0534
- Ploner, M., Lee, M. C., Wiech, K., Bingel, U., and Tracey, I. (2011). Flexible cerebral connectivity patterns subserve contextual modulations of pain. *Cereb. Cortex* 21, 719–726. doi: 10.1093/cercor/bhq146
- Raja, S. N., Carr, D. B., Cohen, M., Finnerup, N. B., Flor, H., Gibson, S., et al. (2020). The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises. *Pain* 161, 1976–1982. doi: 10.1097/j.pain.0000000000001939
- Salomons, T. V., Iannetti, G. D., Liang, M., and Wood, J. N. (2016). The “pain matrix” in pain-free individuals. *JAMA Neurol.* 73, 755–756. doi: 10.1001/jamaneurol.2016.0653
- Sargolini, F., Fyhn, M., Harfint, T., McNoughton, B. L., Witter, M. P., Moser, M. -B., et al. (2006). Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science* 312, 758–762. doi: 10.1126/science.1125572
- Savelli, F., Yoganarasimha, D., and Knierim, J. J. (2008). Influence of boundary removal on the spatial representations of the medial entorhinal cortex. *Hippocampus* 18, 1270–1282. doi: 10.1002/hipo.20511
- Serrahima, C., and Martínez, M. (2023). The experience of dysmenorrhea. *Synthese* 201:173. doi: 10.1007/s11229-023-04148-9
- Solstad, T., Boccara, C. N., Kropff, E., Moser, M. B., and Moser, E. (2008). Representation of geometric borders in the entorhinal cortex. *Science* 322, 1865–1868. doi: 10.1126/science.1166466
- Su, Q., Qin, W., Yang, Q., Yu, C., Qian, T., Mouraux, A., et al. (2019). Brain regions preferentially responding to transient and iso-intense painful or tactile stimuli. *NeuroImage* 192, 52–65. doi: 10.1016/j.neuroimage.2019.01.039
- Taube, J. S., Muller, R. U., and Ranck, J. B. (1990). Head-direction cells recorded from the postsubiculum in freely moving rats: I. Description and quantitative analysis. *J. Neurosci.* 10, 420–435. doi: 10.1523/JNEUROSCI.10-02-00420.1990
- Tiemann, L., Hohn, V. D., TaDinh, S., May, E. S., Nickel, M. M., Gross, J., et al. (2018). Distinct patterns of brain activity mediate perceptual and motor and autonomic responses to noxious stimuli. *Nat. Commun.* 9:4487. doi: 10.1038/s41467-018-06875-x
- Tracey, I., and Mantyh, P. W. (2007). The cerebral signature for pain perception and its modulation. *Neuron* 55, 377–391. doi: 10.1016/j.neuron.2007.07.012
- Tracey, I., Ploghaus, A., Gati, J. S., Clare, S., Smith, S., Menon, R. S., et al. (2002). Imaging attentional modulation of pain in the periaqueductal gray in humans. *J. Neurosci.* 22, 2748–2752. doi: 10.1523/JNEUROSCI.22-07-02748.2002
- Wang, C., Chen, X., Lee, H., Deshmukh, S. S., Yoganarasimha, D., Savelli, F., et al. (2018). Egocentric coding of external items in the lateral entorhinal cortex. *Science* 362, 945–949. doi: 10.1126/science.aau4940
- Wang, H., Guo, Y., Tu, Y., Peng, W., Lu, X., Bi, Y., et al. (2023). Neural responses responsible for the translation of sustained nociceptive inputs into subjective pain experience. *Cereb. Cortex* 33, 634–650. doi: 10.1093/cercor/bhac090
- Wang, S., Su, Q., Qin, W., Yu, C., and Lian, M. (2023). Fine-grained and coarse-grained spatial patterns of neural activity measured by functional MRI show preferential encoding of pain in the human brain. *NeuroImage* 272:120049. doi: 10.1016/j.neuroimage.2023.120049
- Westlin, C., Theriault, J. E., Katsumi, Y., Nieto-Castanon, A., Kucyi, A., Ruf, S. F., et al. (2023). Improving the study of brain-behavior relationships by revisiting basic assumptions. *Trends Cogn. Sci.* 27, 246–257. doi: 10.1016/j.tics.2022.12.015
- Wiech, K., and Tracey, I. (2009). The influence of negative emotions on pain: behavioral effects and neural mechanisms. *NeuroImage* 47, 987–994. doi: 10.1016/j.neuroimage.2009.05.059
- Woo, C. W., Roy, M., Buhle, J. T., and Wager, T. D. (2015). Distinct brain systems mediate the effects of nociceptive inputs and self-regulation on pain. *PLoS Biol.* 13:e1002036. doi: 10.1371/journal.pbio.1002036
- Yao, D., Chen, Y., and Chen, G. (2023). The role of pain modulation pathway and related brain regions in pain. *Rev. Neurosci.* 34, 899–914. doi: 10.1515/revneuro-2023-0037



OPEN ACCESS

EDITED BY

Xerxes D. Arsiwalla,
Wolfram Research, Inc., United States

REVIEWED BY

Adam Safron,
Johns Hopkins University, United States

*CORRESPONDENCE

Scott Andersen
✉ scott.andersen89@gmail.com

RECEIVED 16 February 2023

ACCEPTED 07 February 2024

PUBLISHED 22 March 2024

CITATION

Andersen S (2024) The maps of meaning
consciousness theory.
Front. Psychol. 15:1161132.
doi: 10.3389/fpsyg.2024.1161132

COPYRIGHT

© 2024 Andersen. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

The maps of meaning consciousness theory

Scott Andersen^{1,2*}

¹United States Department of Homeland Security, Washington, DC, United States, ²Liberty University,
Lynchburg, VA, United States

In simple terms, consciousness is constituted by multiple goals for action and the continuous adjudication of such goals to implement action, which is referred to as the maps of meaning (MoM) consciousness theory. The MoM theory triangulates through three parallel corollaries: action (behavior), mechanism (morphology/pathophysiology), and goals (teleology). (1) An organism's consciousness contains fluid, nested goals. These goals are not intentionality, but intersectionality, via the Darwinian byproduct of embodiment meeting the world, i.e., Darwinian inclusive fitness or randomization and then survival of the fittest. (2) These goals are formed via a gradual descent under inclusive fitness and are the abstraction of a "match" between the evolutionary environment and the organism. (3) Human consciousness implements the *brain efficiency hypothesis*, genetics, epigenetics, and experience-crystallized efficiencies, not necessitating best or objective but fitness, i.e., perceived efficiency based on one's adaptive environment. These efficiencies are objectively arbitrary but determine the operation and level of one's consciousness, termed as *extreme thrownness*. (4) Since inclusive fitness drives efficiencies in the physiologic mechanism, morphology, and behavior (action) and originates one's goals, embodiment is necessarily entangled to human consciousness as it is at the intersection of mechanism or action (both necessitating embodiment) occurring in the world that determines fitness. (5) Perception is the operant process of consciousness and is the *de facto* goal adjudication process of consciousness. Goal operationalization is fundamentally efficiency-based via one's unique neuronal mapping as a byproduct of genetics, epigenetics, and experience. (6) Perception involves information intake and information discrimination, equally underpinned by efficiencies of inclusive fitness via extreme thrownness. Perception is not a 'frame rate' but Bayesian priors of efficiency based on one's *extreme thrownness*. (7) Consciousness and human consciousness are *modular* (i.e., a scalar level of richness, which builds up like building blocks) and *dimensionalized* (i.e., cognitive abilities become possibilities as the emergent phenomena at various modularities such as the stratified factors in factor analysis). (8) The meta dimensions of human consciousness seemingly include intelligence quotient, personality (five-factor model), richness of perception intake, and richness of perception discrimination, among other potentialities. (9) Future consciousness research should utilize factor analysis to parse modularities and dimensions of human consciousness and animal models.

KEYWORDS

consciousness, animal model, human model, prospection, personality, intelligence quotient, perception, inhibition

Introduction

Science progresses through both theory and experiment. Moreover, theory precedes experiment *a priori* and guides empirical research (Seth and Bayne, 2022). This article seeks to propose a theory of consciousness for empirical research based on a first-principles conceptualization of the notion of consciousness. This theory seeks to address these notions that are insufficiently covered in many current conceptualizations of consciousness:

- Evolutionary biology acted as the mechanics for the development of consciousness, i.e., random genetic variation, the organism's embodiment meeting the world to subsequently determine fitness and then survival of the fittest in a constantly changing environment.
- *Occam's Razor*: Just because the effects of something may be profound, neither the explanation nor the mechanics behind such effects need to be profound (Stanford Encyclopedia of Philosophy, 2022a). To my knowledge, this notion is commonly lacking in the discussions of consciousness, as consciousness is assumed by some to be a magical phenomenon and/or unique to the complexity of the human mind. I seek to enlighten the use of these notions.
- *Hume's Dilemma*: One cannot derive an *ought* from an *is* (Stanford Encyclopedia of Philosophy, 2022b). It seems that most of the discourse on consciousness imposes an *ought* on human consciousness, as some unique spark of the divine exclusive to humanity. This perception perhaps could simply be attributed to Kahneman's (automatic) System 1 thinking (Kahneman, 2011) or McGilchrist's predominance of left-brain thinking societally, generally in the current times (McGilchrist, 2019), and I seek to clarify this notion.

Equally, it is important to keep in mind the appropriate level of empiricism associated with the scientific, well-defined notion of a *theory*. A theory is broadly conceptualized as a construct or system of ideas that collectively seek to explain a phenomenon in the world, independent of the other phenomena. Essentially, one could think of a theory as a unique, emergent phenomenon from a certain series of ideas (Robson and McCartan, 2016). For example, Darwin's theory of natural selection consisted of the ideas of random genetic variations and how the matching of those random variations to an environment leads to *fitness*, whereby such a fitness leads to increased offspring production compared to those random genetic variations lacking that fitness, which could be referred to as the *survival of the fittest*. Overall, currently, random genetic variation, with a subsequent fitness to the environment and then ultimately leading to the survival of the fittest, is the theory of *natural selection*.

Consciousness defined

In simple terms, consciousness should be seen as the irrevocable (and unexceptional) byproduct of having multiple goals for action in the world and the process by which one continuously adjudicates across such goals to implement action continuously in the world.

Key concepts

Fluid hierarchy of goals

- The fundamental constitution of consciousness involves having multiple goals and continuously adjudicating across such goals so as to operationalize action continuously in the world.
- o One can think of an organism's consciousness as containing a fluid, but a nested, hierarchy of goals.
- o These goals are not a matter of intentionality, but intersectionality, via the Darwinian byproduct of embodiment meeting the world. It is a matter of inclusive fitness, i.e., random variation and then survival of the fittest.
- For all intents and purposes, the MoM consciousness theory teleologically conceptualizes these goals as representations of a "match" of inclusive fitness to an adaptive evolutionary environment, past or present, though apomorphy, which may equally originate from the genetic fringes.
- These goals have a corollary action in the world and a corollary mechanism among the organism (e.g., the need for energy for a human has the corollary action of obtaining food in the world and is activated mechanistically and biochemically via the hormone Ghrelin, among other mechanisms and complexities).
- Some goals have simple corollary actions and mechanisms, while other goals have complex corollary actions and mechanisms. The continuous process of adjudicating one's nested hierarchy of goals makes it possible for the organism to manifest any of its goals, at any time, along with the subsequent action as mediated mechanistically under the right conditions.
- o Some corollaries of goals, mechanisms, and/or actions are fractals among the organism, i.e., the same phenomena, but manifesting differently at multiple levels of analysis.
- This is a phenomenon demonstrated beyond human models and beyond reproach by the work of Dr. Michael Levin and his lab (Tufts University, 2023) and Dr. Josh Bongard and his lab (Kudithipudi et al., 2022), referred to as *multilevel competencies*.
- o These multilevel competencies (and subsequent goals) begin at the most fundamental levels of biology (Levin, 2023a, 2023b).
- The most fundamental of the organism's goals of addressing entropy subsequently appears for humans, fractally at the individual neuron and the whole of the brain, which fundamentally crystallizes prediction-to-outcome matches in the world (i.e., prospection) so as to reduce entropy and free energy of thinking and behaviors of the organisms in the future (Carhart-Harris et al., 2014).
- Moreover, from the very start of the earthly phylogenetic tree, the need for energy manifests intrinsically among organisms, nested upon the presuppositions of both homeostasis and replication, with the corollary action of movement primitively (Kumar and Philominathan, 2010; Swiecicki et al., 2013). Then, the complex physiological mechanism later manifests as, for example, hunger in humans.
- Numerous examples of this phenomenon are observed; for instance, in human pathophysiology, the renin-angiotension system is a higher order representation of mechanistic osmotic equilibration at the cellular level.

- The organism's goals, hierarchy of goals, and means by which to adjudicate across such goals continuously (goal adjudication conceptualized in this theory as "perception") are the modest byproducts of the experiential existence in which an organism or an organism's consciousness finds itself.
- o Heidegger referred to the randomness of the time and place in which your consciousness finds itself as *thrownness* (Heidegger, 2008).
- o Equally, one should conceptualize that the very specific time and place in which consciousness finds itself also determines how that very consciousness develops and the meanings it endows on the world as such, i.e., genetics, epigenetics, and experience further shape one's goals, Bayesian priors, neuronal architecture, and subsequently consciousness. The MoM consciousness theory terms this concept *extreme thrownness*.

Consciousness entangles embodiment

- One can think of an organism's consciousness as containing a fluid, but nested, hierarchy of goals.
- o These goals are not a matter of intentionality, but intersectionality, via the Darwinian byproduct of embodiment meeting the world. It is a matter of inclusive fitness, i.e., random variation and then survival of the fittest.
- For all intents and purposes, the MoM consciousness theory teleologically conceptualizes these goals as representations of a 'match' of inclusive fitness to an adaptive evolutionary environment, past or present.
- Consciousness at its core is the mapping of our embodiment onto the world, which forms the very goals and hierarchy of the goals of consciousness because it is the environment that determines which Darwinian random variations survive to become the fittest.
- o One might argue if embodiment is *ipso facto* required for consciousness, and validly so, but it is beyond reproach to conceptualize earthly and subsequently human consciousness as anything other than what is necessarily entangled to embodiment.
- This theory speaks purely to the notion of earthly, evolutionarily derived consciousness, as instantiated particularly in human experience or Heidegger's concept of *dasein* (Heidegger, 2008), often referred to in the research body as *phenomenological consciousness* (Carruthers, 2019, p. 41).
- o The entanglement of consciousness and embodiment as such makes it beyond the scope of this theory to apply this conceptualization of consciousness to artificial intelligence. While, evidently, the MoM consciousness theory should apply to artificial intelligence, the latter is not subject to the same notion of embodiment or its direct mapping onto the world with the very mechanism of goal orientation occurring via natural selection.
- One could equally (perhaps necessarily) argue for an empirical study of the consciousness of artificial intelligence utilizing the MoM consciousness theory.
- o Additionally, if one can conceptualize *gradual descent* or *gradient descent* (i.e., continuous complexification from optimization to a changing environment) as integral to and the very underpinnings of the evolutionary-derived and evolutionary-directed process of inclusive fitness, a brain or likely even a nervous system, which, ergo, is not required for complex behaviors to be associated with consciousness (Ryan and Grant, 2009).

- For instance, Botton-Amiot et al. (2023) demonstrate that the anemone sea starlet species *N. vectensis* is able to form associative memory when subjected to classical conditioning, i.e., this is a simple sea organism absent of the central nervous system, demonstrating learned behavior and memory. Botton-Amiot et al. (2023) state "these results root associative learning before the emergence of [*nervous system*] centralization in the metazoan lineage and raise fundamental questions about the origin and evolution of cognition in brainless animals" (p. 1).
- Levin (2023a, 2023b) equally and empirically demonstrates goals and the Bayesian priors of associative learning from these goals as much more fundamentally stored among every organism than purely among the centralized nervous system.

Perception is goal adjudication

- The fundamental constitution of consciousness involves having multiple goals and continuously adjudicating across such goals so as to operationalize action continuously in the world.
- o Perception is *de facto* the operant process of consciousness and is the very process of goal adjudication.
- The process of goal adjudication through consciousness (i.e., perception) involves components of both information intake and information inhibition/discrimination (Carruthers, 2019).
- o The intake of information for perception involves not just raw information intake (e.g., vision) but also systems of value judgments that are patterns based on genetics and epigenetics (e.g., IQ and personality) and experience (e.g., learned patterns and behaviors) of the organism or organism's consciousness that equally narrows the process of raw information intake to the simplest operable conceptualizations of understanding.
- For example, an individual's understanding of a helicopter is sufficient for how they personally act in the world, though they likely could not fly, nor fix, nor explain in detail the mechanics of how a helicopter works.
- Perception should likely be conceptualized as a key factor or a *dimension* of consciousness, which is comprised of multiple sub-dimensions of consciousness as organisms maintain multiple constructs of sensory perception (Birch et al., 2020), as the very goal-oriented decisions that organisms make (to include humans) and is evolutionarily derived and evolutionarily directed toward an implicit notion of value judgment or goal rank-ordering in perception. See Birch et al. (2020) for an example of a dimensionalized construct of animal consciousness.
- o For example, why do you not stare at one single molecule of one single object, endlessly for the entirety of one's life, as that singular thing contains an infinite amount of complexity that could never exhaust one's perception during one's lifespan?
- The process of the brain seeking to be as efficient as possible in information discrimination is known as the *brain efficiency hypothesis*, a well-established finding in neuroscience (Basu et al., 2022).
- o These very attempts at efficiency are a byproduct of *extreme thrownness*, i.e., these efficiencies from the organism's goals, hierarchy of goals, and means by which to adjudicate across such goals

continuously (known as perception) are a product of the environment in which an organism or an organism's consciousness finds itself.

Evolutionary derived and evolutionarily directed

- The fundamental constitution of consciousness involves having multiple goals and continuously adjudicating across such goals so as to operationalize action continuously in the world.
- o These goals are the modest byproducts over time of the optimization strategies that organisms utilize in various environments, with a tendency for increasing complexity over time due to continuous optimization to a changing environment. The trend of increasing complexity over time from continuous optimization is a phenomenon known as *gradient descent* or *gradual descent* (Theodoridis, 2015).
- o This operant and differing goal-seeking and adjudication of consciousness, being most fundamentally evolutionarily derived and evolutionarily directed, scales from the most basic goal of matter in combating entropy (Shcherbakov, 2005) and the most basic goals of survival at the cellular and sub-cellular levels.
- Operationalizing the competing goal of survival quintessentially and cellularly further stratifies into the survival mechanisms of homeostasis (during states of low hospitability) and replication (during states of high hospitability), which stand opposite to the perspectives of action in the world (Sinclair and LaPlante, 2019).
- o As organisms scale to various higher levels of complexity, the organism's goals also scale to more nested, more complex differing goals, creating highly complex, highly nested rank-orders of differing goals in the world for the organism (though all fundamentally nested toward addressing entropy, further stratified into the tensioned, cellular goals of homeostasis and replication).
- These goals form not as a matter of intentionality, but intersectionality, via the Darwinian by-product of embodiment meeting the world via inclusive fitness, i.e., random variation and then survival of the fittest.
- o While some goals in the hierarchy are more readily operationalizable or useful, human consciousness is thoroughly filled with “ghost in the machine” goals, which may become operationalized when the appropriate threshold of mechanistic conditions arise
- exemplified by the mismatch of human taste preferences to the modern environment (Breslin, 2013),
- exemplified by the duality of the human mind, which seeks efficiency preeminently (Basu et al., 2022), manifested through the mode of thinking of simplification of the world (Kahneman, 2011; McGilchrist, 2019), and replete with cognitive heuristics, biases, and fallacies.
- empirically exemplified by Schaffner et al. (2023) who demonstrate that sensory perception and its Bayesian encoding of priors so as to tune perception as a matter of fitness maximization.

Modular and dimensionalized

- Human consciousness is centered around the central nervous system (i.e., the brain), and the functional unit of the brain

is the neuron. The neuron's goal, corollary to the mechanism, is the prediction of outcomes followed by the process of plasticization based on the predication-to-outcome match or mismatch in the world (Wacongne et al., 2011; Pitts et al., 2018; Carruthers, 2019; Pereira et al., 2021). This concept is commonly referred to in the literature as *prospexion* (Carruthers, 2019).

- o The process of *prospexion* and subsequent plasticization in one's brain builds networks (mechanistically) and patterns of behavior (the corollary action) for operationalization based on how one's brain develops. These networks and patterns may not necessarily be what is best, right, objective, or most useful. We operate an internally and experientially built model of the world, we operate *maps of meaning*.
- o Pioneering neuroscientist Sokolov (2001) wrote decades ago about this, with one's orienting reflex acting as a kind of adjudication mechanism as such between our internal model(s) of the world and the actuality of the world around us.
- o Additionally, the recent work of Cazettes et al. (2023) in the neuroscientific study of mice found that whatever behavior a mouse implemented and subsequently, regardless of the explanatory process the mice utilized to implement the said behavior, the secondary motor cortex of such mice still encoded the full set of possible behaviors for the situation, i.e., the mice essentially simulated all the behaviors, encoded such into memory, and then acted in the manner they best saw fit.
- As various complexities of goals and goal hierarchies form, along with subsequently more and more complex processes of perception, the human brain equally develops more and more complexity of neuronal networks (the mechanism) and cognitive abilities (the corollary action).
- o Braddick (2001) and Sapolsky (2005) annotate these neuronal networks, as they go from inner layers to outer layers and encode more specific information, which could equally be described as attributed to more complex goal schemas and Bayesian priors.
- As certain levels of complexity of consciousness and subsequently the human brain form, termed *modularities* in the MoM consciousness theory, certain cognitive abilities simply emerge as the emergent phenomena (Gruber and Voneche, 1977; Carruthers, 2019; Birch et al., 2020).
- o For example, Piaget discovered that the cognitive ability of *conservation* emerges (the action) around in children when their brains have reached the corollary level of complexity (mechanistically via *prospexion*) (Gruber and Voneche, 1977).
- Even if Piaget's conservation is learned, a finding not exactly known or most likely currently, it still stands that a pre-requisite level of cognitive complexity must be met so as to “learn” conservation, i.e., emergence is not pre-determined, but as a set of possibilities at the necessary pre-requisite complexity. This is the *neural pre-requisites* line of thinking within the research body.
- These emergent phenomena that simply emerge at varying modularities of consciousness are termed *dimensions* in the MoM consciousness theory.
- o Dimensions of human consciousness have sub-dimensions, and these sub-dimensions have further sub-dimensions, and so on.

- o One may think of consciousness and human consciousness as building blocks. Human consciousness has simply built up to a high level of complexity or *modularity* of consciousness, with a vast number of *dimensions* (or corollary goals via cognitive abilities) as a result of the complexification of goals and corollary mechanisms via evolution with a gradual descent.
- Each instantiation of a species' consciousness must be recognized as a rough type and *sui generis* (Carruthers, 2019) due to the Darwinian landscape in which said consciousness developed. However, since there are genetic, epigenetic, and experiential components to the development of consciousness (via prospection mechanistically), it stands that each organism across a species (e.g., each individual human) additionally may vary in their modularity (and subsequent dimensions) of consciousness.
- o Each unique instantiation of consciousness across a species may have been subtly nuanced, and different modularities and/or dimensions of consciousness are *causa sui* of their differing goals, their particular embodiment (and as mapped onto the world), and their unique neural plasticity, all of which have a context of *extreme thrownness*.
- Alexander Luria demonstrated how human individuals in rural societies lacked certain abstraction abilities possessed by human individuals in industrialized societies, i.e., their dimensions of consciousness (and modularity) may have been lower due to the extreme thrownness of the time, place, and experiences of how their consciousness developed. He notably discovered this finding through the use of IQ tests, particularly the sub-test of Raven's Progressive Matrices (Epstein, 2018).
- Conceptually, factors or *dimensions* of consciousness within the research body have been suggested to include, but are not limited to, memory (LeDoux and Lau, 2020), self-awareness, or selfhood (Brown et al., 2019), or attention (Pitts et al., 2018), among other potential factors.
- The MoM consciousness theory instead posits the most metacognitive dimensions of human consciousness likely as richness of perception (intake), richness of perception inhibition or discrimination, and some existing measures in psychology, with perhaps others yet to be determined.
- o Intelligence quotient (IQ) is the single most reliable and valid measure of individual differences and human life outcomes (Roberts et al., 2007; Ritchie, 2016; Jarrett, 2021), with IQ being a singular measure of a construct of human cognitive abstraction abilities (referred to as "g").
- IQ is a well-established, stratified construct where lower levels and series of cognitive abilities presupposes higher levels and higher series of cognitive abilities, as exemplified by the aforementioned *three stratum theory of cognitive abilities* from Carroll (2005).
- Not only does the construct of IQ evince the modularity and dimensionalization of consciousness of the MoM theory, its predictive validity and Alexander Luria's work suggest the construct as a key component of the human instantiation of consciousness. One could think of IQ as *the measure of richness of one's prospection abilities*.
- o Psychometric measures of personality, particularly as measured via the five-factor model of personality or

colloquially *the big five*, is another well-established, stratified construct of individual differences in psychology with high reliability and validity.

- The stratification of *the big five* is exemplified by the Big Five Aspect Scale (DeYoung et al., 2007).
- Personality is essentially the measure of the metacognitive mosaics of human value judgments and as such seems likely a dimension of human consciousness. One might consider personality as *the measure of the extreme thrownness of one's consciousness*.

Future research

- The seemingly right approach to studying consciousness is factor analysis. Factor analysis is a statistical analysis methodology, common in research, that conceptually identifies the unique, big picture concepts among data sets.
- o For example, factor analysis is the means by which various notions such as intelligence quotient and measures of personality psychology have been derived among various constructs of various fields.
- o Factors, essentially termed *dimensions* in the MoM theory, of human consciousness may include but are not limited to IQ, personality (specifically the five-factor model of personality), perceptual richness, and perceptual inhibition/discrimination richness.
- Ample opportunities are available to study both animal models and human models of consciousness factor analytically.
- o For instance, the organism *C. elegans* with a fully mapped nervous system of 302 neurons and extensive research of its epigenetic mechanisms provides unique opportunity to parse a specific modularity (and subsequently dimensions) of consciousness along with the corollary goals, actions, and mechanisms.
- *C. elegans* is uniquely suited for consciousness research as presupposed upon the MoM consciousness theory due to the work of Oded Rechavi and others in detailing the epigenetics of *C. elegans*.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Conflict of interest

SA was employed by the United States Department of Homeland Security (DHS), this paper reflects solely the opinions of the author, and is neither a representation of the opinions of DHS nor constitutes work performed by DHS. SA was a doctoral student currently at Liberty University, this paper reflects solely the opinions of the author, and is neither a representation of the opinions of Liberty University nor constitutes work performed for Liberty University.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Basu, K., Appukuttan, S., Manchanda, R., and Sik, A. (2022). Difference in axon diameter and myelin thickness between excitatory and inhibitory callosally projecting axons in mice. *Cereb. Cortex* 33, 4101–4115. doi: 10.1093/cercor/bhac329
- Birch, J., Schnell, A. K., and Clayton, N. S. (2020). Dimensions of animal consciousness. *Trends Cogn. Sci.* 24, 789–801. doi: 10.1016/j.tics.2020.07.007
- Botton-Amiot, G., Martinez, P., and Sprecher, S. G. (2023). Associative learning in the cnidarian *Nematostella vectensis*. *PNAS* 120:e2220685120. doi: 10.1073/pnas.2220685120
- Braddick, O. (2001). "Neural basis of visual perception" in *International encyclopedia of the social & behavioral sciences*. eds. N. J. Smelser and P. B. Baltes (Oxford: Pergamon), 16269–16274.
- Breslin, P. A. S. (2013). An evolutionary perspective on food and human taste. *Curr. Biol.* 23, R409–R418. doi: 10.1016/j.cub.2013.04.010
- Brown, R., Lau, H., and LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends Cogn. Sci.* 23, 754–768. doi: 10.1016/j.tics.2019.06.009
- Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., et al. (2014). The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Front. Hum. Neurosci.* 8:20. doi: 10.3389/fnhum.2014.00020
- Carroll, J. B. (2005). "The three-stratum theory of cognitive abilities" in *Contemporary intellectual assessment: theories, tests, and issues*. eds. D. P. Flanagan and P. L. Harrison (New York City: The Guilford Press), 69–76.
- Caruthers, P. (2019). *Human and animal minds: The consciousness questions laid to rest*. Oxford: Oxford University Press
- Cazettes, F., Mazzucato, L., Murakami, M., Morais, J. P., Augusto, E., Renart, A., et al. (2023). A reservoir of foraging decision variables in the mouse brain. *Nat. Neurosci.* 26, 840–849. doi: 10.1038/s41593-023-01305-8
- DeYoung, C. G., Quilty, L. C., and Peterson, J. B. (2007). Between facets and domains: 10 aspects of the big five. *J. Pers. Soc. Psychol.* 93, 880–896. doi: 10.1037/0022-3514.93.5.880
- Epstein, D. (2018). *Range: Why generalists triumph in a specialized world*. London: Penguin Publishing
- Gruber, H. E., and Voneche, J. J. (1977). *The essential Piaget*. New York City: Basic Books Inc.
- Heidegger, M. (2008). *Being and time*. New York City: Harper Perennial Classics
- Jarrett, C. (2021). *Be who you want: Unlocking the science of personality change*. New York: Simon & Schuster
- Kahneman, D. (2011). *Thinking fast and slow*. Farrar, Straus and Giroux, New York
- Kudithipudi, D., Aguilar-Simon, M., Babb, J., et al. (2022). Biological underpinnings for lifelong learning machines. *Nat. Mach. Intell.* 4, 196–210. doi: 10.1038/s42256-022-00452-0
- Kumar, M. S., and Philominathan, P. (2010). The physics of flagellar motion of *E. coli* during chemotaxis. *Biophys. Rev.* 2, 13–20. doi: 10.1007/s12551-009-0024-5
- LeDoux, S., and Lau, H. (2020). Seeing consciousness through the lens of memory. *Curr. Biol.* 30, R1018–R1022. doi: 10.1016/j.cub.2020.08.008
- Levin, M. (2023a). Darwin's agential materials: evolutionary implications of multiscale competency in developmental biology. *Cell. Mol. Life Sci.* 80:142. doi: 10.1007/s00018-023-04790-z
- Levin, M. (2023b). Bioelectric networks: the cognitive glue enabling evolutionary scaling from physiology to mind. *Anim. Cogn.* 26, 1865–1891. doi: 10.1007/s10071-023-01780-3
- McGilchrist, I. (2019). *The master and his emissary: The divided brain and the making of the western world*. New Haven, CT: Yale University Press
- Pereira, M., Megevand, P., Tan, M. X., Chang, W., Wang, S., Rezai, A., et al. (2021). Evidence accumulation relates to perceptual consciousness and monitoring. *Nat. Commun.* 12:3261. doi: 10.1038/s41467-021-23540-y
- Pitts, M. A., Lutsyshyna, L. A., and Hillyard, S. A. (2018). The relationship between attention and consciousness: an expanded taxonomy and implications for 'no-report' paradigms. *Philos. Trans. R. Soc. B* 373:20170348. doi: 10.1098/rstb.2017.0348
- Ritchie, S. (2016). *Intelligence: All that matters*. London: Teach Yourself Publishing
- Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., and Goldberg, L. R. (2007). The power of personality: the comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspect. Psychol. Sci.* 2, 313–345. doi: 10.1111/j.1745-6916.2007.00047.x
- Robson, C., and McCartan, K. (2016). *Real world research*. Hoboken, NJ: John Wiley & Sons Ltd.
- Ryan, T., and Grant, S. (2009). The origin and evolution of synapses. *Nat. Rev. Neurosci.* 10, 701–712. doi: 10.1038/nrn2717
- Sapolsky, R. M. (2005). *Biology and human behavior: the neurological origins of individuality (Second edition)*. Chantilly, VA: The Teaching Company
- Schaffner, J., Bao, S. D., Tobler, P. N., Hare, T. A., and Polania, R. (2023). Sensory perception relies on fitness-maximizing codes. *Nat. Hum. Behav.* 7, 1135–1151. doi: 10.1038/s41562-023-01584-y
- Seth, A. K., and Bayne, T. (2022). Theories of consciousness. *Nat. Rev. Neurosci.* 23, 439–452. doi: 10.1038/s41583-022-00587-4
- Shcherbakov, V. P. (2005). Evolution as resistance to entropy: mechanisms of species homeostasis. *Zhurnal Obshchei Biologii* 66, 195–211.
- Sinclair, D. A., and LaPlante, M. D. (2019). *Lifespan: why we age—and why we don't have to*. New York City: Atria Books
- Sokolov, E. N. (2001). "Orienting response" in *International encyclopedia of the social & behavioral sciences*. eds. N. J. Smelser and P. B. Baltes (Oxford: Pergamon), 10978–10981.
- Stanford Encyclopedia of Philosophy (2022a). Simplicity. Available at: <https://plato.stanford.edu/entries/simplicity/>
- Stanford Encyclopedia of Philosophy (2022b). The problem of induction. Available at: <https://plato.stanford.edu/entries/induction-problem/#HumeProb>
- Swiecicki, J. M., Sliusarenko, O., and Weibel, D. B. (2013). From swimming to swarming: *Escherichia coli* cell motility in two-dimensions. *Integr. Biol.* 5, 1490–1494. doi: 10.1039/c3ib40130h
- Theodoridis, S. (2015). *Machine learning: a Bayesian and optimization perspective*. Academic Press, Cambridge, MA
- Tufts University (2023). Levin lab: publications. Available at: <https://as.tufts.edu/biology/levin-lab/publications>
- Wacongne, C., Labyt, E., Wassenhove, V. V., Bekinschteinand, T., Naccache, L., and Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *PNAS* 108, 20754–20759. doi: 10.1073/pnas.1117807108



OPEN ACCESS

EDITED BY

Xerxes D. Arsiwalla,
Wolfram Research, Inc., United States

REVIEWED BY

Michele Farisco,
Uppsala University, Sweden
Robert Prentner,
ShanghaiTech University, China

*CORRESPONDENCE

Ken Mogi
✉ kenmogi@qualia-manifesto.com

RECEIVED 02 January 2024

ACCEPTED 26 April 2024

PUBLISHED 13 May 2024

CITATION

Mogi K (2024) Artificial intelligence, human cognition, and conscious supremacy.
Front. Psychol. 15:1364714.
doi: 10.3389/fpsyg.2024.1364714

COPYRIGHT

© 2024 Mogi. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Artificial intelligence, human cognition, and conscious supremacy

Ken Mogi^{1,2*}

¹Sony Computer Science Laboratories, Shinagawa, Japan, ²Collective Intelligence Research Laboratory, The University of Tokyo, Meguro, Japan

The computational significance of consciousness is an important and potentially more tractable research theme than the hard problem of consciousness, as one could look at the correlation of consciousness and computational capacities through, e.g., algorithmic or complexity analyses. In the literature, consciousness is defined as what it is like to be an agent (i.e., a human or a bat), with phenomenal properties, such as qualia, intentionality, and self-awareness. The absence of these properties would be termed “unconscious.” The recent success of large language models (LLMs), such as ChatGPT, has raised new questions about the computational significance of human conscious processing. Although instances from biological systems would typically suggest a robust correlation between intelligence and consciousness, certain states of consciousness seem to exist without manifest existence of intelligence. On the other hand, AI systems seem to exhibit intelligence without consciousness. These instances seem to suggest possible dissociations between consciousness and intelligence in natural and artificial systems. Here, I review some salient ideas about the computational significance of human conscious processes and identify several cognitive domains potentially unique to consciousness, such as flexible attention modulation, robust handling of new contexts, choice and decision making, cognition reflecting a wide spectrum of sensory information in an integrated manner, and finally embodied cognition, which might involve unconscious processes as well. Compared to such cognitive tasks, characterized by flexible and *ad hoc* judgments and choices, adequately acquired knowledge and skills are typically processed unconsciously in humans, consistent with the view that computation exhibited by LLMs, which are pretrained on a large dataset, could in principle be processed without consciousness, although conversations in humans are typically done consciously, with awareness of auditory qualia as well as the semantics of what are being said. I discuss the theoretically and practically important issue of separating computations, which need to be conducted consciously from those which could be done unconsciously, in areas, such as perception, language, and driving. I propose conscious supremacy as a concept analogous to quantum supremacy, which would help identify computations possibly unique to consciousness in biologically practical time and resource limits. I explore possible mechanisms supporting the hypothetical conscious supremacy. Finally, I discuss the relevance of issues covered here for AI alignment, where computations of AI and humans need to be aligned.

KEYWORDS

conscious supremacy, artificial intelligence, consciousness, large language model, computation

1 Introduction

Recently, large language models (LLMs) have made rapid progress based on the transformer (Vaswani et al., 2017) architecture, exhibiting many skills emulating but perhaps not matching human cognition, which were nonetheless once considered to be beyond the reach of machine intelligence, such as appropriate text generation based on a context, summarizing, searching under instructions, and optimization. With the advent of advanced AI systems such as ChatGPT (Sanderson, 2023), questions are arising regarding the computational significance, if any, of consciousness. Despite some claims that LLMs are either already or soon becoming conscious (Long, 2023), many regard these generative AI systems as doing computation unconsciously, thus forgoing possible ethical issues involved in AI abuse (Blauth et al., 2022). Generic models of consciousness would also suggest the LLMs to be unconscious as a default hypothesis, unless otherwise demonstrated, e.g., by convincing behavior suggesting the presence of consciousness to an external observer or a theoretical reasoning supported by an academic consensus. If LLMs can or come close to pass human-level cognition tests such as the false belief task in the theory of mind (Charman and Baron-Cohen, 1992; Baron-Cohen, 2000), the Turing test (Turing, 1950), and Winograd schema challenge (Sakaguchi et al., 2021) with their unconscious processing, what, if any, is the computational significance of consciousness?

Here, these abilities would not be necessary conditions for consciousness, as newborns are conscious without manifesting these abilities. The existence of these abilities would certainly be regarded as sufficient conditions for consciousness, in the generally accepted view of the human mind.

The theory of mind is related to the function of consciousness in the reportability and social context. The Turing test is tightly coupled with language, semantics in particular, and therefore closely related to consciousness. The Winograd schema challenge is crucial in understanding natural language, which is concerned with the nature of language here and now, locally, independent of the statistical properties dealt with in LLMs. The relation between functions exhibited by LLMs and consciousness is an interesting and timely question, especially when considering that natural language is typically processed when a human subject is conscious, except in the anecdotal and infrequent case of conversation in unconscious states, such as somnoliquy (Reimão and Lefèvre, 1980), hypnosis (Sarbin, 1997), and in a dream (Kilroe, 2016), which is a state distinctive from typical conscious or unconscious states. In an apparent contradiction to the conventional assumption about the necessity of consciousness in typical natural language exchanges, computations demonstrated by LLMs are considered to be done unconsciously. If conversations involving texts partially or totally generated by LLMs virtually pass the Turing test, without computations involving consciousness, what, if any, does consciousness do computationally?

Velmans (1991) analyzed the function of consciousness in cortical information processing, taking into account the role of focus of attention, concluding that it was not clear if consciousness was necessary for cognitive processes, such as perception, learning, and creativity. Velmans elaborated on the complexity of speech production, where the tongue may make as many as 12 adjustments of shape per second, so that “within 1 min of discourse as many as 10–15 thousand neuromuscular events occur” (Lenneberg, 1967). Based on these observations, Velmans suggested that speech production does not

necessarily require consciousness. Such observations would necessitate a more nuanced consideration of the role of conscious and unconscious processes in language.

Apart from the conscious/unconscious divide, language occupies a central position in our understanding of consciousness. Velmans (2012) streamlined the foundations of consciousness studies, pointing out that the default position would be to reduce subjective experiences to objectively observable phenomena, such as brain function. On a more fundamental level, Velmans argued that language is associated with the dual-aspect nature of the psychophysical element of human experience, where language models the physical world only in incomplete ways, limited by the capacities of our senses. The central role of language in our understanding of the world, including consciousness, should be kept in mind when discussing artificial reproductions of language, including, but not limited to, the LLMs.

Many regard the problem of consciousness as primarily in the phenomenological domain, concerned with what is experienced by a subject when he or she is conscious, e.g., properties such as qualia, intentionality, and self-awareness as opposed to physical or functional descriptions of the brain function. There are experimental and theoretical approaches tackling the cognitive implications of consciousness based on ideas, such as neural correlates of consciousness (NCC, Crick and Koch, 1998; Koch et al., 2016), global workspace theory (Baars, 1997, 2005), integrated information theory (Tononi et al., 2016), and free-energy principle (Friston, 2010).

Wiese and Friston (2021) discussed the relevance of the free-energy principle as a constraint for the computational correlates of consciousness (CCC), stressing the importance of neural dynamics, not states. In their framework, trajectories rather than states are mapped to conscious experiences. They propose CCC as a more general concept than the neural correlates of consciousness (NCC), discussing the nature of the correlates as necessary, sufficient, or both conditions for consciousness.

Some, somewhat controversially, consider quantum effects as essential in explaining the nature of consciousness (Hameroff, 1998; Woolf and Hameroff, 2001). Although there have been significant advances made, explaining the hard problem of consciousness (Chalmers, 1995) from such theoretical approaches remains hypothetical at best, even if not cognitively closed (McGinn, 1994), and a scientific consensus has not been reached yet. There are also arguments that hold that the hard problem is not necessarily essential for the study of consciousness. Seth (2021) argued that if we pursue the real problem of accounting for properties of consciousness in terms of biological mechanisms, the hard problem will turn out to be less important.

Given the difficulty in studying the phenomenological aspects of consciousness, with the advancement in artificial intelligence (AI), there is now a unique opportunity to study the nature of consciousness by approaching it from its computational significance. As artificial intelligence systems, such as LLMs, are reproducing and even surpassing human information processing capabilities, the identification of computational elements possibly unique to consciousness is coming under more focused analysis.

At present, it is difficult to give a precise definition of what computations unique to consciousness are. What follows are tentative descriptions adopted in this paper. From the objective point of view, neural computation correlating with consciousness would typically involve large areas of the brain processing information in coherent and

integrated parallel manners, while sensory qualia represent the result of complex processing in compressed forms, as in color constancy (Foster, 2011). Unconscious computation, on the other hand, does not meet these criteria. From the subjective point of view, conscious computation would be accompanied by such properties as qualia, intentionality, and self-consciousness. Unconscious computations do not cause these aspects of experience to emerge.

Artificial intelligence is an umbrella term, and its specific capabilities depend on parameters and configurations of system makeup and dynamics. For now, we would assume that AI systems referred to here are realized on classical computers. AI systems constructed on quantum computers might exhibit broader ranges of computational capabilities, possibly exhibiting quantum supremacy (Arute et al., 2019), which describes the abilities of quantum computers to solve problems any classical computer could not solve in any practical time. Quantum supremacy is not a claim that quantum computers would be able to execute computations beyond what universal Turing machines (Turing, 1936) are capable of. It is rather a claim that quantum computers can, under the circumstances, execute computations that could, in principle, be done by classical computers, but not within any practical period considering the physical time typically available to humans.

Similarly, conscious supremacy can be defined as domains of computation that can be conducted by conscious processes but cannot be executed by systems lacking consciousness in any practical time. Since the science of consciousness has not yet developed to reach the same level as quantum mechanics, it is difficult to give a precise definition of what conscious supremacy is at present. What follows is a tentative definition adopted in this article. Out of all the computations done in the neural networks in the brain, conscious supremacy refers to those areas of computation accompanied by consciousness, which are done in efficient and integrated ways compared to unconscious computation. Given the limits of resources available in the brain, computations executed in conscious supremacy would be, in a practical sense, impossible to execute by unconscious computation in any meaningful biological time. However, in principle, they could be done. Thus, there are no distinctions between computations belonging to conscious supremacy and other domains in terms of computability in principle. The practical impossibility of non-conscious systems to execute computations belonging to conscious supremacy would have been one of the adaptive values of consciousness in evolution.

The relationship between quantum supremacy and conscious supremacy will be discussed later.

As of now, quantum supremacy remains controversial (McCormick, 2022). The merit of introducing the perhaps equally debatable concept of conscious supremacy is that we can hope to streamline aspects of computation conducted by conscious and unconscious processes.

Abilities to play board games, such as chess, shogi, and go, are no longer considered to be unique to human cognition after AI systems, such as Deep Blue (Campbell et al., 2002) and AlphaZero (Schrittwieser et al., 2020), defeated human champions. After the success of LLMs in executing a large part of natural language tasks, cognitive abilities once considered unique to humans, e.g., the theory of mind, Turing test, and Winograd schema challenge, might not be considered to be verifications of the ability of artificial intelligence systems to perform cognitive tasks on par with humans. It should

be noted that the attribution of the theory of mind to LLMs remains controversial (Aru et al., 2023), and the exact nature of cognitive functions related to natural language, if any, in LLMs is an open question. However, it does seem legitimate to start considering the exclusion of certain computations from the set of those unique to consciousness based on computational evidence. While such exclusion might reflect cognitive biases on the part of humans to raise the bar unfavorably for AI systems, in an effort to solve cognitive dissonance (Aronson, 1969) about the relative superiorities of AI and humans, such considerations could serve as a filter to fine-tune domains of cognitive tasks uniquely executed by human cognition, conscious, and unconscious.

As artificial intelligence systems based on deep learning and other approaches advance in their abilities, tasks considered to be uniquely human would gradually diminish in the spectrum of functionalities. Specifically, the set X of computations considered unique to humans would be the complement of the union of the set of computations executed by artificial intelligence systems A_1, A_2, \dots, A_N under consideration. Namely, $X = A^c$, where $A = A_1 \cup A_2 \cup \dots \cup A_N$ (Figure 1), where the whole set represents the space of possible computations conducted by humans. As the number of artificial intelligence systems increases, the uniquely human domain of computation would ultimately become $X_\infty = A_\infty^c$, where $A_\infty = \lim_{N \rightarrow \infty} A_1 \cup A_2 \cup \dots \cup A_N$.

Needless to say, such an argument is conceptual in nature, as it is difficult to draw a clear line between what could and could not be done by artificial intelligence systems at present. Among computations unique to humans, some would be executed consciously, while some might be a combination of conscious and unconscious computation, involving processes which lie either inside or outside the neural correlates of consciousness (Crick and Koch, 1998; Koch et al., 2016). Theoretically, there could also be computations unique to humans executed unconsciously, although not of central interest in the context adopted here.

Penrose suggested that consciousness is correlated with the quantum mechanical effect, possibly involving quantum gravity (Penrose, 1996). Penrose went on to collaborate with Stuart Hameroff. Penrose and Hameroff together suggested, in a series of papers (Hameroff and Penrose, 1996; Hameroff and Penrose, 2014), that quantum mechanical processes in microtubules were involved in conscious processes, which went beyond the algorithmic capabilities of computability for the classical computer. Specifically, it was

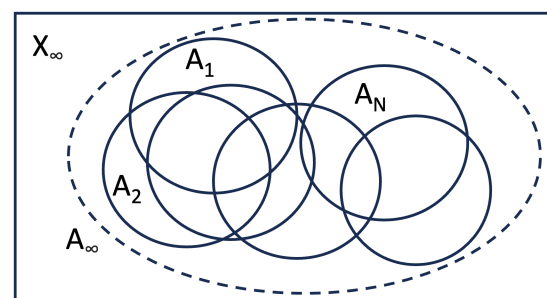


FIGURE 1

The analysis of AI capabilities would help focus the computational domain unique to consciousness (X), which can be defined in terms of instances of AI systems. As the number of AI systems increases, computations unique to consciousness will be more finely defined.

postulated that a process named “Orchestrated objective reduction” (Orch OR) was responsible for the generation of proto-consciousness in microtubules, a hypothesis independent from conventional arguments on quantum computing. One of the criticisms directed to such quantum models of consciousness was based on the fact that temperatures in biological systems are typically too high for quantum coherence or entanglement to be effective (Tegmark, 2000).

2 Possibilities and limits of artificial intelligence systems

Artificial General Intelligence (AGI; Goertzel, 2014) is purported to execute all tasks carried out by a typical human brain and beyond. Proposed tasks to be executed by AGI include the Turing test, coffee making or Wozniak test (Adams et al., 2012), college enrollment test (Goertzel, 2014), employment test (Scott et al., 2022), and the discovery of new scientific knowledge (Kitano, 2016).

In identifying possible areas for uniquely human cognition and potential candidates for conscious supremacy, it is useful to discuss systemic potentials and limits of artificial intelligence, which are currently apparent.

Some LLMs have started to show sparks of general intelligence (Bubeck et al., 2023) beyond abilities for linguistic processing. Such a potential might be explained by the inherent functions of language. The lexical hypothesis (Crowne, 2007) states that important concepts in fields, such as personality study and general philosophy, would be expressible by everyday language. The ability of natural language to represent and analyze a wide range of information in the environment is consistent with the perceived general ability of LLMs to represent various truths about this world, without necessarily being conscious, thus suggesting the central importance of representation in the analysis of intelligence.

What is meant by representation is a potentially controversial issue. In the conventional sense of psychology and philosophy of mind, a representation refers to the internal state that corresponds to an external reality (Marr, 1982). In the constructivist approach, representation would be an active construct of an agent’s knowledge, not necessarily requiring an external reality as a prior (Von Glasersfeld, 1987). Representations in artificial intelligence systems would be somewhere in between, taking inspiration from various lines of theoretical approaches.

One of the problems with LLMs, such as ChatGPT, is the occurrence of hallucination (Ji et al., 2023) and the tendency to produce sentences inconsistent with accepted facts, a term criticized by some researchers as an instance of anthropomorphism. Although humans also suffer from similar misconceptions, subjects typically are able to make confident judgments about their own statements (Yeung and Summerfield, 2012), while methods for establishing similar capabilities in artificial intelligence systems have not been established. Regarding consciousness, metacognitive processes associated with consciousness (Nelson, 1996) might help rectify potential errors in human cognition.

Behaviorist ways of thinking (Araiba, 2019) suggest that human thoughts are ultimately represented in terms of bodily movements. No matter how well developed an intelligent agent might be, manifestations of its functionality would ultimately be found in its objective courses of action in the physical space. From this perspective,

the intelligence of an agent would be judged in terms of its external behavior, an idea in AI research sometimes called instrumental convergence (Bostrom, 2012).

The possibilities and limits of artificial intelligence systems would be tangibly assessed through analysis of behavior. In voluntary movement, evidence suggests that consciousness is involved in vetoing a particular action (free won’t) when it is judged to be inappropriate within a particular context (Libet, 1999).

Thus, from robust handling of linguistic information to streamlining of external behavior, metacognitive monitoring and control would be central in identifying and rectifying limits of artificial intelligence systems, a view consistent with the idea that metacognition plays an essential role in consciousness (Nelson, 1996).

3 Computations possibly unique to conscious processing

As of now, the eventual range of computational capabilities of artificial intelligence is unclear. Employing cognitive arguments based on the observation of what subset of computation is typically done consciously, in addition to insights on the limits of artificial intelligence, would help narrow down possible consciousness-specific tasks. In that process, the division of labor between conscious and unconscious processes could be made, as we thus outline heterogeneous aspects of cognition.

Acquiring new skills or making decisions in novel contexts would typically require the involvement of conscious processing, while the execution of acquired skills would proceed largely unconsciously (Solomon, 1911; Lisman and Sternberg, 2013) in terms of the accompanying phenomenological properties, such as qualia, intentionality, and attention. Any cognitive task, when it needs to integrate information analyzed across many different regions in the brain, typically requires consciousness, reflecting the global nature of consciousness in terms of cortical regions involved (Baars, 2005). The autonomous execution of familiar tasks would involve a different set of neural networks compared to the minimum set of neural activities (neural correlates, Koch et al., 2016) required for the sustaining of consciousness.

It is interesting to note here that some self-learning unsupervised artificial intelligence systems seem to possess abilities to acquire new skills and make decisions in novel contexts (Silver et al., 2017; Schrittwieser et al., 2020). As the ability of artificial intelligence systems approaches the level purported for AGI (Goertzel, 2014), the possibility of the emergence of consciousness might have to be considered.

The global neural workspace (GNW) theory (Dehaene et al., 1998; Mashour et al., 2020) addresses how the neural networks in the brain support a dynamic network where relevant information can be assessed by local networks, eventually giving rise to consciousness. The multimodal nature of the GNW theory has inspired various theoretical works, including those related to deep learning networks (LeCun et al., 2015; Bengio, 2017).

In evolution, one of the advantages of information processing involving consciousness might have been decision-making reflecting a multitude of sensory inputs. Multimodal perception typically subserves such a decision-making process. Since the science of decision-making is an integral part of AI alignment (Yudkowsky, 2015), the difference between conscious and unconscious, as well as human and AI

decision-making processes, would shed much light on the parameters of systems supporting the nature of conscious computation.

Technological issues surrounding self-driving cars (Badue et al., 2021) have emerged as one of the most important research themes today, both from theoretical and practical standpoints. Driving cars involves a series of judgments, choices, and actions based on multimodal sensory information. Judgments on how to drive a vehicle often must be done within limited time windows in *ad hoc* situations, affected by the unpredictability of other human drivers, if any, and there are still challenges toward realizing fully self-driving vehicles (Kosuru and Venkitaraman, 2023). Moral dilemmas involved in driving judgments require sorting out situations concerned with conflicting choices for safety, known collectively as the trolley problem (Thomson, 1985), which is often intractable even when presented with clear alternative schemes (Awad et al., 2018). In real-life situations, there would be perceptual and cognitive ambiguities about, for example, whether you can really save five people by sacrificing one. In the face of such difficulties, fully self-driving cars without conscious human interventions might turn out to be impossible (Shladover, 2016).

The language is a series of micro-decisions, in that words must be selected, depending on the context, as follow-up sequences on what has been already expressed. The apparent success of LLMs in reproducing salient features of embedded knowledge in the language (Singhal et al., 2023) is impressive. However, it might still fall short of executing situated or embodied choice of words, as required, for example, in the college enrollment and employment tests. A linguistic generative AI might nominally pass the Turing test in artificial and limited situations. However, when an AI system implemented in a robot interacts with a human in real-life situations, there might be a perceived uncanny valley (Mori, 2012) linguistically, where negative emotions, such as uneasiness and repulsion, might be hypothetically induced in a human subject as the performance comes nearer to the human level.

4 Possible mechanisms for conscious supremacy

It is possible that there are computations uniquely executed by conscious processes, and there could be some similarities between conscious and quantum computations, independent of whether consciousness actually involves quantum processes in the brain. There could be similarities between postulated quantum supremacy and conscious supremacy, without underlying common mechanisms being necessarily implicated. It is worth noting here that just as it is in principle possible to simulate quantum computing on classical computers, it might be possible to simulate conscious computing, regardless of its nature, on classical computers, e.g., in terms of connectionist models representing neural networks in the brain.

There are several algorithms that demonstrate the superiority of quantum computing. For example, Shor's algorithm (Shor, 1994) can find prime factors of large numbers efficiently. Given a large number N , Shor's algorithm for finding prime factors can run in polynomial time in terms of N , compared to sub-exponential time on optimal algorithms for a classical computer.

In conscious visual perception, the binding problem (Feldman, 2012) questions how the brain integrates visual features, such as colors

and forms, into coherent conscious percepts. The challenge of combinatorial explosion (Treisman, 1999), in which all possible combinations of features, such as the yellow (color) Volkswagen Beetle car (form), must be dealt with, becomes essential there. Given the fact that forms (Logothetis et al., 1995) and colors (Zeki and Marini, 1998) are represented by distributed circuits in the brain, sorting through the possible combinations of forms and colors has similarities with the factoring problem addressed by Shor's algorithm (Figure 2).

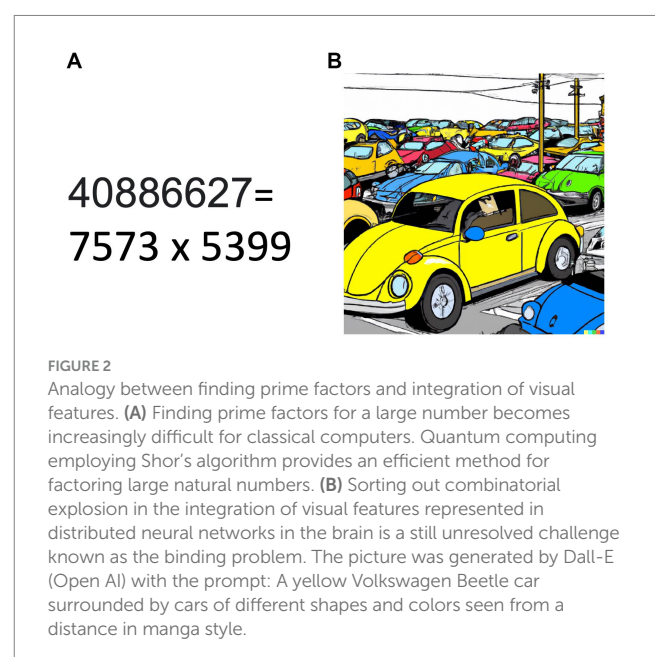
In quantum computing (Deutsch, 1985; Feynman, 1985), quantum superposition and entanglement are ingeniously employed to conduct algorithms effectively impossible for classical computers to execute in realistic time frames. In a quantum computing process, decoherence would introduce noise, and in order to execute on a large scale, a process called quantum error correction (QEC; Cai and Ma, 2021) is essential.

In conscious computing discussed here, similar mechanisms might be at play. For example, the contrast between the noisy neural firings and the apparently Platonic phenomenology of qualia suggests a process in which the variabilities due to noise in neural firings are rectified, named here conscious error correction (CEC). At present, the plausibility or the details of such an error-rectifying scheme is not clear. The possible relationships (if any) between QEC and CEC remain speculative at best at the moment. Despite these reservations, the involvement of error-correcting mechanisms in consciously conducted computation would be a line of thought worth investigating.

5 Implications for AI alignment

As artificial intelligence systems make progress, it is becoming important to align them with humans, an area called AI alignment (Russell and Norvig, 2021).

The elucidation of computations uniquely executed by consciousness and the possible existence of conscious supremacy, i.e., computations specifically and uniquely executed by neural processes correlating with consciousness, would put a constraint on AI alignment schemes.



Specifically, it would be an efficient alignment strategy to develop AI systems with capabilities other than uniquely conscious computations, while leaving computation involving conscious supremacy to humans.

It is interesting to consider the implications of such divisions of labor between AIs and humans for AI safety (Zhang et al., 2021). It would be impractical to require AI systems to carry out tasks better left to humans. Expecting AIs to execute tasks belonging to conscious supremacy would significantly disrupt AI safety.

Eliezer Yudkowsky's conceptualization of Friendly AI (Yudkowsky, 2008) is based on the importance of updating the system in accordance with humans (Russell and Norvig, 2021). Reinforcement learning from human feedback (RLHF; Stiennon et al., 2020), a technique often used in the development of artificial intelligence systems, can be considered to be an instance of developing Friendly AI and an attempt at the division of labor between conscious (human) and unconscious (AI) computations.

Alignment of AIs with humans, in the context of AI safety in particular, would depend on an effective division of labor between cognition unique to humans centered on conscious supremacy and computation conducted by computers, in a way similar to the interaction between conscious and unconscious processes in the human brain. In this context, artificial intelligence systems can be regarded as extensions of unconscious processes in the brain. Insights on cortical plasticities from tool use (Iriki et al., 1996) could provide relevant frameworks for discussion. It is important to note that limiting the functions of artificial intelligence systems to non-conscious operations does not necessarily guarantee robust alignment. Alignment would also depend on parameters that are dependent on the developers and stakeholders in the ecosystem of artificial intelligence. It would be important to discuss various aspects concerning alignment, including those put forward here.

Finally, the development of artificial consciousness (Chrisley, 2008), whether theoretically or practically feasible or not, might not be an effective strategy for AI alignment. From the point of view of the division of labor, computational domains belonging to conscious supremacy would be better left to humans. Artificial intelligence systems would do a better job of alignment by trying to augment computations unique to consciousness, which are to be reasonably executed by humans, rather than by replacing them from scratch.

6 Discussion

I have addressed here the possibility of characterizing conscious processes from a computational point of view. The development of artificial intelligence systems provides unique opportunities to explore and focus more deeply on computational processes unique to consciousness.

At present, it is not clear whether consciousness would eventually emerge from present lines of research and development in artificial intelligence. It would be useful to start from the null hypothesis of the non-existence of consciousness in artificial intelligence systems. We would then be able to narrow down what consciousness uniquely computes.

References

- Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., et al. (2012). Mapping the landscape of human-level artificial general intelligence. *AI Mag.* 33, 25–41. doi: 10.1609/aimag.v33i1.2322
- Araiba, S. (2019). Current diversification of behaviorism. *Perspect. Behav. Sci.* 43, 157–175. doi: 10.1007/s40614-019-00207-0

I have proposed the concept of conscious supremacy. Although this is speculative at present, it would be useful to think in terms of computational contexts apart from the hard problem of the phenomenology of consciousness. The presence of conscious supremacy would be connected to the advantages the emergence of consciousness has provided in the history of evolution. Elucidating the nature of conscious supremacy would help decipher elements involved in consciousness, whether it is ultimately coupled with quantum processes or not.

The value of arguments presented in this paper is limited, as it has not yet specifically identified computations unique to consciousness. The efforts to characterize computations unique to consciousness in terms of conscious supremacy presented here would hopefully help streamline discussions on this issue, although, needless to say, much work remains to be done.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

KM: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author declares that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

Author KM was employed by Sony Computer Science Laboratories.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Aronson, E. (1969). "The theory of cognitive dissonance: a current perspective" in *Advances in experimental social psychology*, vol. 4 (Academic Press), 1–34.

- Aru, J., Labash, A., Corcoll, O., and Vicente, R. (2023). Mind the gap: challenges of deep learning approaches to theory of mind. *Artif. Intell. Rev.* 56, 9141–9156. doi: 10.1007/s10462-023-10401-x

- Arute, F., Arya, K., Babbush, R., Bacon, D., Bardin, J. C., Barends, R., et al. (2019). Quantum supremacy using a programmable superconducting processor. *Nature* 574, 505–510. doi: 10.1038/s41586-019-1666-5
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., et al. (2018). The moral machine experiment. *Nature* 563, 59–64. doi: 10.1038/s41586-018-0637-6
- Baars, B. J. (1997). In the theatre of consciousness. Global workspace theory, a rigorous scientific theory of consciousness. *J. Conscious. Stud.* 4, 292–309.
- Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Prog. Brain Res.* 150, 45–53. doi: 10.1016/S0079-6123(05)50004-9
- Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. B., Forechi, A., et al. (2021). Self-driving cars: a survey. *Expert Syst. Appl.* 165:113816. doi: 10.1016/j.eswa.2020.113816
- Baron-Cohen, S. (2000). Theory of mind and autism: a review. *Int. Rev. Res. Mental Retardat.* 23, 169–184. doi: 10.1016/S0074-7750(00)80010-5
- Beckman, D., Chari, A. N., Devabhaktuni, S., and Preskill, J. (1996). "Efficient networks for quantum factoring" (PDF). *Phys. Rev. A* 54, 1034–1063. doi: 10.1103/PhysRevA.54.1034
- Bengio, Y. (2017). The consciousness prior. *arXiv:1709.08568*. doi: 10.48550/arXiv.1709.08568
- Benioff, P. (1980). The computer as a physical system: a microscopic quantum mechanical Hamiltonian model of computers as represented by Turing machines. *J. Stat. Phys.* 22, 563–591. doi: 10.1007/BF01011339
- Blauth, T. F., Gstrein, O. J., and Zwitter, A. (2022). Artificial intelligence crime: an overview of malicious use and abuse of AI. *IEEE Access* 10, 77110–77122. doi: 10.1109/ACCESS.2022.3191790
- Bostrom, N. (2012). The superintelligent will: motivation and instrumental rationality in advanced artificial agents. *Mind. Mach.* 22, 71–85. doi: 10.1007/s11023-012-9281-3
- Bray, D. (1995). Protein molecules as computational elements in living cells. *Nature* 376, 307–312. doi: 10.1038/376307a0
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). Sparks of artificial general intelligence: early experiments with gpt-4. *arXiv preprint 2303.12712*. doi: 10.48550/arXiv.2303.12712
- Cai, W., and Ma, Y. (2021). Bosonic quantum error correction codes in superconducting quantum circuits. *Fundamental Res.* 1, 50–67. doi: 10.1016/j.fmr.2020.12.006
- Campbell, M., Hoane, A. J. Jr., and Hsu, F. H. (2002). Deep Blue. *Artif. Intell.* 134, 57–83. doi: 10.1016/S0004-3702(01)00129-1
- Chalmers, D. (1995). Facing up to the problem of consciousness. *J. Conscious.* 2, 200–219.
- Charman, T., and Baron-Cohen, S. (1992). Understanding drawings and beliefs: a further test of the metarepresentation theory of autism: a research note. *J. Child Psychol. Psychiatry* 33, 1105–1112. doi: 10.1111/j.1469-7610.1992.tb00929.x
- Chrisley, R. (2008). Philosophical foundations of artificial consciousness. *Artif. Intell. Med.* 44, 119–137. doi: 10.1016/j.artmed.2008.07.011
- Crick, F., and Koch, C. (1998). Consciousness and neuroscience. *Cereb. Cortex* 8, 97–107. doi: 10.1093/cercor/8.2.97
- Crowne, D. P. (2007). *Personality theory*. Oxford: Oxford University Press.
- Dehaene, S., Kerszberg, M., and Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl. Acad. Sci.* 95, 14529–14534. doi: 10.1073/pnas.95.24.14529
- Deutsch, D. (1985). Quantum theory, the church–Turing principle and the universal quantum computer. Proceedings of the Royal Society of London. *A. Math. Phys. Sci.* 400, 97–117.
- Feldman, J. (2012). The neural binding problem. *Cogn. Neurodyn.* 7, 1–11. doi: 10.1007/s11571-012-9219-8
- Feynman, R. P. (1985). Quantum mechanical computers. *Optics News* 11, 11–20. doi: 10.1364/ON.11.2.000011
- Foster, D. H. (2011). Color constancy. *Vis. Res.* 51, 674–700.
- Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787
- Goertzel, B. (2014). Artificial general intelligence: concept, state of the art, and future prospects. *J. Artif. Gen. Intell.* 5, 1–48. doi: 10.2478/jagi-2014-0001
- Hameroff, S. (1998). Quantum computation in brain microtubules? The Penrose–Hameroff 'Orch OR' model of consciousness. *Philos. Trans. R. Soc. London, Ser. A* 356, 1869–1896.
- Hameroff, S. R., and Penrose, R. (1996). Conscious events as orchestrated space-time selections. *J. Conscious. Stud.* 3, 36–53.
- Hameroff, S., and Penrose, R. (2014). Consciousness in the universe: a review of the 'Orch OR' theory. *Phys Life Rev* 11, 39–78. doi: 10.1016/j.plrev.2013.08.002
- Iriki, A., Tanaka, M., and Iwamura, Y. (1996). Coding of modified body schema during tool use by macaque postcentral neurones. *Neuroreport* 7, 2325–2330. doi: 10.1097/00001756-199610020-00010
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.* 55, 1–38. doi: 10.1145/3571730
- Kilroe, P. A. (2016). Reflections on the study of dream speech. *Dreaming* 26, 142–157. doi: 10.1037/drm0000016
- Kitano, H. (2016). Artificial intelligence to win the nobel prize and beyond: creating the engine for scientific discovery. *AI Mag.* 37, 39–49. doi: 10.1609/aimag.v37i1.2642
- Koch, C., Massimini, M., Boly, M., and Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nat. Rev. Neurosci.* 17, 307–321. doi: 10.1038/nrn.2016.22
- Kosuru, V. S. R., and Venkitaraman, A. K. (2023). Advancements and challenges in achieving fully autonomous self-driving vehicles. *World J. Adv. Res. Rev.* 18, 161–167. doi: 10.30574/wjarr.2023.18.1.0568
- Lau, H., and Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends Cogn. Sci.* 15, 365–373. doi: 10.1016/j.tics.2011.05.009
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539
- Lenneberg, E. H. (1967). *Biological foundations of language*, vol. 2. New York: Wiley, 59–67.
- Libet, B. (1999). Do we have free will? *J. Conscious. Stud.* 6, 47–57.
- Lisman, J., and Sternberg, E. J. (2013). Habit and nonhabit systems for unconscious and conscious behavior: implications for multitasking. *J. Cogn. Neurosci.* 25, 273–283. doi: 10.1162/jocn_a_00319
- Logothetis, N. K., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* 5, 552–563. doi: 10.1016/S0960-9822(95)00108-4
- Long, R. (2023). Introspective capabilities in large language models. *J. Conscious. Stud.* 30, 143–153. doi: 10.53765/20512201.30.9.143
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: W. H. Freeman and Company.
- Mashour, G. A., Roelfsema, P., Changeux, J. P., and Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105, 776–798. doi: 10.1016/j.neuron.2020.01.026
- McCormick, K. (2022). Race not over between classical and quantum computers. *Physics* 15:19. doi: 10.1103/Physics.15.19
- McGinn, C. (1994). The problem of philosophy. *Philos. Stud.* 76, 133–156. doi: 10.1007/BF00989821
- Mori, M. (2012). The uncanny valley. *IEEE Robot. Automat.* 19, 98–100. doi: 10.1109/MRA.2012.2192811
- Nelson, T. O. (1996). Consciousness and metacognition. *Am. Psychol.* 51, 102–116. doi: 10.1037/0003-066X.51.2.102
- Penrose, R. (1996). On gravity's role in quantum state reduction. *Gen. Relativ. Gravit.* 28, 581–600. doi: 10.1007/BF02105068
- Reimão, R. N., and Lefèvre, A. B. (1980). Prevalence of sleep-talking in childhood. *Brain Dev.* 2, 353–357. doi: 10.1016/S0387-7604(80)80047-7
- Russell, S. J., and Norvig, P. (2021). *Artificial intelligence: A modern approach*. 4th Edn. London: Pearson.
- Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. (2021). Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM* 64, 99–106. doi: 10.1145/3474381
- Sanderson, K. (2023). GPT-4 is here: what scientists think. *Nature* 615:773. doi: 10.1038/d41586-023-00816-5
- Sarbin, T. R. (1997). Hypnosis as a conversation: 'believed-in imaginings' revisited. *Contemp. Hypn.* 14, 203–215. doi: 10.1002/ch.105
- Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., et al. (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature* 588, 604–609. doi: 10.1038/s41586-020-03051-4
- Scott, A. C., Solórzano, J. R., Moyer, J. D., and Hughes, B. B. (2022). The future of artificial intelligence. *Int. J. Artif. Intell. Mach. Learn.* 2, 1–37. doi: 10.51483/IJAIML.2.1.2022.1-37
- Seth, A. (2021). *Being you: A new science of consciousness*. New York: Penguin.
- Shladover, S. E. (2016). The truth about "self-driving" cars. *Sci. Am.* 314, 52–57. doi: 10.1038/scientificamerican0616-52
- Shor, P. W. (1994). "Algorithms for quantum computation: discrete logarithms and factoring" in *Proceedings 35th annual symposium on foundations of computer science* (Washington, DC: IEEE Computer Society Press), 124–134.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017). Mastering the game of go without human knowledge. *Nature* 550, 354–359. doi: 10.1038/nature24270
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., et al. (2023). Large language models encode clinical knowledge. *Nature* 620, 172–180. doi: 10.1038/s41586-023-06291-2
- Solomon, J. (1911). The philosophy of Bergson. *Mind* XX, 15–40. doi: 10.1093/mind/XX.77.15

- Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., et al. (2020). Learning to summarize with human feedback. *Adv. Neural Inf. Proces. Syst.* 33, 3008–3021. doi: 10.48550/arXiv.2009.01325
- Tegmark, M. (2000). Importance of quantum decoherence in brain processes. *Phys. Rev. E* 61, 4194–4206. doi: 10.1103/PhysRevE.61.4194
- Thomson, J. J. (1985). The trolley problem. *Yale Law J.* 94, 1395–1415. doi: 10.2307/796133
- Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nat. Rev. Neurosci.* 17, 450–461. doi: 10.1038/nrn.2016.44
- Treisman, A. (1999). Solutions to the binding problem: progress through controversy and convergence. *Neuron* 24, 105–125. doi: 10.1016/S0896-6273(00)80826-0
- Turing, A. (1936). On computable numbers, with an application to the Entscheidungsproblem. *J. Math* 58, 345–363.
- Turing, A. (1950). Computing machinery and intelligence, mind. *LIX LIX*, 433–460. doi: 10.1093/mind/LIX.236.433
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Proces. Syst.* 30, 6000–6010. doi: 10.48550/arXiv.1706.03762
- Velmans, M. (1991). Is human information processing conscious? *Behav. Brain Sci.* 14, 651–669. doi: 10.1017/S0140525X00071776
- Velmans, M. (2012). Reflexive monism psychophysical relations among mind, matter, and consciousness. *J. Conscious. Stud.* 19, 143–165.
- Von Glasersfeld, E. (1987). “Learning as a constructive activity” in *Problems of representation in the teaching and learning of mathematics* (Mahwah, NJ, USA: Lawrence Erlbaum Associates), 3–17.
- Wiese, W., and Friston, K. J. (2021). The neural correlates of consciousness under the free energy principle: from computational correlates to computational explanation. *Philos. Mind Sci.* 2:9. doi: 10.33735/phimisci.2021.81
- Woolf, N. J., and Hameroff, S. R. (2001). A quantum approach to visual consciousness. *Trends Cogn. Sci.* 5, 472–478. doi: 10.1016/S1364-6613(00)01774-5
- Yeung, N., and Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 1310–1321. doi: 10.1098/rstb.2011.0416
- Yudkowsky, E. (2008). “Artificial intelligence as a positive and negative factor in global risk” in *Global Catastrophic Risks*. eds. N. Bostrom and M. M. Cirkovic, 308–345.
- Yudkowsky, E. (2015). *Rationality-from AI to zombies*. Berkeley, CA, USA: Machine Intelligence Research Institute.
- Zeki, S., and Marini, L. (1998). Three cortical stages of colour processing in the human brain. *Brain J. Neurol.* 121, 1669–1685. doi: 10.1093/brain/121.9.1669
- Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M. C., and Dafoe, A. (2021). Ethics and governance of artificial intelligence: evidence from a survey of machine learning researchers. *J. Artif. Intell. Res.* 71, 591–666. doi: 10.1613/jair.1.12895



OPEN ACCESS

EDITED BY

Antonino Raffone,
Sapienza University of Rome, Italy

REVIEWED BY

Ken Mogi,
Sony Computer Science Laboratories, Japan
Roland Mayrhofer,
University of Regensburg, Germany

*CORRESPONDENCE

Marieta Pehlivanova
✉ mp8ce@uvahealth.org

RECEIVED 08 March 2024

ACCEPTED 24 May 2024

PUBLISHED 13 June 2024

CITATION

Pehlivanova M, Weiler M and Greyson B (2024)
Cognitive styles and psi: psi researchers are
more similar to skeptics than to lay believers.
Front. Psychol. 15:1398121.
doi: 10.3389/fpsyg.2024.1398121

COPYRIGHT

© 2024 Pehlivanova, Weiler and Greyson. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Cognitive styles and psi: psi researchers are more similar to skeptics than to lay believers

Marieta Pehlivanova*, Marina Weiler and Bruce Greyson

Division of Perceptual Studies, Department of Psychiatry and Neurobehavioral Sciences, University of Virginia School of Medicine, Charlottesville, VA, United States

Introduction: Belief in psi, which includes psychic phenomena such as extra-sensory perception and post-mortem survival, is widespread yet controversial. According to one of the leading and perhaps most tested hypotheses, high belief in psi can be explained by differences in various aspects of cognition, including cognitive styles. Most of this research has been conducted with lay individuals. Here, we tested the hypothesis that academic researchers who investigate psi may exhibit different cognitive styles than lay individuals interested in psi, and are more similar to skeptics.

Methods: We measured two cognitive styles—actively open-minded thinking (AOT) and the need for closure (NFC)—and assessed differences among four heterogeneous groups regarding belief in psi and involvement in related research. Specifically, our study included academic psi researchers ($N = 44$), lay individuals who believe in psi ($N = 32$), academics who are skeptics of psi ($N = 35$), and lay individuals who are skeptics ($N = 33$).

Results: We found group differences in AOT ($p = 0.003$) but not in NFC scores ($p = 0.67$). *Post hoc* tests showed no significant difference in AOT scores between academics who conduct psi research (4.5 ± 0.3) and academic skeptics (4.5 ± 0.3 ; $p = 0.91$) or lay skeptics (4.5 ± 0.4 ; $p = 0.80$). The lay psi group had significantly lower AOT scores (4.2 ± 0.4) than the other three groups ($ps: 0.005–0.04$), indicating a decreased willingness to consider a range of evidence when forming an opinion, including evidence that challenges their beliefs. AOT was negatively associated with psi belief in the two skeptic groups combined ($r = -0.29$, $p = 0.01$), but not in the psi groups ($r = -0.03$, $p = 0.78$).

Discussion: Our research shows that academics who work with psi differ from lay psi individuals, but not from skeptics, in actively open-minded thinking. In other words, despite their high belief in psi phenomena, psi researchers demonstrate a commitment to sound reasoning about evidence that is no different from that of skeptics.

KEYWORDS

paranormal belief, actively open-minded thinking, need for closure, scientific thinking, reasoning

Introduction

Psi phenomena, also known as psychic phenomena, have long captivated the interest and curiosity of humanity. Psi can be defined as experiences of “information or energy transfers” that are not currently explained by known cognitive, neural, or physiological processes (Bem, 2011). Examples of psi include extra-sensory perception (ESP—the ability to perceive information without using one’s physical senses) and psychokinesis (PK—the purported influence of mental processes on physical systems). Belief in psi remains relatively high among the general population, with 41% of Americans believing in ESP, 31% in telepathy, 26% in clairvoyance (the latter two both being types of ESP), 20% in

reincarnation, and 73% endorsing at least one of ten purported psi phenomena (Moore, 2005). According to a more recent YouGov poll, which was representative of the US population, 63% of respondents believed they have had at least one paranormal experience (Orth, 2022).

However, despite the substantial number of individuals who hold beliefs in psi, these beliefs are often met with skepticism and dismissed as irrational and unscientific, particularly among academics (Rouder and Morey, 2011; Wagenmakers et al., 2011). A mere 4% of National Academy of Sciences members expressed favorable beliefs in ESP and PK, as revealed by a poll conducted by McConnell and Clark (1991), with academics in the physical and chemical sciences as well as psychology endorsing the most skeptical views. Similarly, McClenon (1982) reported that only 4% of American Association for the Advancement of Science members considered ESP to be an “established fact,” with 25% viewing it as a “likely possibility.” Recent data from a convenience sample of scientists, engineers, and some academics from top universities appear more favorable toward psi (Wahbeh et al., 2023). Wahbeh et al. (2023) characterized 49% of respondents as “believers” in post-mortem survival, while only 19% were “non-believers” (with the rest “uncertain”). However, the survey’s framing around “consciousness surviving bodily death,” rather than as a survey of “elite scientists” (McClenon, 1982), may have skewed responses. Nevertheless, academic research on psi phenomena dates back to the nineteenth century and continues today, yielding some studies published in psychology and neuroscience journals (Bösch et al., 2006; Storm et al., 2010; Bem, 2011; Cardeña, 2018; Freedman et al., 2023). In stark contrast to academics more generally, most researchers in the field of psi appear to endorse the reality of psi, estimating its likelihood at an average of 79% (Irwin, 2014).

More recently, divergent attitudes toward psi among academics have been revealed in responses to the actual or attempted publication of psi research in scientific outlets beyond parapsychology journals. The peer-reviewed publication of a series of experiments, purportedly demonstrating modest but significant effects of the future on participants’ present responses, in a high-impact psychology journal and by a highly-cited social psychologist (Bem, 2011) elicited strong reactions from many academics (Wagenmakers et al., 2011; Cardeña, 2015). These findings and their publication were variously called a “faulty result,” “an assault on science and rationality,” a failure of the peer-review process (Helfand, 2011), “crazy” and a violation of “deep belief” in science (Hofstadter, 2011), as well as a search “for the impossible” (Reber and Alcock, 2020). Based on the prevailing physicalist view of modern science, psi phenomena are deemed implausible if not impossible. This is a common criticism levied against psi research and forms the basis of its rejection as a default position of many skeptical academics (McConnell and Clark, 1991; Alcock, 2010; Reber and Alcock, 2020). Accordingly, some of these commenters, along with others, have called for censorship of such research and findings. In turn, academics engaged in psi research have described instances of academic suppression (Cardeña, 2015; Weiler et al., 2022), while calling for the open and non-dogmatic study of psi phenomena (Cardeña, 2014). Despite prevailing physicalist views, a growing number of scholars are proposing alternative non-physicalist perspectives, which could

accommodate the possibility of psi phenomena (Kelly et al., 2007, 2015; Kelly and Marshall, 2021).

The controversy surrounding psi has spurred considerable research into the factors contributing to people’s belief in psi. Individual differences in psi belief are associated with different factors related to demographics, personality, cognition, and culture (French, 1992; Irwin, 1993; Kennedy, 2005; Gray and Gallo, 2016; Dean et al., 2022). Compared to the other categories of predictors of, or contributing factors to belief in psi, those related to cognition stand out as particularly important. Namely, cognitive factors probe specific reasons that people may choose to interpret certain experiences as paranormal, as well as their general ability and motivation to evaluate arguments for or against the reality of psi. In addition, some cognitive factors, such as critical thinking, are more malleable compared to personality or culture. Thus, they may be more amenable to training, which could, in turn, influence psi beliefs (Wilson, 2018). As Gray and Gallo (2016) also point out, cognitive influences on psi beliefs are salient because these beliefs feature a “metacognitive component” as they “require thinking about the cognitive abilities and limitations of the human mind” (p. 242).

According to one of the leading and perhaps most tested hypotheses in this domain, high belief in psi can be explained by deficits in various aspects of cognition, including critical and scientific thinking, reasoning, and overall cognitive ability (Alcock, 1981; Irwin, 1993). This hypothesis—historically referred to as the “cognitive deficits hypothesis” of psi belief—has received support, although findings have been mixed depending on the cognitive domain, methodology, and the exact population studied (Irwin, 1993; Gray and Gallo, 2016; Dean et al., 2022). A recent study using a large battery of cognitive tasks reported that strong skeptics outperformed strong believers on measures of analytical or logical thinking, but not on memory measures (Gray and Gallo, 2016). These authors also pointed out that individual differences related to psi beliefs may indeed be viewed as differences, rather than deficits, and need not be “good” or “bad,” nor would they necessarily imply differences in overall cognitive ability or potential for success” (Gray and Gallo, 2016). According to a recent systematic review of the decades-long literature on the association between belief in psi and cognitive functioning (Dean et al., 2022), high psi belief is most consistently associated with increased intuitive thinking (usually quick and emotion-based) and bias toward confirmatory evidence. Differences in self-reported cognitive styles—how people perceive and process information—have also been associated with different levels of psi belief (Gray and Gallo, 2016; Dean et al., 2022). In particular, greater belief in psi has been shown to correlate with lower “actively open-minded thinking”—a rational thinking disposition marked by an extensive exploration of alternatives and evidence to find the optimal answer, even if it contradicts one’s beliefs (Stanovich and West, 1997; Pennycook et al., 2020; Rizeq et al., 2021). Collectively, these findings suggest that individuals may endorse psi beliefs at least partially based on emotion and insufficient consideration of conventional explanations for seemingly anomalous occurrences.

One area of inquiry that has remained unexplored is whether the associations between cognitive styles and psi belief extend to researchers engaged in academic research on psi. Based on

a recent systematic review, over 60% of studies investigating the links between cognition and belief in psi have relied on undergraduate samples, and the remainder used predominantly general population samples or combined ones (Dean et al., 2022). Yet, many academic psi researchers are trained scientists and scholars (Cardeña, 2014). Even though they may exhibit a high level of endorsement of the reality of psi (Irwin, 2014), they likely differ in cognitive characteristics from the general population of lay believers. Importantly, within this group, high endorsement of psi phenomena, which would manifest as high scores on standardized measures of psi belief, may be strongly influenced by researchers' assessment of the published experimental evidence on psi (Irwin, 2014).

Cognitive styles related to evaluating evidence and reaching conclusions are of particular relevance to the controversial nature of psi, as they may contribute to how researchers (whether they are proponents or skeptics of psi) and lay individuals form beliefs about psi or engage with psi research. The literature on the “cognitive deficits hypothesis” of psi belief generally views deficient cognitive characteristics as responsible for (or at least associated with) strong psi beliefs. However, Cardeña (2011), among others, has argued that both staunch believers and skeptics who take an absolutist stance—fully endorsing or rejecting psi—have in common “intolerance for complexity and ambiguity” and unwillingness to consider other perspectives. In addition to actively open-minded thinking (AOT)—extensively investigated in relation to psi beliefs—another important albeit unexplored in this context cognitive style is the “need for cognitive closure,” often shortened as “need for closure” (NFC). NFC captures individual differences in the motivation to seek closure during information processing when faced with a decision or judgment (Webster and Kruglanski, 1994). Specifically, NFC measures the tendency to quickly settle on an answer, even if it is not correct or optimal, to end further information processing, indicating a preference for any answer, as compared with confusion and ambiguity (Webster and Kruglanski, 1994; Neuberg et al., 1997). Individuals who score high on measures of NFC tend to be more “closed-minded, resistant to information inconsistent with their firm opinions, and reluctant to have their knowledge challenged” (Roets et al., 2015).

In this study, we investigated differences in cognitive styles (AOT and NFC) among four heterogeneous groups regarding belief in psi and attitudes toward and involvement in related research: academic psi researchers, lay psi believers, academic skeptics, and lay skeptics. This research sought to shed light on two main questions: (1) Are psi researchers different from lay believers in how they approach knowledge, evidence, and ambiguity? (2) Are psi researchers—who engage in this research as a legitimate scientific pursuit which can yield observations incompatible with physicalist views—different than skeptics with similar academic and scholarly training in terms of considering inconsistent evidence and their motivation to search for the “correct” answer? We assessed AOT, NFC, as well as psi beliefs and psi experiences via self-report questionnaires, and examined differences between groups. We hypothesized that psi researchers would demonstrate high psi belief akin to lay believers, yet cognitive styles more similar to those of academic skeptics than lay believers. This is because psi researchers (a) typically are academics trained in the principles of scientific

inquiry and rigor, including critical evaluation of hypotheses; and (b) they likely developed their views on psi through a different process—e.g., evaluating the outcomes of research, including their own—than lay believers.

Materials and methods

Participants and recruitment

The study included four participant groups: 44 individuals who have engaged in academic psi research (academic psi group); 32 individuals who identify as psi believers or enthusiasts but are not engaged in academic psi research (lay psi group); 35 individuals who are academic or professional skeptics of psi (academic skeptic group); and 33 individuals who are skeptics of psi but not academics (lay skeptic group).

The academic psi group was recruited from mailing lists dedicated to parapsychology (e.g., “Survival Net,” an invitation-only international electronic mailing list for discussion of survival of consciousness, non-local consciousness, and related topics) and institutions focusing on related research (e.g., the Institute of Noetic Sciences and the Windbridge Research Center). In addition, we emailed the study invitation to psi researchers who may not be members of these lists or organizations. Fifty-three individuals consented to and finished the questionnaire within the academic psi group. The final analysis sample consisted of 44 academic psi researchers, excluding 7 respondents who have not conducted psi research and two repeat responses. Among this group, 81.8% identified as Caucasian; 6.8% as Asian; 6.8% as Hispanic; and 11.5% as other (participants could make multiple selections). Additional demographic characteristics for this and other groups are provided in Table 1.

The lay psi group was recruited from large Facebook groups of interest in paranormal topics and through organizations with a focus on psi phenomena and/or psi research (e.g., the Monroe Institute). All 32 respondents in this group who consented to the study finished the questionnaire. Among lay believers, 90.6% identified as Caucasian; 6.2% as Asian; 3.1% as Hispanic; and 3.1% as Other.

Academic skeptics were recruited by personalized email invitation to Fellows of the Committee for Skeptical Inquiry (CSI), who are elected to this position by the organization's Executive Council “for their distinguished contributions to science and skepticism” (Skeptical Inquirer, 2021). Specifically, election requires “outstanding contributions” to (1) a scientific discipline (2) the “communication of science and/or critical thinking,” and (3) to the skeptical movement (Skeptical Inquirer, 2021). In addition, we invited by email some academics who were not part of this list but who have been active contributors against psi research. All 35 respondents in this group who consented to the study finished the questionnaire. Among academic skeptics, 94.3% identified as Caucasian; 2.9% as Hispanic; and 5.7% as other.

Participants in the lay skeptic group were recruited via email invitations to some individuals who have contributed to the Skeptical Inquirer blog (<https://skepticalinquirer.org/>)—a magazine published by the CSI—who also forwarded the study

TABLE 1 Demographic characteristics, psi beliefs and experiences, and cognitive styles by participant group.

| | | Academic psi N = 44 | Lay psi N = 32 | Academic skeptics N = 35 | Lay skeptics N = 33 | Test statistics | P-values |
|------------------|-----------------------------------|---------------------------|---------------------------|--------------------------------|---------------------------|------------------------|----------|
| | | Mean/Median ± SD, or % | Mean/Median ± SD, or % | Mean/Median ± SD, or % | Mean/Median ± SD, or % | | |
| Demographics | Sex | | | | | $\chi^2_{(3)} = 7.9$ | 0.048* |
| | Woman | 22.7% | 50% | 25.7% | 42.4% | | |
| | Man | 72.7% | 50% | 74.3% | 54.5% | | |
| | Decline to answer | 4.5% | 0% | 0% | 3.0% | | |
| | Age [^] | 60.0/59 ± 14.4 | 51.3/50 ± 12.7 | 65.5/66 ± 11.8 | 53.7/58 ± 13.1 | $F_{(3,138)} = 8.1$ | <0.0001 |
| | Education [#] | | | | | $\chi^2_{(9)} = 109.2$ | <0.0001 |
| | High school/ less than college | 0% | 43.8% | 2.9% | 24.2% | | |
| | College/some graduate studies | 4.6% | 28.1% | 8.6% | 51.5% | | |
| | Master's degree | 7.0% | 25.0% | 8.6% | 24.2% | | |
| | Doctoral degree | 88.4% | 3.1% | 80.0% | 0.0% | | |
| NEBS | Beliefs | 77.0/80 ± 18.7 | 89.1/91 ± 9.7 | 8.8/6.6 ± 6.9 | 9.6/7 ± 10.8 | $F_{(3,140)} = 387.4$ | <0.0001 |
| | Experiences | 47.1/45 ± 22.1 | 61.7/66 ± 28.5 | 7.4/6.2 ± 6.8 | 8.1/7 ± 9.8 | $F_{(3,140)} = 72.0$ | <0.0001 |
| Cognitive styles | AOT [^] | 4.5/4.5 ± 0.3 | 4.2/4.2 ± 0.4 | 4.5/4.5 ± 0.3 | 4.5/4.6 ± 0.4 | $F_{(3,138)} = 4.8$ | 0.003 |
| | NFC | 3.0/3.1 ± 0.8 | 3.2/3.4 ± 0.7 | 3.1/3.1 ± 0.7 | 3.2/3.1 ± 0.7 | $F_{(3,140)} = 0.5$ | 0.67 |

SD, Standard deviation; *Statistical test after excluding “decline to answer” observations; [^]N = 42 for the academic psi group; [#]N = 43 for the academic psi group; NEBS, Noetic Experiences and Beliefs Scale.

invite to fellow skeptics. In addition, we recruited participants through a Facebook group focused on skepticism.¹ In the middle of recruitment efforts, we also posted a call for participants on the Skeptical Inquirer blog with the assistance of the magazine. Interested individuals were able to sign up for the study, after endorsing inclusion criteria, and were informed that some will be selected at random to participate. Subsequently, we discovered that most of the randomly selected individuals who endorsed being skeptics provided answers about psi beliefs that resembled those of believers, possibly influenced by the promised gift card. The final sample consisted of 33 lay skeptics, after excluding one respondent who consented to but did not finish the questionnaire and 5 “fake” skeptics. Although we do not report these analyses here for brevity, treating these “fake” skeptics as lay believers did not substantially change the results reported in this article. Among lay skeptics, 87.9% identified as Caucasian; 12.1% as Hispanic; and 3.0% as Asian.

Inclusion criteria common to all groups included being at least 18 years of age and fluent in English. In both skeptic groups, participants were explicitly asked to endorse being “a skeptic of the paranormal and fringe science.” Participants in the lay psi group were asked to endorse being a “believer in the paranormal or psi

enthusiast.” Potential participants in the academic psi group were asked whether they are psi or parapsychology researchers who are producing or have produced empirical or theoretical psi research that would be publishable in an academic journal. Participants in all groups were convenience samples from the respective populations, with a desired minimum sample of 30 in each group. The sample size was chosen based on population limitations, particularly in the two academic groups, and to achieve sufficient numbers for the central limit theorem to relax distributional assumptions.

It is important to clarify conceptual distinctions in psi research engagement between academic psi researchers and academic skeptics. Academic psi researchers typically view psi research as a legitimate scientific pursuit, conducting research to document and understand purported psi phenomena. In contrast, academic skeptics who are fellows of the CSI promote scientific skepticism, which generally takes the position that psi phenomena do not exist and considers investigations into such phenomena to be “pseudoscience.” Aligned with scientific skepticism, one could engage in psi-related research to disprove or debunk psi phenomena or the merit of such scientific pursuits or specifically to investigate beliefs in psi as irrational and people who hold such beliefs as cognitively deficient.²

1 The academic skeptic sample was recruited almost exclusively through the pool of CSI fellows. The lay skeptic sample was recruited partially through CSI affiliates or connections. Participants in the LS group are individuals who have pursued their interest in skepticism through avenues such as the Skeptical Inquirer, connections with other skeptics, or Facebook skepticism groups, but have not contributed to scientific skepticism through “distinguished” academic and/or communication efforts.

2 Among the academic skeptic group, there were 7 respondents who endorsed some involvement in psi-related research. As an example, one respondent listed their long-term involvement with an international committee investigating paranormal claims as “pseudoscience.” Another had been involved in psi research for decades before openly stating skeptical views, questioning both the reality of the phenomena and the value and validity of the research itself.

Procedure

Each group of participants completed a single online questionnaire administered via Qualtrics (Provo, Utah, USA), a secure survey platform with a site license provided by the University of Virginia. The study protocol was approved by the University of Virginia's Institutional Review Board for Social and Behavioral Sciences (protocol #3926). Participants provided consent electronically at the beginning of the survey. Each participant was offered a 10 USD Amazon gift card for completing the survey and asked to provide an email if they were interested in receiving the compensation.

Online questionnaire

The online questionnaire consisted of 62 items, not including the consent question, control questions about eligibility, and the gift card question. Six of these questions were shown conditionally based on answers about education and involvement in academic research (psi or other). Forty-five of the questions pertained to the three self-report measures described in the next subsection. We inquired about participants' socio-demographic characteristics, including standard questions about age, gender, race, country of residence, education, employment status, and religious preference or affiliation. In addition, the questionnaire included items about participants' professional involvement in psi research, including length of involvement, affiliations, and number of published scholarly articles. Respondents were given the opportunity to provide open-ended comments at the end of the survey. The questionnaire allowed completion in multiple sittings and going back to previous items.

Measures

Noetic experiences and beliefs scale

The Noetic Experiences and Beliefs Scale (NEBS) is a novel 20-item self-report questionnaire assessing psi beliefs and psi experiences as separate constructs (Wahbeh et al., 2020). The questionnaire consists of questions about 10 anomalous or extraordinary domains, each evaluated for the respective degree of belief and experience of the participant. For each domain, questions are asked as follows: "I believe that my consciousness is not limited by my physical brain or body" (an example of a belief question) and "I have personally had this experience" (for experience). Responses are reported on a visual analog scale ranging from "disagree strongly" to "agree strongly" for beliefs, and "never" to "always" for experiences, with numerical equivalents between 0 and 100. In the original validation study, the NEBS demonstrated excellent internal consistency (Cronbach's α : 0.90 and 0.93 for belief and experience subscales, respectively), good test-retest reliability at 1 month ($r = 0.83$ and $r = 0.77$ for belief and experience subscales, respectively) and the latent two-factor structure of beliefs and experiences was supported via confirmatory factor analysis (Wahbeh et al., 2020). Cronbach's α in this sample was 0.98 for the belief and 0.96 for the experience subscales, which

may suggest possible redundancy of some of the survey items. When examined separately, the skeptic groups show lower α values (0.78 for beliefs and 0.76 for experiences) than the psi groups (0.92 for beliefs and 0.93 for experiences).

Actively open-minded thinking

To assess actively open-minded thinking as a dispositional cognitive trait, we used a 10-item self-report AOT scale, suggested as the most valid and reliable version by the Society for Judgment and Decision Making at the time of study design in November 2020 [<http://sjdm.org/>; Baron et al. (2015) used an 8-item version; Baron (2019) used an 11-item version]. A composite scale measuring AOT was originally developed by Stanovich and West (1997), based on a conceptualization of the trait by Baron (1985). In the following decades, the measurement of AOT has undergone significant changes, as outlined by Stanovich and Toplak (2023), including adding items tapping into additional facets of AOT, shortening the scale, and refining questions to minimize bias. The version of the scale used here includes items such as "Willingness to be convinced by opposing arguments is a sign of good character" and "Changing your mind is a sign of weakness" (reverse-scored), rated on a five-point scale ranging from "1 = completely disagree" to "5 = completely agree" and including a "3 = neutral" option. Higher scores on this scale indicate greater actively open-minded thinking. The AOT scale demonstrated adequate internal consistency in this sample with a Cronbach's α of 0.73.

Brief need for closure scale

To assess the need for closure as a dispositional trait, we used a brief 15-item Need for Closure Scale (Roets and Van Hiel, 2011). This self-report scale was developed and validated as an abridged version of a modified NFC scale (Roets and Van Hiel, 2007). Even though the revised scale incorporated all five original facets of Order, Predictability, Ambiguity, Closed-mindedness, and Decisiveness, it was validated via principal component analysis as a one-dimensional measure tapping a unitary construct (Roets and Van Hiel, 2007). The brief NFC scale showed good internal consistency (Cronbach's $\alpha = 0.87$), adequate test-retest reliability at 1 month ($r = 0.79$), and good convergent and divergent validity, showing psychometric properties that were similar to those of the revised full scale (Roets and Van Hiel, 2011). The brief NFC includes items such as "I dislike questions which could be answered in many different ways" and "I do not usually consult many different opinions before forming my own view," rated on a six-point scale ranging from "1 = strongly disagree" to "6 = strongly agree," without a neutral option. Higher scores on this scale indicate a greater need for closure. Internal consistency in this sample was also good, with a Cronbach's α of 0.83.

Statistical analysis

Descriptive statistics were presented as means, medians, and standard deviations for continuous variables and percentages within groups for categorical variables. ANOVA was used to test

for group differences in psi beliefs/experiences, cognitive styles, and participants' age, without adjusting for covariates. Analysis of covariance (ANCOVA) was used to test for these group differences while including covariates as additional independent variables in the models. Specifically, given previous findings of age and education associations with AOT and NFC (Kossowska et al., 2012; Chen, 2015; Edgcumbe, 2022), and differences in these demographic variables between groups in this study, we assessed group differences while adjusting for age and education as an ordinal variable.

Pairwise differences after a significant ANOVA/ANCOVA group effect were assessed via Tukey-adjusted *post hoc* tests. Effect sizes for the main effects of ANOVA/ANCOVA were presented as eta squared and partial eta squared. Pearson's chi-squared tests were used to test the association between group and categorical variables like sex and education. Pearson's correlation coefficients were used for all bivariate correlation analyses. All data management and statistical analyses were conducted using SAS 9.4 (SAS Institute, Cary, NC).

Power analysis

Given the sample size limitations in this study, we conducted a *post hoc* sensitivity power analysis using G*Power Version 3.1.9.7 (Faul et al., 2007). A one-way between-subjects analysis of variance with 144 participants and four groups would be sensitive to effects of $\eta^2 = 0.07$ or $f = 0.28$ (conventionally, a medium effect size), assuming 80% power and an alpha of 0.05. In other words, the study would not be able to reliably detect effects smaller than $\eta^2 = 0.07$. Note that G*Power outputs effect sizes in Cohen's f , which has been converted to η^2 according to Cohen (2009).

To our knowledge, there are currently no established benchmarks in the particular groups included in this study for effect sizes or expected mean levels for the cognitive styles under examination. However, some prior research may inform reasonable estimates of group differences in AOT that are associated with objective measures of argument evaluation. Stanovich and West (1997) administered an argument evaluation test and various cognitive style measures to a large group of participants. The authors developed an index of one's ability to evaluate the quality of an argument independently of one's prior beliefs about an issue. Classifying participants into groups based on their high or low reliance on argument quality when evaluating a proposition, Stanovich and West (1997) reported that the high reliance group showed significantly higher disposition toward AOT compared to the low reliance group. Using descriptive statistics from the article, we calculated that the effect size of this difference approximates a Cohen's d of 0.51 (equivalent to f of 0.25 or η^2 of 0.06). For the NFC scale, we could not identify studies directly addressing associations with relevant objective measures. However, associations between NFC and measures relevant to evidence processing, such as intolerance for ambiguity, need for cognition, and dogmatism, fall in the range of 0.58–0.60 in terms of Cohen's d (f : 0.29–0.30 or η^2 around 0.08) (Webster and Kruglanski, 1994). The magnitude of such AOT and NFC effects are in line with what our sample allows us to detect.

Results

Demographic characteristics

Due to the nature of the participant groups, differences in demographics were expected. In terms of education, the two groups consisting primarily of academics—the academic psi and academic skeptic groups—had achieved higher education, on average, than the lay believers and skeptics (Table 1). In addition, the academic groups differed from the lay groups in terms of sex and age. Notably, the sex ratio among the academic psi group exactly mirrors previously published estimates (Mayer et al., 2022). Academic skeptics were the oldest, on average, and differed significantly from both the lay psi group ($p = 0.0001$) and the lay skeptic group ($p = 0.002$), but not from the academic psi group ($p = 0.27$). Participants in the academic psi group were significantly older than those in the lay psi group ($p = 0.03$) but not the lay skeptic group ($p = 0.17$).

Group differences in psi beliefs and experiences

As anticipated, there were differences between the groups on both psi beliefs ($p < 0.0001$, $\eta^2 = 0.89$) and experiences ($p < 0.0001$, $\eta^2 = 0.61$; Table 1, Figure 1), as measured by the NEBS. *Post hoc* tests revealed that the academic psi and lay psi groups have significantly higher psi belief scores than both skeptic groups ($ps < 0.0001$ for all four comparisons). Psi belief scores did not differ significantly between the two skeptic groups ($p = 0.99$). While both psi groups showed high levels of belief, participants in the academic psi group had significantly lower belief scores, on average, than those in the lay psi group ($p = 0.0005$), and showed higher variability in their beliefs.

The pattern of group differences in experience scores was identical to that for belief scores. The academic psi and lay psi groups had significantly higher psi experience scores than both skeptic groups ($ps < 0.0001$), but differed from each other, with the academic group reporting lower levels of psi experiences than the lay group ($p = 0.007$). Psi experience scores did not differ significantly between the academic and the lay skeptic group ($p = 0.99$).

Scores on the psi beliefs and psi experiences subscales were significantly correlated with each other in all four groups, with the weakest correlation occurring in the academic psi group ($r = 0.48$, $p = 0.001$) and the strongest in the lay skeptic group ($r = 0.78$, $p < 0.0001$).

Group differences in cognitive styles

ANOVA revealed group differences in AOT scores ($p = 0.003$, $\eta^2 = 0.09$), but not in NFC ($p = 0.67$, $\eta^2 = 0.01$). *Post hoc* tests showed no significant difference in AOT between the academic psi and academic skeptic groups, which lines up with our original hypothesis ($p = 0.91$). The academic psi group was also not significantly different in AOT scores from lay skeptics ($p = 0.80$).

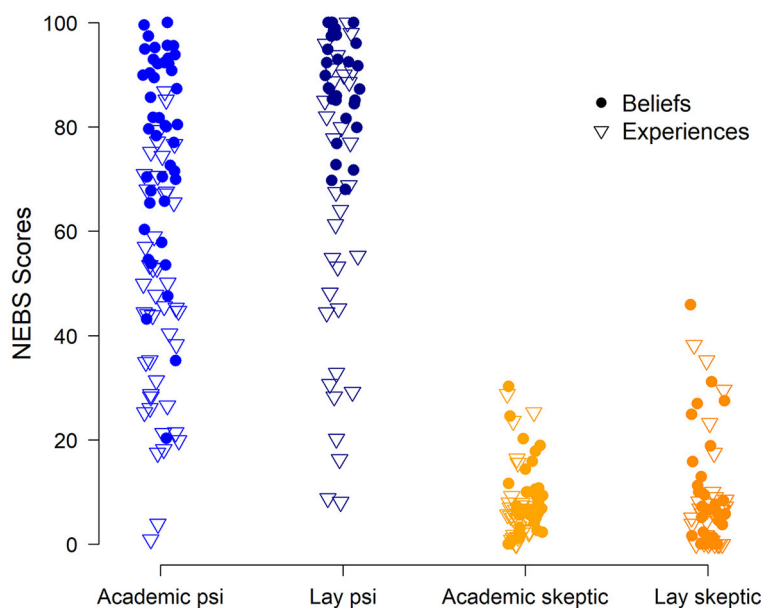


FIGURE 1

Dot plot showing Noetic Experiences and Beliefs Scale (NEBS) scores (Y-axis) by study group (X-axis). Academic psi ($N = 44$) and lay psi ($N = 32$) groups showed higher belief and experience scores than the academic skeptic ($N = 35$) and lay skeptic ($N = 33$) groups.

The lay psi group had significantly lower AOT scores than the academic psi ($p = 0.04$), academic skeptic ($p = 0.01$), and lay skeptic ($p = 0.005$) groups.

ANCOVA adjusting for age and education revealed group differences in AOT [$F_{(3,135)} = 3.68, p = 0.01, \eta^2 = 0.08$], but not in NFC [$F_{(3,136)} = 0.58, p = 0.63, \eta^2 = 0.01$]. Similarly to unadjusted analyses, *post hoc* tests showed no significant difference in AOT between the academic psi and the academic skeptic ($p = 0.98$) or lay skeptic ($p = 0.26$) groups. The lay psi group had significantly lower AOT scores than the lay skeptic group ($p = 0.009$) but was no longer significantly different from the academic psi ($p = 0.94$) and academic skeptic ($p = 0.82$) groups.

Correlations between psi beliefs and cognitive styles

Next, we examined the relationship between cognitive styles and belief in and experience with psi across the entire sample. Belief and experience scores were significantly negatively correlated with AOT ($r = -0.24, p = 0.004$; $r = -0.22, p = 0.01$, respectively; Figure 2 for belief scores), such that higher NEBS scores are associated with lower endorsement of AOT principles. However, belief and experience scores were not correlated with NFC ($r = -0.04, p = 0.62$; $r = -0.14, p = 0.10$). When examining the effect separately for psi vs. skeptic groups, it appears that the significant associations between AOT and psi belief scores are driven by the skeptics, at the lower range of belief and experience scores. Specifically, AOT and psi beliefs and experiences were significantly correlated in the two skeptic groups combined ($r = -0.29, p = 0.01$; $r = -0.27, p = 0.02$, respectively), but not in the two psi groups ($r = -0.03, p = 0.78$; $r = -0.04, p = 0.75$).

Narrative data

Although not necessarily representative, certain comments by participants help contextualize differences and similarities between the groups. Some researchers in the academic psi group commented on the appropriateness of asking about *belief* in psi presumably as the basis of one's interest in purported psi phenomena. For example, one PhD-level psychologist involved in the research for 5–10 years wrote: “For me, it is not about my ‘beliefs’ it is about the evidence.” Another respondent wrote: “This survey was oddly worded if the target audience was research scientists. I don’t ‘believe’ things. I take a flexible position that is constantly reevaluated based on the available data.”

Despite their differences in assessments of psi compared to psi researchers, some academic and lay skeptics stated an openness to the possibility of psi if the right evidence or explanation is presented. One neuroscientist wrote: “I am open to the idea that there are aspects of the physical world that we don’t understand [...], but once those were explicated they would then be understood, modeled, reproducible and would fall into the category of physical world. Thus my statement that ‘I have no belief in the non-physical’.” The idea of openness and the necessity of evidence of psi being reproducible was echoed by others, exemplified by this comment from a lay skeptic: “As a skeptic, I need to have an open mind to all possible answers. [...] I am open to new evidence but it needs to be valid and reproducible evidence.” Notably, several skeptics suggested that personal experience cannot be construed as evidence: “I would need some pretty indisputable evidence, even if I thought something may have happened to myself.”

Compared to the other three groups, participants in the lay psi group were most likely to mention specific psi experiences they may have had—sometimes detailing their different types and duration—

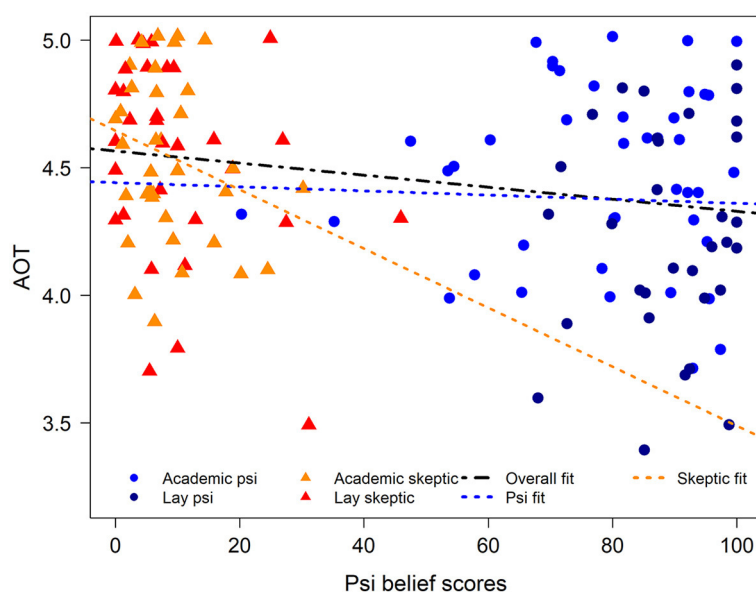


FIGURE 2

Scatterplot for the relationship between psi belief scores and actively open-minded thinking (AOT). Scores for each of the four groups are shown in different colors. Lines of “best fit” for the relationship are shown separately for the total sample (overall fit), the psi groups combined (psi fit), and the skeptic groups combined (skeptic fit). A small amount of jitter was added to values on both axes to facilitate visualization of overlapping points. AOT is negatively correlated with psi belief in the total sample and among the skeptic groups, but not the psi groups.

as well as how those experiences directly influenced their beliefs in psi. Some participants in this group specifically commented on the role of logic and evidence in their perceptions: “*Myself, I’m very logical and what I experience of the energetic and spiritual world to me does not defy the science or contradict logic. If I can’t understand the spiritual and energy logically, I wouldn’t be involved in it.*” Another wrote: “*Paranormal Investigation is all about making sure there’s concrete evidence.*” Additionally, several respondents in this group commented on how they and people in general can develop the ability to experience psi phenomena.

Discussion

In this manuscript, we aimed to test the hypothesis that academic psi researchers may exhibit different cognitive styles compared to lay individuals interested in psi, but similar to skeptics. We compared two cognitive styles relevant to evidence processing and judgments—actively open-minded thinking and the need for closure—between heterogeneous groups in terms of belief in psi and attitudes toward and involvement in psi research. Specifically, we included two groups of academics—psi researchers and skeptics—as well as two lay groups of participants who either believe in psi or are skeptics of it.

Comparing the academic psi and academic skeptic groups

A primary focus of this investigation was to compare academics and researchers who are engaged in studying psi and those who take a skeptical position toward this field and its underlying

phenomena. Not surprisingly given their different engagement with psi, researchers in the field reported significantly greater belief in and perceived experience with psi phenomena compared to academic skeptics, echoing prior findings (Blackmore, 1989; Irwin, 2014). However, as hypothesized, psi researchers and academic skeptics showed no difference in the cognitive styles of AOT and NFC. Together, these findings suggest that these two groups that are philosophically and empirically at odds with each other regarding evidence for psi phenomena nonetheless do not differ in their endorsement of the principles of “good” thinking about evidence (Baron et al., 2015). These encompass actively seeking out evidence that contradicts one’s beliefs, being willing to update one’s beliefs in light of new evidence, and being comfortable with ambiguity (Stanovich and Toplak, 2023). Additionally, the two groups did not differ in the extent to which they form opinions quickly to avoid ambiguity (Roets and Van Hiel, 2011).

Supporting the notion that these two groups are not entirely dissimilar, a previous survey comparing the views of psi researchers and skeptics revealed several areas of agreement (Blackmore, 1989). Among those were the acknowledgment of contributions of psi research to other fields (including psychology, statistics, and philosophy of science), potential concerns about lack of replicability in the field, and general “open-mindedness and doubt” when evaluating evidence (Blackmore, 1989). On the other hand, Blackmore (1989) highlighted an important difference between the two groups in their interpretation of research that aims to establish the reality of psi. Namely, skeptics indicated that they considered only laboratory experiments relevant as evidence of psi, and even then, they found them unconvincing. In contrast, psi researchers indicated that they found the totality of psi research—including experiments and spontaneous cases (e.g., near-death experiences)—to be relevant and convincing.

A more recent survey with members of the Parapsychological Association (PA) substantiated this, revealing that overall they deemed the cumulative experimental psi evidence most persuasive (79% combined for “strongly” or “extremely” persuasive) (Irwin, 2014). However, PA members also viewed spontaneous cases as well as personal experience as persuasive, though to a lesser extent (Irwin, 2014). This divergence was also reflected in our narrative data, where skeptics (both academic and lay) singled out the importance of reproducible, experimental evidence for psi, which they consider to be lacking, and discounted the relevance of personal experience.

Despite historic disagreement and even vitriol, members of the two groups have previously conducted successful and informative “skeptic-proponent collaborations” (Hyman and Honorton, 1986; Schlitz et al., 2006), highlighting areas of agreement including methodological improvements for future psi studies. These collaborative efforts have been acknowledged as valuable to the field by the psi researcher community (Roe, 2017; Parapsychological Association, 2023). Over time, such engagements have contributed to a shift in the nature of the disagreement, moving from disputes about “the existence of [anomalous] effects to their interpretation” (Hyman and Honorton, 1986; Honorton, 1993).

Ultimately, our goal here is not to debate the merits of psi research and evidence. Their interpretation and value have generated significant and long-lasting debates between psi researchers and skeptics (Krippner and Friedman, 2010). The data we present suggest that, despite these differences and the perception of psi researchers as “poor thinkers” and of skeptics as uninformed dogmatists (Roe, 2017), psi researchers and skeptics may not differ considerably in their thinking styles, as is commonly expected.

In the context of these potential similarities, it is interesting to consider what drives psi researchers to engage in this research, even though our study was not designed to directly answer this question. We observed that academic psi researchers endorsed significantly higher psi beliefs, as well as psi experiences, compared to academic skeptics. These experiences attributed to psi processes can serve as a possible motivator of research interests in psi, as 53% of members of the PA found personal experience “strongly” or “extremely” persuasive as a source of evidence for the reality of psi (Irwin, 2014). Indeed, scientists’ own extraordinary and spiritual experiences have in some cases prompted significant career changes, including shifting one’s work toward exploring the nature of consciousness (Woollacott and Shumway-Cook, 2023).

Speculatively, it is conceivable that factors beyond cognitive ones, such as personality, may influence researchers’ inclination to investigate psi phenomena. One possible contributing factor is openness to experience, which is positively correlated with both psi beliefs (Chauvin and Mullet, 2021) and psi experiences (Zingrone et al., 1998–1999). This dimension of personality can be accompanied by unconventional attitudes and interest in novel ideas (McCrae, 1993). Among scientists, openness to experience is specifically associated with conducting “boundary-spanning” and perhaps riskier research (Bateman and Hess, 2015), which undoubtedly applies to psi research.

Comparing the academic psi and lay psi groups

Another important and purposeful comparison in this study was to assess cognitive style differences between academic psi and non-academic lay psi individuals. Although most psi researchers would identify as “believers” (Blackmore, 1989; Irwin, 2014), these groups are fundamentally distinct in terms of their academic interest in purported psi phenomena, their familiarity and involvement with psi research, and their ability to engage with it. As anticipated, both groups showed high levels of belief in and experience with psi compared to skeptics, with the lay psi group nonetheless scoring significantly higher than the academic psi group. In contrast, the academic psi group showed greater levels of AOT compared to the lay group, indicating a greater willingness to consider a range of evidence when forming opinions, including evidence that contradicts their beliefs. This difference did not hold after accounting for educational and age differences between the groups. Nonetheless, we contend that these differences, especially in education, are defining features of the two groups. As such, they are relevant and should not be fully eliminated, for a fair comparison of differences between academic psi researchers and lay psi believers. Notably, after accounting for age and education, the lay psi group also did not differ from the academic skeptic group in terms of AOT.

Despite both groups exhibiting high belief in and perceived experience with psi, they may ultimately differ in how these beliefs originated or strengthened. Beliefs and experiences were positively correlated in both psi groups, but this correlation was stronger in the lay psi group. Notably, many lay individuals explicitly commented that their belief was influenced by their personal experiences, whereas the academic psi group did not highlight a connection. A previous survey with psi researchers did reveal that personal experience was seen as persuasive for establishing the reality of psi, but less so than the cumulative experimental psi evidence (Irwin, 2014).

Overall, these findings suggest a distinction between individuals actively engaged in academic psi research and those who are not but have a strong interest and belief in psi. Although this distinction is rarely or never made in research that focuses on believers’ cognition, including cognitive styles (Gray and Gallo, 2016), it is an important one for both the proponents of psi research and its skeptics. Psi researchers rightfully view their public image as one of the major hurdles facing their field (Irwin, 2014). Thus, any evidence challenging the “deficit hypothesis” as it relates to their own cognition about the legitimacy of psi phenomena should be highlighted. On the other hand, skeptics’ engagement with psi research, which is increasingly finding its way into psychology and related journals (Bösch et al., 2006; Bem, 2011; Cardena, 2018; Freedman et al., 2023), will benefit from viewing psi researchers as fellow academics who may disagree rather than individuals prioritizing belief over evidence (Reber and Alcock, 2020).

Association between belief in psi and actively open-minded thinking

Across our entire sample encompassing diverse groups in terms of belief in psi and involvement in related research, AOT showed small-to-medium inverse correlations with psi belief and experiences. The direction of this relationship suggests that people who endorse beliefs in psi are less likely to endorse the principles of good thinking about evidence, including willingness to seek out evidence that contradicts their beliefs, to update their beliefs with new evidence, and to be comfortable with ambiguity. This association has been demonstrated previously, using heterogeneous measures of AOT and psi belief, in both undergraduate and adult samples (Pennycook et al., 2020; Rizeq et al., 2021; Newton et al., 2023). Notably, our participant selection differed not only in terms of demographics but also with the purposeful sampling at the ends of the psi belief spectrum. Relatedly, in our data, this association appears to be driven by the skeptic groups and is even stronger among them, but is virtually null within the psi groups. This suggests that the inverse relationship between actively open-minded thinking and belief in psi may not be universal, particularly among individuals with strong psi beliefs, which may be influenced by other factors.

Limitations

Our study has limitations that are worth noting, including some pertaining to the selection of participants. The samples of academic psi and academic skeptic individuals are likely representative of their underlying populations. However, participants in the lay groups may be different from non-selected individuals from the general population who may hold belief or skepticism toward psi, as the former were recruited through venues where they actively pursued their interests and appear to be highly educated compared to the general population. Additionally, participants in the different groups were not matched on demographics, but we also presented group comparisons that took into account differences in demographics. Finally, we acknowledge limitations in the variability of some of our measures. In terms of AOT, on average, participants in all groups generally “agreed” with the principles of good thinking about evidence. In terms of psi beliefs, most scores clustered at the ends of the belief spectrum, reflecting the selection criteria for our study groups. Despite the limited ranges of these measures, we identified significant associations and differences.

The findings of this study should be interpreted in the context of the effect sizes we are able to detect with our sample. Our power analysis indicated that we can reliably detect effect sizes in the medium range or higher. When comparing the cognitive styles of psi researchers to those of skeptics, particularly academic ones, we did not find significant differences within that detectable range. It is possible that differences of a smaller magnitude exist between the groups. However, prior research has shown that differences in cognitive styles that are associated with more objective reasoning measures are typically of medium magnitude (Stanovich and West, 1997). Therefore, if small differences in

cognitive styles do exist between the groups, they are unlikely to be of practical significance.

Conclusion and future directions

Here we presented a unique comparison of cognitive styles among groups that differ in belief in psi and involvement in psi research. Our study contributes to a more nuanced understanding of the role that cognitive styles, particularly actively open-minded thinking and the need for closure, may play in the formation of psi beliefs. Additionally, it explores related differences and similarities between researchers and academics who are engaged in psi research and (1) lay believers or (2) skeptics. The cognitive styles explored here measure dispositions toward good thinking and they are markers, but not direct measures, of the ability to think critically. Future investigations could probe deeper into other aspects of cognition (including task-based) to fully examine the range of potential differences among groups, especially between academic psi researchers and academic skeptics. In addition to cognitive differences between them, other influences on psi beliefs should be explored further, as scientists, just like humans in general, have personal and sometimes strong beliefs that may impact their opinions, in addition to empirical and theoretical considerations (Coll and Taylor, 2004). Finally, given the null association found between actively open-minded thinking and psi belief at high levels of belief, future studies could investigate this relationship along the entire range of these variables. Additionally, exploring the reasons behind this differential finding could provide further insights into the development and maintenance of psi beliefs.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the University of Virginia's Institutional Review Board for Social and Behavioral Sciences (protocol #3926). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

MP: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. MW: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology. BG: Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was supported by a grant from the Bial Foundation/Fundação Bial (No. 212/2020).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Alcock, J. E. (1981). *Parapsychology, Science or Magic?: A Psychological Perspective*, 1st Edn. New York, NY: Pergamon Press.
- Alcock, J. E. (2010). *Attributions About Impossible Things. Debating Psychic Experience: Human Potential or human Illusion?* London: Praeger, 29–41.
- Baron, J. (1985). *Rationality and Intelligence*. Cambridge: Cambridge University Press.
- Baron, J. (2019). Actively open-minded thinking in politics. *Cognition* 188, 8–18. doi: 10.1016/j.cognition.2018.10.004
- Baron, J., Scott, S., Fincher, K., and Emlen Metz, S. (2015). Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *J. Appl. Res. Memory Cognit.* 4, 265–284. doi: 10.1016/j.jarmac.2014.09.003
- Bateman, T. S., and Hess, A. M. (2015). Different personal propensities among scientists relate to deeper vs. Broader knowledge contributions. *Proc. Nat. Acad. Sci.* 112, 3653–3658. doi: 10.1073/pnas.1421286112
- Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *J. Pers. Soc. Psychol.* 100, 407–425. doi: 10.1037/a0021524
- Blackmore, S. (1989). What do we really think? A survey of parapsychologists and skeptics. *J. Soc. Psych. Res.* 55, 251–262.
- Bösch, H., Steinkamp, F., and Boller, E. (2006). Examining psychokinesis: the interaction of human intention with random number generators—A meta-analysis. *Psychol. Bull.* 132, 497–523. doi: 10.1037/0033-2909.132.4.497
- Cardeña, E. (2011). On wolverines and epistemological totalitarianism. *J. Parapsychol.* 75, 3–14.
- Cardeña, E. (2014). A call for an open, informed study of all aspects of consciousness. *Front. Hum. Neurosci.* 8:17. doi: 10.3389/fnhum.2014.00017
- Cardeña, E. (2015). The unbearable fear of psi: On scientific suppression in the 21st century. *J. Sci. Exp.* 29, 601–620.
- Cardeña, E. (2018). The experimental evidence for parapsychological phenomena: a review. *Am. Psychol.* 73, 663–677. doi: 10.1037/amp0000236
- Chauvin, B., and Mullet, E. (2021). Individual differences in paranormal beliefs: the differential role of personality aspects. *Curr. Psychol.* 40, 1218–1227. doi: 10.1007/s12144-018-0047-9
- Chen, V. (2015). “There is no single right answer”: The potential for active learning classrooms to facilitate actively open-minded thinking. *Collected Essays Learning Teach.* 8, 171–180. doi: 10.22329/celt.v8i0.4235
- Cohen, J. (2009). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edn. London: Psychology Press.
- Coll, R. K., and Taylor, N. (2004). Probing scientists’ beliefs: How open-minded are modern scientists? *Int. J. Sci. Educ.* 26, 757–778. doi: 10.1080/0950069032000138860
- Dean, C. E., Akhtar, S., Gale, T. M., Irvine, K., Grohmann, D., Laws, K. R., et al. (2022). Paranormal beliefs and cognitive function: a systematic review and assessment of study quality across four decades of research. *PLoS ONE* 17: e0267360. doi: 10.1371/journal.pone.0267360
- Edgumbe, D. R. (2022). Age differences in open-mindedness: from 18 to 87-years of age. *Exp. Aging Res.* 48, 24–41. doi: 10.1080/0361073X.2021.1923330
- Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146
- Freedman, M., Binns, M. A., Meltzer, J. A., Hashimi, R., and Chen, R. (2023). Enhanced mind-matter interactions following rTMS induced frontal lobe inhibition. *Cortex* 22:16. doi: 10.1016/j.cortex.2023.10.016
- French, C. C. (1992). Factors underlying belief in the paranormal: do sheep and goats think differently. *The Psychol.* 5, 295–299.
- Gray, S. J., and Gallo, D. A. (2016). Paranormal psychic believers and skeptics: a large-scale test of the cognitive differences hypothesis. *Memory Cognit.* 44, 242–261. doi: 10.3758/s13421-015-0563-x
- Helfand, D. (2011). *ESP, and the Assault on Rationality. The New York Times*. Available online at: <https://www.nytimes.com/roomfordebate/2011/01/06/the-esp-study-when-science-goes-psychic/esp-and-the-assault-on-rationality> (accessed January 7, 2011).
- Hofstadter, D. (2011). *A Cutoff for Crazyiness. The New York Times*. Available online at: <https://www.nytimes.com/roomfordebate/2011/01/06/the-esp-study-when-science-goes-psychic/a-cutoff-for-crazyiness> (accessed January 7, 2011).
- Honorton, C. (1993). Rhetoric over substance: the impoverished state of skepticism. *J. Parapsychol.* 57, 191–214.
- Hyman, R., and Honorton, C. (1986). A joint communiqué: the psi ganzfeld controversy. *J. Parapsychol.* 50, 351–364.
- Irwin, H. J. (1993). Belief in the paranormal: a review of the empirical literature. *J. Am. Soc. Psych. Res.* 87, 1–39.
- Irwin, H. J. (2014). The views of parapsychologists: a survey of members of the Parapsychological Association. *J. Soc. Psych. Res.* 78, 85–101.
- Kelly, E. F., Crabtree, A., and Marshall, P. (Eds.). (2015). *Beyond Physicalism: Toward Reconciliation of Science and Spirituality*. Lanham, MD: Rowman and Littlefield.
- Kelly, E. F., Kelly, E. W., Crabtree, A., Gauld, A., Grosso, M., Greyson, B., et al. (Eds.). (2007). *Irreducible Mind: Toward a Psychology for the 21st Century*. Lanham, MD: Rowman and Littlefield.
- Kelly, E. F., and Marshall, P. (Eds.). (2021). *Consciousness Unbound: Liberating Mind from the Tyranny of Materialism*. Lanham, MD: Rowman and Littlefield.
- Kennedy, J. E. (2005). Personality and motivations to believe, misbelieve, and disbelieve in paranormal phenomena. *J. Parapsychol.* 69, 263–292.
- Kossowska, M., Jaśko, K., and Bar-Tal, Y. (2012). Need for closure and cognitive structuring among younger and older adults. *Polish Psychol. Bull.* 43, 40–49. doi: 10.2478/v10059-012-0005-6
- Krippner, S., and Friedman, H. L. (Eds.). (2010). *Debating Psychic Experience: Human Potential or Human Illusion? (Illustrated edition)*. London: Praeger.
- Mayer, G., Leverett, C., and Zingrone, N. (2022). Women and parapsychology 2022: an online survey. *J. Anmol.* 22, 465–498.
- McClendon, J. (1982). A survey of elite scientists: their attitudes toward ESP and parapsychology. *J. Parapsychol.* 46, 127–152.
- McConnell, R. A., and Clark, T. K. (1991). National Academy of Sciences’ opinion on parapsychology. *J. Am. Soc. Psych. Res.* 85, 333–365.
- McCrae, R. R. (1993). Openness to experience as a basic dimension of personality. *Imag. Cognit. Pers.* 13, 39–55. doi: 10.2190/H8H6-QYKR-KEU8-GAQ0
- Moore, D. W. (2005). *Three in Four Americans Believe in Paranormal. Gallup*. Available online at: <https://news.gallup.com/poll/16915/Three-Four-Americans-Believe-Paranormal.aspx> (accessed June 16, 2005).

- Neuberg, S. L., Judice, T. N., and West, S. G. (1997). What the need for closure scale measures and what it does not: Toward differentiating among related epistemic motives. *J. Pers. Soc. Psychol.* 72, 1396–1412. doi: 10.1037/0022-3514.72.6.1396
- Newton, C., Feeney, J., and Pennycook, G. (2023). On the disposition to think analytically: four distinct intuitive-analytic thinking styles. *Pers. Soc. Psychol. Bull.* 12:01461672231154886. doi: 10.1177/01461672231154886
- Orth, T. (2022). *Two-thirds of Americans Say They've Had a Paranormal Encounter*. YouGov. Available online at: <https://today.yougov.com/society/articles/44143-americans-describe-paranormal-encounters-poll> (accessed October 20, 2022).
- Parapsychological Association (2023). *The Role of Skepticism in Parapsychology*. Available online at: <https://www.parapsych.org/section/49/skepticism.aspx> (accessed December 28, 2023).
- Pennycook, G., Cheyne, J. A., Koehler, D. J., and Fugelsang, J. A. (2020). On the belief that beliefs should change according to evidence: implications for conspiratorial, moral, paranormal, political, religious, and science beliefs. *Judgm. Dec. Making* 15, 476–498. doi: 10.1017/S1930297500007439
- Reber, A. S., and Alcock, J. E. (2020). Searching for the impossible: parapsychology's elusive quest. *The Am. Psychol.* 75, 391–399. doi: 10.1037/amp0000486
- Rizeq, J., Flora, D. B., and Toplak, M. E. (2021). An examination of the underlying dimensional structure of three domains of contaminated mindware: paranormal beliefs, conspiracy beliefs, and anti-science attitudes. *Thinking Reason.* 27, 187–211. doi: 10.1080/13546783.2020.1759688
- Roe, C. A. (2017). PA presidential address 2017: withering skepticism. *J. Parapsychol.* 81, 143–159.
- Roets, A., Kruglanski, A. W., Kossowska, M., Pierro, A., and Hong, Y. (2015). “Chapter four - the motivated gatekeeper of our minds: new directions in need for closure theory and research,” in *Advances in Experimental Social Psychology*, Vol. 52, eds J. M. Olson and M. P. Zanna (London: Academic Press), 221–283.
- Roets, A., and Van Hiel, A. (2007). Separating ability from need: clarifying the dimensional structure of the need for closure scale. *Pers. Soc. Psychol. Bull.* 33, 266–280. doi: 10.1177/0146167206294744
- Roets, A., and Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Pers. Ind. Diff.* 50, 90–94. doi: 10.1016/j.paid.2010.09.004
- Rouder, J. N., and Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychon. Bull. Rev.* 18, 682–689. doi: 10.3758/s13423-011-0088-7
- Schlitz, M., Wiseman, R., Watt, C., and Radin, D. (2006). Of two minds: sceptic-proponent collaboration within parapsychology. *Br. J. Psychol.* 97, 313–322. doi: 10.1348/000712605X80704
- Skeptical Inquirer. (2021). *Committee for Skeptical Inquiry Names Ten New Fellows | Skeptical Inquirer*. Available online at: <https://skepticalinquirer.org/2021/01/committee-for-skeptical-inquiry-names-ten-new-fellows/> (accessed January 5, 2021).
- Stanovich, K. E., and Toplak, M. E. (2023). Actively open-minded thinking and its measurement. *J. Int.* 11:27. doi: 10.3390/jintelligence11020027
- Stanovich, K. E., and West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *J. Educ. Psychol.* 89, 342–357. doi: 10.1037/0022-0663.89.2.342
- Storm, L., Tressoldi, P. E., and Risio, L. D. (2010). A meta-analysis with nothing to hide: reply to Hyman (2010). *Psychol. Bull.* 136, 491–494. doi: 10.1037/a0019840
- Wagenmakers, E.-., J., Wetzels, R., Borsboom, D., and van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). *J. Pers. Soc. Psychol.* 100, 426–432. doi: 10.1037/a0022790
- Wahbeh, H., Delorme, A., and Radin, D. (2023). Rating the persuasiveness of empirical evidence for the survival of consciousness after bodily death: a cross-sectional study. *J. Anomal. Exp. and Cognit.* 3:125. doi: 10.31156/jaex.24125
- Wahbeh, H., Yount, G., Vieten, C., Radin, D., and Delorme, A. (2020). Measuring extraordinary experiences and beliefs: a validation and reliability study. *F1000Research* 8:741. doi: 10.12688/f1000research.20409.3
- Webster, D. M., and Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *J. Pers. Soc. Psychol.* 67, 1049–1062. doi: 10.1037/0022-3514.67.6.1049
- Weiler, M., Casseb, R. F., and Moreira-Almeida, A. (2022). A possible case of censorship of submissions on the nature of consciousness. *J. Anomal. Exp. Cognit.* 2:2. doi: 10.31156/jaex.24121
- Wilson, J. A. (2018). Reducing pseudoscientific and paranormal beliefs in university students through a course in science and critical thinking. *Sci. Educ.* 27, 183–210. doi: 10.1007/s11191-018-9956-0
- Woollacott, M., and Shumway-Cook, A. (2023). Spiritual awakening and transformation in scientists and academics. *Explore* 19, 319–329. doi: 10.1016/j.explore.2022.08.016
- Zingrone, N. L., Alvarado, C. S., and Dalton, K. (1998-1999). Psi experiences and the “big five”: relating the NEO-PI-R to the experience claims of experimental subjects. *Eur. J. Parapsychol.* 14, 31–51.



OPEN ACCESS

EDITED BY

Luca Simone,
UNINT - Università degli studi Internazionali
di Roma, Italy

REVIEWED BY

Lieven Decock,
VU Amsterdam, Netherlands
Zoran Josipovic,
New York University, United States

*CORRESPONDENCE

Sascha Benjamin Fink
✉ sascha.fink@fau.de

RECEIVED 07 December 2023

ACCEPTED 26 March 2024

PUBLISHED 27 June 2024

CITATION

Fink SB (2024) How-tests for consciousness
and direct neurophenomenal structuralism.
Front. Psychol. 15:1352272.
doi: 10.3389/fpsyg.2024.1352272

COPYRIGHT

© 2024 Fink. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

How-tests for consciousness and direct neurophenomenal structuralism

Sascha Benjamin Fink^{1,2*}

¹Centre for Philosophy and AI Research, Institute for Science in Society, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany, ²Centre for the Study of Perceptual Experience, University of Glasgow, Glasgow, Scotland

Despite recent criticism, the search for neural correlates of consciousness (NCCs) is still at the core of a contemporary neuroscience of consciousness. One common aim is to distinguish merely statistical correlates from “NCCs proper”, i.e., NCCs that are uniquely associated with a conscious experience and lend themselves to a metaphysical interpretation. We should then distinguish between NCCs as data and NCCs as hypotheses, where the first is just recorded data while the second goes beyond any set of recorded data. Still, such NCC-hypotheses ought to be testable. Here, I present a framework for so-called “sufficiency tests.” We can distinguish four different classes of such tests, depending on whether they predict creature consciousness (which systems are conscious), state consciousness (when a system is conscious), phenomenal content (what a system is conscious of), or phenomenal character (how a system experiences). For each kind of test, I provide examples from the empirical literature. I also argue that tests for phenomenal character (How-Tests) are preferable because they bracket problematic aspects of the other kinds of tests. However, How-Tests imply a metaphysical tie between the neural and phenomenal domain that is stronger than supervenience, delivers explanations but does not close the explanatory gap, uses first-person methods to test hypotheses, and thereby relies on a form of direct neurophenomenal structuralism.

KEYWORDS

consciousness, neural correlate of consciousness (NCC), phenomenal content, phenomenal character, supervenience, explanatory correlates of consciousness (ECCs)

Highlights

- Explanatory correlates of consciousness hint at explanations by predicting and thereby accounting for phenomenal features.
- What is presented as neural correlates of consciousness are often hypotheses that generalize beyond recorded data and thereby ought to be considered testable.
- In sufficiency tests for NCCs, neural data are used to make predictions about consciousness.
- There are at least four different kinds of sufficiency tests for NCC-hypotheses: Testing for creature conscious (Which-Test), for consciousness at a moment in time (When-Test), for conscious content (What-Test), or phenomenal character (How-Test).
- How-Tests require a systematic connection between the phenomenal and neural domains, thereby entailing a form of neuro-phenomenal morphism. Interpreted metaphysically, it motivates a direct neurophenomenal structuralism.

1 Introduction

The search for neural correlates of consciousness (NCCs) is central to the contemporary neuroscience of consciousness. But how can we know that we found an NCC? Or, at least, know that we are getting closer? If these questions are reasonable, they reveal that there are two ways of thinking about NCCs: as data or as hypotheses.

If we think of NCCs as data, we look at actual data sets and find correlations between neural and phenomenal variables by statistical means, i.e., whether some neural activation does correlate to some degree with some conscious experience in this finite set of data points. Because correlation is gradable, we will find NCCs in any data set unless we restrict correlation to a degree of relevance. Generally, NCCs here are “read off” actual data sets.

In contrast, if we consider NCCs as hypotheses, we go beyond any actual data set and instead generalize. That is, we presume that the occurrence of some type of neural event will *always* (at least, under some conditions) correlate with some conscious experience because it is, in a strong sense, sufficient for consciousness, as per Chalmers’ definition of an NCC (Chalmers, 2000). It is then a matter of cunning extrapolation, generalization, and theory-building to come to a reasonable hypothesis about what characterizes that type of neural event that perfectly correlates with some type of conscious experience (see also Fink, 2016). If there is such a type-NCC, it cannot be “read off” any finite set of data. Finite data sets can only be ground for hypothesizing about such a type-NCC. Instead, such type-NCCs should hold for a hypothetical set of *all possible* data sets attainable by empirical means.

Most neuroscientific “theories of consciousness” entail an NCC-hypothesis. For example, prefrontalists suggest that all NCCs involve the prefrontal cortex and thereby disagree with recursive processing theorists, who do not only focus on the prefrontal cortex but on any neural event involving recursive processing (Lamme, 2004), while apical amplification theorists argue that “apical amplification enables conscious perceptual experience” (Marvan et al., 2021), and so on. All use NCC-data as support for NCC-hypotheses, which are sometimes associated with more ambitious “theories of consciousness” (which could include additional hypotheses about the function of consciousness, its phylogenetic origins, and so on).

If an NCC-hypothesis is well enough established, we may treat it as a reliable neural *indicator* of consciousness. We then infer conscious experience from neural data. But if these inferences fail (esp. if consciousness is missing or is of the wrong kind), then this can be seen as speaking against that generalization and, thereby, a specific NCC-hypothesis. This is, in effect, a test. It is what distinguishes viewing NCCs *as data* from viewing NCCs *as hypotheses*: NCCs, viewed as data, are not testable because we do not make claims beyond the finite data set. One may doubt the methodological soundness of how the data set was assembled, but one does not put the data set to the test. Only NCCs, viewed as hypotheses, are testable because they generalize beyond any finite data set: For any neural event of type *N*, consciousness of type *C* occurs. Such generalization might succeed or fail. Whether an NCC-hypothesis fails or succeeds depends on whether the relevant neural goings-on do co-occur with the relevant kind of consciousness under the relevant circumstances.

The call for testability has already been baked into a prominent elucidation of what an NCC should be: Seth and Edelman (2009) asked for *explanatory correlates of consciousness* (see also Seth, 2009). To be explanatory, neural correlates of consciousness must

be “experimentally testable and [...] account for key properties of conscious experience” (Seth and Edelman, 2009, p. 1440).

Here, I focus on the question on this desideratum that NCCs must “account for key properties of conscious experience.” I argue that there is a specific kind of test, which I call How-Test, that leads us directly to such explanatory correlates of consciousness. In addition, such How-Tests presume a mapping of phenomenal structures (i.e., structures of experience) to neural structures. So, in an outlook, I elucidate their connection to structural approaches to consciousness.

I start with Seth and Edelman’s account and how we might interpret it (section 2) before characterizing how sufficiency tests for NCC-hypotheses work generally (section 3). I then differentiate four different kinds of sufficiency-tests for NCC-hypotheses—Which-, When-, What-, and How-Tests—before discussing their individual shortcomings and what they presuppose (section 4). How-Tests have several advantages and also maximize explanatoriness in the sense of Seth and Edelman. How-Tests are therefore preferable. However, How-Tests rest on some not-so-trivial conditions and suggest a kind of *direct neurophenomenal structuralism*, all of which I discuss in the final section 5.

2 NCCs beyond statistics: explanatory correlates in context

Seth and Edelman (2009) argued that neural correlates of consciousness (NCC) must be “experimentally testable and [...] account for key properties of conscious experience” (Seth and Edelman, 2009, p. 1440). Here, facilitating explanations is meant as an additional constraint, a constraint beyond statistical constraints (like significance) or logical constraints (like sufficiency of the neural for the phenomenal).

Such additional, non-statistical constraints on correlation are needed because, otherwise, finding correlations is cheap, and it may trivialize the endeavor of finding NCCs. Why? At least for two reasons.

First, because correlation is ubiquitous: At its core, it is just a measure of the degree of dependence between the values of two variables. Traditionally, in the neuroscience of consciousness, we “treat consciousness as a variable” (Baars, 1997) and inquire which variable in our neuroscientific data is co-dependent on it. However, any two variables correlate statistically to some degree, even if only slightly in some random samples (such as individual data sets).¹ In science, the way to avoid triviality is to only *report* correlations that are significant, suggestive, etc. What makes these significant, suggestive, etc., is that the degree of dependence exceeds some numerical cutoff point. Technically, however, there is still a correlation between variables below these thresholds, but to a degree where we find it uninformative. This is illustrated by the fact that, historically and contextually, the

1 Rodgers and Nicewander (1988) diagnose 13 different ways of assessing correlation coefficients between the values of variables, all of which are gradable, e.g., the (Galton-) Pearson product-moment correlation coefficient (Pearson, 1895; Stigler, 1989), Spearman’s or Kendall’s rank correlation coefficient (Spearman, 1904; Kendall, 1938; Kruskal, 1958), or Székely’s distance correlation measure (Székely et al., 2007).

cutoff point can vary. Correlation, unconstrained by such cutoffs, is ubiquitous and therefore trivial to find.

Second, because correlation is “metaphysically promiscuous” (Fink and Lin, 2022), different positions on how the mind relates to the body—even positions contradicting each other!—are still compatible with systematic correlations between mental and bodily events. This has a great advantage: If we know that x and y correlate, we can largely bracket the question of *how* they relate, e.g., whether neural and phenomenal goings-on are identical (Place, 1956) or are two distinct but co-occurring properties (Chalmers, 2003), whether one supervenes on the other (Kim, 1979) or emerges from the other (Silberstein, 2001), whether they are two aspects of the same (Spinoza, 1677) or merely in pre-stabilized harmony (Leibniz, 1720), etc.² Empirical NCC researchers focus on finding out *which* neural goings-on correlate with *which* phenomenal goings-on. They focus on the relata, while metaphysicians theorize about the relation. But no matter what metaphysicians converge on at the end of the day (if they converge at all), their answer will be compatible with a correlation between what is given by neuroscientific means and what is given in introspection or phenomenology.³ Indeed, that has been one of the motivating factors behind focusing on correlates rather than something else: Crick and Koch (1998, p. 97) forcefully asserted that they “think that most of the philosophical aspects of the problem should, for the moment, be left on one side, and that the time to start the scientific attack is now.” Focusing on correlation, which is promiscuous to many forms of metaphysics, allows for this beneficial division of labor.

However, some researcher may still want to contribute to metaphysics by finding where consciousness has its foothold in the physical world, i.e., by identifying the neural substrate of conscious experience. To differentiate it from merely statistical NCCs, call this the *NCC proper*: The NCC proper is that NCC which lends itself to metaphysical interpretations (such as identification and realization), even though it does not force a specific one.

However, we can never be sure that there is *any* metaphysical relation between measured correlates. Even if we add statistical thresholds, there may still be significant correlations without any underlying connection, which Pearson (1897) called “spurious correlations.” To sieve these out, we need additional constraints on correlation.

Which constraints on correlations should we accept? Some of these are already motivated by statistical considerations. Beyond the statistical constraints, we find, e.g., the ability to account for phenomenal features (Seth, 2009; Seth and Edelman, 2009),

synchronous occurrence with the phenomenal experience (Aru et al., 2012), being systematically entailed by a theory (Hohwy and Seth, 2020), being necessary and sufficient (Crick, 1995), or—most prominently—being minimally sufficient (Chalmers, 2000). These non-statistical constraints on correlation are motivated by special goals or interests and therefore are not universally accepted or adequate. Synchronicity, for example, would be a detrimental constraint on NCCs if our goal is to *avoid* the occurrence of consciousness, e.g., during surgery: Anaesthesiologists would rather like to know neural precursors to an experience in order to have enough time to intervene and thereby prevent the awakening of a patient. Or consider that a demand for being systematically entailed by a theory may be ill-motivated at the beginning of a research program when theories are missing, are rudimentary, or cannot yet be fleshed out in neural terms (compare Overgaard and Kirkeby-Hinrup, 2021).⁴ There would be no place for NCC research to start if entailed-by-theory were a universal constraint.⁵ Therefore, most non-statistical constraints on NCCs are only reasonable in context—and the same holds for the demand to be explanatory in the proposal by Seth and Edelman (2009).

There are at least two reasons why we might be equally skeptical about NCCs being explanatory.

First, no NCC could fulfill the requirement of facilitating explanations if an explanatory gap persists (Levine, 1983). Accepting an explanatory gap does not automatically make us anti-materialists, as Papineau (1993, p. 180) and Levine point out: Even if phenomenal goings-on are indeed identical to neural goings-on, we cannot explain that identity. Identities just are. Water just is H_2O . Asking “But why?” is futile. This is one likely ingredient of the meta-problem of consciousness (Chalmers, 2020).

Second, explanatory correlates may very well pick out merely statistical correlates because explanations are not always indicators of truth. In one prominent view, they are reason to accept a fact, an answer to a *why*-question (van Fraassen, 1980, ch. 5): This x is so *because of* y . The best explanations certainly are true, but the history of science is full of false answers to *why*-questions.⁶ However, we can hardly deny that even faulty attempts are nevertheless explanations, just not good ones. It makes sense to distinguish between successful and faulty attempts to explain where the first one tracks truth and the second does not—but this requires dissociating explanation from tracking truth. As a matter of fact, humans accept something as an explanation if they *accept* its explanans as *true*, not if the explanans is

2 Ward (1911, 600–602), one of the first to use the phrase “neural correlates of consciousness,” advocated for a *methodological parallelism*: “We reject materialism, accordingly, while still maintaining this *psychoneural parallelism* to be a well-established fact. From this we must distinguish a second sense of parallelism founded on the disparity just mentioned as pertaining to the psychical and neural correlates. We may call this *physiologico-psychological*, or, more briefly, *methodological parallelism*. It disclaims as illogical the attempt to penetrate to psychical facts from the standpoint of physiology [...]. It also forbids the psychologist to piece out his own shortcomings with tags borrowed from the physiologist. The concepts of the two sciences are to be kept distinct [...].”

3 The only exceptions are variants of eliminativism.

4 Overgaard and Kirkeby-Hinrup (2021) attest that most theories of consciousness are only loosely connected to neural implementations. Therefore, finding the NCC will not solve all problems concerning which theories of consciousness is the right one.

5 Other constraints (such as necessity, sufficiency, or minimality) are worrisome for interdisciplinary projects: If something neural must be considered as necessary for an experience, then NCC research cannot inform (or be combined with) research on artificial consciousness, mind-uploading, or embodied or extended approaches. Minimality might be problematic if we ponder distributed systems with parts that are already conscious, like the United States (Clark, 2010). Mere sufficiency might not be acceptable if we want to keep identity theory as a candidate (Polák and Marvan, 2018).

6 For example, the uptake of phlogiston was used by Rutherford to explain why plants burn so well (Conant, 1964)—but there is no phlogiston.

in fact true.⁷ Similarly, some candidates for an NCC proper might lend themselves to explaining phenomenal features—but actually lack any metaphysical connection. Grush (2006) criticized proposals for the NCC regarding the phenomenal flow of time by Varela (1999) and Lloyd (2002). Each *explains* those phenomenal features of the slightly extended “saddle back” of the felt moment, but each fails to be a proper NCC for other reasons.

For these two reasons, the demand for being explanatory might not only filter out those neural activations *to which experiences are identical* as proper neural correlates, but it might also favor merely statistical correlates if they, e.g., have similar features to a coincidentally co-occurring phenomenal experience. Therefore, we might want to reject explanatoriness, despite being desirable, as a universal constraint.

Seth and Edelman continue with two constraints that have the potential for being universal constraints, namely that we should search for correlates that are “experimentally testable and [...] account for key properties of conscious experience” (Seth and Edelman, 2009, p. 1440). Each can be dissociated from explanation even though each facilitates explanations.

To be testable, we should interpret “accounting for key features” as facilitating certain predictions: Use the neural to predict conscious features. NCC-hypotheses would be testable by how well they allow us to predict phenomenality. In the next section, I will focus more generally on testing NCC-hypotheses before distinguishing four kinds of tests in section 4. Of those, the so-called How-Test maximizes “accounting for key features.”

3 Testing NCC-hypotheses

I argued that we need non-statistical constraints on correlation and that the explanatoriness of an NCC is, by itself, not necessarily a universal constraint. However, explanatoriness is a desirable feature if we aim for a neuroscientific account of consciousness, where goings-on in the brain are used to account for the presence of some form of consciousness. However, “accounts for” need not be read as “explains.”

Another way to read Seth and Edelman’s notion of “accounts for” is as *prediction*: If neural goings-on truly accounts for phenomenal goings-on, we should be able to *predict* consciousness based on neural data. Successful prediction of consciousness’s features based on neural data is then an indicator of proper “accounting.” It is also a general and necessary constraint on NCC-hypotheses: If a candidate for an NCC fails to fit incoming data, we ought to reject it. This interpretation emphasizes how close accountability is to testability.

Testing NCCs is not too different from testing in other areas. Generally, we can expect three stages: In the first stage (data collection), we gather data. In the second stage (hypothesizing), we come up with more general hypotheses (e.g., by proposing models, theories, laws). In

the third stage (testing), we test our hypotheses against new data. How does this apply to the neuroscience of consciousness?

In the first stage, we gather data about which individual neural events correlate with which phenomenal events. Fink (2016) calls such a tuple a *token-NCC* because it concerns non-repeatable particulars in specific subjects at specific moments under specific circumstances.⁸ Here, constraints come into play to arrive at a more refined set of data that reduces possible noise in the data.

In the second stage, the goal is to find unifying principles among heterogeneous sets of tuple-NCCs by choosing specific features shared by them. It is worth hypothesizing that these common features are *NCC-makers*: We suggest that all (and only) neural events that have those features will co-occur with consciousness. If hypothesis *H* is true, its associated NCC-makers constitute the *type-NCC*. The hypothesis is that any neural token that has these features will also correlate with experience.⁹

However, not all features shared by token-NCCs in the data set will be suitable NCC-makers because some will not contribute to a neural event’s status as an NCC at all. For example, features like the weight of the activated area, its color, or its distance to the left eye can likely be ignored. Other features are preferable candidates for being NCC-makers, e.g., an area’s location in the overall structure of the nervous system, its interconnections to other areas, its role in neural processing, and so on.¹⁰

This picture sketches mainly a *bottom-up* approach to theorizing. Therefore, spelling out NCC-makers in the language of neuroscience is preferable, even if this *prima facie* limits our NCCs to neural systems. This limitation, however, is only *prima facie*, as the NCC-making features might also occur in non-neural systems as well (e.g., recursive processing). However, in this approach, these abstract features must be grounded in neural data to be considered as NCC-makers instead of being motivated by conceptual reasoning (as in, e.g., higher-order thought theory) or phenomenological reflection (as in, e.g., integrated information theory).

Such bottom-up motivated type-NCC-hypotheses allow for predictions because (a) they are general and (b) they specify neural events as being sufficient for a conscious experience: Any of the competing hypotheses claim that neural events with *these* features will correlate with consciousness. If events with these hypothesis-specific features do not correlate with consciousness, then that hypothesis apparently did not pick the right bunch of features. It loses credibility. If such events do correlate with consciousness, it gains credibility. By

⁷ This is a reason to reject the ontic account of explanation as brought forward by, e.g., Craver (2014). There, the facts in the world do the explaining. However, then, to spot an explanation, we would need know which facts pertain before we can know whether some speech act amounts to an explanation or not. The ontic account conflates whether some speech act is an explanation with whether an explanation is true.

⁸ Thus, data points in NCC research are not between neural and phenomenal states because states are repeatable (see Steward, 1997). Instead, they are events.

⁹ There might also be partial type-NCCs, i.e., types that capture some token-NCCs (e.g., in non-pathological humans), but cannot be generalized to encompass *all* token-NCCs (e.g., all humans but not all animals). For example, it might be that some, but not all NCC, are marked by thalamic activation (see, e.g., Young, 2012). Then, thalamic activation might be a partial NCC-making feature, a partial type-NCC. In the following, I will focus on universal type-NCCs when I speak of type-NCCs, i.e., NCC-makers that pick out all NCCs.

¹⁰ Ward (1911, p. 602) already mentioned that *morphological* features are likely not as relevant as physiological features for NCCs.

such predictions, type-NCC-hypotheses are testable insofar as the chosen features are detectable.¹¹

In the third stage, we can put universal type-NCC hypotheses to the test. We do so by looking for a neural event *e* that has the relevant NCC-making features. We then see whether *e* comes with consciousness. (Admittedly, this might be the hardest methodological challenge, as the discussion concerning access vs. phenomenal consciousness illustrates.) If *e* does not come with consciousness, this undermines the fact that the chosen NCC-making features are sufficient for consciousness. These are, therefore, tests of sufficiency, not necessity (see Fink, 2016, for tests of necessity).

This framework allows us to interpret Seth and Edelman's demand that neural correlates should be "experimentally testable and [...] account for key properties of conscious experience" (Seth and Edelman, 2009, p. 1440) in terms of *prediction* rather than *explanation*. In contrast to explanation, prediction is a more universal constraint in that it appears to be more compatible with different metaphysics or preconceptions about the problems that might remain at the end of the day (e.g., the explanatory gap). Additionally, even the best explanation must be abandoned if it fails to fit new data. Prediction therefore trumps explanation as a mark of quality. In this sense, reading "accounts for" as "predicts" emphasizes its role in testing, an emphasis Seth and Edelman themselves made.

Additionally, testing is now a core duty in NCC research. While explanation is mainly a *post-hoc* activity, one we can only do *after* data are collected and analyzed or *after* tests are done, prediction is an

ante-hoc activity, one we do *before* the relevant data are collected or analyzed, *before* we test. Only already gathered data need explanation—it comes at the dusk of a research project; prediction, instead, motivates further data gathering—it comes at the dawn of new research. Explanations may suggest further tests, but only so far as they also engender predictions. Predicting is therefore often more fundamental than explaining.¹²

However, even if we could perfectly predict from neural data *when* an experience occurs, we might still fail to account for this experience's features or "key properties," as Seth and Edelman demand. Mainly because a prediction of occurrences is not a prediction of features. A *linea negra* allows us to predict the occurrence of a birth in the following months, but it does not account for the baby's features, e.g., its hair color.

Luckily, explanation and prediction are not exclusive: Our best universal type-NCC-candidate might allow us to predict *and* explain. The question is: Is there a kind of test that *maximizes* "accounting for phenomenal features" in both the sense of prediction and explanation without each one's shortcomings?

To answer this question, I distinguish four kinds of tests in the next section. The tests are characterized by what they predict. For each, I present examples and discuss their shortcomings. One of these, the How-Test, seems to strike a nice balance between prediction and explanation. It is, in my view, the kind of test best suited to finding meaningful and relevant NCCs. The How-Test, however, has interesting implications, which I discuss in the last section.

4 Four kinds of tests in NCC research

I argued above that we can view what is often called "NCCs" either as data or as hypotheses. "NCCs", understood as data, refer to sets of measured data points (i.e., sets of token-NCCs), while "NCCs", understood as hypotheses, go beyond measured data. Here, we aim at characterizing general NCC-makers, i.e., features that make any neural event with these features correlate with consciousness. NCC-hypotheses therefore aim to capture type-NCCs. Because of their generality, these NCC-hypotheses are testable. But how do we test?

In an NCC-sufficiency-test, we aim to find out whether a chosen set of measurable features *F* is a NCC-maker (for experiences of a type *C*). In other words: Do *all* neural activations that have *F* correlate with consciousness (of type *C*) or not? If yes, then *F* counts as sufficient for consciousness. If not, then *F* is not sufficient. If *F* is not sufficient, then *F* does not constitute a type-NCC. Therefore, the hypothesis that picked *F* as an NCC-maker is less likely to be true.

A test can be either supportive or undermining to be informative. In both, I focus here on sufficiency, which is prominent in defining NCCs as being *minimally sufficient* for consciousness (Chalmers, 2000).¹³ In *supportive* tests, we aim to show that if the chosen

11 While I focused on bottom-up theorizing, the same holds for type-NCC-hypotheses that are derived top-down: Sometimes, NCC-making features are not derived primarily from neural data, but from a theory—what Hohwy and Seth (2020) call systematic NCCs. This process is not always straightforward because many available theories of consciousness relate only loosely to neuroscience (Overgaard and Kirkeby-Hinrup, 2021; Schlicht and Dolega, 2021). So, here, we first need to translate the non-neural posits of a theory (e.g., higher order thoughts, dynamic cores, fame in the brain, etc.) into neural terms. Then, these neural analogs are picked as NCC-making features. Again, such top-down type-NCC-hypotheses allow for prediction and testing. Here, however, immunization is too easy: If we find a mismatch between incoming data and prediction, then this does not necessarily speak against the theory of consciousness. Instead, the mismatch could be due to a failed translation of its posits into neuroscience. For example, most neuroscientists favor prefrontal activation as the neural equivalent of higher-order thoughts, but one might also consider areas with specific activation triggered reliably by input from lower sensory areas as being a seat of higher-order representations. This loose relation between non-neural theories of consciousness and neural events makes testing such theories tricky. For example, IIT's Φ might be an NCC-making feature, but is hardly measurable in complex systems such as human brains. It is unclear to which degree approximations of Φ really allow us to test IIT itself. For any failed test, critics can always see the mistake in the approximation, not in the theory. If we want to increase scientific progresses by systematic falsification of theories—as both Popper, experimentum crucis tests, and null-hypothesis testing suggest—then we minimize experimental ambiguity. Thus, direct detectability of the NCC-making features is an advantage. This favors capturing NCC-makers on the implementational rather than the algorithmic level. Neural correlates first, computational correlates of consciousness second (contra Wiese and Friston, 2021).

12 This illustrates why projects such as COGITATE are such an important step forward in the discipline.

13 Fink (2016) focuses on comparative tests where we pitch NCC-hypotheses against each other such that the results of a test are at the same moment supporting one *and* undermining the other. This is the underlying rationale of

NCC-making feature-set F is present in a neural event, so is the relevant kind of consciousness. In *undermining* tests, we show that a neural event that has the relevant features-set F fails to correlate with the relevant kind of consciousness. So, we show that these features are *not* sufficient for consciousness. Notably, this differs from similarly common tests of *necessity*, featuring prominently in the battery of tests by the COGITATE project (Melloni et al., 2023). Here, the failure of some neural features to occur even though a person was conscious in the relevant way is supposed to speak against a hypothesis. Here, however, one goes beyond the classical understanding of an NCC because one tests whether a neural type is *necessary* for consciousness.

In contrast, all of the four kinds of tests discussed here are tests of sufficiency, not tests of necessity.

NCC-tests that focus on sufficiency use neural data to motivate a prediction about consciousness: Given such-and-such neural facts, we expect such-and-such conscious facts. Thus, all predictions in these tests only concern phenomenality. (Note that as soon as we predict specific neural event types based on phenomenality, we enter into necessity tests).

Unfortunately, phenomenality is itself not directly accessible “from the outside.” So, strictly speaking, what is predicted are often *indicators* of phenomenal change. For example, we may predict a specific psychophysical performance indicating a change in the magnitude of an illusion for a given individual. Or we might predict a specific type of verbal report indicating a change in experience.¹⁴ However, we should not mistake such indicators of phenomenal change for what is predicted: Different methods of assessing phenomenal change (e.g., introspective report, psychophysical performance, a gaze shift, etc.) may all indicate *the same change in phenomenality*. What is predicted is, first of all, the phenomenal change. How this change in experience affects observable indicators is secondary. Unless one defends a behavioristic theory of consciousness, what is predicted are phenomenal features first and foremost.

What distinguishes the four tests is the kind of prediction they focus on. Predictions can concern creature consciousness, state consciousness, phenomenal content, or phenomenal character. That is, roughly, (i) *which* systems can be conscious (creature consciousness), (ii) *when* systems are conscious (state consciousness), (iii) *what* a system is conscious of (phenomenal content), and (iv) *how* a system that is conscious is experiencing this state (phenomenal character). For each test, I present a paradigmatic example from empirical literature, and discuss the problems that are associated with it. Of the four, the How-Test avoids most problems plaguing the others.

adversarial collaboration such as COGITATE (Melloni et al., 2023), which should be considered a leap forward for the field. However, this approach already presupposed that we have to go beyond Chalmers’s definition of an NCC, as Fink (2016) points out: On the level of type-NCCs, we have to presume that some features are *necessary*, such that all neural events that correlate with consciousness will share these features. In this article, however, we do not need to go so far: We can focus on sufficiency tests.

¹⁴ For example, we may predict what you report yourself as thinking about during a daydreaming episode. We might even predict a phenomenology, i.e., we predict how the change of a deep structure of experience is captured in a specific phenomenological theory (e.g., Husserlian, Merleau-Pontyian, Sartrean, Heideggerian, or otherwise).

4.1 Which-Tests

First, the Which-Test. Here, the predictions concern the kinds of organisms that can be conscious, given their neural architecture. The prediction has the form:

Which-Test: If an organism o with a neural system s is capable of neural events with features F_1, \dots, F_i , then o is capable of conscious experiences.

Which-Tests are therefore tests for *creature consciousness* (Rosenthal, 1986).¹⁵ As such, it is a question about a capability: Not “Is this thing conscious?” but “Can it be conscious?”

A paradigmatic example is the discussion on whether fish can feel pain (see Braithwaite, 2010; Michel, 2019, for an overview). If, for example, thalamo-cortical loops are a requirement for consciousness (see, e.g., Bachmann et al., 2020), fish cannot feel pain because they have no cortex and their brain is therefore incapable of thalamo-cortical loops. However, fish could be conscious if local recurrent processing were sufficient for consciousness (Lamme, 2004, 2006). If we know whether fish are capable of feeling pain, then we can decide whether we should rather accept thalamo-cortical loops or recurrent processing as proper type-NCCs. Another currently prominent example is the discussion about AI consciousness.

There is, however, a fundamental problem with the Which-Test: Consciousness is, unfortunately, largely private. As external observers, we cannot directly observe its presence in others, especially in non-humans.

If consciousness is private, we have to rely on indirect measures and indicators. However, for nearly any indicator, its sensitivity, reliability, accuracy, or significance has been questioned (at least by illusionists, see Frankish, 2016). Each indicator for consciousness can likely be gamed, as discussions on AI consciousness illustrate. Even for humans—organisms of which we are most certain that they are capable of consciousness—the reliability of behavioral markers is seriously questioned: Blocking behavior does not block consciousness, as anaesthetic awareness illustrates.

Doubts about the sensitivity, reliability, or accuracy expand even to cognitive indicators, at least as long as we cannot reject the distinction between *access* and *phenomenal* consciousness (Block, 1997): If the phenomenal features of an event are (or: can be) accessed by other neural subsystems—i.e., if these phenomenal features influence their processing (e.g., is used in guiding action, belief, deliberation, evaluation, affect, etc.)—then this event is access conscious. If it feels like something is in that state (i.e., if it has phenomenal features), then it is phenomenally conscious—independently of whether these features are also accessed. The distinction, which was first introduced as a conceptual distinction (Block, 1995), has drawn a lot of discussion and criticism, but it has not been ruled out yet. In fact, several neuroscientists accept it (e.g., Lamme, 2004; Koch and Tsuchiya, 2007). Later, Block (2005) argued that the distinction between access and phenomenal consciousness is not merely conceptual but truly picks out different neural processes.

¹⁵ But see McBride (1999) for a critique.

If the distinction between access and phenomenal consciousness cannot be ruled out, then what we can observe in others or gather from their reports can only count as indicators of access consciousness. This leaves open whether what is accessed were phenomenal or non-phenomenal states. If so, none of the behavioral or cognitive indicators for the presence of consciousness can count as absolutely reliable. More so, it also leaves open whether some phenomenal features we predicted but failed to measure were merely *unaccessed*. In principle, we might be correct in our predictions but lack the means to show that. So even in humans, ascriptions of consciousness outside non-pathological middle-aged subjects (e.g., vis-à-vis fetuses or comatose patients) are therefore open to reasonable doubt. This holds *a fortiori* if we go outside the species of *homo sapiens*. This contestability is a severe drawback of any Which-Test.

Which-Tests are helpful to illustrate that two theories about NCC-makers are not co-extensional (because they attribute consciousness to different organisms). However, it is far from being an uncontentious test for NCC candidates themselves due to the lack of direct external access to the phenomenal correlate. Any indirect indicator relies heavily on calibration in non-pathological middle-aged subjects (Goldman, 1997). Therefore, they become more and more dubitable and untrustworthy the further we stray from this group.

A solution to this problem is to focus on individuals where doubts about their ability to be conscious are minimal, namely middle-aged humans.

4.2 When-Tests

In a When-Test, researchers focus on organisms where we can be reasonably certain that they are conscious: If they are not conscious, then neither are the researchers. This often means adult *homo sapiens*.

However, not anything that *can be* conscious *is* conscious. In some phases of our life—deep sleep? stupor? anaesthesia?—we are usually considered to be *unconscious*. The prediction in When-Tests has the form:

When-Test: If an organism *o* with a neural system *s* is in a state *n* with features F_1, \dots, F_i at *t*, then *o* is conscious at *t*.

When-Tests are therefore tests for *state consciousness*: We predict when a system is in a conscious state. Not “Can this thing be conscious?” but “Is it conscious *now*?”

A paradigmatic example comes from research into dream consciousness. A classical view was that we are conscious during REM sleep phases but lose consciousness in NREM phases (Aserinsky and Kleitman, 1953). Crick and Mitchison (1983) even equate dream sleep with REM sleep. Looking at the differences in neural activation between REM- and NREM-phases (understood as dreaming and non-dreaming phases) could then be used for tracking down NCC-makers.¹⁶ Another

case might be anaesthesia: While we are usually conscious, humans are considered to be unconscious under anaesthesia. Several common anaesthetics are antagonists of the NMDA-receptor. Flohr (2000) can be read as suggesting that the functioning of the NMDA-receptor complex is a candidate for a universal type-NCC.

However, both sleep consciousness and anaesthesia also illustrate core problems with When-Tests. They also relate to the privacy of consciousness: During certain phases of our lives, it is hard to assess from the outside whether someone is conscious or not.

Again, if the distinction between access and phenomenal consciousness cannot be ruled out, then certain phases might only come with diminished *access* to our phenomenal goings-on rather than diminished phenomenality itself. This means that it could be missed *even by the experiencers themselves*. Most of the phases that come into focus for a When-Test—anaesthesia, sleep, stupor, dementia, coma, and so on—are already marked by diminished cognitive and behavioral abilities. So, it is not out of the question that our third-person methods for externally assessing the presence of consciousness as well as second- and first-person methods simply fail to keep track of phenomenality during these episodes. At the very least, there is a non-negligible uncertainty about whether an absence of evidence for phenomenality should count as evidence for the absence of phenomenality itself. In dream research, for example, REM was early on associated with dream sleep mainly because subjects reported most often and most detailed when awakened from such phases. However, now, we do have enough evidence of dreams during NREM-phases (see, e.g., Suzuki et al., 2004). Being able to report after awakening is then not necessarily a condition for dream experiences.¹⁷ Similarly, most anaesthetic cocktails do not only block muscle movement but also inhibit the formation of memories—something that might even be desirable (Ghoneim, 2000). That the absence of evidence for consciousness was no evidence for its absence became obvious when anaesthesiologists themselves provided reports from experiences under such chemical influences (Topulos et al., 1993). An extreme conclusion from this research would be: We never lose phenomenal consciousness, but at most lose access to it.

Again, we may use the When-Test to show that two hypotheses differ: If hypothesis *A* makes different predictions than hypothesis *B* concerning phases of unconsciousness, then they are not co-extensional. Ideally, such predictions can be used empirically. However, any When-Test is hardly uncontentious due to the limitations on accessing phenomenality from the outside.

A solution to this problem is to focus on episodes where accessibility is less controversial. The following two types of tests, What- and How-Tests, therefore only concern such phases of uncontested access.

¹⁶ This is the route suggested by, e.g., Nir and Tononi (2010, p. 92): “In principle, studying mental experiences during sleep offers a unique opportunity to explain how changes in brain activity relate to changes in consciousness

[...]. In fact, if it were not for sleep, when consciousness fades in and out on a regular basis, it might be hard to imagine that consciousness is not a given, but depends on the way in which the brain is functioning.”

¹⁷ For a different view, see Malcolm (1959).

4.3 What-Tests

In the What-Test, we do not focus on contentious organisms (such as fishes or embryos), nor do we pick contentious episodes (such as deep sleep, dizziness, intoxications, anaesthesia, or coma). Instead, we focus on predicting the content of an experience. Not “Can this thing be conscious?” or “Is it conscious *now*?” but “What is it conscious of?” The prediction in What-Tests has the following form:

What-Test: If an organism o 's neural system s is in a state n with features F_1, \dots, F_i at t , then o is conscious at t of x .

Because the What-Test focuses on the contents of experiences, it is closer to “accounting for phenomenal features” than the other two tests, which did not predict features of consciousness itself but the presence of consciousness *per se*.

An interesting example of a What-Test comes from Horikawa et al. (2013). The team used a pattern classifier combined with a semantic net trained on fMRI data to predict the content of dream reports. If dream reports are seen as reflecting the contents of dream experiences, then the neural features used for this classification are good candidates for being NCC-makers of this specific conscious content. If the pattern classifier makes predictions about dream content *beyond the training set*, one can assess the accuracy of such predictions.¹⁸ Such What-Tests have the advantage that we circumvent the Which-Test's problem of contentious organisms and the When-Test's problem of contentious conscious episodes (although not in this specific case).

However, there are problems with What-Tests too. First, there are quite a number of competing theories on how a mental state gains its content, i.e., theories of what determines that it has *this* content rather than any other. But we need to decide on one to perform a What-Test. Therefore, we would be reliant on three separate assumptions for each What-Test: (i) an NCC hypothesis we wanted to test, namely which neural features makes a specific content *conscious*; (ii) a theory about the circumstances that determine the content of a neural event; and (iii) a theory about where the content-carrying vehicles are located in the brain (if we abstract from location: a theory of how the brain codes for content). The focus is on testing (i), but in a What-Test, we are reliant on (ii) and (iii) as well. The latter become additional and independent *variables*. If a type-NCC-hypothesis fails a What-Test, then the result is ambiguous: One can hardly decide whether this speaks against a specific theory about the *location* of content-carrying vehicles, against a specific theory of what determines content for a located neural vehicle, or against a theory of what makes content conscious, i.e., a hypothesis about NCC-makers. This is an unfortunate ambiguity.

Second, in some cases, an individual may not be able to tell what the content of their conscious mental state is. Consider, as examples, hypnagogic imagery, visual hallucinations in a Ganzfeld, or phantasms under psychedelics: Individuals themselves are puzzled concerning what exactly it is that they are experiencing. They might be able to draw something resembling their visuals—even to a degree where they

can print it on a T-shirt—but they may still be unable to say what this drawing represents. There might be a principled reason for this: Wollheim (1987) distinguished between representational and configurational aspects of an image. In some cases, we may only grasp the configurational aspects while the representational aspects are inaccessible, maybe even inexistent.

There is even an open debate on whether all phenomenal states have content or whether there are some that have phenomenal features that are not grounded in content, i.e., mental paint or mental latex (Block, 1996). Psychedelic visuals and similar states could be cases of this: They could be states with configurational aspects but without (accessible) representational aspects. If so, then What-Tests are limited in their application.

Even in cases where subjects can access their conscious contents perfectly, they may lack the conceptual or expressive capacities to *convey* the content accurately to external researchers, either by language or other means. So, could the Horikawa paradigm be executed with someone with amnesia, aphasia, anomia, and an incapability to draw? Hardly. They could not provide dream reports, verbal or otherwise. But would this mean that this person does not dream? Hardly.

So, again, we need a way to assess the content of a conscious experience *externally*. This would be unproblematic if we go with externalist theories of content fixing, where external circumstances determine the content of a mental state. However, most representational theories of consciousness arguably focus on *narrow* content, which can be adequately appreciated by the experiencing subjects and with subject-internal conditions for content-determination. Only for narrow content does it make sense to locate the vehicle of specific content *inside* a brain. For non-narrow content, the same localisable neural vehicle may carry different contents, depending on external circumstances (Burge, 1979). So, no neural vehicle alone could count as sufficient for a specific content. This hardly squares with the definition of NCCs where neural states are considered to be minimally sufficient for consciousness. If we search for neural correlates for conscious contents in Chalmers' sense, phenomenal content must be narrow.¹⁹

This suggests a tension: externally accessible content fixers would allow us to override the subject and make content externally assessable, but they do not lend themselves to neural correlates of conscious content because the correlation of content would extend beyond the brain. Therefore, internally accessible content fixers are currently the most prominent candidates for conscious content that is fully introspectable. However, narrow content will sometimes be ineffable²⁰ or fail to be externally assessable. The What-Test, to me, seems to steer us into this unattractive dilemma.

¹⁹ An additional problem is created for non-narrow theories where what a person says about the content of her mental state diverges from what the content truly is. For example, in teleofunctionalism, the evolutionary history of one's species determines the content of one's mental states. Then, our own attributions of contents (e.g., I see a woman with clean skin) may diverge from what could be the actual content of the mental state (e.g., I see a woman with genes for parasite resistance).

²⁰ This ineffability is not one of principle, but a contingent one: Would the person have had the conceptual capacities, they may have conveyed it to external observers. But, as a matter of fact, they lacked the conceptual capacities. The ineffability of content is here capacity-relative.

¹⁸ The unfortunate disadvantage of that study is that it does not rest on a specific hypothesis about NCCs, but rather shows that pattern classifiers for the content of dream reports can be trained on fMRI data.

A third problem for What-Tests is that they rely on contents being systematically and rigidly associated with their neural vehicles: If we do not assume such a systematic and rigid association, we cannot predict any kind of content given only neural data. However, there is no such strong relation between contents and vehicles: The content *red* can be represented by ink on paper, sound waves, chiseled lines in stone, chalk on a blackboard, certain neurones firing, etc. Certain contents may put *constraints* on which neural architectures can implement them (arguably, temporal retention and protention are contents of this kind; see Grush, 2005, 2006). However, even if contents motivate constraints on neural architecture, these will not be so strong that we end up with a one-to-one relation between contents and architectures, but likely one-to-many: The same content can still be found in many architectures. Me, a squid, and a robot may all represent “danger.” Vice versa, the content “*and*” (conjunction) may need a specific wiring, but this does not mean that all wirings of that kind on any scale of the neural system necessarily represent “*and*.” Therefore, we cannot infer from a specific set-up of a neural vehicle what its content is—or whether it has content at all.

We could say, as representationalists do, that representational features—what is being represented where and in what format—are indeed NCC-makers. However, such representational features should currently count as additional *non-neural* contributing factors that make neural events an NCC. We do not know if such representational features reduce solely to neural features or reduce at all. Even if they are reducible to neural features, it is not obvious to which neural features they reduce to because, currently, no reductive theory of representation is universally accepted. Under these conditions, we cannot expect to capture what makes an NCC solely in neural terms if the NCC-maker is representational.

If the same content can be represented across different neural (and non-neural) systems, then theories of content determination must count as additional assumptions. Consider two neural events *a* and *b* of the same type: one may have and the other may lack specific representational features if non-neural factors co-determine content. In that case, neural data hardly suffices for predictions of conscious content. This is illustrated in the study by Horikawa et al. (2013): The pattern classifiers is *trained for individuals* because we lack a neural theory of content attribution fine-grained enough for interindividual predictions of content.

There is no connection between contents and their vehicle constrained enough to predict content from vehicles without contentious additional auxiliary hypotheses.

Even though What-Tests could be among the most promising tests for NCC hypotheses, they will hardly be decisive.

4.4 How-Tests

How-Tests rely on the distinction between phenomenal character (roughly, how something feels like) and phenomenal content (roughly, what we are conscious of).²¹ This mirrors the distinction between

representational and configurational aspects introduced for paintings (Wollheim, 1987) and later extended to aesthetic perception and representational seeing (Nanay, 2005). If accepted, we can remain open to what Block (1996) calls mental paint or mental latex—experiences that either lack representational content (latex) or where phenomenal character is not determined by content (paint). Even if the distinction between content and character is only conceptual, How-Tests predict character itself from neural data—without a detour via content. Its predictions have the following form:

How-Test: If an organism *o*’s neural system *s* is in a state *n* with features F_1, \dots, F_i at *t*, then the organism *o* is conscious at *t* (of *x*) in a *y*-way.

For How-Tests, we neither ask “Can this thing be conscious?” nor “Is it conscious *now*?” nor “What is it conscious of *now*?” but only “How does it feel under these conditions?”

The character of a mental event is introspectable (at least in so far as it is accessible). The content of a mental event (at least if externally co-determined) may only be partially introspectable. Additionally, while content can be shared across individuals to allow for communicable thought, character likely differs across individuals even under the same conditions (Hohwy, 2011; Fink, 2018).

How-Tests exploit this possibility of phenomenal variations under the same conditions across individuals. They focus on *inter-individual differences*: Under the same external conditions, two individuals may have different experiences. For example, presented with the same version of the Ebbinghaus illusion (two circles *a* and *b*, where each is surrounded by an array of circles, making *a* and *b* appear larger or smaller than they are), I might see circles *a* and *b* as being equal in size while you see one internal circle as being slightly larger (Schwarzkopf et al., 2010). Or when we are bombarded with photons of 550 nm wavelength, you may see them most often as red while I see them most often as green (Hofer et al., 2005). Such differences will show themselves, e.g., in psychophysical test, where we want to see which differences in a physical stimulus are registered by an individual over a large number of trials.

In How-Tests, we predict such differences in experiences based on differences in the neural makeup of individuals. We predict *phenomenal* inter-individual differences based on underlying *neural* inter-individual differences. Given some NCC-hypothesis *H*, certain differences in an *H*-relevant neural area or feature ought to lead to phenomenal differences.

How can we make an inference from variations in neural features to specific variations in phenomenal features? The presupposition is that there must be some morphism between neural structures and phenomenal structures: There is a mapping from phenomenal domains onto the neural domain (i.e., brain matter and what it does) that preserves the relations that reign in and among phenomenal experiences. Fink et al. (2021) call this the *structural similarity constraint* (see also Clark, 2000; Papineau, 2015; Gert, 2017).²² They

²¹ It might be that it either depends on the other in, e.g., representationalism (phenomenality depending on content) or phenomenalism (content depending on phenomenality). Only then would every How-Test be a What-Test and vice versa. But this is an open issue. As long as the distinction is only *prima facie*

plausible, it motivates differentiating predictions of content from predictions of character.

²² Another isomorphism-presupposition has been brought forward by Palmers (1999, 2003). Palmer argued that if two individuals have the same

argue that all phenomenal structures have a correspondence with neural structures, but not all neural structures have a correspondence in phenomenality.²³ If this holds for all phenomenal relations, then differences in phenomenal relations (e.g., whether a color caused by a photon is closer to this or that color, whether two circles appear to be the same or not) map onto differences in neural relations. Thus, if we know which structures in the brain phenomenal structures map onto—their structural NCCs—we can predict structural differences in experiences from the differences in the neural structures that phenomenal structures correspond to.

What is a neural structure? A structure can be understood as the net of relations in a domain. Here, the domain is defined by neuroscience, i.e., is constituted by the entities that neuroscience focuses on and, more specifically, the relations between these entities as captured with established neuroscientific methods. Examples of neuroscientific entities are neurones, synapses, Brodmann areas, neurotransmitters, spikes, and so on; examples of neural relations are neural connections, spike rhythms, the size of a neural area, increases or decreases in activation, and so on; examples of neuroscientific methods are EEG, fMRI, PET, and so on. However, we should leave this list open as neuroscience is still in development: New entities are still being introduced—like the default mode network, recently introduced by Raichle et al. (2001)—and new methods are under development. Our understanding of neural structures therefore will develop in step with the developments in neuroscience, its theories, and methods. A fortiori, different methods capture different neural structures, sometimes as part of a trade-off. EEG signals, for example, are well-suited to capture the temporal dynamics of neural activation, i.e., the relations between temporally located neural events, but fail to capture fine spatial details. In contrast, CT is much better suited to capture the spatial distribution of neural matter but fails to capture fast changes. Each method, present or future, could capture a structure relevant to the structural similarity constraint. What matters is that the focus is on the relations that these methods reveal in considering which structures account for the fine structure of phenomenal consciousness. The How-Test is therefore open to such developments.

Several studies have employed How-Tests: Genc et al. (2015) predicted specific differences in the individual speed of the traveling wave in binocular rivalry²⁴ based on the individual surface area of a person's V1. Genc et al. (2011) predicted the same from the diverging diffusion properties of the corpus callosum connections between V1 in the right and left hemispheres. Previously, Schwarzkopf et al. (2010) predicted the extent

of a specific configuration of a stimulus for size illusions (Ebbinghaus and Ponzo) based on the individual surface area of a person's V1.

These How-Tests can be easily confused with something that is not a test for an NCC-hypothesis. For example, Haynes and Rees (2005), Miyawaki et al. (2008), and Haynes (2009) made predictions about phenomenality from neural data. However, unlike a How-Test, these predictions were based on a trained pattern classifier, not on hypotheses about which phenomenal structure—e.g., the distribution in the visual field—is systematically related to which neural structures. In a How-Test, however, we need an explicit hypothesis *ante experimentum*. In Genc et al. (2015), the underlying hypothesis is that V1 is the NCC for the distribution in the visual field. So, the smaller V1, the harder it is to experience two different-sized shapes as being different without interference. Thus, we expect a larger Ebbinghaus effect in small cortices. Similarly, the larger a person's V1, the longer it will take a signal from one end to be transmitted to the other. Thus, we expect a longer traveling wave in a larger V1. Such underlying hypotheses *ante experimentum* are missing in studies that employ pattern classifiers, even though they indeed show that *somehow* phenomenal specifics can be predicted from brain data.

In short, the basics of How-Tests are established by comparative psychophysics, where we learn that people sometimes experience the same stimulus differently. It presupposes that there is a morphism between the phenomenal and a part of the neural realm. NCC-hypotheses that pick out neural structures that correspond to phenomenal structures can be How-tested. The goal then is to predict differences in psychophysical performance (indicative of differences in the judged phenomenal experiences) based on measures of relevant neural differences. The credibility of an NCC-hypothesis is lowered if the neural features it picks out can change without any corresponding change in consciousness.

How-Tests avoid most of the shortcomings of other tests. In contrast to Which-Tests, we need not concern ourselves with non-human (or even non-biotic) beings. In contrast to When-Tests, we need not concern ourselves with circumstances where the presence of consciousness is contestable. In contrast to What-Tests, we are not reliant on denying mental latex or accepting specific theories of content-determination or vehicle-location. This, I believe, makes How-Tests the strongest contenders for putting NCC-candidates to the test. (There might, however, be some limits as they focus mainly on differences *in* experience, not the difference between consciousness and unconsciousness, but see Fink and Kob, 2023.)

How-Tests also fulfill the explanatoriness constraint *directly*: It is the neural itself, not the neural *in virtue of being a vehicle for representation*, that allows us to account for phenomenal features.

Additionally, morphisms that allow for predictions often hint at explanations: Why does the traveling wave take longer in larger visual cortices rather than smaller ones? Because it takes longer in a larger visual cortex for an activation associated with, e.g., a house-experience to propagate through to the other side of the visual cortex if the rate of signal propagation is stable across brains and brain areas. This stable propagation rate could be tied to general biological constraints on single neurons and their interactions. Note that such an explanation does not close Levine's explanatory gap: These are not explanations of why this or that neural event is associated with consciousness at all, but merely why this or that neural change leads to this or that phenomenal change. Thereby, How-Tests bracket the explanatory gap because they already focus on non-contentious episodes in consciousness, not the consciousness-unconsciousness-distinction. Instead, How-Test explanations are explanations of why consciousness

structure relating their various experiences (e.g., of color), then the two will behave the same. In the *How*-test, this is given a neural twist: If two individuals have the same structure relating their various experiences (e.g., of color), then they will have the same structural relations in their neural correlates. If they differ the relevant neural structure, we should expect differences in phenomenal structures. But due to these neural differences, they will not only experience differently but also behave differently. However, it is the difference in experience what we predict based on an NCC-hypothesis. This phenomenal difference explains the behavioral differences across a broad range of behavioral tests.

²³ Additionally, phenomenal structures might be multiply realized in the same brain.

²⁴ Roughly: if we projected an image into one eye and simultaneously another image into the other, how long does it take for one to switch to the other in experience.

has this or that feature. Not consciousness itself, but its features are explained bottom-up. The explanatory gap is neither bridged nor touched, but rather ignored (or, if one is so inclined, accepted).

In this section, I argued that How-Tests avoid shortcomings and problems of other tests. If How-Tests are truly the best contenders for arriving at *explanatory correlates* of consciousness, then this has some interesting implications, as I will illustrate in the next section.

5 The How-Test and direct neurophenomenal structuralism

In the last section, I argued that How-Tests are least problematic in comparison to other tests: (i) They do not deal with systems where it is contentious whether they can be conscious or not; (ii) they do not deal with episodes where it is contentious whether a system is conscious during these phases or not; (iii) they do not rely on further hypotheses of content fixing; and (iv) they do not rely on representationalism and allows one to be bracket discussions about mental paint and mental latex, i.e., cases where some character cannot be reduced to content. In the end, How-Tests are also excellent candidates for arriving at *explanatory correlates of consciousness*, in the sense of Seth and Edelman (2009, p. 1440) because they focus on whether an NCC-hypothesis is experimentally testable by accounting for key properties of conscious experience.

How-Tests work. Some of the most trail-blazing experiments in the neuroscience of consciousness already use them. However, if we accept them as adequate tests, they also have some interesting implications, especially concerning (a) metaphysics, (b) the individuation of experience types, and (c) the status of first-person methods. These, together, are suggestive of a position we may call *direct neurophenomenal structuralism* (dNPS). If How-Tests are acceptable, dNPS is a suitable foundation for contemporary consciousness science. Let me first reflect on three implications of the How-Test before sketching dNPS as a foundation for consciousness studies in section 5.4.

5.1 Metaphysics and the How-Test

Note that How-Tests require *systematic* relations between neural and phenomenal features: Specific differences in neural makeup map onto specific differences in a person's experience. This systematicity exceeds the demands required for *supervenience*, sometimes sold as “near-enough physicalism” (Kim, 2005): *A supervenes on B* if any change in *A* requires a change in *B*. *A* is then fully dependent in its dynamics on *B*. No change in *A* without a change in *B*. However, supervenience leaves open whether the change is *systematic*. In principle, supervenience leaves open the possibility that a just noticeable difference (say, a change from an experience as of *red-41* to one as of *red-42*) requires massive changes in brain activation. For supervenience, any change will do—even those that appear unsystematic. Supervenience therefore is silent on the nature of the change in the supervenience base required for a change in the supervening. In How-Tests, however, the change is required to be systematic: Not any change will do. A specific change *here* must come with a specific change *there*. We can motivate this phenomenologically: We can experience smooth changes from one color to the next, which are more likely to be achieved if the underlying neural substrate has to change only marginally, thereby mirroring similarity relations between colors in the similarity between the neural states coding for colors (see esp. Brouwer

and Heeger, 2009). The requirements for How-Tests are therefore stricter than supervenience.

Instead of supervenience, How-Tests are suggestive of *grounding* (Schaffer, 2009; Fine, 2012; Correia and Skiles, 2019)—which mirrors the “accounts for” relation in Seth and Edelman's explanatory correlates. Still, the fact that phenomenal features are grounded in neural features does not necessarily mean that one explains the other (Wilson, 2014), leaving room for explanatory gaps.

5.2 Individuation of phenomenal character

How-Tests need to be able to individuate types of phenomenal character, i.e., what specific kind of experience a subject currently has. In addition, they must do so systematically and via an experience's phenomenal structure. This points to an underlying “phenomenal structuralism”: Relations can be used to individuate phenomenal character. The neural domain also has its own things going on, but it also preserves some features of phenomenality, namely structural features, which Fink et al. (2021) have called the *structural similarity constraint*. How-Tests rely on this idea. This goes beyond a first-order mapping where *features* of one domain can be mapped into *features* of another domain. This has been the old game of reducing “qualia” i.e., the atomic properties of experience (like *redness*), to neural activation.

For a How-Test, we map relations onto relations. While features can be one-place (unary) predicates, relations are necessarily many-place. This allows us to map distances and dimensions in phenomenality onto distances and dimensions in the neural domain. We map structures and relations rather than relata or non-relational properties. Only then can we say that a specific degree of change in a neural domain comes with a specific degree of change in the phenomenal domain, which results in our prediction in a How-Test.

However, this means that we leave “qualia” behind, which were introduced by Lewis (1929) as intrinsic and non-relational properties of the mental and thereby not relations or dimensions. The morphisms required for a How-Test are then much closer to those envisioned by Fink et al. (2021) in their take on *neurophenomenal structuralism*. This view is motivated by the success of structuralism in the sciences more generally, e.g., biology shedding species-intrinsicism for patterns of inheritances (Hull, 1989). Leaving qualia behind may then be no loss, but instead overcoming a superfluous relic of metaphysics, namely consciousness as an assemblage of intrinsic, unary properties.

5.3 The role of first-person methods

Interestingly, How-Tests give first-person methods a decisive role in the neuroscience of consciousness. In general, first-person methods are hard to do without in any inquiry into consciousness, despite criticism of its alleged privileges: An individual token experience—my pain now—is in principle not a phenomenon that is directly accessible in its character by everyone equally. Only I can feel the painfulness of me stubbing my toe, while others can only come to notice it via observing my behavior in combination with some form of “mind reading.” Therefore, we will have to employ first-person methods to some degree in some stage of the neuroscience of consciousness or else go *ignoramus et ignorabimus* (Du Bois-Reymond, 1872). However, to what degree, in what stage and what kind of first-person methods ought to be used is a matter of ongoing debate.

What role can first-person methods play in a natural science of consciousness? At the start, first-person methods can deliver the explananda, what is to be explained, for the neuroscience of consciousness. However, this comes with a version of the meta-problem of consciousness (Chalmers, 2018): Do we need to explain consciousness or, instead, need to explain what people *believe* about consciousness? If we want to avoid eliminativism, first-person methods must be given an explicit place in the process of scientifically investigating consciousness itself, not merely in delivering something to investigate.

Instead of merely motivating an explanandum, philosophers such as Gallagher (2003) have suggested *front-loaded phenomenology*. Here, phenomenological insights steer experimental design. Thereby, phenomenological theories themselves become testable hypotheses as they turn into auxiliary presuppositions used in experimental set-up.²⁵

How-Tests propose a different approach on how to incorporate first-person methods. Note that in a How-Test, we are aiming at the specificities of a single individual's consciousness. These are not targeted by classical Phenomenology—the school that pertains to studying the essences of consciousness (its *Wesenheiten*). Phenomenology never understood itself as targeting individual subjectivity but subjectivity *per se*. It therefore rejects the label of a “first-person method.”²⁶ So How-Tests deviate from Phenomenology: Individual reports and psychophysical performances of single subjects are interpreted as indicating phenomenal changes *in that one person*.

In contrast to Gallagher's proposal, these first-person methods are not front-loaded: They do not steer experimental design. Nor are they, strictly speaking, establishing explananda. Instead, they are used to investigate whether some NCC-hypotheses really pick out explanatory NCCs or not.

In How-Tests, first-person methods are therefore used to *test* a neuroscientific hypothesis: Are all neural events with these features NCCs? Thereby, first-person methods can be seen as integral to every stage of the neuroscience of consciousness: They deliver explananda, they can steer experimental design, they are data for correlation, and they are used to evaluate neuroscientific NCC-hypotheses. One cannot escape first-person methods in this picture.

Notably, this does not solve the problem of how to deal with the unreliability, inaccuracy, insensitivity, and all the other shortcomings of first-person methods. However, luckily, these are largely gradable features. They may thereby be minimized in certain experimental settings, e.g., when we use stimuli above the threshold in rested individuals with no distractors. Exactly, this is the case in the How-Tests of Schwarzkopf et al. (2010), Genc et al. (2011, 2015), and so on.

5.4 Direct neurophenomenal structuralism

How-Tests, understood in this way, hint at a specific foundational position on how phenomenality is grounded in neural activation (compare 5.1): *direct neurophenomenal structuralism* (dNPS). It is based on two basic tenets proposed by Fink et al. (2021). The first concerns relational individuation

(compare 5.2): Types of phenomenal experiences can be individuated by their relations (esp. of graded similarity and difference) to other types of phenomenal experiences, i.e., by their location in a network of intra-phenomenal relations. The experience of a specific shade of red, for example, is what it is because of its graded dissimilarity to any other shade of color experience. The second concerns neuro-phenomenal mapping: There is a systematic mapping of phenomenal structures to a subset of neural structures. In getting to the phenomenal structures that we aim to map to neural structures, we cannot do so without some form of first-person access, however indirect or messy (compare 5.3). Otherwise, we would lack access to one correlatum and therefore could not find a correlation. However, to predict one from the other, phenomenal structures must relate to neural structures in a systematic way, such that the first are grounded in the second. Therefore, such a neuro-phenomenal structural mapping is the foundation on which How-Tests are built.

Note that the relation between phenomenal and neural structures needs to be *direct* to differentiate the How- from the What-Test: We can go directly from neural structure to phenomenal structure. This type of structuralism underlying the How-Test therefore deviates from the forms of structuralism presented by Lyre (2022), Lau et al. (2022), or, in some interpretation, Chalmers (1997). Each subscribes to a systematic mapping of phenomenal structures to neural structures, but *indirectly*, i.e., by a detour via some intermediary. Lyre (2022) suggests perceptual content, Lau et al. (2022) suggest mnemonic content, Chalmers (1997) points out the coherence between phenomenal and cognitive structures. Any reductive strategy built on these views is indirect: To reduce consciousness, one first reduces phenomenality to the intermediary, then reduces the intermediary to the neural.

These forms of *indirect* neurophenomenal structuralism have two major disadvantages. First, to be general, they require each phenomenal experience to inherit the features of the intermediary domain: Each phenomenal experience must have, e.g., content or function. However, why commit to this before all the research is done? Why rule out mental paint or mental latex *a priori*, or instances where a mental state's character is not determined by its function, as these forms of structuralism seem to do? If at all, these should be ruled out *a posteriori*, as such associations between character and cognitive processes are, if at all, contingently true. Second, such *indirect* neurophenomenal structuralists require auxiliary hypotheses to test their theories neuroscientifically: They must answer how character relates to the intermediary domain *and* how the intermediary then relates to neural or behavioral goings-on.

Why take a detour when there is a direct route? In How-Tests, we *directly* predict phenomenal character from the neural structure without some intermediary. So, there is no need for any auxiliary commitments on how other domains (of content, of functions, etc.) relate to the neural. In addition, we need not commit to consciousness necessarily having additional features, such as content or function. But, indeed, in direct neurophenomenal structuralism, it can turn out *a posteriori* that there is no such thing as mental latex or phenomenal experiences without cognitive function. However, there is no need for an *a priori* leap of faith: Contingently, the neural structure N' that a phenomenal structure S maps onto could either be the same or differ from the neural structure N'' that the structure of the cognitive domain maps onto (see Figure 1). So the more prudent and theoretically conservative presupposition would be a *direct* neurophenomenal structuralism,

²⁵ This is thereby a strong deviation from what Husserl imagined phenomenology to be, namely a non-empirical *Wesensschau*.

²⁶ Gallagher's account of front-loaded phenomenology is therefore not really a way to incorporate first-person methods into a science, but of incorporating theorizing about first-person phenomena in phenomenological terms into the science.

Two Varieties of Neurophenomenal Structuralism

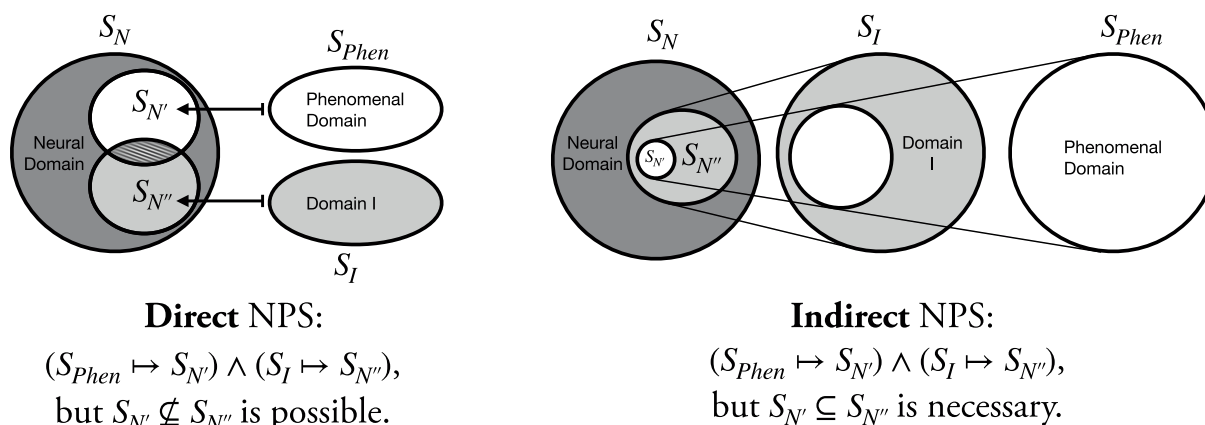


FIGURE 1

One can differentiate *direct* from *indirect* neurophenomenal structuralism (NPS). In direct NPS, a phenomenal structure is mapped directly onto a neural structure. In indirect NPS, a phenomenal structure is first mapped into a domain I (e.g., the domain of mental content, of cognitive functions or states, etc.) and I's structure is subsequently mapped onto a neural structure. Direct and indirect NPS only become indistinguishable if the neural structure onto which the structure of a phenomenal domain is mapped is indeed a subset of the neural structure that I's structure is mapped onto. But, in principle, the two can come apart. Additionally, they make different *a priori* presuppositions. In direct NPS, one can, in principle, (a) deny the existence of I – e.g., there are no representations – or (b) accept the existence of I but hold that the structures of I and phenomenality map into different neural structures, i.e., structures that fail to fully overlap. In contrast, in indirect NPS the existence of I must be accepted and the neural structure that phenomenality's structure is mapped onto must be a subset of the neural structure I's structure is mapped onto. Thereby, direct NPS comes with less theoretical commitments compared to indirect NPS (see also Fink and Kob, 2023).

which could function more broadly as part of a foundation for the neuroscience of consciousness.

Let me summarize: I am strongly in favor of searching for explanatory correlates of consciousness if, as I argued in section 2, the emphasis is on neural correlates that account for phenomenal features and are experimentally testable. Explanation is, in this picture, secondary. In the introduction, I distinguished NCC as data (i.e., sets of token-NCCs) from more general hypotheses about type-NCCs. I presented four sufficiency tests in section 4: Which-, When-, What-, and How-Tests. I argued that How-Tests avoid severe shortcomings of the other three tests. How-Tests rely on the idea that certain changes in the neural domain can account systematically for certain changes in the phenomenal domain. Additionally, it may also deliver correlates that are explanatory—not necessarily of consciousness *per se*, but at least of its specificities. This leaves the classical explanatory gap untouched, but mainly concerning the consciousness-unconsciousness distinction, not concerning the relations between phenomenal characters.

In this last section, I argued that How-Tests, because they are successful, have interesting implications: First, the metaphysical relation between the neural and the phenomenal goes beyond supervenience. Second, if there is a neuroscience of *consciousness* (not of beliefs about consciousness), it needs to incorporate first-person methods at every stage of the scientific process. Third, the morphism needed for How-Tests will concern structures and therefore does not address qualia but instead is more suggestive of some kind of neurophenomenal structuralism. Fourth, such a neurophenomenal structuralism will not be indirect—as commonly suggested—but direct. No need for detours. Future research should then be dedicated to the potential and limits of such a *direct neurophenomenal structuralism*.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

SF: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The author is grateful for the funding in the following two projects: BMBF-project PSYCHEDELSI “Ethische, rechtliche und soziale Implikationen der psychedelischen Renaissance, TP3: Philosophie” (01GP2214C) and DFG/AHRC-project SENSOR “Sensory Engineering: Investigating Altered and Guided Perception and Hallucination” (527947799 & FI 2369/3-1).

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

References

- Aru, J., Bachmann, T., Singer, W., and Melloni, L. (2012). Distilling the neural correlates of consciousness. *Neurosci. Biobehav. Rev.* 36, 737–746. doi: 10.1016/j.neubiorev.2011.12.003
- Aserinsky, E., and Kleitman, N. (1953). Regularly occurring periods of eye motility, and concomitant phenomena, during sleep. *Science* 118, 273–274. doi: 10.1126/science.118.3062.273
- Baars, B. J. (1997). *In the theater of consciousness*. Oxford: Oxford University Press.
- Bachmann, T., Aru, J., and Suzuki, M. (2020). Dendritic integration theory: a thalamocortical theory of state and content of consciousness. *Philos. Mind Sci.* 1. doi: 10.33735/phimisci.2020.II.52
- Block, N. (1995). On a confusion about a function of consciousness. *Behav. Brain Sci.* 18, 227–247. doi: 10.1017/S0140525X00038188
- Block, N. (1996). Mental paint and mental latex. *Philos. Issues* 7, 19–49. doi: 10.2307/1522889
- Block, N. (2005). Two neural correlates of consciousness. *Trends Cogn. Sci.* 9, 46–52. doi: 10.1016/j.tics.2004.12.006
- Braithwaite, V. (2010). *Do fish feel pain?* Oxford: Oxford University Press.
- Brouwer, G. J., and Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *The Journal of Neuroscience* 29, 13992–14003. doi: 10.1523/JNEUROSCI.3577-09.2009
- Burge, T. (1979). Individualism and the mental. *Midwest Stud. Philos.* 4, 73–121. doi: 10.1111/j.1475-4975.1979.tb00374.x
- Chalmers, D. (2018). The meta-problem of consciousness. *J. Conscious. Stud.* 25, 6–61.
- Chalmers, D. (2020). How can we solve the meta-problem of consciousness? *J. Conscious. Stud.* 27, 201–226.
- Chalmers, D. J. (1997). *The conscious mind: in search of a fundamental theory*. Oxford: Oxford University Press: Oxford Paperbacks.
- Chalmers, D. (2003). “Consciousness and its place in nature” in *Blackwell guide to the philosophy of mind*. eds. S. Stich and T. A. Warfield (Oxford: Blackwell), 102–141.
- Chalmers, D. (2000). “What is a neural correlate of consciousness?” in *Neural correlates of consciousness: empirical and conceptual questions*. ed. T. Metzinger (Cambridge, MA: MIT Press), 17–39.
- Clark, A. (2000). *A theory of sentience*. New York: Oxford University Press.
- Clark, A. (2010). Spreading the joy? Why the machinery of consciousness is (probably) still in the head. *Mind* 118, 963–993. doi: 10.1093/mind/fzp110
- Conant, J. B. (1964). “The overthrow of the phlogiston theory” in *Case 2: the overthrow of the phlogiston theory (the chemical revolution of 1775–1789)*. ed. J. B. Conant (Cambridge, US: Harvard University Press).
- Correia, F., and Skiles, A. (2019). Grounding, essence, and identity. *Philos. Phenomenol. Res.* 98, 642–670. doi: 10.1111/phpr.12468
- Craver, C. F. (2014). “The ontic account of scientific explanation” in *Explanation in the special sciences: the case of biology and history*. eds. M. I. Kaiser, O. R. Scholz, D. Plenge and A. Hüttemann (Dordrecht: Springer Netherlands), 27–52.
- Crick, F. (1995). *The astonishing hypothesis: the scientific search for the soul*. New York: Scribner.
- Crick, F., and Koch, C. (1998). Consciousness and neuroscience. *Cereb. Cortex* 8, 97–107. doi: 10.1093/cercor/8.2.97
- Crick, F., and Mitchison, G. (1983). The function of dream sleep. *Nature* 304, 111–114. doi: 10.1038/304111a0
- Du Bois-Reymond, E. (1872). *Über die Grenzen des Naturerkennens*. Leipzig: Veit & Comp.
- Fine, K. (2012). “Guide to ground” in *Metaphysical grounding*. eds. F. Correia and B. Schnieder (Cambridge University Press), 37–80.
- Fink, S. B. (2016). A deeper look at the ‘neural correlate of consciousness’. *Front. Psychol.* 7:1044. doi: 10.3389/fpsyg.2016.01044
- Fink, S. B. (2018). Introspective disputes deflated: the case for phenomenal variation. *Philos Stud.* 175:3165–3194. doi: 10.1007/s11098-017-1000-8
- Fink, S. B., and Kob, L. (2023). “Can structuralist theories be general theories of consciousness?” in *Conscious and unconscious mentality: examining their nature, similarities, and differences*. eds. J. Hvorecký, T. Marvan and M. Polák (Milton Park: Routledge).
- Fink, S. B., Kob, L., and Lyre, H. (2021). A structural constraint on neural correlates of consciousness. *Philos. Mind Sci.* 2, 1–23. doi: 10.33735/phimisci.2021.79
- Fink, S. B., and Lin, Y.-T. (2022). Progress and paradigms in the search for the neural correlates of consciousness: editorial introduction to the special issue the neural correlates of consciousness. *Philos. Mind Sci.* 2, 1–7. doi: 10.33735/phimisci.2021.103
- Flohr, H. (2000). “NMDA receptor-mediated computational processes and phenomenal consciousness” in *Neural correlates of consciousness: empirical and conceptual questions*. ed. T. Metzinger (Cambridge, MA: MIT Press), 245–258.
- Frankish, K. (2016). Illusionism as a theory of consciousness. *J. Conscious. Stud.* 23, 11–39.
- Gallagher, S. (2003). Phenomenology and experimental design. *J. Conscious. Stud.* 10, 85–99.
- Genc, E., Bergmann, J., Singer, W., and Kohler, A. (2015). Surface area of early visual cortex predicts individual speed of traveling waves during binocular rivalry. *Cereb. Cortex* 25, 1499–1508. doi: 10.1093/cercor/bht342
- Genc, E., Bergmann, J., Tong, F., Blake, R., Singer, W., and Kohler, A. (2011). Callosal connections of primary visual cortex predict the spatial spreading of binocular rivalry across the visual hemifields. *Front. Hum. Neurosci.* 5:161. doi: 10.3389/fnhum.2011.00161
- Gert, J. (2017). Quality spaces: mental and physical. *Philos. Psychol.* 30, 525–544. doi: 10.1080/09515089.2017.1295303
- Ghoneim, M. M. (2000). Awareness during anesthesia. *Anesthesiology* 92:597. doi: 10.1097/0000542-200002000-00043
- Goldman, A. I. (1997). Science, publicity, and consciousness. *Philos. Sci.* 64, 525–545. doi: 10.1086/392570
- Grush, R. (2005). Internal models and the construction of time: generalizing from state estimation to trajectory estimation to address temporal features of perception, including temporal illusions. *J. Neural Eng.* 2, S209–S218. doi: 10.1088/1741-2560/2/3/S05
- Grush, R. (2006). How to, and how not to, bridge computational cognitive neuroscience and Husserlian phenomenology of time consciousness. *Synthese* 153, 417–450. doi: 10.1007/s11229-006-9100-6
- Haynes, J.-D. (2009). Decoding visual consciousness from human brain signals. *Trends Cogn. Sci.* 13, 194–202. doi: 10.1016/j.tics.2009.02.004
- Haynes, J.-D., and Rees, G. (2005). Predicting the stream of consciousness from activity in human visual cortex. *Curr. Biol.* 15, 1301–1307. doi: 10.1016/j.cub.2005.06.026
- Hofer, H., Singer, B., and Williams, D. R. (2005). Different sensations from cones with the same photopigment. *J. Vis.* 5, 444–454. doi: 10.1167/5.5.5
- Hohwy, J. (2011). Phenomenal variability and introspective reliability. *Mind Lang.* 26, 261–286. doi: 10.1111/j.1468-0017.2011.01418.x
- Hohwy, J., and Seth, A. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philos. Mind Sci.* 1, 1–35. doi: 10.33735/phimisci.2020.II.64
- Horikawa, T., Tamaki, M., Miyawaki, Y., and Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science* 340, 639–642. doi: 10.1126/science.1234330
- Hull, D. (1989). *The metaphysics of evolution*. Albany: State University of New York Press.
- Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika* 30, 81–93. doi: 10.1093/biomet/30.1-2.81
- Kim, J. (1979). Causality, identity and supervenience in the mind-body problem. *Midwest Stud. Philos.* 4, 31–49. doi: 10.1111/j.1475-4975.1979.tb00372.x
- Kim, J. (2005). “Physicalism, or something near enough” in *Physicalism, or something near enough* (Princeton University Press), 149–174. Available at: <http://www.jstor.org/stable/j.ctt7snrs.10>
- Koch, C., and Tsuchiya, N. (2007). Attention and consciousness: two distinct brain processes. *Trends Cogn. Sci.* 11, 16–22. doi: 10.1016/j.tics.2006.10.012
- Kruskal, W. H. (1958). Ordinal measures of association. *J. Am. Stat. Assoc.* 53, 814–861. doi: 10.1080/01621459.1958.10501481
- Lamme, V. A. F. (2004). Separate neural definitions of visual consciousness and visual attention: a case for phenomenal awareness. *Neural Netw.* 17, 861–872. doi: 10.1016/j.neunet.2004.02.005
- Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends Cogn. Sci.* 10, 494–501. doi: 10.1016/j.tics.2006.09.001

- Lau, H., Michel, M., LeDoux, J. E., and Fleming, S. M. (2022). The mnemonic basis of subjective experience. *Nat. Rev. Psychol.* 1, 479–488. doi: 10.1038/s44159-022-00068-6
- Leibniz, G. W. (1720). *Lehrsätze über Die Monadologie, Ingleichen von Gott Und Seiner Existenz, Seinen Eigenschaften Und von Der Seele Des Menschen Etc. Wie Auch Dessen Letzte Vertheidigung Seines Systematis Harmoniae Praestabilitae Wider Die Einwürffe Des Herrn Bayle*. Meyers sel: Witwe Buchhandlung.
- Levine, J. (1983). Materialism and qualia: the explanatory gap. *Pac. Philos. Q.* 64, 354–361. doi: 10.1111/j.1468-0114.1983.tb00207.x
- Lewis, C. I. (1929). *Mind and the world order: outline of a theory of knowledge*. New York, NY: Charles Scribener's Sons.
- Lloyd, D. (2002). Functional MRI and the study of human consciousness. *J. Cogn. Neurosci.* 14, 818–831. doi: 10.1162/089892902760191027
- Lyre, H. (2022). Neurophenomenal structuralism. A philosophical agenda for a structuralist neuroscience of consciousness. *Neurosci. Conscious.* 2022:nia012. doi: 10.1093/nc/nia012
- Malcolm, N. (1959). *Dreaming*. London: Routledge & Kegan Paul.
- Marvan, T., Polák, M., Bachmann, T., and Phillips, W. A. (2021). Apical amplification—a cellular mechanism of conscious perception? *Neurosci. Conscious.* 2021:niab036. doi: 10.1093/nc/niab036
- Mcbride, R. (1999). Consciousness and the state/transitive/creature distinction. *Philos. Psychol.* 12, 181–196. doi: 10.1080/095150899105864
- Melloni, L., Cogitate Consortium Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., et al. (2023). *An adversarial collaboration to critically evaluate theories of consciousness*, Preprint (Version 1). Available at Research Square.
- Michel, M. (2019). Fish and microchips: on fish pain and multiple realization. *Philos. Stud.* 176, 2411–2428. doi: 10.1007/s11098-018-1133-4
- Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M., Morito, Y., Tanabe, H. C., et al. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron* 60, 915–929. doi: 10.1016/j.neuron.2008.11.004
- Nanay, B. (2005). Is twofoldness necessary for representational seeing? *Br. J. Aesthet.* 45, 248–257. doi: 10.1093/aesthj/ayi034
- Nir, Y., and Tononi, G. (2010). Dreaming and the brain: from phenomenology to neurophysiology. *Trends Cogn. Sci.* 14, 88–100. doi: 10.1016/j.tics.2009.12.001
- Overgaard, M., and Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philos. Mind Sci.* 2, 1–16. doi: 10.33735/phimisci.2021.37
- Palmers, S. E. (1999). *Vision science: photons to phenomenology*. Cambridge, MA: MIT Press.
- Palmers, S. E. (2003). “Consciousness and isomorphism” in *Essential sources in the scientific study of consciousness*. eds. B. J. Baars, J. B. Newman and W. P. Banks (Cambridge, MA: MIT Press), 186–200.
- Papineau, D. (1993). Physicalism, consciousness and the antipathetic fallacy. *Austr. J. Philos.* 71, 169–183. doi: 10.1080/00048409312345182
- Papineau, D. (2015). “Can We Really See a Million Colours?” in *Paul Coates, and Sam Coleman (eds), Phenomenal Qualities: Sense, Perception, and Consciousness* (Oxford University Press, 2015). doi: 10.1093/acprof:oso/9780198712718.003.0010
- Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proc. R. Soc. Lond.* 58, 240–242. doi: 10.1098/rspl.1895.0041
- Pearson, K. (1897). Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc. R. Soc. Lond.* 60, 489–498. doi: 10.1098/rspl.1896.0076
- Place, U. T. (1956). Is consciousness a brain process? *Br. J. Psychol.* 47, 44–50. doi: 10.1111/j.2044-8295.1956.tb00560.x
- Polák, M., and Marvan, T. (2018). Neural correlates of consciousness meet the theory of identity. *Front. Psychol.* 9:1269. doi: 10.3389/fpsyg.2018.01269
- Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proc. Natl. Acad. Sci. U. S. A.* 98, 676–682. doi: 10.1073/pnas.98.2.676
- Rodgers, J. L., and Alan Nicewander, W. (1988). Thirteen ways to look at the correlation coefficient. *Am. Stat.* 42, 59–66. doi: 10.2307/2685263
- Rosenthal, D. M. (1986). Two concepts of consciousness. *Philos. Stud.* 49, 329–359. doi: 10.1007/BF00355521
- Schaffer, J. (2009). “On what grounds what” in *Metametaphysics: new essays on the foundations of ontology*. eds. D. Manley, D. J. Chalmers and R. Wasserman (Oxford: Oxford University Press), 347–383.
- Schlicht, T., and Dolega, K. (2021). You can't always get what you want: predictive processing and consciousness. *Philos. Mind Sci.* 2:8. doi: 10.33735/phimisci.2021.11.80
- Schwarzkopf, D. S., Song, C., and Rees, G. (2010). The surface area of human V1 predicts the subjective experience of object size. *Nat. Neurosci.* 14, 28–30. doi: 10.1038/nn.2706
- Seth, A. (2009). Explanatory correlates of consciousness: theoretical and computational challenges. *Cogn. Comput.* 1, 50–63. doi: 10.1007/s12559-009-9007-x
- Seth, A. K., and Edelman, G. M. (2009). “Consciousness and complexity” in *Encyclopedia of complexity and systems science*. ed. R. A. Meyers (New York: Springer), 1424–1443.
- Silberstein, M. (2001). Converging on emergence. Consciousness, causation and explanation. *J. Conscious. Stud.* 8, 61–98.
- Spearman, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.* 15, 72–101. doi: 10.2307/1412159
- Spinoza, B. (1677). *Ethica, ordine geometrico demonstrata*. Amsterdam: Jan Rieuwertsz.
- Steward, H. (1997). *The ontology of mind: events, processes, and states*. Oxford: Clarendon Press.
- Stigler, S. M. (1989). Francis Galton's account of the invention of correlation. *Stat. Sci.* 4, 73–79. doi: 10.1214/ss/1177012580
- Suzuki, H., Uchiyama, M., Tagaya, H., Ozaki, A., Kuriyama, K., Aritake, S., et al. (2004). Dreaming during non-rapid eye movement sleep in the absence of prior rapid eye movement sleep. *Sleep* 27, 1486–1490. doi: 10.1093/sleep/27.8.1486
- Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Stat.* 35, 2769–2794. doi: 10.1214/009053607000000505
- Topulos, G. P., Lansing, R. W., and Banzett, R. B. (1993). The experience of complete neuromuscular blockade in awake humans. *J. Clin. Anesthesiol.* 5, 369–374. doi: 10.1016/0952-8180(93)90099-Z
- van Fraassen, B. C. (1980). *The scientific image*. Oxford: Oxford University Press.
- Varela, F. J. (1999). Present-time consciousness. *J. Conscious. Stud.* 6, 111–140.
- Ward, J. (1911). “Psychology” in *Encyclopedia Britannica*. ed. 11th ed, vol. XXII (Cambridge: Cambridge University Press), 547–604.
- Wiese, W., and Friston, K. J. (2021). The neural correlates of consciousness under the free energy principle: from computational correlates to computational explanation. *Philos. Mind Sci.* 2:9. doi: 10.33735/phimisci.2021.81
- Wilson, J. M. (2014). No work for a theory of grounding. *Inquiry* 57, 535–579. doi: 10.1080/0020174x.2014.907542
- Wollheim, R. (1987). *Painting as an art*. Princeton: Princeton University Press.
- Young, B. (2012). Stinking consciousness! *J. Conscious. Stud.* 19, 223–243.



OPEN ACCESS

EDITED BY

Xerxes D. Arsiwalla,
Wolfram Research, Inc., United States

REVIEWED BY

Ken Mogi,
Sony Computer Science Laboratories, Japan

*CORRESPONDENCE

Robert Prentner
✉ robert.prentner@amcs.science

RECEIVED 08 May 2024

ACCEPTED 19 June 2024

PUBLISHED 15 July 2024

CITATION

Prentner R and Hoffman DD (2024) Interfacing
consciousness. *Front. Psychol.* 15:1429376.
doi: 10.3389/fpsyg.2024.1429376

COPYRIGHT

© 2024 Prentner and Hoffman. This is an
open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Interfacing consciousness

Robert Prentner^{1,2*} and Donald D. Hoffman³

¹Institute of Humanities, ShanghaiTech University, Shanghai, China, ²Association for Mathematical
Consciousness Science, Munich, Germany, ³Department of Cognitive Sciences, University of
California, Irvine, Irvine, CA, United States

The current stage of consciousness science has reached an impasse. We blame the physicalist worldview for this and propose a new perspective to make progress on the problems of consciousness. Our perspective is rooted in the theory of conscious agents. We thereby stress the fundamentality of consciousness outside of spacetime, the importance of agency, and the mathematical character of the theory. For conscious agent theory (CAT) to achieve the status of a robust scientific framework, it needs to be integrated with a good explanation of perception and cognition. We argue that this role is played by the interface theory of perception (ITP), an evolutionary-based model of perception that has been previously formulated and defended by the authors. We are specifically interested in what this tells us about the possibility of AI consciousness and conclude with a somewhat counter-intuitive proposal: we live inside a simulation instantiated, not digitally, but in consciousness. Such a simulation is just an interface representation of the dynamics of conscious agents for a conscious agent. This paves the way for employing AI in consciousness science through customizing our interface.

KEYWORDS

conscious agent theory (CAT), interface theory of perception (ITP), conscious realism, AI consciousness, agency, computation, spacetime, simulation hypothesis

1 The current impasse in the science of consciousness

There is a large consensus in the scientific community, according to which consciousness is somehow a product of information processing in the brain. There exist many different theories in the field (Signorelli et al., 2021), which have produced impressive new insights, such as discovering a range of candidates for the neural correlates of consciousness.¹ However, these theories fail to also explain these correlations: why do they exist in the first place? To a physicist working on high-energy particle physics, it would surely seem very disappointing if the standard model were simply a list of correlations, say, between particle motions and detector values. Even if we were able to “furnish systematic correlations” (Seth and Bayne, 2022), this wouldn’t provide much relief.

Yet, we believe there is a need to abandon the consensus view. We need new theories that actually do have the potential to explain, not just list or predict, these correlates. One such theory is the conscious agent theory (CAT; Hoffman and Prakash, 2014; Fields et al., 2018; Hoffman et al., 2023). CAT is presented as a theory of consciousness on its own terms, not a theory of consciousness as it arises from physical processes in the brain or elsewhere. One could forget everything one knew about physics, and still engage in CAT. But it would be a mistake to conclude from this that CAT is not a mathematically precise theory. On the contrary, it starts with a minimal but rigorously defined set of assumptions (Hoffman and Prakash, 2014):

¹ Although the full story is not quite as straightforward, cf. Signorelli et al. (2021); Lepauvre and Melloni (2021).

1. Consciousness exists. We represent this by a (possibly infinite) set X of experiences. In Hoffman et al. (2023), this set was interpreted as an agent's potential to have experiences.
2. An agent could have this experience (e.g., seeing red), rather than that one (seeing green). The mathematical way to represent this is to say that the set of conscious experiences is measurable² enabling us to state a probability to undergo any specific experience. An agent not only has the potential for conscious experiences but there are specific experiences that it undergoes at any given moment.

Unlike many other theories of consciousness, CAT takes agency as a fundamental ingredient. Only agents are conscious, and it is via their actions that they affect the world. Whereas experience reflects the private, first-personal aspect of consciousness, action consequences amount to its publicly observable, third-personal aspect. In CAT, this is formalized via conditional probabilities:

3. Consciousness makes a difference to the agent. There is a conditional probability that expresses how likely it is for a conscious agent to act in a certain way, given that it undergoes a specific prior experience.³ In CAT, this is called the “decision” of an agent. Consciousness also makes a difference to the world and its future perception by an agent. What is true for decisions, is also true for the execution of actions.

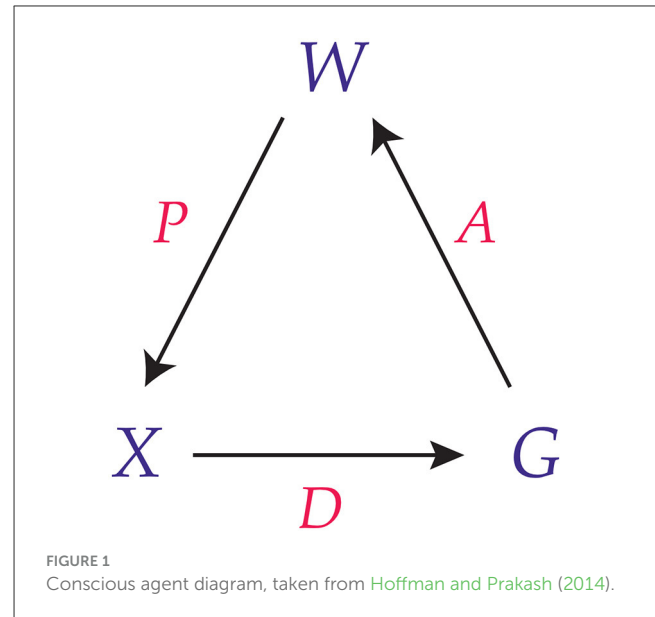
Together, this results in a tripartite structure that is shown in Figure 1. Other theories in consciousness studies use conditional probabilities as well, chiefly among them the “Integrated Information Theory of Consciousness” (Tononi et al., 2016; Albantakis et al., 2023). However, the difference is that integrated information theorists use conditional probabilities to specify a *physical substrate* of consciousness, whereas in CAT conditional probabilities are used to specify the dynamics of consciousness itself. Conditional probabilities have also long been suspected to play a crucial role in the computational approach to perception, e.g. according to a Bayesian model (Knill and Richards, 1996; Hoffman et al., 2015). Increasingly, this perspective gets adopted in predictive processing theories of consciousness too (Seth and Bayne, 2022). However, in CAT these are typically not seen as uncertainties about the perception of a physical world but as probabilistic elements inherent to consciousness.

It seems suggestive now to build networks of conscious agents that could account for many (or all) of the processes described by cognitive science (Fields et al., 2018). The idea here is not that consciousness is one more process built on top of many other supporting processes (such as learning, memory, representation, decision, etc.), but that consciousness provides the basis from which these processes emerge in the first place. More speculatively even, it has been proposed that physics itself arises from the combination and fusion of conscious agents (Hoffman et al., 2023).

Abbreviations: CAT, Conscious agent theory; ITP, Interface theory of perception.

² Technically, we need to endow the set X with a sigma-algebra \mathcal{X} which defines possible “events” based on the underlying set. This sigma-algebra could be interpreted as our cognitive representation (Hoffman et al., 2023).

³ More formally, this is done by defining a Markovian kernel over the set of experiences and actions of an agent.



We further believe that CAT has the resources to integrate a range of subjects from physics to AI. AI consciousness only starts to make sense once we abandon a physicalist worldview.

2 The interface theory of perception

On its own, the theory of conscious agents seems to be somewhat removed from the empirical day-to-day research in the scientific studies of cognitive (neuro)science. But to the avail of CAT, a recent proposal has been defended in the literature that provides an account of the formative processes underlying perception. The so-called interface theory of perception (ITP; Hoffman et al., 2015; Prentner, 2021) is deeply rooted in evolutionary theory and thus lies within the bounds of conventional scientific discourse. According to ITP, the things that we perceive (both objects and structures) arise as solutions to the problem of representing the world in a way that allows an agent to choose actions that increase its fitness. Fitness payoffs are agent-dependent values mapped from a domain that includes world states, the classes and states of agents, and their available action classes. Hence the relevant payoff functions are generically not homomorphic to the structure of the agent-independent world “out there” (Prakash et al., 2020). If I see an apple in front of me, what is the probability that there really is an apple in front of me, irrespective of the way I observe it? Almost certainly zero. If I see symmetries in the world, what is the probability that there really are symmetries in the world, irrespective of the way I can act on the objects I perceive? Almost certainly zero. If I perceive any structure at all, what is the probability that there really are those structures, irrespective of the way observers exist in the world? Almost certainly zero.

Our perceptions, do not mirror the world in any deeper sense apart from their consequences for fitness (Prakash et al., 2021). Rather than giving us an insight into the nature of reality, perception can be compared to a desktop interface. It allows an agent to successfully interact with its world, very much like

dragging and dropping icons on a computer desktop allows us to move and delete files in the computer. This might sound similar to theories of the embodied mind (Chemero, 2009), sensori-motor contingency (O'Regan and Noë, 2001), or active inference (Clark, 2017; Parr et al., 2022). But other than those theories, ITP goes one step further and seeks to undermine our belief in physical objects that serve as embodiments, as substrates of sensory and motor processes, or as basis for inference.

Still, ITP leaves open an important question: if perception is an interface, what does it interface with?

3 Conscious realism

According to conscious realism, the whole universe can be represented as a network of conscious agents (Hoffman, 2008; Hoffman and Prakash, 2014). Hence, interfaces are used by conscious agents to represent networks of conscious agents — consciousness self-reflectively represents itself via interfaces. Thereby, agency is a fundamental concept. Many things can be said to exist in the universe. Among them are physical events in spacetime and subjective experiences. We propose that space and time can be derived from the network of conscious agents in terms of a representation by which agents, in order to act, make sense of the hyper-dimensional dynamics of consciousness.

But also our subjective experiences, such as our experience of the arrow of time can be recovered from an (unchanging) network of conscious agents. We typically think of our actions in terms of sequences of events in time. But time, in the theory of conscious agents, is a mere artifact of projection (Hoffman et al., 2023; Hoffman, 2024). One might note at this point that the intent of CAT is to re-conceptualize our view of the world and to serve as “theory of theories” that non-reductively links various areas of the natural world such as those studied by fundamental physics, evolution by natural selection, or cognitive science. Such a re-conceptualization is not only needed to explain the neural correlates of consciousness, namely as necessary correlations between a network of conscious agents and its (interface) representation, but to make sense of reality more generally.

Since CAT does not start by stipulating, from the outset, many of the typical features of our subjective experience such as selfhood or the experience of an arrow of time, it seems prudent to call the kind of minimal consciousness invoked by CAT a *non-dual*⁴ variety of consciousness. Indeed, as we saw in the basic definition of CAT reviewed in the first section, all that CAT is premised on is the idea that we have (a potentially infinite number of) experiences that can be individuated probabilistically and evolve in terms of conditional probabilities—if an agent were to experience x now, it will, with some positive probability, experience y later.⁵ At this stage, nothing yet has been said about the subjective/objective dichotomy, the objective structure of the world, or any quasi-axiomatization of subjective experience. By contrast, CAT is a relational theory from which one could recover different interface representations of

the subjective experience of the agent in question (ideally with mathematical precision). But the experiences of many agents might be utterly unlike our own subjective experience.

4 Interfaces to consciousness

4.1 Spacetime

It is very unlikely that our species-specific interface bears any similarity with whatever lies underneath it. If the interface theory is right also on a fundamental level, the probability that this deeper reality is spatiotemporal in nature is close to zero. Although exotic at first sight, such a view seems to align well with recent findings from fundamental physics—at least if one lets go of the assumption that our classical (perceptual) model of reality is somehow approximating ground truth. Many physicists now believe that spacetime is not a fundamental entity. This is independent of the particular approaches endorsed by researchers such as Smolin (2001), Rovelli (2004), Gross (2005), or Arkani-Hamed (2010). Of course, it is still an open question what would replace spacetime, but all approaches agree that spacetime has to go eventually (see also Musser, 2017). Hoffman et al. (2023) advised to heed those physicists and link spacetime to the asymptotic dynamics of conscious agents, as it can be classified via the notion of a “decorated permutation.” Still, this is very counter-intuitive. After all, it certainly looks as if space and time are fundamentally real. But looks can be deceiving. And this is exactly what ITP tells us. Moreover, one might worry that the fact that we can do science of any kind presupposes space and time. But while the interface theory seems to imply that we should not take space and time as being there when no one looks, it still cautions us to take them seriously. And this dissolves the worry. ITP invites us to think of space and time as real for most practical purposes, but not simpliciter.

4.2 Agency and life

Consciousness is deeply linked to agency. Hence, one would perhaps expect to see the first glimpses of consciousness in living beings, which are — according to our present state of knowledge — the first instances of embodied agents that we can observe in the world. Yet, this merely reflects our ignorance of the fact that also the world underneath organisms might be rich in agency (a claim suggested by some interpretations of quantum mechanics such as QBism; von Bayer, 2016). Prebiotic agency normally stays invisible to us. But this could be a mere artifact of our (limited) interface. According to Nagel (1974): “if one travels too far down the phylogenetic tree, people gradually shed their faith that there is experience there at all.” We do not see any logical reason why this should stop at the living. However, what is different at the level of non-living beings is that it becomes harder to ascribe true agency there. It is in living beings that consciousness appears to us for the first time. But it appears in the form of *embodied* agents, not agency itself. Sometimes these embodiments give us more insights into consciousness (in the case of living beings), sometimes less (in the case of dead matter). Again, taking agency to be an exclusive

⁴ = non-objective but also non-subjective.

⁵ Technically, as stated in Hoffman et al. (2023), this amounts to a “qualia-kernel” that would integrate over all possible actions and external states of the network.

property of living beings might be valid for most practical purposes, but not simpliciter.

4.3 Computation

In the theory of conscious agents, “computation” is not merely a concept that could be usefully employed to describe a certain empirical matter (as, for example, when we say “the brain computes”). It is inherent to the theory itself. At the moment, it is still unclear what non-computable functions a conscious agent network could implement. Yet, it is relatively straightforward to show that networks of conscious agents are computationally universal (Hoffman and Prakash, 2014), i.e., they could simulate other architectures known to be computationally universal (such as certain cellular automata or Turing machines). This fact was also exploited by Fields et al. (2018), who aimed to show how networks of conscious agents could implement various cognitive (read: computational) mechanisms. Given a purely formal definition of information (Cover and Thomas, 2006), CAT defines information processing in terms of (conditional) probabilities. In addition, conscious realism proposes that physics can be recovered from networks of conscious agents as an interface representation. Together, these claims would indicate that, contra Rolf Landauer’s mantra “information is physical” (Landauer, 1999), the dynamics of consciousness fully accounts for a substantial notion of information processing. “Computation” would be one of many ways to describe the dynamics of consciousness as it appears on the interface of perception. Our claim is then that physics is information that comes from consciousness: IT from BIT from CIT.⁶

5 Consciousness and AI

5.1 A new paradigm

The question of whether artificial intelligence can become conscious currently gets much attention from scholars and media (Chalmers, 2022; Association for Mathematical Consciousness Science, 2023). According to the consensus view mentioned at the beginning of this article, one should expect computers to become conscious as soon as they implement the right computations (for example, mimicking the processes happening in our brains, Butlin et al., 2023). Yet, if consciousness is fundamental, it is inscrutable how computation could give rise to it. This appears to put us in a position that denies the possibility of AI consciousness. But this overlooks a crucial idea, which has to do with ITP. Accordingly, “computation” is just the name for an interface representation of the dynamics of consciousness. An interface hides and simplifies what lies beyond it. Yet, with the power of AI, we can custom-tailor our interface. Put differently, while we do not create consciousness in the process, we can use technology to help us get new insights into the (pre-existing) realm of conscious agents, similar to how we could use AI to get new insights in physics (Krenn et al., 2022). Yet, consciously experiencing these insights (including understanding them) is something that we need to do.

⁶ For the Indian doctrine of citta-mātra see Westerhoff (2018).

5.2 A simulation in consciousness

In his now-famous simulation argument, Bostrom (2003) proposed the following argument to show that we are “almost certainly living in a computer simulation”:

1. In the future, enormous computational resources will be available to a post-human society. One thing that members of this society will do is run computer simulations about their ancestors (i.e., us),
2. If you run the right computations, then the programs instantiating those computations will be conscious,
3. It is then (statistically) prudent to assume that we are just among those simulated beings, rather than being part of the original race that conceived the simulation.

Much has been written about the simulation argument. In particular, the claim of computationalism about consciousness strikes many as wrong, who are immersed in the scientific study of consciousness (Hoffman, 2019; Seth, 2021). A physicalist objection is that this wrongly assumes a strong notion of “substrate independence” (Prentner, 2017), the claim that the computations underlying consciousness can be instantiated in all kinds of substrates—no matter whether they are biological or artificial. But the objection can be easily countered by noting that advanced simulations will be fine-grained enough to simulate any physical system, and consciousness could then just run on such a “virtual machine.” By contrast, conscious realism accepts a variety of the simulation argument but with an important caveat: the simulation we are in is a simulation instantiated in consciousness! After all, consciousness—unlike a physical or biological system—is not a substrate that could itself be simulated. The reasons why the simulation argument (as stated by Bostrom and followers) is incorrect is not because it is not sufficiently physicalist, but because it is not sufficiently idealist.

6 Discussion

Conscious realism is the claim that the universe consists entirely of conscious agents. ITP says that we interact with this reality not directly but through a perceptual interface. These claims provide us with a new agenda for consciousness science in the future, resolving some challenges, but opening up others. Those challenges pertain to the nature of spacetime (it is not fundamental), agency (it is not limited to biological systems), and computation (it is not physical). Instead, CAT ultimately re-conceives these concepts as arising from the dynamics of conscious agents as we see them through an interface. In this light, to say that we live inside a simulation means that the simulation is what conscious agents are doing, as another conscious agent would perceive it. This paves the way for employing AI in consciousness science through customizing our perceptual interface.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

RP: Conceptualization, Writing – original draft, Writing – review & editing. DH: Conceptualization, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We thank ShanghaiTech University and the University of California, Irvine, for institutional support.

References

- Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., et al. (2023). Integrated information theory (IIT) 4.0: formulating the properties of phenomenal existence in physical terms. *PLoS Comp. Bio.* 19:e1011465. doi: 10.1371/journal.pcbi.1011465
- Arkani-Hamed, N. (2010). The future of fundamental physics. *Daedalus* 141, 53–66. doi: 10.1162/DAED_a_00161
- Association for Mathematical Consciousness Science (2023). *The Responsible Development of AI Agenda Needs to Include Consciousness Research*. Available online at: <https://amcs-community.org/open-letters/> (accessed July 5, 2024).
- Bostrom, N. (2003). Are you living in a computer simulation? *Philosop. Quat.* 53, 243–255. doi: 10.1111/1467-9213.00309
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., et al. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.
- Chalmers, D. J. (2022). Could a large language model be conscious? *arXiv preprint arXiv:2303.07103*.
- Chemero, A. (2009). *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/8367.001.0001
- Clark, A. (2017). “How to knit your own markov blanket,” in *Philosophy and Predictive Processing*, eds. T. Metzinger, and W. Wiese (Frankfurt am Main: MIND Group).
- Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory*. Hoboken, NJ: Wiley.
- Fields, C., Hoffman, D. D., Prakash, C., and Singh, M. (2018). Conscious agent networks: formal analysis and application to cognition. *Cogn. Syst. Res.* 47, 186–213. doi: 10.1016/j.cogsys.2017.10.003
- Gross, D. (2005). Einstein and the search for unification. *Curr. Sci.* 89, 2035–2040. doi: 10.1142/978981272718_0001
- Hoffman, D. D. (2008). Conscious realism and the mind-body problem. *Mind Matter* 6, 87–121.
- Hoffman, D. D. (2019). *The Case Against Reality. How Evolution Hid the Truth from our Eyes*. New York: W.W. Norton.
- Hoffman, D. D. (2024). Spacetime is doomed: time is an artifact. *Timing Time Percept.* 12, 189–191. doi: 10.1163/22134468-bja10096
- Hoffman, D. D., and Prakash, C. (2014). Objects of consciousness. *Front. Psychol.* 5:577. doi: 10.3389/fpsyg.2014.00577
- Hoffman, D. D., Prakash, C., and Prentner, R. (2023). Fusions of consciousness. *Entropy* 25:129. doi: 10.3390/e25010129
- Hoffman, D. D., Singh, M., and Prakash, C. (2015). The interface theory of perception. *Psychon. Bull. Rev.* 22, 1480–1506. doi: 10.3758/s13423-015-0890-8
- Knill, D. C. and Richards, W. (1996). *Perception as Bayesian Inference*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511984037
- Krenn, M., Pollice, R., Guo, S. Y., Aldeghi, M., Cervera-Lierta, A., Friederich, P., et al. (2022). On scientific understanding with artificial intelligence. *Nat. Rev. Phys.* 4, 761–769. doi: 10.1038/s42254-022-00518-3
- Landauer, R. (1999). Information is a physical entity. *Physica A* 263, 63–67. doi: 10.1016/S0378-4371(98)00513-5
- Lepauvre, A., and Melloni, L. (2021). The search for the neural correlate of consciousness: progress and challenges. *Philos. Mind Sci.* 2:87. doi: 10.33735/phimisci.2021.87
- Musser, G. (2017). “Spacetime is doomed,” in *Space, Time and the Limits of Human Understanding*, eds. S. Wuppuluri, and G. C. Ghirardi (Cham: Springer), 217–227. doi: 10.1007/978-3-319-44418-5_17
- Nagel, T. (1974). What is it like to be a bat? *Philos. Rev.* 83, 435–450. doi: 10.2307/2183914
- O'Regan, J. K., and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behav. Brain Sci.* 24, 939–973. doi: 10.1017/S0140525X0100115
- Parr, T., Pezzulo, G., and Friston, K. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/12441.001.0001
- Prakash, C., Fields, C., Hoffman, D. D., Prentner, R., and Singh, M. (2020). Fact, fiction, and fitness. *Entropy* 22:514. doi: 10.3390/e2205014
- Prakash, C., Stephens, K. D., Hoffman, D. D., Singh, M., and Fields, C. (2021). Fitness beats truth in the evolution of perception. *Acta Biotheor.* 69, 319–341. doi: 10.1007/s10441-020-09400-0
- Prentner, R. (2017). Consciousness: a molecular perspective. *Philosophies* 2, 26–32. doi: 10.3390/philosophies2040026
- Prentner, R. (2021). Dr Goff, Tear Down This Wall! The interface theory of perception and the science of consciousness. *J. Conscious. Stud.* 28, 91–103. doi: 10.53765/20512201.28.9.091
- Rovelli, C. (2004). *Quantum Gravity*. New York: Cambridge University Press. doi: 10.1017/CBO9780511755804
- Seth, A. K. (2021). *Being You. A New Science of Consciousness*. London: Dutton Books.
- Seth, A. K., and Bayne, T. (2022). Theories of consciousness. *Nat. Rev. Neurosci.* 23, 439–452. doi: 10.1038/s41583-022-00587-4
- Signorelli, C. M., Szczotka, J., and Prentner, R. (2021). Explanatory profiles of models of consciousness - towards a systematic classification. *Neurosci. Conscious.* 2021:niab021. doi: 10.1093/nc/niab021
- Smolin, L. (2001). *Three Roads to Quantum Gravity*. New York: Basic Books.
- Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nat. Rev. Neurosci.* 17, 450–461. doi: 10.1038/nrn.2016.44
- von Bayer, H. C. (2016). *QBism. The Future of Quantum Physics*. Cambridge, MA: Harvard University Press. doi: 10.4159/9780674545342
- Westerhoff, J. C. (2018). *The Golden Age of Indian Buddhist Philosophy*. Oxford: Oxford University Press. doi: 10.1093/oso/9780198732662.001.0001

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



OPEN ACCESS

EDITED BY

Luca Simone,
UNINT—Università Degli Studi Internazionali
di Roma, Italy

REVIEWED BY

April Hargreaves,
National College of Ireland, Ireland
Massimo Tusconi,
University of Cagliari, Italy

*CORRESPONDENCE

Aleš Oblak
✉ ales.oblak@psih-klinika.si

RECEIVED 07 February 2024

ACCEPTED 18 July 2024

PUBLISHED 06 August 2024

CITATION

Oblak A, Kuclar M, Horvat Golob K,
Holnhaner A, Battelino U, Skodlar B and
Bon J (2024) Crisis of objectivity: using a
personalized network model to understand
maladaptive sensemaking in a patient with
psychotic, affective, and
obsessive-compulsive symptoms.
Front. Psychol. 15:1383717.
doi: 10.3389/fpsyg.2024.1383717

COPYRIGHT

© 2024 Oblak, Kuclar, Horvat Golob,
Holnhaner, Battelino, Skodlar and Bon. This
is an open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other forums
is permitted, provided the original author(s)
and the copyright owner(s) are credited and
that the original publication in this journal is
cited, in accordance with accepted academic
practice. No use, distribution or reproduction
is permitted which does not comply with
these terms.

Crisis of objectivity: using a personalized network model to understand maladaptive sensemaking in a patient with psychotic, affective, and obsessive-compulsive symptoms

Aleš Oblak^{1*}, Matic Kuclar², Katja Horvat Golob¹,
Alina Holnhaner¹, Urška Battelino³, Borut Škodlar^{1,2} and
Jurij Bon^{1,2}

¹Laboratory for Cognitive Neuroscience and Psychopathology, University Psychiatric Clinic Ljubljana, Ljubljana, Slovenia, ²Department of Psychiatry, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia, ³Faculty of Slovenian and International Studies, New University, Ljubljana, Slovenia

Introduction: Psychiatric comorbidities have proven a consistent challenge. Recent approaches emphasize the need to move away from categorical descriptions of symptom clusters towards a dimensional view of mental disorders. From the perspective of phenomenological psychopathology, this shift is not enough, as a more detailed understanding of patients' lived experience is necessary as well. One phenomenology-informed approach suggests that we can better understand the nature of psychiatric disorders through personalized network models, a comprehensive description of a person's lifeworld in the form of salient nodes and the relationships between them. We present a detailed case study of a patient with multiple comorbidities, maladaptive coping mechanisms, and adverse childhood experiences.

Methods: The case was followed for a period of two years, during which we collected multiple streams of data, ranging from phenomenological interviews, neuropsychological assessments, language analysis, and semi-structured interviews (Examination of Anomalous Self Experience and Examination of Anomalous World Experience). We analytically constructed a personalized network model of his lifeworld.

Results: We identified an experiential category "the crisis of objectivity" as the core psychopathological theme of his lifeworld. It refers to his persistent mistrust towards any information that he obtains that he appraises as originating in his subjectivity. We can developmentally trace the crisis of objectivity to his adverse childhood experience, as well as him experiencing a psychotic episode in earnest. He developed various maladaptive coping mechanisms in order to compensate for his psychotic symptoms. Interestingly, we found correspondence between his subjective reports and other sources of data.

Discussion: Hernan exhibits difficulties in multiple Research Domain Criteria constructs. While we can say that social sensorimotor, positive valence, and negative valence systems dysfunctions are likely associated with primary deficit

(originating in his adverse childhood experience), his cognitive symptoms may be tied to his maladaptive coping mechanisms (although, they might be related to his primary disorder as well).

KEYWORDS

schizoaffective disorder, personalized network model, enactive psychiatry, psychiatric comorbidity, obsessive compulsive disorder, qualitative phenomenology, RDoC

1 Introduction

A young man enters a fast-food restaurant. The waitress recognizes him as a regular, but he nonetheless avoids her, stepping aside, pretending to be interested in the packets of mustard. He is buying time, preparing for the social interaction. When he orders his sandwich, he does not look at the waitress' face. He feels terrified, self-conscious, embarrassed. He knows he is wrong. He knows he is a child of a mother with schizophrenia. He experienced a psychotic episode himself. He is haunted by the memories of monsters that prey on at night, of the voices commanding him to peel off his skin, of his mirror image that took on a life of its own and started to mock him. The interaction with the waitress is pleasant enough. Still, he gives her an unreasonably large tip as an apology because she had to endure his wrongness. The knowledge of his wrongness makes him unable to cross a road as he is not sure that he will not walk into traffic, or ascend a staircase knowing that he will not suddenly decide to simply throw himself down the stairs. To alleviate the wrongness, he drinks, he smokes weed.

Psychiatric disorders are commonly associated with a heterogeneous clinical presentation, such as the one described above, which makes both treatment and research challenging (Allsopp et al., 2019). Two main classifications of mental health disorders, the Diagnostic and Statistical Manual of Psychiatric Disorders (DSM) and International Classification of Diseases (ICD), are criticized for classifying disorders based on clinical descriptions as there is overlap among the symptoms and biological features of disorders (Lilienfeld and Treadway, 2016). Different pathophysiological mechanisms may therefore lead to the same diagnosis (Cuthbert and Kozak, 2013; Sanislow, 2016). Descriptive classification of psychiatric disorders also fails to consider heterogeneity within each condition for different persons and time course (Wardenaar and De Jonge, 2013; Feczko et al., 2019). The problem of disorder classification partially contributes to the phenomenon of psychiatric comorbidity. As many as 45 % of patients satisfy the criteria for more than one disorder in a year (Allsopp et al., 2019). In fact, diagnostic systems DSM-III to DSM-5 have been criticized for simplifying diagnostic categories, which can result in overlooking a correct diagnosis or falsely diagnosing multiple comorbidities. One of the general criticisms of DSM categorization system that we find in the phenomenological literature is also that it treats the various symptoms as disparate parts rather than an unified whole (Stanghellini and Mancini, 2017).

Nordgaard et al. (2023) provide a conceptual analysis of psychiatric comorbidity. In somatic medicine, comorbidity refers to the co-occurrence of two distinct disease entities, each of which

having a known etiology or pathology (Kaplan and Feinstein, 1974). However, in psychiatry, etiology is typically unknown and symptoms that are unique to only one disorder are rare. Nordgaard et al. (2023) further point out that - specifically in research - the notion of psychiatric comorbidity assumes independence of disease entities. What is commonly omitted is the idea of diagnostic hierarchy (i.e., the idea that certain diagnoses should not be made in the presence of other specific disease entities; Pincus et al., 2004; Ghaemi, 2018; Kotov et al., 2018). Nordgaard et al. (2023) acknowledge that true comorbidity may not even exist in psychiatry, while diagnosing comorbidities is associated with several problems, such as polypharmacy (Pjevac and Korošec Hudnik, 2023; Korošec Hudnik et al., 2024), higher risk of misdiagnosis, and misinterpretation of empirical findings in research where inclusion criteria are too liberal.

Several methodological frameworks have been proposed in order to tackle the problem of symptom heterogeneity. The Research Domain Criteria (RDoC) is an investigative framework that proposes to identify endophenotypes mediating psychiatric disorders (Insel and Cuthbert, 2009). Endophenotype refers to the intermediate level of description between genetics, and the behavioral and phenomenological signs of psychiatric disorders (Cuthbert and Kozak, 2013). RDoC claims that psychiatric disorders are deviations of otherwise normal dimensions of human psychological functioning. These dimensions include cognition, positive affect, negative affect, physiological arousal, and social psychology (Cuncic, 2020). RDoC subscribes to explanatory pluralism, that is considering different levels of description (e.g., genetic, electrophysiological, behavioral, phenomenological) as different insights into the same process rather than one of them being epistemologically superordinate (Cuthbert and Insel, 2013). The transition towards dimensional models, however, has been criticized from the perspective of phenomenological psychopathology, citing poor conceptual clarification of psychiatric disease entities (Parnas, 2014).

Recently, one approach towards a more comprehensive account of psychiatric disorders was put forth by De Haan (2020). She argues that the central property of psychiatric disorders are maladaptive patterns of sensemaking. De Haan (2020) proposes that methodologically, we might be able to address sensemaking in psychiatric disorders by constructing personalized network models (PNM). PNMs are depictions of different aspects of a patient's functioning that influence and modulate each other, leading to changes in psychiatric symptoms. PNMs represents phenomena in terms of nodes (a relevant variable) and edges (connections between them). De Haan (2020) outlines four domains that are

to be included in a PNM: biological, social, experiential, and existential (i.e., a person's attitude towards their broader situation).

So far, only one empirical study demonstrated the use of PNM. Larsen et al. (2022a,b) used it to investigate the relationship between psychosis and cannabis use in a longitudinal dual case study. They conducted six interviews per patient. Analytically, they used the collected qualitative material to construct a web of relationships between salient aspects of the patients' lives, in particular in relation to their cannabis use. For one patient, stopping cannabis use proved challenging due to his proximity to the drug (e.g., frequent use by his partner, engagement in the local crime scene). For the second patient, the cessation of cannabis was related to an intricate feedback loop wherein desensitization (which initially proved to be a useful coping mechanism) made her environment less salient. Further, she was concerned with obesity which was maintained through increased appetite under the acute effect of cannabis.

However, De Boer et al. (2022) have criticized network models, arguing that they suffer from a boundary problem; that is, the question of what domains to include in the network analysis of a given patient, as well as how can we even differentiate different domains from one another? One of their arguments is that network analysis favors perspectivism: epistemological position that claims that the construction of scientific bodies of knowledge must take into account the researchers' perspectives as well. A patient may feel that the core of their problems is, for example, them struggling in school. Taking this belief seriously, by necessity constrains our potential interventions, and as such we may miss out on the optimal change for that specific patient. The second major critique of network analysis refers to the relevancy of the patterns emerging from it. Is the overall knowledge about the psychopathology constructed by a network analysis relevant for the patient, the clinician or the broader psychiatric community?

To recapitulate: Not only is the phenomenon of psychiatric comorbidity relevant and difficult to tackle, its proposed solutions are fraught with problems as well. The present paper attempts to contribute to these discussions by refining the PNM approach using contemporary methods in qualitative phenomenology. A case report is presented. The case was chosen for two reasons: a) the patient presents with several symptom constellations and b) was interested in participating in a longitudinal study. The present paper has three main research goals:

- (1) It presents a proof of principle of how PNMs could be integrated into qualitative phenomenological methodology;
- (2) It evaluates whether using novel frameworks in psychopathology (RDoC and PNMs) can assist us in psychiatric diagnosis;
- (3) It presents a novel phenomenological category (what we term crisis of objectivity) that was identified with PNMs.

Due to the complexity of developing novel methodological frameworks, we opted for demonstrating our understanding of PNMs (as originally developed by De Haan, 2020) with a case study. In doing so, we are following similar approaches in phenomenological psychopathology wherein data from single, highly engaged patients are used to resolve technical or conceptual issues (de Haan and Fuchs, 2010; Henriksen et al., 2010;

Stanghellini and Rosfort, 2013; Luhrmann and Marrow, 2016; Wigand et al., 2018; Englebert et al., 2019). Thus, rather than focusing on the clinical relevance of the case itself, we wanted to use the data from this patient to demonstrate how the conceptual framework of PNMs can be integrated with modern methods in qualitative research and phenomenological psychopathology to better understand psychiatric comorbidity.

2 The case

The patient was recruited from an ongoing project of testing the efficacy of online psychotherapy. Thus, contact with him was conducted over video conference. The patient, who we will anonymize as Hernan, is a 27-year-old man of Western European origin living abroad. Formally trained as a journalist, he has recently left his journalistic job to pursue a career as an online content creator. He lives with his husband. Hernan reports a life-long history of psychosis-like experiences, ranging from auditory and visual hallucinations and common periods of dissociation. He frequently experiences nighttime hallucinations and parasomnias during which he talks, screams, and attacks others. His father was imprisoned for drug-related crimes. Hernan believes that his mother suffers from schizophrenia and therefore questions his own perception of reality. Because of imperative auditory hallucinations, he often scratched his skin to the point of injury. He experiences frequent obsessive compulsive symptoms (OCS). These consist of him continuously checking his environment in potentially dangerous situations.

At seventeen, Hernan was diagnosed with schizoaffective disorder and prescribed aripiprazole, which he declined taking. He has not received any psychiatric treatment later on, citing high costs of medical services as the main reason. In the past, he attended psychotherapy sessions for OCS, which ended after a few seasons for the same reason. Prior to being included in this study, he has again started with psychoanalytic psychotherapy. Between the ages of ten and fifteen, he attempted suicide several times and had occasional outbursts of anger and heteroaggressive thoughts. Hernan still has suicidal thoughts and egodystonic intrusive thoughts about buying a gun, but does not intend to commit suicide. He agreed to suicide prevention contract during the therapy.

He remembers being a sad, quiet, angry and lonely child who was discouraged from showing emotions by his mother. Together with his brothers she often ridiculed him. His mother could not accept him being gay, so Hernan pretended to be confused about his sexuality to avoid conflicts. When he broached the subject again at nineteen, his family reacted with aggression. Following a domestic altercation, Hernan left in a hurry and was homeless for nine months.

After turning 18, Hernan smoked marijuana multiple times a day and has taken methylenedioxymethamphetamine (MDMA) and lysergic acid diethylamide (LSD) many times. At present he regularly takes delta-8-tetrahydrocannabinol (delta-8 THC) (200 mg/2 weeks). His psychotic symptoms precede the beginning of the use of recreational drugs. He admits to drinking alcohol excessively during social events in the past, but denies regular use. In the past five years he abstains from alcohol. This has

been confirmed by his partner in a conversation with the principal investigator.

3 Materials and Methods

Qualitative material was collected using in-depth phenomenological interviews derived from multiple methodological frameworks, predominantly micro-phenomenology (Petitmengin, 2006), interpretative phenomenological analysis (Smith et al., 2022), and constructivist grounded theory (Charmaz, 2014). Hernan's symptoms were additionally assessed using the Examination of Anomalous World Experience (EAWE; Sass et al., 2017) and Examination of Anomalous Self Experience (EASE, Parnas et al., 2005) semi-structured interviews. EASE and EAWE items were scored as 0 (absent or questionably present) or 1 (definitely present, covering all severity levels).

The interviews followed a funnel-shaped structure, wherein we started with a general discussion on a specific aspect of Hernan's life. After identifying specific aspects of experience (e.g., a symptom), we followed the guidelines of micro-phenomenological interview to examine the episodes in detail. We additionally collected descriptions of Hernan's baseline experiences in more stressful periods of his life (e.g., when he spent nine months being homeless).

In total, 19 interviews were conducted with Hernan over an 18-month period. Two of the sessions were dedicated to the EAWE and EASE interviews. 16 interviews (including one session for the EAWE interview and one for a debriefing of the project) were conducted by a researcher with several years of experience in conducting phenomenological interviews (AO). EASE was conducted by a clinician trained in this method (AH). The medical history was taken by a psychiatry resident (KHG) in two interview sessions. Throughout the duration of the study, Hernan received supportive psychotherapy from a licensed psychotherapist (MK).

Following the RDoC approach, Hernan was evaluated on multiple psychological domains, using the adjusted version of the miniRDoC battery (first presented in Förstner et al., 2022), consisting of questionnaires and a cognitive task. Positive affect was evaluated using the *drive*, *fun-seeking* and *reward responsiveness* subscales of the Behavioral Inhibition System and Behavioral Approach System scale (BIS/BAS; Carver and White, 1994), and the *positive subscale* of the Positive and Negative Affect Schedule (PANAS; Crawford and Henry, 2004). Negative affect was evaluated using the *behavioral inhibition system* subscale of BIS/BAS, the *negative* subscale of PANAS, and the *phobic anxiety* subscale of the Symptom Checklist (SCL-90; Maruish, 2000). His social cognition was evaluated using the *getting along* and *participation* subscales of the World Health Organization Disability Assessment Schedule (WHODAS 2.0; Ustun et al., 2010), and *interpersonal sensibility* and *anger/hostility* subscales of the SCL-90. His sensorimotor cognition was evaluated using the *somatization* subscale of SCL and *mobility* subscale of WHODAS. His hot cognition was evaluated using the Emotion Regulation Questionnaire (ERQ; Gross and John, 2003) consisting of two

dimensions: *expressive suppression* and *cognitive reappraisal*; the *cognition* subscale of WHODAS, and the Barratt Impulsiveness Scale (Barratt, 1965). Finally, cold cognition was evaluated using a verbal 2-back task. The 2-back task consisted of a letter appearing in the middle of the computer screen for 2.0 seconds. Hernan had to indicate, by button press, whether the letter was equal to or different from the letter that appeared on screen two trials previous. Performance accuracy (correct VS incorrect) was collected.

3.1 Analysis

The interviews were transcribed verbatim.¹ The qualitative material was analyzed according to interpretative phenomenological analysis (Smith et al., 2022) in two phases. The first phase consisted of inductive-deductive coding of interview transcripts. In qualitative research, coding refers to the process of assigning more general, descriptive tags to sections of raw texts (Charmaz, 2014). For inductive-deductive approach we analyzed the text according to preexisting concepts from the (psychopathological, psychiatric) literature (e.g., *hallucination*), while at the same time, paying attention to the data that may question or re-examine existing concepts (e.g., *crisis of objectivity*) (Flick and Flick, 2011). A codebook was constructed in which all the relevant categories are described according to the following elements: (a) telling name; (b) relationship to other categories; (c) meaningful examples; and (d) potential additional comments (Nelson, 2017). The codebook is made available at: <https://osf.io/dj8pt/>.

In the second phase of analysis, we identified the relationships between different experiential categories in order to construct a PNM. PNM refers to a network of all the salient aspects of the patient's life as well as the explanation of the connections between them. For the construction of the PNM, we started with the experiential categories yielded by inductive-deductive coding (forming the nodes of the PNM). Then, the interview transcripts were re-analyzed so as to identify the relationships between individual categories. Due to the novelty of the PNM approach, we adopted a simplifying assumption, wherein we searched for two types of relationships between categories: *upregulation* and *downregulation*. If category A upregulates category B, category B becomes more expressed. If category A downregulates category B, category B becomes less expressed. We assumed a rhizomatic structure to Hernan's PNM: All categories could, in principle, be connected to any other category. Further, the relationships between categories were assumed to be directed. Thus, category A could regulate category B, but category B could also regulate category A. Relationships between the categories had to be grounded in the data in order to be considered valid. Further, we only considered those relationships admissible that were well-grounded: that is, that occurred in multiple interview sessions. Relationships between categories that

¹ To ensure Hernan's anonymity, the transcripts will only be made available to researchers upon reasonable request.

could only be grounded in a single experiential episode were discarded.

We divided Hernan’s PNM into five domains: sensemaking (i.e., the stance he takes towards his own existential situation), symptoms, developmental factors, biological factors, and social factors. An important caveat has to be made regarding these domains. Since the only source of data that was directly accessible to us was phenomenological, only relationships pertaining to the experiential level of description were explicated. For example, Hernan’s mother was diagnosed with schizophrenia. As such, there is likely a genetic component to his disorder. However, this relationship remained unspecified, as we only had access to how Hernan reflectively makes sense of his family background.

We additionally analyzed the transcripts using keyword analysis. We extracted openly available interview data from two qualitative phenomenological studies on the sense of realness (Oblak et al., 2021, 2022). In Oblak et al. (2021) the sense of realness was explored in the normative population. In Oblak et al. (2022), a transdiagnostic sample of “altered” experiences of realness was collected (ranging from mystical, psychedelic, to psychopathological experiences). The transcripts of Hernan’s interviews were searched for the root morphemes of five keywords that we commonly observed in his reports: *rationality*, *truth*, *objectivity*, *reality*, and *fact*. A visual inspection was performed to remove instances where these words were used by the interviewers. Each root morpheme was expressed as an average occurrence of the word per interview. To validate this keyword analysis, we performed a statistical analysis on the data derived from the two groups. The data were tested for parametric assumptions. For all keywords, Shapiro-Wilk’s test revealed that normality was violated. Mann-Whitney rank test was used to estimate the difference between the two groups. FDR correction for multiple comparisons was used. The data from the two groups were then merged (i.e., we obtained a transdiagnostic sample). The percentile rank for Hernan’s word use was then determined.

Hernan’s symptoms were evaluated using the Schizotypal Personality Questionnaire (SPQ-32; Davidson et al., 2016), consisting of the *ideas of reference*, *suspiciousness*, *no close friends*, *constricted affect*, *eccentric behavior*, *social anxiety*, *magical thinking*, *odd speech*, and *unusual perception* subscales; the Yale-Brown Obsessive Compulsiveness Scale (Y-BOCS; Storch et al., 2015), consisting of *obsessions* and *compulsions* subscales; the Patient Health Questionnaire (PHQ-9; Kroenke et al., 2001), examining depression, and General Anxiety Disorder (GAD-7; Swinson, 2006), examining anxiety, Rumination Response Scale (RRS; Nolen-Hoeksema, 2000), consisting of *brooding* (maladaptive self-related cognition) and *reflective pondering* (adaptive self-related cognition). SCL subscales for *OCD*, *depression*, *anxiety*, *psychoticism* and *paranoid ideation* were used. Childhood Trauma Screener (CTS; Glaesmer et al., 2013) was used to test for adverse childhood experience. Finally, the total score on WHODAS was used to evaluate the general degree of Hernan’s functional impairment. Norms were derived from the patients from our clinical practice. The exceptions are norms for ERQ, which were obtained from Barrios and Olalde-Mathieu (2021), Y-BOCS, which were obtained from Fink (2018), and SPQ-32, which

were obtained from <https://faculty.lsu.edu/asap/normative-data.php>.

4 Results

4.1 Symptoms: clinical scales

Hernan’s symptoms are summarized in Table 1. Hernan’s Y-BOCS scores correspond to mild OCS. His PHQ score corresponds to the presence of moderate depressive symptoms. His GAD score corresponds to severe anxiety. Analyses of the factor structure of RRS suggest that rumination can be subdivided into *reflective pondering* (adaptive response), and *brooding* (maladaptive response). On reflective pondering, Hernan ranks in the 74th percentile, whereas on brooding, he ranks in the 90th percentile

TABLE 1 Hernan’s symptoms as assessed by Schizotypal Personality Questionnaire (SPQ-32), Yale-Brown Obsessive Compulsiveness Scale (Y-BOCS), Patient Health Questionnaire (PHQ-9), Generalized Anxiety Disorder Assessment (GAD-7), Symptom Checklist (SCL-90), WHO Disability Assessment Schedule (WHODAS 2.0), Rumination Response Scale (RRS), and Childhood Trauma Screener (CTS).

| Clinical scale | Hernan’s score (percentile rank) |
|----------------------|----------------------------------|
| SPQ-32 | |
| Ideas of reference | 7 (16) |
| Suspiciousness | 9 (38) |
| No close friends | 12 (93) |
| Constricted affect | 11 (87) |
| Eccentric behavior | 16 (99) |
| Social anxiety | 4 (20) |
| Magical thinking | 0 (16) |
| Odd speech | 9 (55) |
| Unusual perception | 12 (100) |
| YBOCS | |
| YBOCS (Obsessions) | 15 (NA) |
| YBOCS2 (Compulsions) | 14 (NA) |
| PHQ-7 | 12 (66) |
| GAD-9 | 19 (74) |
| RRS | |
| Reflective pondering | 11 (75) |
| Brooding | 19 (90) |
| SCL | |
| OCD | 15 (84) |
| Depression | 26 (96.2) |
| Anxiety | 20 (67.8) |
| Psychoticism | 15 (97.8) |
| Paranoid ideation | 4 (55.8) |
| CTS | 112 (99) |
| WHODAS - Sum score | 98 (91.6) |

TABLE 2 Comparison of epistemological keywords between Hernan, participants from the normative population, and participants who had experienced an altered sense of realness.

| Keyword | Hernan [N/interview, percentile rank] | Normative experience (N = 30) [mean, SD] (Oblak et al. 2021) | Altered experience (N = 14) [mean, SD] (Oblak et al. 2022) | P-value (FDR-adjusted), significance level |
|-------------|---|--|--|--|
| Rationality | 6.89 (98) | 0.14 [0.43] | 1.7 [2.21] | 0.002, *** |
| Truth | 5.42 (98) | 1.17 [1.38] | 1.97 [2.34] | 0.356, NS |
| Fact | 6.95 (93) | 0.5 [1.19] | 4.88 [5.83] | 0.002, *** |
| Logic | 6.11 (100) | 0.17 [0.41] | 0.65 [1.2] | 0.057, NS |
| Objectivity | 4.52 (100) | 0.04 [0.15] | 0.55 [0.78] | 0.002, *** |

NS, non-significant; *** $p < 0.005$.

(based on the group of all participants in our lab who had completed RRS; $N = 275$).

4.2 Research domain criteria perspective

Hernan exhibits scores in the middle range for sensorimotor, negative valence, and positive valence systems. Notably, he demonstrates high scores in behavioral approach but consistently low scores in social processes, reflecting feelings of alienation and isolation. Although he scores in the 99th percentile for anger/hostility (i.e., suggesting lack of these feelings), he may have presented socially desirable responses. The cognitive systems domain is ambiguous, with reasonably high performance on the 2-back task (83rd percentile), high expressive suppression, and moderately low impulsivity. However, he faces challenges in everyday cognitive functioning and employs cognitive reappraisal less frequently.

4.3 Hernan’s speech and conduct during interviews

In interviews, Hernan was polite but avoided eye contact, often clutching a pillow for comfort and occasionally scratching himself. Apart from one instance of dissociation, interviews took place without complications. However, Hernan often expressed a desire for researchers’ approval. Given his tendency for socially desirable answers in the early sessions, we placed less importance on those interviews in the analysis. Hernan has a divergent style of thinking with occasionally disorganized speech. He is noticeably preoccupied with his metacognition and sense of reality. We noticed that he commonly uses terms related to epistemology, which we confirmed through quantitative keyword analysis.

Table 2 summarizes how Hernan’s use of epistemological terms compares to participants in two qualitative phenomenological studies investigating the sense of realness. Firstly, we see that there is a significant difference in the use of terms relating to rationality, fact and objectivity between a group recruited from the normative population and participants who experience an altered sense of presence. Hernan is in the 98th percentile for the use of the keyword rationality, in the 93rd percentile for the keyword fact, and 100th percentile for the keyword objectivity

Hernan’s focus on objectivity is reflected in his tendency for axiomatic language, phrasing his experience in terms of natural

laws: “The present as well as the future can only exist upon the foundations of the past, meaning that each state of being is just a conclusion, a natural conclusion of the states of being that came before it.” Consider the following as well: “imagination and memory are going through the same pathways.”

4.4 Symptoms: phenomenology

The core aspect of Hernan’s psychopathology relates to his childhood maltreatment. His mother suffers from schizophrenia. Hernan reports his mother encouraging bullying among his siblings. At the beginning of his studies at university, upon coming out as homosexual to his family, his mother threw him out of the family’s apartment. This resulted in him living at the homeless shelter for nine months. Hernan is ambivalent about his period of displacement; he reports psychotic symptoms being diminished during this time, while experiencing homelessness as traumatizing.

Hernan experiences anxiety as the most detrimental for his everyday functioning (an assessment confirmed by his score on GAD-9; see Table 1). His anxiety is mostly related to his social life, and is severe enough that it commonly results in self-isolation. He quit his job as a journalist and started working from home because he did not have to interact with others. He relates his self-isolation to a decrease in depressive and anxious symptoms. Hernan developed maladaptive coping skills (e.g., almost ritual-like methodical approach for easing the discomfort of uncertainty). The symptoms as a whole and his upbringing contribute to his specific pattern of sensemaking that we call the crisis of objectivity (see below).

4.4.1 Psychotic symptoms and anomalous experience of self and the world

Hernan’s EASE and EAW scores are summarized in Table 3. Hernan scored 14 points on the EASE semi-structured interview. On EAW, he scores 59 points, or 28 if we only account for schizophrenia-exclusive categories (captivation of attention by isolated details, loss of social common sense, alienated scrutinizing of others’ behavior, algorithmic approach to social understanding/interaction; intrusiveness of the gaze of another; tangential responding; disinclination for human society; adherence to abstract, intellectualistic, and/or autonomous rules; pervasive disbelief, skepticism, curiosity re the obvious/taken-for-granted; static quality, stillness, or morbid intellectualism; Sass et al., 2017). During the EASE interview, Hernan reported on the

TABLE 3 Scores on EAWE and EASE dimensions.

| Scale | Score (w/o schizophrenia-exclusive categories) |
|---------------------------------------|--|
| EAWE | 59 (28) |
| Space and objects | 16 (11) |
| Time and events | 5 (1) |
| Other persons | 16 (8) |
| Language | 2 (2) |
| Atmosphere | 8 (4) |
| Existential orientation | 3 (2) |
| EASE | 14 |
| Cognition and stream of consciousness | 4 |
| Self-awareness and presence | 6 |
| Bodily experience | 4 |
| Demarcation | 0 |
| Existential reorientation | 0 |

For EAWE, scores are separated into total scores and those without symptoms present in other disorders than schizophrenia (in parentheses).

anomalous experience of self-awareness and presence. For example, he described the experience of diminished presence with a sense of barrier, a “filter” between himself and the external world:

I will feel pushed back inside of my own mind and I would still receive the information from my senses, touch, taste, vision, etc. But I’ll feel it like muffled, like there was a filter between the information and me. I will still receive it [the information] but not feel like it was generally perceived by me.

He also described anomalous experience of his own embodiment in a sense of psycho-physical split, describing his body as a “meat suit”.

4.4.2 “I used to be on fire, now only the ashes remain”: making sense of a history of mental disorder

A central aspect of Hernan’s self-narrative is that he is a person suffering from a mental disorder. He had been diagnosed with schizoaffective disorder and experienced at least one psychotic episode:

[O]ut of nowhere, my reflection in the mirror stopped, like, moving and talking. [...] I couldn’t escape the thrall of the mirror. [...] [T]he entity in the mirror just kept talking to me and was extremely nasty and telling me how I’m worthless and I deserve everything that happened to me and that I should kill myself.

Hernan regularly experiences simple visual pseudohallucinations, most typically appearing as anomalous textures, as well as imperative auditory hallucinations. Notably, he hears a voice saying the word “peel,” which prompts him to

self-harm by scratching his skin. He also frequently experiences parasomnias, associated with falling asleep or waking up. The parasomniac hallucinations are veridical and emotionally congruent. Upon waking up, in the middle of the night, he often fights his husband, whom he mistakes for a hallucination:

It was a monster or a home invader that I was seeing. [...] And biologically, the body is [] geared up to react to a threat. [...] So, I find myself in the middle of what could be the most stressful situation any person can find himself themselves in, which is fighting for what I believe to be my survival, except that I find myself thrust in the middle of that fight, in the middle of that stress, without experiencing consciously all the steps leading up to that fight.

4.4.3 Falling into the sunken place: symptoms of anxiety

Hernan himself refers to dissociation in social settings as “episodes”, and compares them to the Sunken Place from the movie *Get Out* (Peele, 2017). It is an exclusively negatively valenced experience that takes place in social settings and has a typical temporal dynamic: *rising phase*, *peak phase*, and *break*. During the rising phase, Hernan has insight into what is happening:

Stuff [is] pulling away from me and I can see and feel myself entering this state. [...] I want to get out of here. I need to get out of here. It’s bad to be here. I’m afraid. I’m uncomfortable.

The gradual transition into an episode allowed him to develop protective behaviors in public settings (e.g., while he was still a journalist, he was able to read out prepared questions from a list). During the peak phase, Hernan experiences a detachment from his surroundings. This is associated with perceptual anomalies (e.g., tunnel vision, fading of color intensity, spatial distortions). While Hernan is sensorially connected to his environment, he feels as if his surroundings are no longer accessible to him. Hernan feels that lived time breaks down (e.g., he experiences a sense of eternity) and has to interfere with it in order for it to stop. While in the peak phase, Hernan feels as if verbal strategies (e.g., commanding himself to snap out of it) are not effective regulators. Hernan experiences decreased insight and the episodes are not amenable to conscious reflection, and he no longer comprehends the words that are being spoken². Break occurs either through sensory deprivation or a change in the level of consciousness (e.g., sleep).

4.4.4 “There is no spoon in the microwave”: obsessive compulsive symptoms

Finally, Hernan exhibits clear-cut signs of OCS. OCS occur when he is engaged in a potentially dangerous situation (e.g., leaving a spoon in a microwave). The presence of a negative outcome (e.g., being hit by a car, falling down the stairs) prompts him to start thinking about the danger. He then to mistrust his own cognition (e.g., he is unable to rely on his working memory,

2 Due to this, Falling Into the Sunken Place, as an experiential category, should be understood critically. We constructed it analytically, and not as a phenomenon directly reported by Hernan.

informing him of having looked both ways before crossing the road):

I put [the soup] in a bowl inside of the microwave. [...] I close the door. And then I opened the door again [...] It's the feeling [that] the only thing that's left is my memory of it being the way it was. And so I wonder, *Did I leave the spoon inside and reopen?* Check. *No spoon.* I close it. I reopen it. No spoon. [...] It's fear that something's not right. [...] I must have spent so much time of my life just opening and closing the stupid microwave.

Hernan is aware that his behavior is unreasonable, however, he is unable to stop himself: "I am [...] rationally aware that this is not a normal fear to have—that is, there is no point that I'm being ridiculous and well, I'm being ridiculous." Hernan experiences compulsions (e.g., making a small movement with his fingers above the bowl of soup in the microwave) as a form of irrationality. Yielding to them is a form of defeat for him.

4.5 Crisis of objectivity

The core of Hernan's experience is what we term the *crisis of objectivity* (CoO). CoO is a pattern of sensemaking wherein Hernan mistrusts his own cognition. Hernan himself experiences CoO as the baseline aspect of consciousness: "In general, that's an undercurrent. Not trusting myself is [...] the catchphrase of my life at this point." We can relate CoO to his trauma of growing up with a parent with schizophrenia: "My own mother is completely confused in her head. [...] And I said I didn't want to be like them, and to not be like them, I needed to be [pause] intellectually better."

In his youth, Hernan's only source of solace was the media. He notes that obtaining an education represented finding a better life away from his family. Striving for objectivity was also closely tied to his journalistic profession: "My job is literally to be as accurate as possible because other people rely on me to have access to objective reality." Thus, for Hernan, objectivity is the paramount value:

Even if it's something that requires a lot of time because like having to ponder and take apart the thread of a very difficult knot of, of threads that are stuck together, it can take months of, you know, regularly taking time just to myself, staring at the ceiling, laying down, doing nothing, just thinking for hours and hours and hours until I find an answer. Because an answer exists. There's always an answer.

Having had problems with his mental health in general, but psychotic experience specifically, Hernan exists with the awareness that his judgments might be false. He distrusts himself not only when experiencing intrusive thoughts — what he calls the "call of the void" — but in every moment of reflection:

[T]here's always the thought in the back of my head [...] [t]here's always a thought that my body is not just mine. Something could happen at any point [...] [that] is a genuine threat to myself. [...] [E]ach new experience not just reinforces this

knowledge that my body is not trustworthy [...] It is objectively true no matter how much you think things through.

While Hernan generally believes his senses, he has had experiences where this perceptual security was put into question. Thus, his baseline experience is that of the constant probability that his senses and cognition are false:

I have a better imagination than I do a memory. [...] It's being able to tell to myself: *I just checked [the microwave]. There is no, there is no spoon.* And I'm able [...] to remember that I did [check it]. [...] [T]he problem is that even though I have the ability to have a visual projection of the memory, it's worthless because of the nature of the memory. The memory is not true. [...] It will mean nothing because I know what the memory is and I know that every aspect of the memory is artificial.

Hernan is unable to trust his instincts. Rather, he has to reason about everything:

Because I hold myself to that standard that I don't want my mind to become a barren wasteland no matter what genetic or what predisposition I have. [...] I force myself to never, you know, go with my instinct, do what my mother would do.

CoO was apparent both in his reports on his daily life experience as well as his voluntary comments when performing the working memory task. Despite scoring in the 83rd percentile, Hernan continuously vocalized doubts about his task performance.

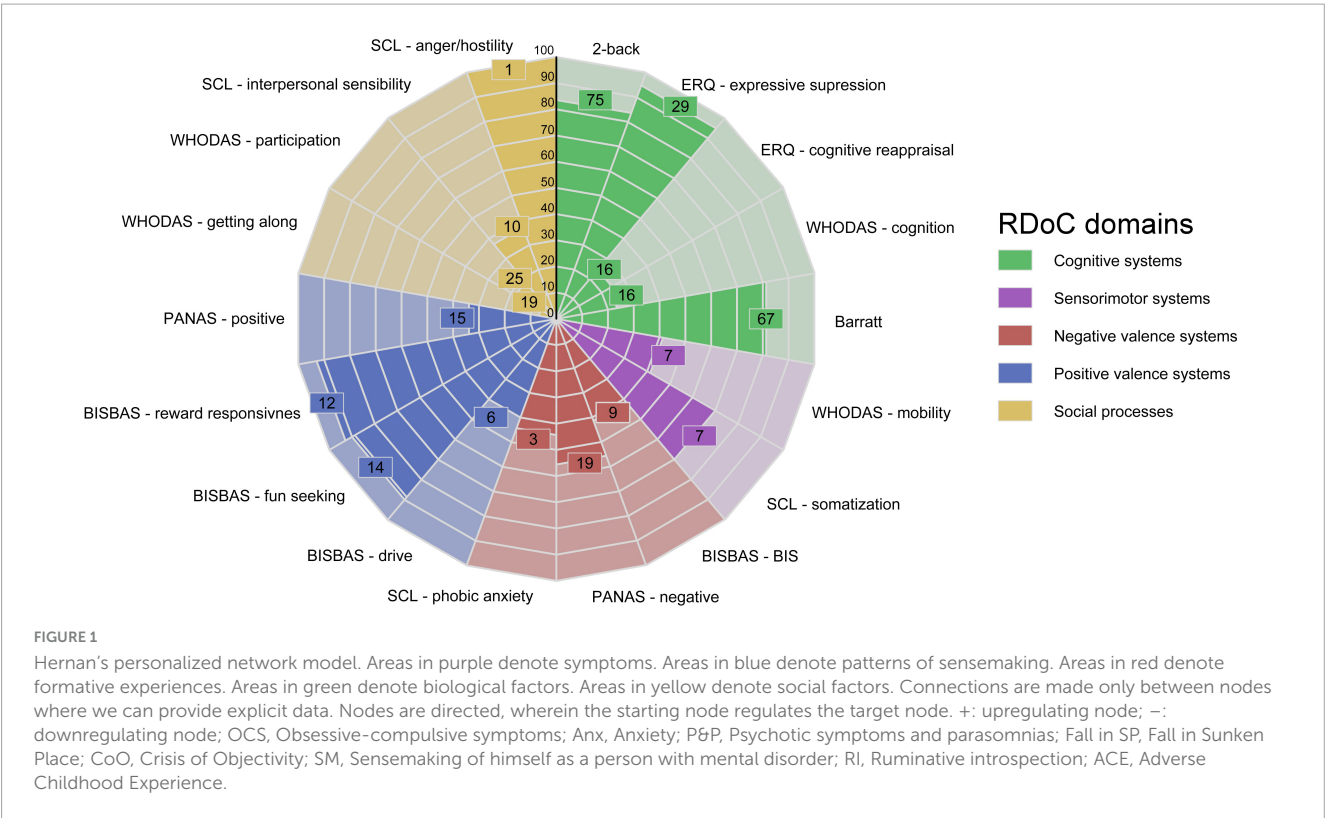
4.6 Ruminative introspection

Hernan engages in the process of *ruminative introspection* (RI). In response to a stressful event, Hernan reflects on this event in a highly ritualized fashion. His RI follows a specific set of steps:

- (1) Hernan isolates himself and tries to minimize the presence of sensory stimuli, in particular, bodily sensations;
- (2) RI is associated with a change in the level of consciousness;
- (3) Hernan reflects on various events:
 - (4.1) Hernan allows new ideas, related to the situation that he is reflecting on, to arise;
 - (4.2) Once Hernan zeroes in on the topic, his rumination starts to deepen:

I try to force myself to empathize as if someone else was observing me and had all the correct information. [...] What would they think? How would they judge me? If that person was not me, if that person did not have all of my biases, and all of my trauma and all of my problems.

Hernan relates subjective events to objective facts, so he is able to allocate responsibility in difficult social interactions:



[S]ince a lot of my socialization is done online, it's been very practical for me, because I can literally go back into the text conversation and read the history of the conversations. [...] I say: *did I do the right thing? Did the other person do the right thing?* [...] One of the elements that led to our falling out [with best friend] was that I was doing [a gift] for him. I didn't do [it] in time [...] and he got really angry [...] So, first I called back to memory, the events that occurred. Meaning every conversation, every interaction, when the topic was brought up by my best friend or me. [...] When did I start working on it? How many hours of work did I get done? [...] I was working two full-time jobs at the same time. And I just didn't have the time. So, this is a fact. As objective a fact can be. [...] [I]t can take weeks to go through every question that I want to ask myself. And go through all the facts and all the beliefs and all the biases and all the perceptions that are relevant to the topic.

Whenever Hernan is ruminating on specific social interactions, he is attempting to evaluate the statements that interlocutors had made according to formal logic:

[M]y filter is made out of knowledge, logical fallacies, about my own abilities. [...] If I can say: *oh, this is, erm, appeal to authority.*

(5) Hernan reifies conclusions of RI (i.e., phrases it as a statement or a syllogism) and commits it to memory:

So what I do is I try to logically spell out for myself, Um, it's like having a word document inside my head. I keep adding notes and I remember all the notes I've taken and I keep adding more and more and trying to put structure and logic in the way I think.

Hernan associates reaching a positive result of RI with positive feelings. Hernan noted that he might one day write a book of all of his realizations, specifically as a form of revenge against his mother.

4.7 Hernan's personalized network model

We observed stable relationships between different aspects of Hernan's experience. Specifically, we analyzed which experiential categories *upregulate* others (i.e., make them more pronounced), and which *downregulate* them (make them less pronounced). Hernan's personalized network model is outlined in Figure 1.

Hernan's PNM is rhizomatic (as per our assumptions): all nodes are connected but no one node is connected to all others. When outlining Hernan's PNM, it became apparent that CoO represents the central, maladaptive pattern of sensemaking, as it is present in the highest number of connections ($N = 7$). The mistrust into his own cognition was brought about by a) his traumatic experience of being a child of a patient with schizophrenia; and b) having himself suffered a psychotic experience in earnest. CoO is then reinforced by his parasomniac and anxious symptoms. The CoO forms the

basis of his OCS, and is subsequently reinforced by what Hernan perceives as “yielding to irrationality” inherent in compulsions.

Interestingly, Hernan reports on CoO being beneficial when dealing with his anxiety:

I want to avoid the mess and the clutter [and] stupidity that I've seen in adults when I was a kid. [...] And my whole life I've been trying to build myself structure, a mental structure, to rely on where my instinct might fail me.

Further, the process of RI downregulates both his feelings of anxiety and *Falling into the Sunken Place*. For an example of the former consider this:

When I am deep in this thinking, I am a lot less anxious. [...] Just spending time by myself and thinking very hard about things, um, because the dark period was triggered by the outside events.

For an example of the latter, consider the following:

It dissipated after [...] looking around myself and getting a sense of where I am [...] [T]here are all the coping mechanisms that come with realizing something is in my head that we start with having to dissociate myself [...] I also determine that everything I feel and everything I believe at the moment is fake. [...] And so as soon as I can determine something is in my head, there is an entire process [...] [I am] taking a moment to address one by one the actual feelings that I'm going through and taking long breaths, sitting down, looking around myself a lot to make sure that there really is nothing there.

5 Discussion

We presented a case study of Hernan, a patient who exhibits various symptoms precluding straightforward diagnosis and psychotherapeutic treatment. The goal of this paper was to see whether contemporary approaches in psychopathology (phenomenological psychopathology, PNMs, RDoC) can aid in diagnosis; to demonstrate how PNMs could be integrated into qualitative phenomenological research; and to illustrate a novel aspect of experience identified by PNM (*crisis of objectivity*). In the discussion, we will contextualize our findings within these three approaches, discussing Hernan's condition as a disorder of the self, disorder of sensemaking, as well as how his condition fits within domains of functioning, described by RDoC.

5.1 The phenomenological perspective on self-disorders

The distinction between the minimal self (or ipseity) and narrative self is crucial in phenomenological psychopathology. The minimal self relates to personal experiences, while the narrative self encompasses one's identity through stories and beliefs

(Hutto, 2016). In psychotic spectrum disorders the minimal self is disrupted, evident in symptoms like impaired sense of agency and extracampine hallucinations (Chan and Rössler, 2002).

While schizoaffective disorder lacks robust evidence for anomalous self-experience, growing data suggest overlaps in phenomenology, biology, and genetics between schizophrenia and bipolar disorder, challenging their distinct diagnostic boundaries (Keshavan et al., 2011). Hernan's EASE score was 14, compared to averages of 20.7 for schizophrenia and 6.3 for bipolar disorder as per Henriksen et al. (2021). Hernan exhibits disturbances of minimal self, which are not present equally in all relevant domains. This places Hernan on a milder end of the schizophreniform spectrum, which is consistent with his diagnosis of schizoaffective disorder (Henriksen et al., 2021).

Further, Hernan has difficulties trusting information that he appraises as being *subpersonal*. In the 20th century, mind sciences revealed that much of our psychology operates below conscious awareness. For example, in cognitive science, intuition relies on implicit learning of statistical patterns (Damassio, 1996). Neuroscience similarly reveals that the insula in the brain processes information from internal organs (Seth, 2013). While some of this leads to conscious awareness (interoception), only a fraction reaches the temporoparietal junction for conceptual processing of bodily schemas (Quesque and Brass, 2019). Thus, Hernan's *Crisis of Objectivity* reflects a disturbance in what we could refer to as “cognitive” self, mistrusting subpersonal processes such as intuition. This echoes Cartesian doubt and a loss of perceptual safety:

If I'm carrying something, no matter the item, like just carrying groceries or, or a roll of toilet paper in my hand, I'm going to not trust myself to continue holding that item. Even if there are no consequences to it, I will still feel afraid and stressed at the thought of dropping it. [...] The mere notion of me not trusting myself is a source of stress.

The pervasive sense of doubt has been identified as a central aspect of obsessive-compulsive psychopathology. Already in Pierre Janet's description of OCS, doubt and feelings of uncertainty play a predominant role in understanding this disease (Pitman, 1987). These observations were confirmed by recent studies. Samuels et al. (2017) report on a large-scale survey that demonstrates pervasive doubting is associated with checking symptoms of OCS, as well as depression and anxiety. Chiang and Purdon (2023) report on semi-structured interviews that demonstrate OCS is associated with one's doubting whether they performed certain tasks well, as well as more profound questioning of their own memory and perception (what we termed CoO).

From a phenomenological perspective, CoO may be linked to an underlying self-disorder. In the following quotation, we see how the dynamics of CoO is established. CoO could be also interpreted as an experience, leading into *lack of natural evidence* and subsequent *hyperreflexivity*, as it is described in self-disorders (Parnas and Sass, 2001). Furthermore, we could suspect that at least some of his rumination processes, described in the category of RI, seem to be of secondary nature as a consequence of underlying hyperreflexivity. Hyperreflexivity is one of the fundamental components of ipseity disturbances. Consequently,

aspects of oneself are experienced as a kind of external object (Sass and Parnas, 2003).

Hernan presents with OCS as well as psychotic symptoms. The co-occurrence of both pathologies was recognized as early as the first descriptions of schizophrenia (Bleuler, 1911). Stengel (1945) also speculated that in patients with schizophrenia the psychotic reaction was kept under control with the aid of obsessional symptoms. From the phenomenological perspective, the feature that is usually thought to be crucial for differentiating an obsession from a delusion is insight (De Haan et al., 2013). However, individuals with OCS are prone to confusion between reality and possibility; they tend to mistake hypothetical possibilities for real probabilities, which has been termed inferential confusion (Aardema et al., 2005). The crucial aspect of the latter is distrust of the senses, also seen in our patient. Some authors state that the compulsive adherence to the doubt in the obsessive patient could be an equivalent to the absolute certainty in a delusion of the psychotic patient (Dalle Luche and Iazzetta, 2008) and it has even been suggested that OCD could be better characterized as a belief disorder (Ahern et al., 2019).

5.2 RDoC perspective

The problem of symptom heterogeneity can be tackled through the RDoC framework. Hernan displays an idiosyncratic pattern of maladaptive sensemaking. The question is whether this maladaptive pattern of sensemaking constitutes a cognitive dysfunction that could be attributed to psychosis spectrum disorder (e.g., disorganization symptoms). Schizophrenia spectrum disorders present varied symptoms, notably executive function deficits (Torrent et al., 2007; Gotra et al., 2020), affecting processing speed, memory, attention and reasoning. These deficits persist over time and are not attributable to antipsychotic treatments (Green et al., 2019).

Hallucinations, in particular auditory hallucinations, are considered to be disruptions in multiple domains in the RDoC framework. They implicate the cognitive domain at the level of language, perception, declarative memory, and cognitive control. On the level of social domain, they are related to affiliation (RDoC construct that refers to positive interactions with others) and perception and understanding of self (agency). Finally, in the negative valence systems, hallucinations are related to both sustained and acute threat (Ford et al., 2014). Within perception (a cognitive systems domain), dysfunctions of sensory integration are well-documented in psychosis spectrum disorders. A recent RDoC-informed study has demonstrated that visual integration dysfunction is a symptom that is general across psychosis spectrum disorders and not characteristic only of schizophrenia (Grove et al., 2018).

In Hernan, we observed some symptoms that could be linked to motor systems disorders. During interactions, he exhibits twitches, self-harming behavior (scratching) and often interrupts the interviewers. Sensorimotor dysfunctions may serve as a biomarker for psychosis spectrum disorders (Mittal et al., 2017; Hirjak et al., 2018). Patients exhibit varied movement disorders, including velocity scaling, stereotypies, catatonic immobility, and perseveration. Smooth movement execution and motor plan

updating are disrupted, manifesting as poor postural control, motor learning issues, and eye-blink conditioning. Schizophrenia shows psychomotor slowing, affecting emotional and motor regulation. This can result in varied movement patterns and catatonia.

RDoC-based research highlights impaired facial recognition, especially fear and anger detection, in psychosis spectrum disorders like schizophrenia (Gur and Gur, 2016; Gur et al., 2017). Emotion recognition, including gaze perception, is a common transdiagnostic biomarker (Tso et al., 2020). Hernan exhibits issues with eye contact and scores on social cognition scales like WHODAS and SCL indicate challenges in this area.

Discrimination (on the basis of gender, race, creed, etc.) is a risk factor for psychopathology during development, perhaps by putting additional demands on implicit emotional regulation (Vargas and Mittal, 2021). Being homosexual and an immigrant may have played a modifying role in Hernan's psychopathology. Psychotic spectrum disorders show impaired performance in a variety of cognitive control tasks (Sabharwal et al., 2016; Smucny et al., 2018). Finally, rumination, linked to worse psychiatric prognosis and severity, correlates with socioeconomic status (Silveira et al., 2020).

Bayer et al. (2023) propose language production as a reliable measure of thought disorder, linking positive disorder to semantic coherence and negative disorder to syntactic complexity. Psychosis spectrum disorders impact speech and social function but not mood disorders (Cohen et al., 2012). Hernan shows fluent speech with complex ideas but displays positive thought disorder, acknowledging his "meandering" thinking style.

Within the RDoC framework, anxiety aligns with danger detection in the negative valence domain. OCS is viewed as a maladaptive system hyperactivity evolved for detecting danger (Woody et al., 2019). Figee et al. (2016) connect OCS to behavioral addiction and maladaptive reward systems. Common neurobiological mechanisms exist between OCS and anxiety (Gillan et al., 2017). Compulsions are also linked to impulsivity, a component of cognitive control (Moreno-Montoya et al., 2022).

5.3 Methodological considerations

The present paper represents an attempt to put the idea of PNMs, originally introduced in De Haan (2020), into practice. In this study, we employed PNMs as a conceptual framework for qualitative analysis. As such, our approach amounts to a somewhat narrow use of PNMs (which could also be employed within a quantitative framework by using dynamical systems or graph theory). Nonetheless, we found using PNMs as a guiding principle for qualitative analysis useful. In doing so, we made several methodological observations, which may be of use to other researchers.

Firstly, using PNMs as a qualitative framework necessitates that both the nodes and connections be grounded in the data. In our experience, this requires a somewhat larger amount of data collection than is typical for qualitative studies. We were only able to ground all the relevant experiential categories after 14 interview sessions (from our own experience with qualitative phenomenology, on average, 11 participants—assuming one interview per participant—are sufficient to exhaust all the

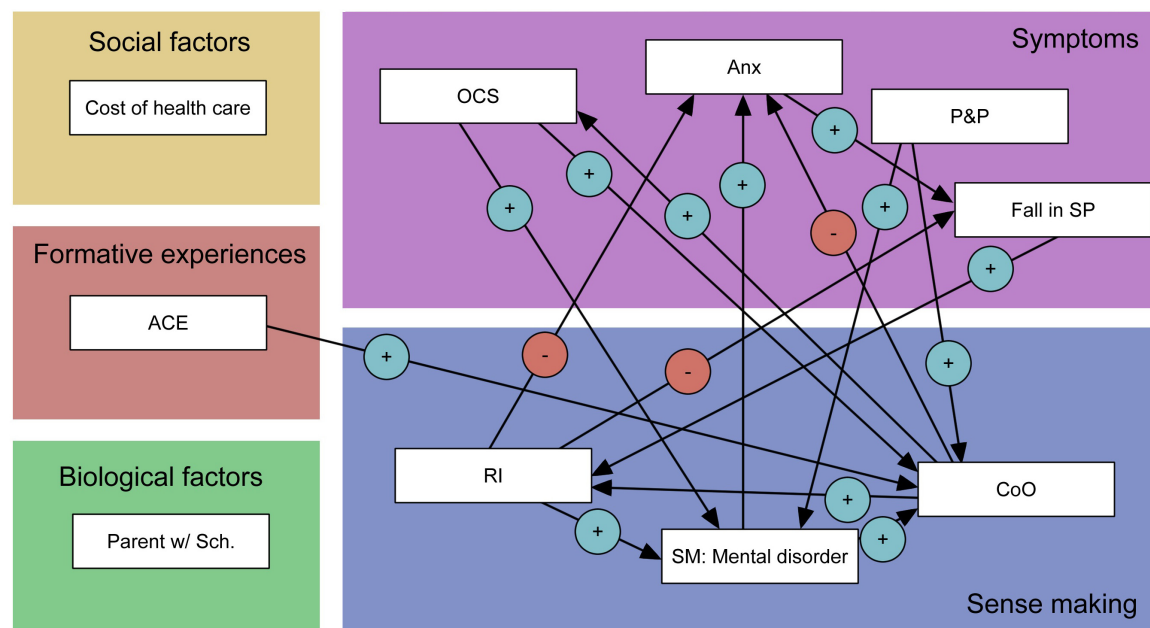


FIGURE 2

Hernan's scores on the mini RDoC battery. Green: cognitive systems; red: negative valence systems; blue: positive valence systems; yellow: social processes. Areas with saturated colors denote Hernan's percentile rank on a given construct. Percentile ranks are represented by colored areas. Hernan's absolute scores on given constructs are marked with white numerals. Higher values represent greater presence of the psychological function in question; thus scores on SCL, WHODAS, and Barratt were inverted to reflect this principle.

relevant qualitative material). Even after 14 interview sessions, two connections were only observed once, and were thus omitted from the final presentation of the data.

Secondly, it became apparent early in the study that grounding all the connections in the qualitative material would be prohibitively complex without adopting the simplifying assumption of them referring to up- and downregulation. It is likely that establishing more complex connections would require either a) further data collection; or b) adopt additional data collection techniques.

Thirdly, as evident in Figure 2, our PNMs primarily consist of data collected at the same level of description: lived experience. Establishing connections with other domains (social and biological) may require non-qualitative data collection strategies (e.g., experimental research design).

An open question remains how we might translate the collected qualitative material into quantifiable data.

6 Limitations

In the present study, we sought to further develop the methodological framework of PNMs, originally proposed by De Haan (2020) and Larsen et al. (2022a,b). In this study, we used the data from a patient, chosen due to the complexity of his clinical presentation and his willingness to participate in the study in the long-term. Our approach has two critical limitations. Firstly, while we collected quantitative measures (i.e., questionnaires) and objective data (i.e., formal linguistic analysis and cognitive task measures) in addition to qualitative phenomenological reports, these are not used in the PNMs. Rather, they serve as objective

validations of patterns observed in the PNMs. Future work on PNMs should focus on integrating quantitative measurements as well as rely more heavily on graph theory in constructing them. Second, Hernan was originally recruited for a study on online psychotherapy. As such, we were unable to perform clinical tests (e.g., bloodwork and neuroimaging). Additionally, as is well-known, clinical interviews that are not conducted in person are of lower quality and validity as they lack intersubjective attunement and implicit countertransference dynamics (Jansson and Nordgaard, 2016).

7 Concluding remarks

Symptom heterogeneity constitutes a challenging problem in psychiatry, both in terms of how best to research specific psychiatric disorders and treat them in a clinical setting. Recently, several frameworks have been proposed to tackle the problem of psychiatric comorbidities. We presented a case study where several approaches were used in order to identify the core of his pathology. We employed an in-depth mixed methods approach to describe his psychopathology. We identified one experiential category—the *crisis of objectivity*—as the core psychopathological theme of his lifeworld. CoO refers to his persistent mistrust towards any information that he obtains that he appraises as originating in his subjectivity. We can developmentally trace CoO to his adverse childhood experience, as well as him experiencing a psychotic episode in earnest. Hernan developed various maladaptive coping mechanisms in order to compensate for his psychotic symptoms. Interestingly, we found correspondence between his subjective reports and other sources of data. Hernan exhibits difficulties

in multiple RDoC domains. While we can say that social, sensorimotor, positive valence, and negative valence systems dysfunctions are likely associated with primary deficit (originating in his adverse childhood experience), his cognitive symptoms may be tied to his maladaptive coping mechanisms. Our multi-method approach demonstrates how multiple sources of data may converge onto novel understanding of symptom clusters by identifying their common core. It would be of interest to see whether such an approach can be further used for personalized treatment of psychiatric disorders and whether it can be scaled up to include enough patients to be representative of relevant populations.

Data availability statement

The datasets presented in this study can be found in online repositories, <https://osf.io/dj8pt/>. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

Ethics statement

The studies involving humans were approved by the University Psychiatric Clinic Ljubljana Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

AO: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software,

Supervision, Visualization, Writing—original draft, Writing—review and editing. MK: Conceptualization, Writing—review and editing. KHG: Investigation, Writing—original draft, Writing—review and editing. AH: Investigation, Writing—original draft, Writing—review and editing. UB: Supervision, Writing—review and editing. BŠ: Supervision, Writing—review and editing. JB: Conceptualization, Funding acquisition, Supervision, Writing—original draft, Writing—review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Slovenian Research and Innovation Agency (ARIS) research grants P5-0110 (JB) and J3-4534 (BŠ).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Aardema, F., Emmelkamp, P. M. G., and O'Connor, K. P. (2005). Inferential confusion, cognitive change and treatment outcome in obsessive-compulsive disorder. *Clin. Psychol. Psychother.* 12, 337–345. doi: 10.1002/cpp.464
- Ahern, C., Fassnacht, D. B., and Kyrios, M. (2019). "Obsessions and phobias," in *The Oxford handbook of phenomenological psychopathology*, eds G. Stanghellini, M. Broome, A. Raballo, A. V. Fernandez, P. Fusar-Poli, and R. Rosfort (Oxford: Oxford University Press), 576–583. doi: 10.1093/oxfordhb/9780198803157.013.61
- Allsopp, K., Read, J., Corcoran, R., and Kinderman, P. (2019). Heterogeneity in psychiatric diagnostic classification. *Psychiatry Res.* 279, 15–22. doi: 10.1016/j.psychres.2019.07.005
- Barratt, E. S. (1965). Factor analysis of some psychometric measures of impulsiveness and anxiety. *Psychol. Rep.* 16, 547–554. doi: 10.2466/pr0.1965.16.2.547
- Barrios, F. A., and Olalde-Mathieu, V. E. (2021). Replication data for: Psychometric properties of the emotion regulation questionnaire [dataset]. *Harvard Dataverse* doi: 10.7910/DVN/7S6FWB
- Bayer, J. M. M., Spark, J., Krcmar, M., Formica, M., Gwyther, K., Srivastava, A., et al. (2023). The SPEAK study rationale and design: A linguistic corpus-based approach to understanding thought disorder. *Schizophr. Res.* 259, 80–87. doi: 10.1016/j.schres.2022.12.048
- Beuler, E. (1911). *Dementia praecox oder gruppe der schizophrenien*. Leipzig: Verlag Franz Deuticke.
- Carver, C. S., and White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *J. Pers. Soc. Psychol.* 67, 319–333. doi: 10.1037/0022-3514.67.2.319
- Chan, D., and Rossor, M. N. (2002). —But who is that on the other side of you?" Extracampine hallucinations revisited. *Lancet* 360, 2064–2066. doi: 10.1016/S0140-6736(02)11998-2
- Charmaz, K. (2014). *Constructing grounded theory*, 2nd Edn. Thousand Oaks, CA: Sage.
- Chiang, B., and Purdon, C. (2023). A study of doubt in obsessive-compulsive disorder. *J. Behav. Ther. Exp. Psychiatry* 80:101753. doi: 10.1016/j.jbtep.2022.101753
- Cohen, A. S., Najolia, G. M., Kim, Y., and Dinzeo, T. J. (2012). On the boundaries of blunt affect/alogia across severe mental illness: Implications for research domain criteria. *Schizophr. Res.* 140, 41–45. doi: 10.1016/j.schres.2012.07.001
- Crawford, J. R., and Henry, J. D. (2004). The positive and negative affect schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *Br. J. Clin. Psychol.* 43, 245–265. doi: 10.1348/0144665031752934
- Cuncic, A. (2020). *Overview of the research domain criteria (RDOC) approach. Psychopathology: Definition, types, and diagnosis*. Available online at: <https://www.verywellmind.com/overview-of-the-research-domain-criteria-4691025> (accessed January 28, 2024).

- Cuthbert, B. N., and Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Med.* 11:126. doi: 10.1186/1741-7015-11-126
- Cuthbert, B. N., and Kozak, M. J. (2013). Constructing constructs for psychopathology: The NIMH research domain criteria. *J. Abnorm. Psychol.* 122, 928–937. doi: 10.1037/a0034028
- Dalle Luche, R., and Iazzetta, P. (2008). When obsessions are not beliefs. *Comprendre* 16, 141–157.
- Damasio, A. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 351, 1413–1420. doi: 10.1098/rstb.1996.0125
- Davidson, C. A., Hoffman, L., and Spaulding, W. D. (2016). Schizotypal personality questionnaire – brief revised (updated): An update of norms, factor structure, and item content in a large non-clinical young adult sample. *Psychiatry Res.* 238, 345–355. doi: 10.1016/j.psychres.2016.01.053
- De Boer, N. S., Kostić, D., Ross, M., De Bruin, L., and Glas, G. (2022). Using network models in person-centered care in psychiatry: How perspectivism could help to draw boundaries. *Front. Psychiatry* 13:925187. doi: 10.3389/fpsyg.2022.925187
- De Haan, S. (2020). *Enactive psychiatry*, 1st Edn. Cambridge: Cambridge University Press. doi: 10.1017/9781108685214
- de Haan, S., and Fuchs, T. (2010). The ghost in the machine: Disembodiment in schizophrenia – two case studies. *Psychopathology* 43, 327–333. doi: 10.1159/000319402
- De Haan, S., Rietveld, E., Stokhof, M., and Denys, D. (2013). The phenomenology of deep brain stimulation-induced changes in OCD: An enactive affordance-based model. *Front. Hum. Neurosci.* 7:653. doi: 10.3389/fnhum.2013.00653
- Englebert, J., Monville, F., Valentini, C., Mossay, F., Pienkos, E., and Sass, L. (2019). Anomalous experience of self and world: Administration of the EASE and EAW scales to four subjects with schizophrenia. *Psychopathology* 52, 294–303. doi: 10.1159/000503117
- Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., and Fair, D. A. (2019). The heterogeneity problem: Approaches to identify psychiatric subtypes. *Trends Cogn. Sci.* 23, 584–601. doi: 10.1016/j.tics.2019.03.009
- Figee, M., Pattij, T., Willuhn, I., Luigjes, J., Van Den Brink, W., Goudriaan, A., et al. (2016). Compulsivity in obsessive-compulsive disorder and addictions. *Eur. Neuropsychopharmacol.* 26, 856–868. doi: 10.1016/j.euroneuro.2015.12.003
- Fink, J. (2018). Dataset: Changing levels of disgust through imagery rescripting and cognitive reappraisal in contamination-based obsessive-compulsive disorder: An experimental study [dataset]. Mendeley. doi: 10.17632/V3HKVFH3GP.1
- Flick, U., and Flick, U. (2011). *An introduction to qualitative research*, 4. Edn. Thousand Oaks, CA: SAGE.
- Ford, J. M., Morris, S. E., Hoffman, R. E., Sommer, I., Waters, F., McCarthy-Jones, S., et al. (2014). Studying Hallucinations Within the NIMH RDoC Framework. *Schizophr. Bull.* 40, S295–S304.
- Förstner, B. R., Tschorn, M., Reinoso-Schiller, N., Maričić, L. M., Röcher, E., Kalman, J. L., et al. (2022). Mapping Research Domain Criteria using a transdiagnostic mini-RDoC assessment in mental disorders: A confirmatory factor analysis. *Eur. Arch. Psychiatry Clin. Neurosci.* 273, 527–539. doi: 10.1007/s00406-022-01440-6
- Ghaemi, S. N. (2018). After the failure of DSM: Clinical research on psychiatric diagnosis. *World Psychiatry* 17, 301–302. doi: 10.1002/wps.20563
- Gillan, C. M., Fineberg, N. A., and Robbins, T. W. (2017). A trans-diagnostic perspective on obsessive-compulsive disorder. *Psychol. Med.* 47, 1528–1548. doi: 10.1017/S0033291716002786
- Glaesmer, H., Schulz, A., Häuser, W., Freyberger, H., Brähler, E., and Grabe, H.-J. (2013). Der childhood trauma screener (CTS)—entwicklung und validierung von schwellenwerten zur klassifikation. *Psychiatr. Praxis* 40, 220–226. doi: 10.1055/s-0033-1343116
- Gotra, M. Y., Hill, S. K., Gershon, E. S., Tamminga, C. A., Ivleva, E. I., Pearson, G. D., et al. (2020). Distinguishing patterns of impairment on inhibitory control and general cognitive ability among bipolar with and without psychosis, schizophrenia, and schizoaffective disorder. *Schizophr. Res.* 223, 148–157. doi: 10.1016/j.schres.2020.06.033
- Green, M. F., Horan, W. P., and Lee, J. (2019). Nonsocial and social cognition in schizophrenia: Current evidence and future directions. *World Psychiatry* 18, 146–161. doi: 10.1002/wps.20624
- Gross, J. J., and John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *J. Pers. Soc. Psychol.* 85, 348–362. doi: 10.1037/0022-3514.85.2.348
- Grove, T. B., Yao, B., Mueller, S. A., McLaughlin, M., Ellingrod, V. L., McInnis, M. G., et al. (2018). A Bayesian model comparison approach to test the specificity of visual integration impairment in schizophrenia or psychosis. *Psychiatry Res.* 265, 271–278. doi: 10.1016/j.psychres.2018.04.061
- Gur, R. C., and Gur, R. E. (2016). Social cognition as an RDoC domain. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* 171, 132–141. doi: 10.1002/ajmg.b.32394
- Gur, R. E., Moore, T. M., Calkins, M. E., Ruparel, K., and Gur, R. C. (2017). Face processing measures of social cognition: A dimensional approach to developmental psychopathology. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 2, 502–509. doi: 10.1016/j.bpsc.2017.03.010
- Henriksen, M. G., Raballo, A., and Nordgaard, J. (2021). Self-disorders and psychopathology: A systematic review. *Lancet Psychiatry* 8, 1001–1012. doi: 10.1016/S2215-0366(21)00097-3
- Henriksen, M. G., Škodlar, B., Sass, L. A., and Parnas, J. (2010). Autism and perplexity: A qualitative and theoretical study of basic subjective experiences in schizophrenia. *Psychopathology* 43, 357–368. doi: 10.1159/000320350
- Hirjak, D., Meyer-Lindenberg, A., Kubera, K. M., Thomann, P. A., and Wolf, R. C. (2018). Motor dysfunction as research domain in the period preceding manifest schizophrenia: A systematic review. *Neurosci. Biobehav. Rev.* 87, 87–105. doi: 10.1016/j.neubiorev.2018.01.011
- Hutto, D. D. (2016). Narrative self-shaping: A modest proposal. *Phenomenol. Cogn. Sci.* 15, 21–41. doi: 10.1007/s11097-014-9352-4
- Insel, T. R., and Cuthbert, B. N. (2009). Endophenotypes: Bridging genomic complexity and disorder heterogeneity. *Biol. Psychiatry* 66, 988–989.
- Jansson, L., and Nordgaard, J. (2016). *The psychiatric interview for differential diagnosis*. New York, NY: Springer International Publishing. doi: 10.1007/978-3-319-33249-9
- Kaplan, M. H., and Feinstein, A. R. (1974). The importance of classifying initial comorbidity in evaluating the outcome of diabetes mellitus. *J. Chron. Dis.* 27, 387–404. doi: 10.1016/0021-9681(74)90017-4
- Keshavan, M. S., Morris, D. W., Sweeney, J. A., Pearson, G., Thaker, G., Seidman, L. J., et al. (2011). A dimensional approach to the psychosis spectrum between bipolar disorder and schizophrenia: The Schizo-Bipolar scale. *Schizophr. Res.* 133, 250–254. doi: 10.1016/j.schres.2011.09.005
- Korošec Hudnik, L., Blagus, T., Redenšek Trampuž, S., Dolžan, V., Bon, J., and Pjevac, M. (2024). Case report: Avoiding intolerance to antipsychotics through a personalized treatment approach based on pharmacogenetics. *Front. Psychiatry* 15:1363051. doi: 10.3389/fpsyg.2024.1363051
- Kotov, R., Krueger, R. F., and Watson, D. (2018). A paradigm shift in psychiatric classification: The hierarchical taxonomy of psychopathology (HiTOP). *World Psychiatry* 17, 24–25. doi: 10.1002/wps.20478
- Kroenke, K., Spitzer, R. L., and Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *J. Gen. Intern. Med.* 16, 606–613. doi: 10.1046/j.1525-1497.2001.016009606.x
- Larsen, J. L., Johansen, K. S., and Mehlsen, M. Y. (2022a). What kind of science for dual diagnosis? A pragmatic examination of the enactive approach to psychiatry. *Front. Psychol.* 13:825701. doi: 10.3389/fpsyg.2022.825701
- Larsen, J. L., Johansen, K. S., Nordgaard, J., and Mehlsen, M. Y. (2022b). Dual case study of continued use vs cessation of cannabis in psychosis: A theoretically informed approach to a hard problem. *Adv. Dual Diagn.* 15, 22–36. doi: 10.1108/ADD-11-2021-0013
- Lilienfeld, S. O., and Treadway, M. T. (2016). Clashing diagnostic approaches: DSM-ICD versus RDoC. *Annu. Rev. Clin. Psychol.* 12, 435–463. doi: 10.1146/annurev-clinpsy-021815-093122
- Luhrmann, T. M., and Marrow, J. (2016). *Our most troubling madness: Case studies in schizophrenia across cultures*. Berkeley, CA: University of California Press.
- Maruish, M. E. (2000). *Handbook of psychological assessment in primary care settings*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Mittal, V. A., Bernard, J. A., and Northoff, G. (2017). What Can Different Motor Circuits Tell Us About Psychosis? An RDoC Perspective. *Schizophr. Bull.* 43, 949–955. doi: 10.1093/schbul/sbx087
- Moreno-Montoya, M., Olmedo-Córdoba, M., and Martín-González, E. (2022). Negative valence system as a relevant domain in compulsivity: Review in a preclinical model of compulsivity. *Emerg. Top. Life Sci.* 6, 491–500. doi: 10.1042/ETLS2022.0005
- Nelson, J. (2017). Using conceptual depth criteria: Addressing the challenge of reaching saturation in qualitative research. *Qual. Res.* 17, 554–570. doi: 10.1177/1468794116679873
- Nolen-Hoeksema, S. (2000). The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. *J. Abnorm. Psychol.* 109, 504–511. doi: 10.1037/0021-843X.109.3.504
- Nordgaard, J., Nielsen, K. M., Rasmussen, A. R., and Henriksen, M. G. (2023). Psychiatric comorbidity: A concept in need of a theory. *Psychol. Med.* 53, 5902–5908. doi: 10.1017/S0033291723001605
- Oblak, A., Boyadzhieva, A., and Bon, J. (2021). Phenomenological properties of perceptual presence: A constructivist grounded theory approach. *Constructiv. Found.* 16, 295–308.
- Oblak, A., Boyadzhieva, A., Caporusso, J., Škodlar, B., and Bon, J. (2022). How things take up space: A grounded theory of presence and lived space. *Qual. Rep.* 27, 2556–2582. doi: 10.46743/2160-3715/2022.5762

- Parnas, J. (2014). The RDoC program: Psychiatry without psyche? *World Psychiatry* 13, 46–47. doi: 10.1002/wps.20101
- Parnas, J., and Sass, L. A. (2001). Self, solipsism, and schizophrenic delusions. *Philos. Psychiatry Psychol.* 8, 101–120. doi: 10.1353/ppp.2001.0014
- Parnas, J., Möller, P., Kircher, T., Thalbitzer, J., Jansson, L., Handest, P., et al. (2005). EASE: Examination of anomalous self-experience. *Psychopathology* 38, 236–258. doi: 10.1159/000088441
- Peele, J. (2017). *Get out*. Universal City, CA: Universal Pictures.
- Petitmengin, C. (2006). Describing one's subjective experience in the second person: An interview method for the science of consciousness. *Phenomenol. Cogn. Sci.* 5, 229–269. doi: 10.1007/s11097-006-9022-2
- Pincus, H. A., Tew, J. D., and First, M. B. (2004). Psychiatric comorbidity: Is more less? *World Psychiatry* 3, 18–23.
- Pitman, R. K. (1987). Pierre janet on obsessive-compulsive disorder (1903): Review and commentary. *Arch. Gen. Psychiatry* 44:226. doi: 10.1001/archpsyc.1987.01800150032005
- Pjevac, M., and Korošec Hudnik, L. (2023). A case report—“When less is more”: Resistant inpatient reduction of anticholinergic burden in a patient with clozapine-resistant schizophrenia. *Front. Psychiatry* 14:1222177. doi: 10.3389/fpsyg.2023.1222177
- Quesque, F., and Brass, M. (2019). The role of the temporoparietal junction in self-other distinction. *Brain Topogr.* 32, 943–955. doi: 10.1007/s10548-019-00737-5
- Sabharwal, A., Szekely, A., Kotov, R., Mukherjee, P., Leung, H.-C., Barch, D. M., et al. (2016). Transdiagnostic neural markers of emotion–cognition interaction in psychotic disorders. *J. Abnorm. Psychol.* 125, 907–922. doi: 10.1037/abn0000196
- Samuels, J., Bienvenu, O. J., Krasnow, J., Wang, Y., Grados, M. A., Cullen, B., et al. (2017). An investigation of doubt in obsessive–compulsive disorder. *Compr. Psychiatry* 75, 117–124. doi: 10.1016/j.comppsyg.2017.03.004
- Sanislow, C. A. (2016). Updating the research domain criteria. *World Psychiatry* 15, 222–223.
- Sass, L. A., and Parnas, J. (2003). Schizophrenia, consciousness, and the self. *Schizophrenia Bull.* 29, 427–444. doi: 10.1093/oxfordjournals.schbul.a007017
- Sass, L., Pienkos, E., Skodlar, B., Stanghellini, G., Fuchs, T., Parnas, J., et al. (2017). EASE: Examination of anomalous world experience. *Psychopathology* 50, 10–54. doi: 10.1159/000454928
- Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends Cogn. Sci.* 17, 565–573. doi: 10.1016/j.tics.2013.09.007
- Silveira, ÉD. M., Passos, I. C., Scott, J., Bristot, G., Scotton, E., Teixeira Mendes, L. S., et al. (2020). Decoding rumination: A machine learning approach to a transdiagnostic sample of outpatients with anxiety, mood and psychotic disorders. *J. Psychiatr. Res.* 121, 207–213. doi: 10.1016/j.jpsychires.2019.12.005
- Smith, J. A., Flowers, P., and Larkin, M. (2022). *Interpretative phenomenological analysis: Theory, method and research*, 2nd Edn. Thousand Oaks, CA: SAGE.
- Smucny, J., Lesh, T. A., Newton, K., Niendam, T. A., Ragland, J. D., and Carter, C. S. (2018). Levels of cognitive control: A functional magnetic resonance imaging-based test of an RDoC domain across bipolar disorder and schizophrenia. *Neuropsychopharmacology* 43, 598–606. doi: 10.1038/npp.2017.233
- Stanghellini, G., and Mancini, M. (2017). *The therapeutic interview in mental health: A values-based and person-centered approach*, 1st Edn. Cambridge: Cambridge University Press, doi: 10.1017/9781316181973
- Stanghellini, G., and Rosfort, R. (2013). Borderline depression a desperate vitality. *J. Conscious. Stud.* 20, 7–8.
- Stengel, E. (1945). A study on some clinical aspects of the relationship between obsessional neurosis and psychotic reaction types. *J. Ment. Sci.* 91, 166–187. doi: 10.1192/bjp.91.383.166
- Storch, E. A., De Nadai, A. S., Conceição, Do Rosário, M., Shavitt, R. G., Torres, A. R., et al. (2015). Defining clinical severity in adults with obsessive–compulsive disorder. *Compr. Psychiatry* 63, 30–35. doi: 10.1016/j.comppsyg.2015.08.007
- Swinson, R. P. (2006). The GAD-7 scale was accurate for diagnosing generalised anxiety disorder. *Evid. Based Med.* 11, 184–184. doi: 10.1136/ebm.11.6.184
- Torrent, C., Martínez-Arán, A., Amann, B., Daban, C., Tabarés-SeisDedós, R., González-Pinto, A., et al. (2007). Cognitive impairment in schizoaffective disorder: A comparison with non-psychotic bipolar and healthy subjects. *Acta Psychiatr. Scand.* 116, 453–460. doi: 10.1111/j.1600-0447.2007.01072.x
- Tso, I. F., Lasagna, C. A., Fitzgerald, K. D., Colombi, C., Sripada, C., Peltier, S. J., et al. (2020). Disrupted eye gaze perception as a biobehavioral marker of social dysfunction: An RDoC investigation. *J. Psychiatry Brain Sci.* 5:e200021.
- Ustun, T. B., Kostanjsek, N., Chatterji, S., and Rehm, J. (2010). In *Measuring health and disability: Manual for WHO disability assessment schedule (WHODAS 2.0)*, eds T. B. Üstün, N. Kostanjsek, S. Chatterji, and J. Rehm (Geneva: World Health Organization), 88.
- Vargas, T. G., and Mittal, V. A. (2021). Testing whether implicit emotion regulation mediates the association between discrimination and symptoms of psychopathology in late childhood: An RDoC perspective. *Dev. Psychopathol.* 33, 1634–1647. doi: 10.1017/S0954579421000638
- Wardenaar, K. J., and De Jonge, P. (2013). Diagnostic heterogeneity in psychiatry: Towards an empirical solution. *BMC Med.* 11:201. doi: 10.1186/1741-7015-11-201
- Wigand, M. E., Lang, F. U., Müller-Stierlin, A. S., Reichardt, L., Trif, S., Schulze, T. G., et al. (2018). Psychosis is mutable over time: A longitudinal psychopathology study. *Psychopathology* 51, 186–191. doi: 10.1159/000486897
- Woody, E. Z., Hoffman, K. L., and Szechtman, H. (2019). Obsessive compulsive disorder (OCD): Current treatments and a framework for neurotherapeutic research. *Adv. Pharmacol.* 86, 237–271. doi: 10.1016/bs.apha.2019.04.003



OPEN ACCESS

EDITED BY

Luca Simone,
UNINT - Università degli studi Internazionali di
Roma, Italy

REVIEWED BY

Ken Mogi,
Sony Computer Science Laboratories, Japan

*CORRESPONDENCE

Zoran Josipovic
✉ zoran@nyu.edu

RECEIVED 17 June 2024

ACCEPTED 08 July 2024

PUBLISHED 23 August 2024

CITATION

Josipovic Z (2024) Reflexivity
gradient—Consciousness knowing itself.
Front. Psychol. 15:1450553.
doi: 10.3389/fpsyg.2024.1450553

COPYRIGHT

© 2024 Josipovic. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC
BY\)](#). The use, distribution or reproduction in
other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Reflexivity gradient—Consciousness knowing itself

Zoran Josipovic^{1,2*}

¹Department of Psychology, New York University, New York, NY, United States, ²Nonduality Institute, Woodstock, NY, United States

Some consider phenomenal consciousness to be the great achievement of the evolution of life on earth, but the real achievement is much more than mere phenomenality. The real achievement is that consciousness has woken up within us and has recognized itself, that within us humans, consciousness knows that it is conscious. This short review explores the reflexivity of consciousness from the perspective of consciousness itself—a non-conceptual nondual awareness, whose main property is its non-representational reflexivity. In light of this nondual reflexivity, different types of reflexivity proposed by current theories can be seen as a gradation of relational or transitive distances between consciousness as the knower and consciousness as the known, from fully representational and dual, through various forms of qualified monism, to fully non-representational and nondual.

KEYWORDS

reflexivity gradient, nondual reflexivity, awareness of awareness, consciousness itself, nondual awareness

Introduction

Much of the current research on consciousness could be summed up by a well-known metaphor: standing outdoors on a sunny day while facing away from the sun, one may see various objects and events in the environment as illuminated by the light reflected off them and mistakenly conclude that they themselves are the original source of that light. Similarly, insisting on explaining consciousness as something other than itself, as a phenomenal content, a cognitive or affective function, a state of arousal, or a conceptual structure, we fail to see consciousness itself. Admittedly, these aspects are parts of conscious experiencing, and a great deal of progress has been made in recent years in understanding them (Koch et al., 2016; Michel et al., 2019; Lepauvre and Melloni, 2021). However, consciousness itself or consciousness as such—a foundational awareness that is distinct from contents, functions, and states—is still insufficiently researched.

Attempts to include it within contemporary discourse on consciousness are slowly gaining traction in neuroscience and the philosophy of mind (Josipovic, 2014, 2019, 2021; Ricard and Singer, 2017; Dunne et al., 2019; Metzinger, 2020, 2024) but are often plagued by misunderstandings of what this conscious is. In this research, as I have done over a number of years (Josipovic, 2014, 2016, 2019, 2021; Josipovic and Miskovic, 2020), I have presented the perspective that consciousness itself is a type of awareness whose main property is its inherent, non-representational reflexivity—it knows that it is aware without needing mediation by mental representations.

Consciousness itself does not rely on mental representations to know either itself or what is present to it; thus, it is a different way of knowing from the usual conceptual mind that is based on mental representations. This awareness is nondual, both within itself and between itself and phenomena. It is nondual within itself because it knows itself without taking itself as an object of this knowing, and it is nondual between itself and phenomena because it knows phenomena without taking itself as a separate conceptually reified subject and phenomena as its objects. This awareness does not fragment experiences into reified dualities of subject vs. object, self vs. other, us vs. them, good vs. bad, and similar. Hence, it has been termed nondual awareness (Williams, 2000; Higgins, 2013; Josipovic, 2014, 2021; Laish, 2015; MacKenzie, 2015).

Although a singular presence, when nondual awareness is explicit, its inherent properties become self-evident, irrespective of whether they are subsequently conceptualized or not. These aspects have been discussed in other studies (Rabjam, 2007; Josipovic, 2019, 2021; Fasching, 2021) and are only listed here in four groups: (1) Being, Presence, Emptiness, and Spaciousness; (2) Cognitive Luminosity, and Reflexivity; (3) Bliss, Ecstasy, Universal Love, and Compassion; and (4) Singularity, Unity, and Self.

Since nondual awareness is singular and uncompounded, its dimensions are not separate elements from which awareness is assembled or from whose relationships or interactions it emerges.

When present explicitly in an experience, nondual awareness appears as distinct from any phenomenal content that is co-present with it, from the functions that create content, and from global states, as well as from unconscious substrate that structures ordinary concept-based experiencing. Nondual awareness appears as that which is, as has been, conscious or aware in any experience; in other words, it appears as consciousness itself or consciousness as such.

However, ordinarily, although present, this awareness is only implicit in an experience but can become explicit under special circumstances or due to certain contemplative and other practices. Therefore, for any experience, there is a gradient of how implicit or explicit nondual awareness is in that experience, and that gradient is orthogonal to the local content and global state (for a detailed discussion see Josipovic, 2021).

When nondual awareness is fully explicit during wakefulness, it is experienced as simultaneously transcendent and immanent in conscious states and contents. It is transcendent, as the silent aware space that pervades and encompasses the entire conscious experience, one's entire perceptual bubble, and it is immanent, as that out of which everything is made, the way water in a glass is both the medium in which ice cubes float and the substance out of which they are made (Josipovic, 2016, 2021).

As stated above, the main property of consciousness itself or nondual awareness is its inherent, non-representational reflexivity—it knows that it is aware without needing mediation by mental representations and without taking itself as an intentional object; hence, it is non-relational. This type of reflexivity can be termed as nondual reflexivity. It is unique to consciousness itself and is that which makes consciousness itself what it is. The implicit–explicit gradient of nondual awareness can be understood as the gradient of how evident nondual awareness is to itself, or in other words, as the gradient of its reflexivity (Josipovic, 2021).


I have discussed in detail the neural correlates of nondual awareness and non-representational nondual reflexivity previously (Josipovic, 2019, 2021). Presently, I will only briefly summarize them in order to further clarify the present discussion.

Although isomorphism between phenomenal and neural levels should not be presumed, neither should it be rejected *a priori*. Since nondual awareness is phenomenally and functionally distinct from attention, monitoring, working memory, evaluation, and decision, i.e., from processes that contribute to constructing perceptual, affective, and cognitive contents and from those that determine global states, its neural correlate could likewise be distinct—a dedicated network with its characteristic dynamics. A neural correlate of nondual awareness needs to be able to function with both low and high levels of arousal and amounts of content and serve as the integrative conscious space within which both intrinsic and extrinsic contents can co-occur.

I have previously proposed that the central precuneus network with a self-sustaining oscillatory resonant dynamic regime is the most likely neural correlate of nondual awareness (Josipovic, 2014, 2019, 2021). This functional network links the central precuneus with the dorsolateral prefrontal cortex, the dorsal anterior cingulate cortex, the dorsomedial prefrontal cortex, and the inferior parietal lobule (Cavanna and Trimble, 2006; Margulies et al., 2009; Cunningham et al., 2017; Buckner and DiNicola, 2019). The central precuneus is unique among different subdivisions of the precuneus as it can functionally connect with both the intrinsic (default mode network) and the extrinsic (dorsal attention network and executive control network) systems (Li et al., 2019). This finding corresponds to a major function of nondual awareness in increasing the integration of intrinsic self-related and extrinsic environment-related aspects of experience (Josipovic et al., 2012; Josipovic, 2014). This role functionally differentiates the neural correlate of nondual awareness from the neural correlate of monitoring, which is associated with networks for salience detection and involuntary attention, whose effect is to induce switching between the intrinsic and extrinsic systems in the brain and increase their functional segregation (Josipovic, 2010, 2013, 2014, 2019). Like other cortical networks, the central precuneus network is reciprocally connected to the subcortical nuclei of the reticular activating system that supply arousal and to the thalamic nuclei that enable its cortical organization (Tomasi and Volkow, 2011). However, these subcortical areas, although necessary, are not sufficient by themselves for nondual awareness.

When nondual awareness is explicit during normal wakefulness and its inherent reflexivity is vividly present, the pre-frontal nodes of its network, the dorsolateral prefrontal cortex in particular, function to add the necessary amplitude and persistence to the network-wide resonance and coherence (Helfrich and Knight, 2016; Schmidt et al., 2018). On the other hand, it is possible, especially at times of minimized phenomenal content, that an increased level of functional integration or recursive feedback in the posterior nodes of the central precuneus network is alone sufficient to establish sustained oscillatory resonance, in which neurons inform each other about their excitation levels, or in other words, their information processing capacity, without processing any other content, which is experienced as inherent, non-representational nondual reflexivity. Furthermore, a neural network informing itself

TABLE 1 Reflexivity gradient—consciousness knowing itself.

| Implicit | Transitional | Explicit |
|---|--|----------------------|
|  | | |
| Dual | Qualified monist | Nondual |
| Representational | Representational or non-representational | Non-representational |
| Conceptual | Conceptual or non-conceptual | Non-conceptual |
| Relational transitive | Relational or non-relational | Non-relational |

Reflexivity gradient with three zones indicating transitive or relational distance: dual, qualified monist, and nondual, corresponding to the three zones of the implicit-explicit gradient of nondual awareness: implicit, transitional, and explicit. Dualistic reflexivity is representational, conceptual, and transitive, with the different-order knower and known. Qualified monist reflexivity can be either representational or non-representational, conceptual or non-conceptual, and relational or non-relational, with the same-order knower and known. Nondual reflexivity is non-representational, non-conceptual, and non-relational, without the knower-known structure.

about its capacity to process information can be instituted in a relatively simple electronic circuit, without any sign of awareness or consciousness. Hence, the biological constraints on a system’s capacity for consciousness apply here and, even more importantly, for nondual awareness that requires a human-level brain (Josipovic, 2021).

In light of nondual reflexivity, different types of reflexivity proposed by current theories can be seen as a gradation of relational or transitive distances between consciousness as the knower and consciousness as the known, ranging from fully representational and dual, through various forms of qualified monism, to fully non-representational and nondual. The types of reflexivity are shown in Table 1.

Reflexivity theories

Theories of reflexivity have been previously grouped broadly into two types, based largely on how they view the nature and the role of representation in consciousness (Siewert, 2022). The first type can be termed the mental representational or cognitive-analytic type, and it holds that, for an experience or a mental state to be conscious, it has to be represented by another state that is different from it (Rosenthal, 2004, 2012; Gennaro, 2013). This idea is known as the transitivity principle (ibid.) and indicates a relationship between two kinds of mental representations, those representing the state itself, that is, what we are conscious of, and those re-representing them, that is, how we are conscious of it. In cases of reflexivity, this transitivity principle necessitates a third-order representation, re-representing the second one (Rosenthal, 2012).

The other type of reflexivity theories are the phenomenological theories, which can be intentional representational, or non-intentional (Zahavi, 2005; Montague, 2016; Gallagher, 2022; Strawson, 2022). Here, representation is understood as phenomenal intention of conscious states, their about-ness. Reflexivity is seen as a more immediate self-knowing that accompanies most, if not all, conscious states and is pre-reflectively “given” with experience, not requiring reflection or mental re-representation; hence, it is

not explicitly transitive. For most phenomenologists, consciousness knowing itself is pre-reflective self-knowing, which is understood as not explicitly conscious but as nevertheless present and enabling all conscious experiences.

From the viewpoint of nondual awareness, these two types of theories can be seen as reflecting the two seemingly contradictory aspects of nondual awareness: its transcendence and its immanence. In their claim that representations that give rise to conscious knowing are of a different order than those that represent what we are conscious of, representational theories reflect the transcendent aspect of nondual awareness, the fact that consciousness itself is distinct from all other aspects of experience, the way space is different from everything in it. In their claim that reflexive knowing is intrinsic to experience, phenomenological theories reflect the immanent aspect of consciousness itself, where nondual awareness appears as the substance out of which everything in experience is made, the way water is the substance out of which ice cubes that float in it are made.

Representational and phenomenological theories also differ in terms of the epistemic distance they propose between the knower and the known. Representational theories are more indirect in that re-representations needed for conscious knowing, in general, and especially for reflexivity, in particular, are of an entirely different order from those needed to represent that which is known. Conscious knowing is seen as a relational property conferred onto the known by these higher-order representations (Lycan, 2023). Phenomenological theories, on the other hand, are more direct. They reject the idea that the states that one is conscious of are objects of consciousness. Likewise, they do not agree that, in reflexivity, consciousness is conscious of itself as its object (Zahavi, 2005, 2018). In other words, they do not accept the epistemic dualism of the subject as the knower and the object as the known. Instead, they propose that what makes an experience, in general, conscious is intrinsic to that conscious experience and that reflexivity is an inherent and even defining property of consciousness itself, requiring neither a higher-order nor same-order representation (Gallagher and Zahavi, 2023).

Dual representational

Strong transparency

Reductive representationalism sees consciousness as entirely reducible to mental representations (Lycan, 2023). When such representations related to the subject and object are strongly reified, phenomenal consciousness can appear to be just an illusion and consciousness knowing itself, an impossibility. The idea that the mind cannot know itself is an old one, appearing first in the Brihadaranyaka Upanishad (Radhakrishnan, 1994) and later in various Buddhist sutras (Luk, 2001). Briefly, it could be said that it applies only to the impossibility of knowing consciousness itself via dualistic conceptual thinking (Sansk. vijñana) but does not hold true for knowing more directly via intuitive awareness (Sansk. prajñana), or in other words, via intrinsic reflexivity of consciousness itself.

In Western philosophy, ideas about the impossibility of the mind knowing itself have been expressed most clearly by Hume (1978) and more recently by Tye (2014) and Dennett (1987), and in the context of cognitive neuroscience, the ideas have been expressed by Frankish (2016). The more recent views are sometimes referred to as strong transparency arguments or as reductive representationalism (Dretske, 1995; Tye, 2014).

According to these, any attempt to introspect consciousness finds only properties of objects in external or internal environments, but not the actual phenomenal qualities of experience, nor the consciousness itself, since on this account, representational processes such as spatial perspective, body-ownership, and agency, that are involved in the minimal or core self and confer these properties to subjective phenomenality, are unconscious and not available to introspection. In the well-known metaphor, they are like a highly transparent windowpane that one looks through, but which one does not itself see (Metzinger, 2010).

The strong transparency thesis has been argued against extensively by many (Zahavi, 2005; Montague, 2016; Chalmers, 2020), so I will not explore those arguments in the present discourse. I will only make a couple of points from the perspective of nondual awareness, and contemplative practice more generally. These offer two different ways to notice the usually unconscious processes involved in constructing minimal self experience or even the homeostatic proto-self identity, in addition to relatively common 'seeing-through' and de-constructing of various extended and narrative self-models. One is through developing ability to sustain focused attention for prolonged periods of time, resulting in various absorption states where, for however brief periods, there is cessation of these minimal self processes, followed by their reappearance once one emerges from the absorption state (Josipovic and Miskovic, 2020; Metzinger, 2024). Alternatively, once nondual awareness is discovered and stabilized, one can, at times, become aware of such processes because this awareness is, phenomenally, the most subtle aspect of conscious experience. These and other processes involved in constructing the self and environment then appear to it as contents in its epistemic space (Josipovic, 2021).

The claimed inability of introspection to find anything other than external phenomenal contents under ordinary dualistic cognition (Tye, 2014; Montague, 2016) is, in part, due to attention being habitually oriented toward finding and attending to an object. In other words, the abovementioned claim is due to the inability to sufficiently turn the attention around to attend to awareness itself (Josipovic and Miskovic, 2020). This "turning around" is not some act of permanent contortion but is meant to instigate a collapse of the dualistic attending and monitoring into just being aware and, in doing so, reveal nondual awareness—consciousness itself—as already present in one's experience.

Dual representational

Higher order

Unlike reductive representation theories, the less reductive or even non-reductive representational theories allow for

the possibility of consciousness knowing itself (Rosenthal, 2004; Carruthers and Gennaro, 2023). In terms of transitivity distance, these theories are, at least in their main forms, also strongly dualistic.

According to higher-order theories, conscious experiencing is possible because the first-order representations of some contents or states, which are themselves unconscious, are re-represented by certain higher-order representations that are different from them (Rosenthal, 2004, 2012). In other words, the first-order representations are the objects to which higher-order representations are directed. Higher-order representations are generally understood as enabling access consciousness or as being equivalent to it (Block, 2007) or as being a function of monitoring (Brown et al., 2019; Lycan, 2023). The phenomenal properties of experience are believed to be the semantic properties of these higher-order representations (Siewert, 2022). The most recent version of a higher-order theory, Brown's higher-order representation of representation theory as applied to emotions by LeDoux (Brown et al., 2019; LeDoux, 2024), points to a hierarchy of higher-order representations, which results in a gradation of conscious experience from pre-conscious to fully conscious, i.e., from anoetic to noetic and autonotic.

Higher-order representations responsible for the conscious state or inner awareness are themselves unconscious, non-inferential, and not available for direct introspection (Rosenthal, 2004, 2012). However, when reflecting on one's experience, such as during confidence judgments (Webb et al., 2023), inferential decisions (Fleming, 2020), or conceptual introspection (Carruthers and Gennaro, 2023), they or their re-representations become fully conscious metacognition. Reflexivity, and especially being aware that one is aware, is then due to a third-order re-representation of those higher-order representations. According to this view, it occurs only in conscious introspection that requires focusing attention on some conscious states (Rosenthal, 2012). From the perspective of nondual awareness, the necessity of a third-order re-representation for awareness of awareness seems like an obvious indication that such conceptual processes cannot be the mechanism of inherent reflexivity of consciousness itself. Since nondual reflexivity is, so to speak, immediate, as an inherent property of awareness, and is non-conceptual and non-transitive, phenomenally, it is very different from attending as a subject to awareness as an object of one's conceptual introspecting (Josipovic, 2019, 2021).

Some higher-order theorists have proposed that, in a transition to conscious metacognitive states, higher-order representations themselves shift from being unconscious to being conscious (Gennaro, 2013). This shift has raised questions of how conscious experience can come from two equally unconscious representations; how an infinite regression of re-representations can be avoided; and what causes higher-order representations to shift from being unconscious to being conscious (Zahavi, 2005; Kriegel, 2009; Montague, 2016).

Different variants of higher-order theories could be seen, in addition to their main differences as also differing in terms of transitivity distance between their higher-order representations and that which they represent. For example, higher-order perception theories for which higher-order representations are

perception-like outputs of internal monitoring could be considered less dualistic than the assertoric meta-thoughts of a higher-order thought theory (Carruthers and Gennaro, 2023). Similarly, a higher-order global state theory for which a higher-order representation is a global self-world representational state, which encompasses first-order representations, could be seen as arguably less dualistic (Van Gulick, 2004). The wide intrinsicity view (Gennaro, 2013; Cole, 2014) in which a higher-order representation is intrinsic to the first-order state it represents is even less dualistic and, in the intrinsicity claim, it begins to resemble those qualified monist theories of reflexivity that reject the higher-order premise altogether.

Qualified monist

Self-representation

Discomfort with the dualistic transitive distance between higher-order representations and what they re-represent can be seen as motivating the same-order representational theories that are both representational and relational but claim that consciousness knowing itself is a special kind of relationship. Hence, they could be termed as qualified monist theories (Kriegel, 2009; Montague, 2016; Strawson, 2022).

Self-representation theory (Kriegel, 2009) claims that a mental state that is conscious represents itself but is one with that self-representation. Conscious experience is then seen as an integration of representations for content properties—what is being conscious, with representations for subjective character—representing-as-occurring-now-in-me (Kriegel, 2024). Self-representation responsible for the reflexive property of experience, also known as inner awareness, is not a type of thought or a type of perception resulting from monitoring, as higher-order theories claim, but a unique kind of representation that is more intimate. In other words, self-representation responsible for the reflexive property of experience is less dualistic and yet still relational as consciousness takes itself as its object. Furthermore, on this view, it is only in virtue of such self-representations that a state is phenomenally conscious (Kriegel, 2024).

Same-order approaches have been criticized on similar grounds as the higher-order ones, that they still contain the problem of how to make the two representations, for content properties and for subjective character, one unified experience (Zahavi, 2018). In addition, intentional representations are seen as not being able to reference themselves reflexively since the direction of intentionality of consciousness, according to phenomenological orthodoxy, is always away from itself and toward something other than itself (Peters, 2013).

Qualified monist

Objectivist

Reflexivity theories in this group are largely phenomenological and perceive reflexivity as representational and relational; unlike dualist or qualified dualist, they think of reflexive act as being

“implicit” in, or given with, any conscious experience (Montague, 2016; Strawson, 2022). Theories in this group hold that reflexivity does not require consciousness to consider itself a separate object of introspective reflection in order to know itself. This further step toward decreasing the distance between consciousness as the knower and consciousness as the known could be seen as a jump to a different level of cognitive intimacy from the previous ones, as here, the two, though still different, are given together within one experience, as a symbiosis of sorts. Hence, these theories can be thought of as different versions of qualified monism.

According to an early version of this view attributed to Brentano (Gallagher, 2022) and other similar same-order representation views, any experience is constituted by two simultaneous components: awareness of its content, whether perceptual, affective, or cognitive; and an awareness that perceiving, feeling, etc., is occurring, or in other words, a reflexive awareness that one is having a particular conscious experience. These two make one mental state and one unified experience. In terms of intentionality, this implies that, within a single experience, intention is divided into two co-occurring and related intentions, one oriented toward the content and the other directed reflexively toward itself. These two intentional targets have been considered the primary and the secondary objects of consciousness, giving these views their characteristic objectivist orientation (Gallagher and Zahavi, 2023). With respect to reflexivity, they also indicate a relational gap between awareness as the knower and awareness as the known.

By defining representation very broadly as intentionality or about-ness with respect to anything that is experienced and side-stepping the older argument over whether intentionality is noetic or noematic, contemporary interpretations of the above view of reflexivity (Montague, 2016) claim that all phenomenal contents are representational, in the sense that, with any conscious experience, there is something that is being experienced. In the same spirit, reflexivity is also seen as representational and relational since, minimally, it is about being aware that one is aware (Montague, 2017). This view then allows for the redefining of the relational gap between awareness as representation and awareness as the one that is being represented. This is something that has posed a problem for early phenomenologists such as Brentano and Husserl (Zahavi, 2005). The claim is then that reflexivity, which is given with any experience, does not contain a subject–object gap between the knower and the known. Nevertheless, by insisting on the relational nature of reflexivity, they could not come any closer to an explanation of it than to restate a view held by the self-representation theory, which attributes this absence of gap to awareness’ relation to itself being somehow special due to the uniqueness of consciousness (Montague, 2017).

From the viewpoint of nondual awareness or consciousness as such, these observations can be regarded as accurate intuitions arising from awareness itself but which are then being distorted by unconscious conceptual reifications and relational concepts. In other words, at a representational level, being conscious that we are consciously experiencing is a derivative of the inherent non-representational reflexivity of consciousness itself. As a result, these theories fall short in explanatory power as they do not yet recognize non-representational nondual awareness as foundational

consciousness or consciousness itself for which reflexivity is not a relation but its intrinsic property.

Without discovering nondual awareness and its non-conceptual and non-representational mode of knowing, it will remain difficult for an objectivist approach to understand how one can know the nature of reflexivity since a more direct reflexivity cannot be conceptually introspected as a separate state or object (Montague, 2016; Josipovic, 2019). Similarly, without stabilizing nondual awareness in an ordinary waking experience, it can be difficult to see how it is possible to experience the properties of awareness itself, including nondual reflexivity, as distinct from the phenomenal properties of perceptual and other contents. Since nondual awareness can, in principle, pervade and encompass all other aspects of experience, including conceptual processes, ordinary introspection can occur within the epistemic space of nondual awareness as just another type of a cognitive event.

Qualified monist

Subjectivist

Reflexivity theories in this group see reflexivity as non-relational and non-intentional, and instead, as a property of conscious subject or some minimal phenomenal self that is present in any experience (Zahavi, 2005, 2018, 2024; Gallagher, 2022; Marchetti, 2024). They argue that any intentional stance toward consciousness is necessarily objectifying and that one is already pre-reflectively self-aware without having to become one's intentional object (Zahavi, 2005; Frank, 2007). Furthermore, any objectifying intentional conscious state is claimed to have an underlying pre-reflective, non-relational reflexivity that makes it possible. It can then be said that these views are basically monist as they view all experience as subjective experience.

Unlike noematic intentionality, reflexivity in these more recent subjectivist theories is non-perspectival and does not require an observational distance and perspective from which a subject is witnessing experience and awareness (Zahavi, 2005). Instead of intentionality, they propose that self-awareness implies identity and that consciousness is intrinsically self-aware.

A question has been raised that, if self-awareness is non-representational, how is it then instantiated (Montague, 2016)? One view sees intrinsic reflexivity not only as non-intentional and non-relational but also, somewhat contradictorily, as dependent on mental representations, for example, as a schema of a system's capacity to represent (Peters, 2013). Another view sees self-awareness as having a unique temporal structure, one that is distinct from that of intentional consciousness and especially from that of fully conceptual reflective introspection. Husserl (1913) termed this structure as an impression-retention-protention structure, indicating something akin to an experience enabled by short-term memory only, which can track, retain, and make predictions or expectations over short time scales (Zahavi, 2003). Pre-reflective self-awareness is, on this view, unified with whatever phenomena appear with it in an experience.

As previously mentioned, these theories correctly intuit the immanence of consciousness itself in and as experience. However,

they do not see its transcendence and, hence, do not understand its space as the unchanging ground of being. This problem is in part due to the implicit serial view of experiencing. Temporal views always involve some, however subtle, subject-object dualities of the attender and the attended, and the observer and the observed, where the observed stream of consciousness unfolds as a series of successive events. Reflexivity, even if only viewed as pre-reflective self-awareness, is seen as a temporal process that unfolds over a time span, however short in duration. In contrast, consciousness itself or nondual awareness is atemporally present, but this should not be understood from a temporal perspective as that would lead to the impossibility of instantaneousness. Rather, the correct perspective here is spatial, as nondual awareness is present to itself all at once, the way space is phenomenally an all-encompassing steady background within which things and events occur (Blackstone, 2007; Josipovic, 2019, 2021). Therefore, in respect to phenomena, it does not have a separate attender who, for example, attends to a melody. Rather, a melody simply occurs within its space. In addition, since nondual awareness knows by merely mirroring, there is no transitive distance between this awareness and the phenomena that appear to it, akin to the way images that are reflected in a mirror are different from the mirror but are not separate from it (Josipovic, 2021). Furthermore, the impression-retention-protention structure indicates a certain degree of mental representations, which are not intrinsic to consciousness itself but can co-occur within it as structures and events within its intrinsically empty epistemic space.

Nondual reflexivity

Nondual reflexivity is the inherent property of consciousness itself and entirely non-representational and non-intentional, or in other words, nondual. Consciousness itself as nondual awareness knows that it is aware without needing mediation by mental representations and without taking itself as an intentional object (Rabjam, 2001; Josipovic, 2019). Nondual reflexivity is the essential property of consciousness itself that makes it what it is (Josipovic, 2019; Josipovic and Miskovic, 2020).

As this awareness is nondual, it cannot not take itself as its object nor can it be something that a separate subject possesses as a capacity. Rather, it knows itself by being itself, through its self-presencing or self-disclosing (Guenther, 1984; Manjusrimitra, 2001). With respect to phenomena, nondual awareness functions like a mirror, merely "mirroring" what is present in experience, without categorizing, labeling, associating, evaluating, deciding, etc. (Rabjam, 2001; Norbu, 2013; Josipovic, 2019; Josipovic and Miskovic, 2020).

Nondual awareness is in itself entirely without both conceptual and non-conceptual representations. Hence, it is entirely silent and unmoving like empty space. It does not make any utterances about itself or anything else. Just as space is more subtle than all things in it, this awareness is also more subtle than all phenomenal contents and global states that co-occur with it (Rabjam, 2001; Josipovic, 2019, 2021; Metzinger, 2024).

The idea that awareness is always an awareness of something, and therefore, necessarily intentional and relational, can be seen

as being based on the unconscious semantic structuring of cognition into a subject and object and on the misidentification of foundational nondual awareness with a conceptually reified subject who is attending to and monitoring contents and states (Higgins, 2013; Josipovic, 2014, 2021). Within the aware epistemic space of nondual awareness, which is non-preferential or choice-less, even very subtle effortless monitoring is an intentionality toward optimization. Similarly, however effortless, attending is a selection. These processes are not intrinsic to nondual awareness itself (for a more detailed discussion see Josipovic, 2019).

For nondual awareness, which is in itself homogenous and singular and merely mirrors any contents and states, the immediacy of its knowing does not make any distance for there to be an intentional relation. Since the reflexivity of this awareness is its inherent property, rather than a result of some function, both its reflexivity and its mirroring of contents and states are single nondual experiencing.

Furthermore, within nondual experiencing, the essential properties of nondual awareness such as being, emptiness, and luminosity also appear as the universal properties of phenomena that co-occur with it, in addition to their specific properties (Rabjam, 2001, 2007; Josipovic, 2019, 2021). In this sense, phenomena are not different from nondual awareness within which they occur, and knowing them does not require for this awareness to be intentionally related to something other than itself or to abandon its reflexivity.

Owing to the nondual nature of its reflexivity, nondual awareness cannot be mistaken about itself. However, an inferential *a posteriori* belief, or any learned *a priori* belief, can be mistaken about it. In particular, one can hold a mistaken belief that one has realized nondual awareness when one has not yet. Since

this awareness is present in every conscious experience even when only implicit, when any less dualistic state is experienced, one can easily have a sense that this is nondual awareness. Then, upon nondual awareness' reflexivity activating clearly, one retrospectively understands that one was mistaken and yet, paradoxically, also knows that this was that which was aware in every experience. With the practice of "abiding as nondual awareness" over time, this awareness becomes revealed in greater depth in terms of its being unique and distinct from the more subtle layers of perceptual, affective, and cognitive constructs and from the various subtle states of consciousness (Manjusrimitra, 2001; Rabjam, 2001; Josipovic, 2019; Josipovic and Miskovic, 2020).

When explicit, the nondual reflexivity of consciousness itself is very subtle and quiet, empty and immediate, and completely intimate, without distance (Rabjam, 2001; Josipovic, 2019; Metzinger, 2024). The progressive loss of this reflexivity has been identified by certain nondual contemplative traditions as the epistemic cause of the sense of a separate self (Rabjam, 2001, 2007; Higgins, 2013). It progresses in a way described in the following: as the luminosity of nondual awareness intensifies, it first becomes very vivid, outshining, and obscuring its other dimensions. With further intensification, it creates a very subtle sense of self as "one who is aware." Then, it develops a subtle duality between itself and what is present within its epistemic space, which then becomes its subtle object. This split contains the seed of the subject-object conceptual structure, which eventually replaces the knowing via mirroring with knowing via mental representations and constitutes the loss of awareness' inherent nondual reflexivity—its capacity to directly know itself. It now mistakes itself as a conceptual subject and phenomena as its conceptualized objects. In ordinary experiencing, this foundational conceptual structure is

TABLE 2 Reflexivity and representation gradient (see text below).


| Nondual awareness | | Reflexivity | Reification | Representation |
|-------------------------------|---|-------------------------------|-------------|---|
| Implicit |  | Dual | Coarse | Higher-order reductive conceptual |
| Transitional | | Qualified Monist objectivist | Subtle | Same-order non-reductive conceptual or non-conceptual |
| | | Qualified Monist subjectivist | Very subtle | Same-order or non-representational |
| Explicit | | Nondual | Empty | Non-representational |
| Radiance of nondual awareness | | | | |

Table depicting the "radiating" of nondual awareness through the layers of representations.

reified with layers of representations, associations, and further re-representations (Guenther, 1984; Metzinger, 2010; Josipovic, 2016, 2021).

Phenomenally, nondual awareness is the most essential self in that it is who or what is conscious or aware in any experience, while the other levels of self, such as proto, core-minimal, and extended self, to the extent that they are conscious, appear to it as its contents. However, because of its nondual way of knowing, this awareness is not a self in the usual sense of self as separate from non-self. Since nondual awareness has no preference for what content or state unfolds within it, it is not the self as the one who is attending, monitoring, or recollecting, prompted by conscious and unconscious motivations. As nondual awareness is complete in itself, its bliss is the final reward, so it has no motivation other than itself. This has been expressed in an old saying that the ultimate goal of all doing is being, and being is fully revealed only in nondual awareness. In encounters with others, nondual awareness mirrors how implicit or explicit this awareness is in another, and expressing this may be experienced by another as helpful, liberating, or merely annoying.

Discussion

Different theories of reflexivity discussed in this research can be seen not only as different mutually exclusive theories but also as different types of reflexivity determined by the gradient of how implicit or explicit nondual awareness is in any experience, the degree to which consciousness itself is self-evident to itself. When nondual awareness is implicit, consciousness knows itself only indirectly through reified representations of the subject and object, and reflexivity is indirect and dualistic. When it is transitional, reflexivity is less representational and different types of reflexivity can occur, with a progressively decreasing relational distance; therefore, these views can be thought of as versions of qualified monism theories. Finally, when nondual awareness is explicit, reflexivity is fully non-representational and non-relational and self-evident as the property of consciousness itself.

On the view that consciousness itself as nondual awareness is always present in an experience irrespective of how implicit or explicit it is, different types of reflexivity discussed in this research could also be seen as a structure with a gradation of conceptual reifications, from coarse dualistic to very subtle monistic close to consciousness itself. Nondual reflexivity, as the non-conceptual primordial knowing, shines through and is reflected in these layers of conceptualizations as different types of reflexivity.

In light of this, Table 1 can be re-organized as Table 2.

When conceptual knowing occurs within explicit nondual awareness, it is both encompassed and pervaded by it, and therefore, it is not different in its essential properties from

awareness itself (Norbu, 1987; Josipovic, 2021). At the same time, it retains its relative properties in the hierarchy of concepts, where different types of concepts have differentiable relations to consciousness itself (Singh, 1989; Pruiett, 2016). This is because both the expressions of the nondual authentic being and the expressions of the dualistic self-other concept structure are equally pervaded by the space of nondual awareness.

Some consider phenomenal consciousness to be the great achievement of the evolution of life on earth but the real achievement is much more than mere phenomenality. It is that consciousness has woken up within us and has recognized itself, that within us humans, consciousness knows that it is conscious.

Author contributions

ZJ: Conceptualization, Writing—original draft, Writing—review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Funding for this research was provided through a private grant to Nonduality Institute.

Acknowledgments

The author wishes to thank all the individuals who inspired this research.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Blackstone, J. (2007). *The Empathic Ground: Intersubjectivity and Nonduality in the Psychotherapeutic Process*. Albany, NY: SUNY Press.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behav. Brain Sci.* 30, 481–548. doi: 10.1017/S0140525X07002786

- Brown, R., Lau, H., and LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *TICS* 9, 754–768. doi: 10.1016/j.tics.2019.06.009
- Buckner, R. L., and DiNicola, L. M. (2019). The brain's default network: updated anatomy, physiology and evolving insights. *Nat. Rev. Neurosci.* 20, 593–608. doi: 10.1038/s41583-019-0212-7
- Carruthers, P., and Gennaro, R. (2023). “Higher-order theories of consciousness,” in *The Stanford Encyclopedia of Philosophy*, eds. E. N. Zalta, and U. Nodelman. Available at: <https://plato.stanford.edu/archives/fall2023/entries/consciousness-higher/> (accessed September 3, 2023).
- Cavanna, A. E., and Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain* 129, 564–583. doi: 10.1093/brain/awl004
- Chalmers, D. (2020). Debunking arguments for illusionism about consciousness. *J. Conscious. Stud.* 27, 258–281.
- Cole, D. (2014). Rocco Gennaro: the consciousness paradox: consciousness, concepts and higher-order thoughts. *Minds Mach.* 24, 227–231. doi: 10.1007/s11023-014-9337-7
- Cunningham, S. I., Tomasi, D., and Volkow, N. D. (2017). Structural and functional connectivity of the precuneus and thalamus to the default mode network. *Hum. Brain Mapp.* 38, 938–956. doi: 10.1002/hbm.23429
- Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA: The MIT Press.
- Dretske, D. (1995). *Naturalizing the Mind*. Cambridge, MA: The MIT Press.
- Dunne, J. D., Thompson, E., and Schooler, J. (2019). Mindful meta-awareness: sustained and non-propositional. *Curr. Opin. Psychol.* 28, 307–311. doi: 10.1016/j.copsyc.2019.07.003
- Fasching, W. (2021). Prakasa. A few reflections on the Advaitic understanding of consciousness as presence and its relevance for philosophy of mind. *Phenomenol. Cogn. Sci.* 20, 679–701. doi: 10.1007/s11097-020-09690-2
- Fleming, S. M. (2020). Awareness as inference in a higher-order state space. *Neurosci. Conscious.* 2020:niz020. doi: 10.1093/nc/niz020
- Frank, M. (2007). Non-objectal subjectivity. *J. Conscious. Stud.* 14, 152–173.
- Frankish, K. (2016). Illusionism as theory of consciousness. *J. Conscious. Stud.* 23, 1–22.
- Gallagher, S. (2022). *Phenomenology*, 2nd Edn. London: Palgrave-Macmillan.
- Gallagher, S., and Zahavi, D. (2023). “Phenomenological approaches to self-consciousness,” in *The Stanford Encyclopedia of Philosophy*, eds. E. N. Zalta, and U. Nodelman. Available online at: <https://plato.stanford.edu/archives/win2023/entries/self-consciousness-phenomenological/> (accessed January 20, 2024).
- Gennaro, R. (2013). Defending HOT theory and the wide intrinsicality view: a reply to Weisberg, Van Gulick, and Seager. *J. Conscious. Stud.* 20, 82–100.
- Guenther, H. V. (1984). *Matrix of Mystery*. Boulder, CO: Shambhala.
- Helfrich, R. F., and Knight, R. T. (2016). Oscillatory dynamics of prefrontal cognitive control. *Trends Cogn. Sci.* 20, 916–930. doi: 10.1016/j.tics.2016.09.007
- Higgins, D. (2013). *The Philosophical Foundations of Classical Dzogchen in Tibet - Investigating the Distinction between Dualistic Mind (sems) and Primordial Knowing (ye shes)*. Wien: Arbeitskreis für Tibetische und Buddhistische Studien.
- Hume, D. (1978). “A treatise of human nature,” in *rev. P.H. Nidditch*, ed. L. A. Selby-Bigg (Oxford: Oxford University Press).
- Husserl, E. (1913). *Ideas: General Introduction to Pure Phenomenology*. Trans. W.R. Boyce, Gibson. New York, NY: Collier Books.
- Josipovic, Z. (2010). Duality and nonduality in meditation research. *Conscious. Cogn.* 19, 1119–1121. doi: 10.1016/j.concog.2010.03.016
- Josipovic, Z. (2013). Freedom of the mind. *Front. Psych.* 4:538. doi: 10.3389/fpsyg.2013.00538
- Josipovic, Z. (2014). Neural correlates of nondual awareness in meditation. *Ann. N. Y. Acad. Sci.* 1307, 9–18. doi: 10.1111/nyas.12261
- Josipovic, Z. (2016). Love and compassion meditation: a nondual perspective. *Ann. N. Y. Acad. Sci.* 1373, 65–71. doi: 10.1111/nyas.13078
- Josipovic, Z. (2019). Nondual awareness: consciousness-as-such as non-representational reflexivity. *Prog. Brain Res.* 244, 273–298. doi: 10.1016/bs.pbr.2018.10.021
- Josipovic, Z. (2021). Implicit-explicit gradient of nondual awareness or consciousness as such. *Neurosci. Conscious.* 2021:niab031. doi: 10.1093/nc/niab031
- Josipovic, Z., Dinstein, I., Weber, J., and Heeger, D. J. (2012). Influence of meditation on anti-correlated networks in the brain. *Front. Hum. Neurosci.* 5:183. doi: 10.3389/fnhum.2011.00183
- Josipovic, Z., and Miskovic, V. (2020). Nondual awareness and minimal phenomenal experience. *Front. Psychol.* 11:2087. doi: 10.3389/fpsyg.2020.02087
- Koch, C., Massimini, M., Boly, M., and Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nat. Rev. Neurosci.* 17, 307–321. doi: 10.1038/nrn.2016.22
- Kriegel, U. (2009). *Subjective Consciousness: A Self-Representational Theory*. Oxford: Oxford University Press.
- Kriegel, U. (2024). “What Is inner awareness?” in *Consciousness and Inner Awareness*, eds. D. Bordini, A. Dewalque, and A. Giustina (Cambridge: Cambridge University Press).
- Laish, E. (2015). Natural awareness: the discovery of authentic being in the rDzogs chen tradition. *Asian Philos.* 25, 34–64. doi: 10.1080/09552367.2015.1016735
- LeDoux, J. E. (2024). Consciousness, the affectome, and human life. *Neurosci. Biobehav. Rev.* 159:105601. doi: 10.1016/j.neubiorev.2024.105601
- Lepauvre, A., and Melloni, L. (2021). The search for the neural correlate of consciousness: progress and challenges. *Philos. Mind Sci.* 2, 1–26. doi: 10.33735/phimisci.2021.87
- Li, R., Utevsy, A. V., Huettel, S. A., Braams, B. R., Peters, S., Crone, E. A., et al. (2019). Developmental maturation of the precuneus as a functional core of the default mode network. *J. Cogn. Neurosci.* 31, 1506–1519. doi: 10.1162/jocn_a_01426
- Luk, C. (2001). *Surangama Sutra*. Sri Lanka: Buddha Dharma Education Association Inc.
- Lycan, W. (2023). “Representational theories of consciousness,” in *The Stanford Encyclopedia of Philosophy*, eds. E. N. Zalta, and U. Nodelman. Available online at: <https://plato.stanford.edu/archives/win2023/entries/consciousness-representational/> (accessed November 3, 2023).
- MacKenzie, M. (2015). Reflexivity, subjectivity, and the constructed self: a Buddhist model. *Asian Philos.* 25, 275–292. doi: 10.1080/09552367.2015.1078140
- Manjusrimitra, L. K. (2001). *Primordial Experience*. Boston, MA: Shambhala.
- Marchetti, G. (2024). The self and conscious experience. *Front. Psychol.* 15:1340943. doi: 10.3389/fpsyg.2024.1340943
- Margulies, D. S., Vincent, J. L., Kelly, C., Lohmann, G., Uddin, L. Q., Biswal, B. B., et al. (2009). Precuneus shares intrinsic functional architecture in humans and monkeys. *Proc. Natl. Acad. Sci. U. S. A.* 106, 20069–20074. doi: 10.1073/pnas.0905314106
- Metzinger, T. (2010). *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. New York, NY: Basic Books.
- Metzinger, T. (2020). Minimal phenomenal experience meditation, tonic alertness, and the phenomenology of pure consciousness. *Philos. Mind Sci.* 1, 1–44. doi: 10.33735/phimisci.2020.1.46
- Metzinger, T. (2024). *The Elephant and the Blind: The Experience of Pure Consciousness: Philosophy, Science, and 500+ Experiential Reports*. Cambridge, MA: The MIT Press.
- Michel, M., Beck, D., Block, N., et al. (2019). Opportunities and challenges for a maturing science of consciousness. *Nat. Hum. Behav.* 3, 104–107. doi: 10.1038/s41562-019-0531-8
- Montague, M. (2016). *The Given: Experience and its Content*. Oxford: Oxford University Press.
- Montague, M. (2017). What kind of awareness is AoA? *Grazer Philosophische Studien* 94, 359–380. doi: 10.1163/18756735-09403004
- Norbu, C. N. (1987). *The Cycle of Day and Night*. Barrytown, NY: Station Hill Press.
- Norbu, C. N. (2013). The mirror: advice on presence and awareness. *Religions* 4, 412–422. doi: 10.3390/rel4030412
- Peters, F. (2013). Theories of consciousness as reflexivity. *Philos. Forum* 44, 341–372.
- Pruett, C. (2016). Shifting concepts: the realignment of Dharmakirti on concepts and the error of subject/object duality in Pratyabhijñāsaiva thought. *J. Indian Philos.* doi: 10.1007/s10781-016-9297-8
- Rabjam, L. (2001). *A Treasure Trove of Scriptural Transmission: A Commentary on the precious Treasure of the Basic Space of Phenomena*. Junction City, KS: Padma Publ.
- Rabjam, L. (2007). *The Precious Treasury of Philosophical Systems: A Treatise Elucidating the Meaning of the Entire Range of Buddhist Teachings*. Junction City, KS: Padma Publ.
- Radhakrishnan, S. (1994). *The Principal Upanishads*. New Delhi: HarperCollins.
- Ricard, M., and Singer, W. (2017). *Beyond the Self: Conversations between Buddhism and Neuroscience*. Cambridge: MIT Press.
- Rosenthal, D. (2012). Higher-order awareness, misrepresentation and function. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 1424–1438. doi: 10.1098/rstb.2011.0353
- Rosenthal, D. M. (2004). “Varieties of higher-order theory,” in *Higher-Order Theories of Consciousness: An Anthology*, ed. R. Gennaro (Philadelphia, PA; Amsterdam: John Benjamins).
- Schmidt, H., Avitabile, D., Montbrio, E., and Roxin, A. (2018). Network mechanisms under-lying the role of oscillations in cognitive tasks. *PLoS Comput. Biol.* 14, 1–24. doi: 10.1371/journal.pcbi.1006430
- Siewert, C. (2022). “Consciousness and intentionality,” in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. Available online at: <https://plato.stanford.edu/archives/sum2022/entries/consciousness-intentionality/> (accessed December 10, 2023).

- Singh, J. (1989). *Abhinavagupta's Trident of Wisdom*. Albany, NY: SUNY Press.
- Strawson, G. (2022). Self-awareness: acquaintance, intentionality, representation, relation. *Rev. Philos. Psychol.* 13, 311–328. doi: 10.1007/s13164-022-00639-9
- Tomasi, D., and Volkow, N. D. (2011). Functional connectivity hubs in the human brain. *Neuroimage* 57, 908–917. doi: 10.1016/j.neuroimage.2011.05.024
- Tye, M. (2014). Transparency, qualia realism, and representationalism. *Philos. Stud.* 170, 39–57. doi: 10.1007/s11098-013-0177-8
- Van Gulick, R. (2004). “Higher-order global states (HOGS): an alternative higher-order model of consciousness,” in *Higher Order Theories of Consciousness*, ed. R. Gennaro (Amsterdam: John Benjamins), 67–92.
- Webb, T. W., Miyoshi, K., So, T. Y., Rajananda, S., and Lau, H. (2023). Natural statistics support a rational account of confidence biases. *Nat. Commun.* 14:3992. doi: 10.1038/s41467-023-39737-2
- Williams, P. (2000). *The Reflexive Nature of Awareness*. New Delhi: Motilal Banarsidass.
- Zahavi, D. (2003). “Inner time-consciousness and pre-reflective self-awareness,” in *The New Husserl: A Critical Reader*, ed. D. Welton (Bloomington, IN: Indiana University Press), 157–180.
- Zahavi, D. (2005). *Subjectivity and Selfhood: Investigating the first-Person Perspective*. Cambridge, MA: The MIT Press.
- Zahavi, D. (2018). Consciousness, self-consciousness, selfhood: a reply to some critics. *Rev.Philos. Psych.* 9, 703–718. doi: 10.1007/s13164-018-0403-6
- Zahavi, D. (2024). Being you or not. *J. Conscious. Stud.* doi: 10.53765/20512201.31.5.206



OPEN ACCESS

EDITED BY

Luca Simone,
UNINT – Università degli studi Internazionali
di Roma, Italy

REVIEWED BY

Chris Percy,
University of Derby, United Kingdom
Robert Prentner,
ShanghaiTech University, China
Alfredo Brancucci,
Foro Italico University of Rome, Italy

*CORRESPONDENCE

Naotsugu Tsuchiya
✉ naotsugu.tsuchiya@monash.edu

RECEIVED 25 March 2024

ACCEPTED 28 May 2024

PUBLISHED 03 April 2025

CITATION

Tsuchiya N, Bruza P, Yamada M, Saigo H and
Pothos EM (2025) Quantum-like Qualia
hypothesis: from quantum cognition to
quantum perception.
Front. Psychol. 15:1406459.
doi: 10.3389/fpsyg.2024.1406459

COPYRIGHT

© 2025 Tsuchiya, Bruza, Yamada, Saigo and
Pothos. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Quantum-like Qualia hypothesis: from quantum cognition to quantum perception

Naotsugu Tsuchiya^{1,2,3*}, Peter Bruza⁴, Makiko Yamada⁵,
Hayato Saigo⁶ and Emmanuel M. Pothos⁷

¹Faculty of Medicine, Nursing, and Health Sciences, School of Psychological Sciences, Turner Institute for Brain and Mental Health, Monash University, Melbourne, VIC, Australia, ²Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology (NICT), Suita-shi, Osaka, Japan, ³Laboratory of Qualia Structure, ATR Computational Neuroscience Laboratories, Kyoto, Japan, ⁴School of Information Systems, Queensland University of Technology, Brisbane, QLD, Australia, ⁵National Institutes for Quantum and Radiological Science and Technology, Chiba, Japan, ⁶Nagahama Institute of Bio-Science and Technology, Nagahama, Japan, ⁷Department of Psychology, City, University of London, London, United Kingdom

To arbitrate theories of consciousness, scientists need to understand mathematical structures of quality of consciousness, or qualia. The dominant view regards qualia as points in a dimensional space. This view implicitly assumes that qualia can be measured without any effect on them. This contrasts with intuitions and empirical findings to show that by means of internal attention qualia can change when they are measured. What is a proper mathematical structure for entities that are affected by the act of measurement? Here we propose the mathematical structure used in quantum theory, in which we consider qualia as “observables” (i.e., entities that can, in principle, be observed), sensory inputs and internal attention as “states” that specify the context that a measurement takes place, and “measurement outcomes” with probabilities that qualia observables take particular values. Based on this mathematical structure, the Quantum-like Qualia (QQ) hypothesis proposes that qualia observables interact with the world, as if through an interface of sensory inputs and internal attention. We argue that this qualia-interface-world scheme has the same mathematical structure as observables-states-environment in quantum theory. Moreover, within this structure, the concept of a “measurement instrument” in quantum theory can precisely model how measurements affect qualia observables and states. We argue that QQ naturally explains known properties of qualia and predicts that qualia are sometimes indeterminate. Such predictions can be empirically determined by the presence of order effects or violations of Bell inequalities. Confirmation of such predictions substantiates our overarching claim that the mathematical structure of QQ will offer novel insights into the nature of consciousness.

KEYWORDS

qualia, quantum cognition, consciousness, attention, similarity, Bell inequality, bistable perception

Highlights

- The recent explosion in theories of consciousness, which aim to link subjectivity and physical substrates, require a better characterization of mathematical structure of quality of consciousness, or qualia.
- In traditional and intuitive models of qualia, a particular quale is assumed to be a point in a high dimensional space.
- Such models assume that qualia exist independent of measurements, but they are incompatible with the findings that qualia are generally affected by measurements.
- To account for how the measurement can affect qualia, a Quantum-like Qualia (QQ) hypothesis proposes a mathematical structure employed in quantum theory.
- We will outline how QQ can be tested with various experimental paradigms, building on the successful quantum cognition framework.

1 Introduction

Research on consciousness has recently entered a new phase. A burst of neuroimaging studies on consciousness since 1990 has produced a huge amount of empirical data, requiring a principled explanation for consciousness and its neuronal substrate (Koch et al., 2016; Mashour et al., 2020; Seth and Bayne, 2022). Over the last 20 years, many of the initial ideas about consciousness and brains were abandoned in the face of empirical data. The remaining theories have retained their core principles in the form of variations that have branched out from these theories. Some theories aspire to make quantitative predictions, a few of which are currently pitted against each other in an adversarial way (Melloni et al., 2021). Through empirical tests of rival theoretical predictions, substantial scientific progress is to be expected, as has happened in other fields, such as physics and experimental psychology (Einstein et al., 1935; Bell, 1964; Freedman and Clauser, 1972; Aspect et al., 1982; Kahneman, 2003).

As the science of consciousness matures, it has become increasingly clear that we lack an understanding of the target phenomenon, namely consciousness. While “consciousness” can mean the level or presence of consciousness, as in the clinical science of coma, general anesthesia, or deep sleep (Casarotto et al., 2016), this article focuses on the issue of quality of consciousness, feelings of what-it-is-like-to-be, or, in short, qualia (Balduzzi and Tononi, 2009; Kanai and Tsuchiya, 2012; Tsuchiya and Saigo, 2021; Tye, 2021; Lyre, 2022). Qualia in consciousness research comes in two senses, broad and narrow. In the broad sense, we use a quale to mean a moment of entire conscious experience across all sensory modalities and thoughts, that is, everything being experienced. Qualia in the narrow sense refers to one aspect of the experience, such as the “redness” of the sunset, the particular flavor and taste of tuna sashimi, and so on (Balduzzi and Tononi, 2009; Kanai and Tsuchiya, 2012). This article embraces both senses of qualia. What is not qualia concerns everything that is not part of our conscious experience.

In this article, Section 2 reviews the popular models of qualia and their deficiencies. To address these deficiencies, Section 3 proposes the Quantum-like Qualia (QQ) hypothesis. Our hypothesis is inspired by the mathematical structure of quantum theory. None of our claims

rests on whether or not microscopic quantum phenomena play a significant role in the brain and/or consciousness. Section 4 focuses on empirical research projects that can test the validity of the QQ hypothesis, followed by the conclusion in Section 5.

2 Traditional qualia models and their deficiencies

Traditional models of qualia are founded on the notion of points in a putative metric space, sometimes called a psychological space, quality space, qualia space, phenomenal space (Clark, 2000; Rosenthal, 2015; Lee, 2021; Figure 1A). These models have been proposed for various modalities, such as color, time, pain, sound and smell (Shepard and Cooper, 1992; Churchland, 2005; Klineciewicz, 2011; Kostic, 2012; Young et al., 2014; Renero, 2014). In the cognitive domain, there are strong arguments that concepts reside in such a space (Gärdenfors, 2000). Thus it seems natural to start with the idea to represent qualia as single points in a high dimensional space. Here, a definite point corresponds to a particular quale (either in the narrow or broad sense). To specify a combination of narrow qualia or a quale in the broad sense, multiple points are often considered as well.¹

In the case of narrow sense qualia, the distance between the two points relates to the “similarity” between the respective qualia (e.g., a red quale and an orange quale are close in similarity, but red and green are dissimilar). Inspired by early work by Shepard, many variants of such similarity models have been proposed (Krumhansl, 1978; Ashby and Perrin, 1988; Nosofsky, 1991), where visualization techniques such as multidimensional scaling (Borg and Groenen, 2005) have played a central role (Figures 1A,B). Under this framework, various types of qualia, e.g., color (Indow, 1988; Shepard and Cooper, 1992; Churchland, 2005; Bujack et al., 2022; Zeleznikow-Johnston et al., 2023), sound (Shepard, 1982; Renero, 2014; Cowen et al., 2020), object (Hebart et al., 2020), emotion (Figure 1B) (Cowen and Keltner, 2017; Nummenmaa et al., 2018), olfaction (Young et al., 2014), art (Graham et al., 2010) etc., have been investigated and visualized based on similarity ratings of pairwise comparisons between the set of qualia under investigation.

Despite widespread use, the psychological space approach to modeling qualia encounters three challenges: the inability to adequately capture indeterminate and dynamic facets of qualia, as well as their intricate interactions with internal mental processes. The following summary briefly covers these three points.

Firstly, as this approach assumes a quale is a definite entity (e.g., a point or points in a space), it is unable to capture the intuition that some qualia appear to be indeterminate entities. The indeterminacy of qualia becomes apparent when one introspects on the border of experience in space or time or the nature of unattended or barely attended experience. To determine

¹ Temporally extended and varying qualia can be represented as either a dynamically moving single point in high-dimensional space or a single point of a very high-dimensional space, where different time points are represented as different dimensions.

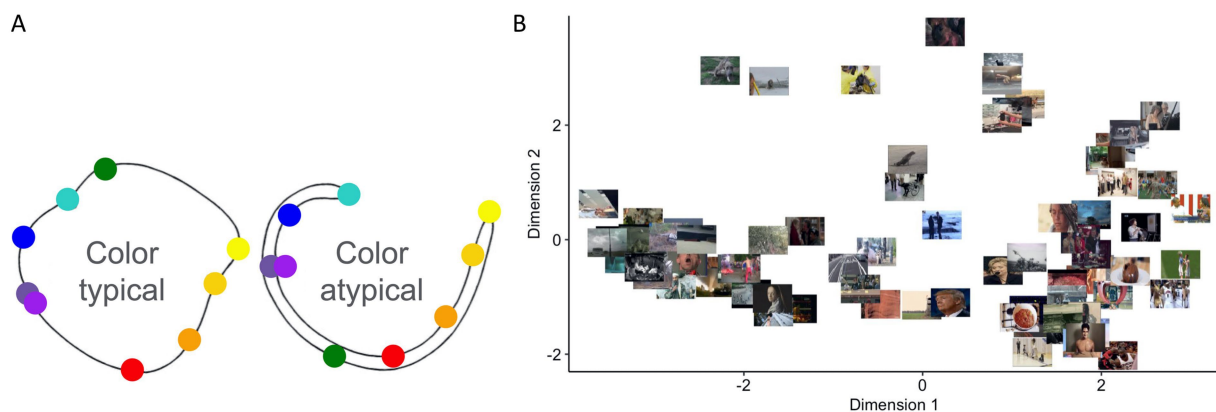


FIGURE 1

Traditional psychological space models (Shepard and Cooper, 1992; Rosenthal, 2015; Lee, 2021) assume each quale occupies a point in space (or a combination of points). “Distances” between two points are assumed to be related to perceived experiential similarity (Krumhansl, 1978; Ashby and Perrin, 1988; Nosofsky, 1991). (A) A classic color hue ring model for the representation of similarity relationships among 9 colors for color-typical and red-green color blind individuals. Modified from Shepard and Cooper (1992). (B) Similar representations (points-in-high dimensional spaces) have been used in other domains of experience, such as emotional experience. Adapted from Lin (2023), which used the emotional movie stimuli in Cowen and Keltner (2017).

the spatial border of experience, one can stretch their arms to estimate the limit of the visual field at the periphery, and experientially confirm that this limit is tenuous. Under complete darkness, it is not clear that any such boundary exists. Time also seems to have an indeterminate character. The start and end times of an event often feel unsure and a moment rarely feels point-like, but is typically experienced as having some duration (Filk, 2013). Even when one is focally attending to qualia, one can sense an uncertainty regarding the phenomenal appearance. Changes in certain aspects of qualia have been psychophysically confirmed. The very act of attending can alter the quality of the experience (Carrasco and Barbot, 2019).

Qualia can be uncertain in two ways. Firstly, the “epistemic” uncertainty of qualia implies that qualia themselves are always determinate, i.e., in a definite state, but measurement processes inject noise so that there is uncertainty about the value of this definite state. Epistemic uncertainty can be captured by modifying the classical model by replacing a point with a cloud of points. However, we suspect that some qualia are “ontologically” indeterminate. Such qualia can be characterized as being in an indefinite “state” whereby properties can only be attributed by means of measuring an ensemble of like qualia. Consequently, indeterminate qualia cannot be modeled or represented as a cloud of dots.

Secondly, the psychological space approach is by default static and does not account for the temporal dynamics of qualia, because it maps sensory inputs into qualia “at a given time” (see also Footnote 1). The temporal dynamics of qualia, however, are one of the most studied aspects of qualia, from very fine time scales using masking and priming (Bachmann, 2000; Breitmeyer and Ogmen, 2007), to larger time scales involving adaptation, expectation (Melloni et al., 2011), and multistability (Maier et al., 2012; Brascamp et al., 2018). If the space itself changes dynamically, the traditional psychological space approach may require substantial updates to account for the spatio-temporal dynamics of qualia.

Thirdly, the psychological space approach is not well developed regarding how qualia interact with internal mental processes, such as attention. As alluded to above, how we attend to sensory inputs appears to significantly alter what we experience (Carrasco and Barbot, 2019), as implied from change blindness and inattention blindness demonstrations (Simons and Rensink, 2005; Pitts et al., 2018). However, before we pay attention, we already experience something at the to-be-attended locations, and that is the reason why we can consciously direct attention there. The psychological space model is similarly unclear about how qualia relate to other internal processes, such as memory and expectation.

Of course, any general framework can be in principle extended. Yet, since the pioneering work by Shepard (1962a,b, 1970, 1980, 1987), subsequent extensions (e.g., concerning dynamics) have not been proposed. It is noteworthy that masking effects have been documented for over a century (Exner, 1868; Breitmeyer and Ogmen, 2007), and despite more than six decades of exploration within high dimensional point models, scant insights into these effects have emerged. We contend that the outlined QQ hypothesis presented here holds promise for explicating such masking phenomena, even without properly fleshed out computational models.²

Thus, the psychological space approach to modeling qualia as points in a dimensional space appears deficient in regard to psychophysically-informed intuitions that qualia are indeterminate, dynamic, and interact with other mental processes. But why do researchers continue to adhere to the psychological-space models? We surmise that this is due to the

² One promising venue is dynamical models of consciousness and qualia (Fekete and Edelman, 2011; Esteban et al., 2018; Moyal et al., 2020). However, so far, such models do not address the issue of how measurements and observations affect qualia, one of the central points of our paper.

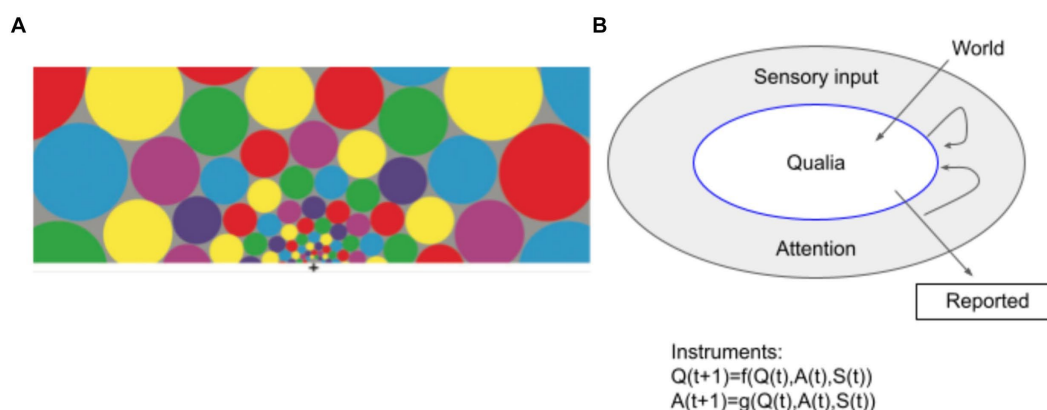


FIGURE 2

Conceptual framework of the QQ hypothesis. **(A)** An exemplar sensory input of many colorful patches with the size of each patch proportional to cortical magnification (Tyler, 2015). While you are fixating on the cross at the bottom center, you see the color of each patch without moving your eyes. However, you may feel your experience changes depending on where you direct your attention. **(B)** The conceptual diagram of QQ. QQ considers Qualia as observables that are properties of a system that can be in principle “measured,” probed and reported. Sensory inputs and Attention act as an interface or a “state” between Qualia and the World. For example, here the state can be “the sensory input as in **(A)** AND attending to a red patch on the right.” Then, we can define and measure a probability that a particular value is assigned to the observable, for example, $\text{Prob}(\text{“color } Q \text{ for the leftmost circle”} = \text{“blue”} \mid \text{the state}) = 0.7$. How Qualia (Q), Attention (A) and Sensory input (S) evolve over time with or without measurement is formalized by the theory of Instruments (Davies and Lewis, 1970; Ozawa and Khrennikov, 2021). Informally, the putative interaction between the world and qualia, qualia and subjective reports, and how reports alter attention and qualia through instruments are depicted by arrows in the panel.

combination of the intuitive appeal of such models and the lack of compelling alternatives.³

Interestingly, a similar situation arose in the field of cognitive science, in particular decision making. In decision making, models based on standard probability theory and logic have been persistently challenged by many (apparently) paradoxical findings in human decision making. Some of these paradoxes in decision making have had fairly natural explanations by means of quantum probability theory, which was introduced in psychology with the quantum cognition framework (Khrennikov, 2010; Busemeyer and Bruza, 2024; Haven and Khrennikov, 2013; Pothos and Busemeyer, 2022).⁴ Notably, analogous qualia-related concerns have been raised in the context of human decision-making. By incorporating the indeterminacy inherent in quantum theory and acknowledging the role of measurement in determining the state within cognitive processes, it has become possible to more effectively model these phenomena, propelling the growth of the quantum cognition field. Consequently, we posit that quantum cognition establishes the conceptual and theoretical foundation of the Quantum-like Qualia hypothesis.

Decision making and other cognitive processes are inextricably linked to perception and sensation (Barsalou, 2010) and also appear to share basic neural processing architectures. Thus, it seems natural to consider the application of quantum probability theory as an alternate mathematical framework for qualia, in order to address the challenges for the psychological space approach.

3 The Quantum-like Qualia hypothesis

The three essential challenges for existing models for qualia (i.e., indeterminacy, dynamics, and interactions) are inherently related with the limitations in “classical” approaches. Classical approaches assume that qualia can be probed, observed, reported or “measured,” without affecting them. To consider a more general mathematical structure, it is useful to start with the assumption that such “measurements” necessarily affect qualia. How much these measurements affect qualia can vary depending on various factors.

Quantum theory offers a mathematical structure that deals with entities whose properties can change upon measurement. As we argue below, such a mathematical structure, proposed as a Quantum-like Qualia (QQ)⁵ hypothesis, attains the three desired features for qualia. QQ states that qualia are like quantum entities, which are inextricably affected by measurement. We first give a broad sketch of QQ (Figure 2), then explain technical concepts with familiar examples from consciousness research. More detailed mathematical formulations will be pursued in future work.

³ For more recent mathematically elaborated models, see Hoffman et al. (2023), Kleiner (2024), Kleiner and Ludwig (2024) and references therein.

⁴ Some studies in quantum cognition are highly relevant to our proposal (Filk, 2009; Khrennikov, 2015, 2021; Atmanspacher and Müller-Herold, 2016). Our Quantum-like Qualia (QQ) hypothesis is quite orthogonal to the Quantum Brain hypothesis, which considers quantum mechanical processes in the brain (Hameroff and Penrose, 2014) and the role of consciousness in quantum collapse (Chalmers and McQueen, 2021) (see also Smolin, 2022). QQ is completely consistent with the possibility that all physical events happening in the brain are purely classical. Our core idea is to utilize the mathematical formalism of quantum theory, as outlined below. For these and other related issues see Atmanspacher (2017).

⁵ This is different from the quantum question (QQ) equality by Wang et al. (2014).

TABLE 1 Conceptual summary of quantum terminologies (columns: observables, states, averaged measurement outcomes) and how they are used in (rows) quantum theory, quantum cognition, and QQ (the Quantum-like Qualia hypothesis).

| | Observables | States | Averaged measurement outcomes |
|---------------------|--------------------------|-------------------------------------|-------------------------------|
| Quantum Theory | \mathcal{A} | Ψ | $\Psi(a), a \in \mathcal{A}$ |
| Quantum Cognition | Response options (fixed) | Mental states (dynamic) | Responses |
| Quantum-like Qualia | Qualia (dynamic) | Sensory inputs, attention (dynamic) | Reportable aspects of qualia |

Each cell entry explains a representative usage of each concept.

3.1 Separating qualia observables from states (of sensory input and attention)

To account for the indeterminacy of qualia, QQ distinguishes each instance of measured value of qualia (say, color qualia $Q = \text{“red”}$) from all possible measurable qualia. Inspired by quantum theory, we call all possible measured outcomes “observables.” Observables are intrinsic properties of a system that can, in principle, be measured. For example, a color qualia observable at the fixation can be a coarse set of color labels, such as $Q = \{\text{“red,” “blue,” “green,” ...}\}$. QQ does not presuppose that all aspects of qualia can be simultaneously measured and reported⁶.

Now consider a situation where you momentarily see many color patches (Figure 2A). Suppose you are attending to the right most red patch. This kind of “sensory input” and “attention” constitute a “state,” separate from “observables.” While each color quale can be indeterminate, under a particular “state,” the expected value of a particular quale (modeled as an observable) is given. Formally, states are like functions that return the expected value for a given quale, when a particular observable is measured.

3.2 Dynamics of qualia observables and states: updates through instruments

In quantum theory, there are three mathematically equivalent ways to consider the dynamics of observables and states (Sakurai and Napolitano, 2020) (see Table 1 for a summary). QQ considers both observables and states to change over time. This interpretation is called an “interaction” picture.

In most quantum cognition studies, observables are possible response options, which are fixed, while (mental) states change dynamically. This idea of fixed-observables and dynamic-states is called the “Schrödinger” picture. In QQ, we consider sensory inputs and attention as “states.” It is not difficult to imagine how these “states” can change measurement outcomes.

In some fields of physics (e.g., particle physics), states are considered to be fixed, while observables change. This dynamic is called the “Heisenberg” picture. In QQ, it is natural to consider changes of qualia observables as a consequence of changes in the brain

through perceptual learning, sensory adaptation, and so on (Song et al., 2017). In this case, even if sensory inputs and attention are fixed, qualia can change.

In this paper, we predominantly consider sensory inputs and internal attention as major foundational elements of states, but other mental elements, such as memories and expectations, can also constitute states. Thus, in this interaction picture, QQ explicitly considers how qualia (observables) interact with states (sensory inputs and attention). Without a state, we cannot consider a particular measurement outcome of any qualia observable.

Finally, to formalize how qualia observables interact with other mental processes, we introduce the concept of a “measurement instrument” (cf. the arrows in Figure 2; Davies and Lewis, 1970). In modern measurement theory, any measurement of the system is described by a mathematical structure called a (measurement) instrument, which offers a generalization of a conditional probability. In standard quantum physics, measurements are considered all-or-nothing. As the theory of quantum measurement matured, researchers arrived at the concept of instruments as the most general form of measurement. The formalism of instruments offers a bridge from nonlinear wave collapse (which is the result of a measurement in standard quantum theory) to the unitary dynamics of an isolated system and ‘unsharp’ or weak measurements. We propose that this generalized formalism to characterize the effects of measurements would be particularly useful when considering the interaction between qualia and attention. Attention may not determine qualia in an all-or-nothing way, but rather in an unsharp or weak way.

Instruments are utilized in modern quantum measurement theory and have started being applied in the field of quantum cognition (Khrennikov, 2015; Ozawa and Khrennikov, 2021). Instruments can describe how qualia observables and states of sensory inputs and attention dynamically develop upon measurements.

While the above descriptions are sufficient to understand the foundations of the QQ hypothesis, we now expand the conceptual framework and provide associated technical details.

3.3 What counts as a system?

We define qualia observables as all possible intrinsic properties of a system. But what is meant by the term “system”? We consider a system minimally as “that which is experiencing the qualia in question.” It would correspond to “the complex” in Integrated Information Theory (Albantakis et al., 2022). Over time, a system itself can change (then observables would change accordingly). Yet the

⁶ Note this statement is about measurement and reports on qualia. We assume that qualia exist before measurements in the same way quantum particles exist before measurement.

system should still need to be identified as a coherent entity or phenomenon. A system has an associated set of qualia observables, which can be measured from the outer environment.

3.4 A state as an interface between qualia and the world

The interrelationship between the system and the environment external to it is represented by the state of the system. In a sense, a state can be considered as an interface. This idea may sound strange at first, but actually it is equally applicable across classical and quantum theory (Ojima, 2005; Saigo et al., 2019; Saigo, 2021). For example, the temperature of water in a cup as an observable needs to be determined in the context (= “state”) of where and how the measurement instruments are placed.⁷

In QQ, such a context would involve at least sensory inputs and attention. In a particular state, call it “ ϕ ,” the expected value of reporting a particular quale, $P(Q = q|\phi)$ can be established. For example, in a state ϕ = “one is sitting at the sunset with the mind wandering,” $P(Q = \text{“seeing the color of red”} | \phi)$ can be established. Or, in a state ϕ = “sensory input to a participant is a weak grating stimulus with masking under a particular attentional instruction,” we may obtain $P(Q = \text{“faint”} | \phi) = 0.7$, when we assume Q as observables with outcomes of {highly visible, less visible, faint, not visible}. Note that in this framework, there is no point in talking about considering a single-trial quale as in $[Q = \text{“faint”}]$ without considering the state. We can consider only an ensemble of measurement outcomes given a particular state.

The notion of an interface between system and environment is an important idea, as discussed in many theories of consciousness. Just to name a few, “interface” in interface theory of consciousness (Hoffman et al., 2015, 2023; Prakash et al., 2020; Prentner, 2021), “background conditions” in the Integrated Information Theory of consciousness (Albantakis et al., 2022), “Markov blanket” in the free energy principle (Kirchhoff et al., 2018), and “mediation” in philosophy (Taguchi, 2019).

Inspired by the mathematical structure of quantum theory, QQ aspires to establish principled associations among observables, states, and their interactions, not at the level of an individual event (or the qualia property at each moment) but at the level of collections of similar events. In fact, for every individual event, the set of all qualia properties would be unique and never identical to the other sets, especially when space and time are considered. Thus, QQ proposes that qualia should not be considered at the level that assumes definiteness of qualia properties for each event. Rather, QQ proposes to consider qualia at the level of ensembles where some “similar”

qualia properties are grouped together (as in the above categorical set of observables). How to construe “similar” is an important question, which the authors have discussed elsewhere, using concepts from category theory (Tsuchiya et al., 2016, 2022, 2023). In category theory, it is quite explicit what one considers as similar is a choice of mathematicians or scientists, not automatically or uniquely ‘given’ by the world (Cheng, 2022). In most theoretical and experimental contexts, qualia are similar as long as they are considered similar in some way by the observing individual, as in the everyday usage of “similar.”

In summary, “state” is an interface that assigns an “average” value to each observable, noting that measurement of a single event may not be possible.

3.5 Instrument formalism for dynamics of qualia and states

Let us now consider the dynamics of qualia. For simplicity, in relation to a discrete time step, denote qualia, sensory input, and attention at time t as $Q(t)$, $S(t)$, and $A(t)$. Their interdependency is illustrated by the arrows in Figure 2. The dynamical update rules are expressed as

$$Q(t+1) = f(Q(t), S(t), A(t)) \text{ and}$$

$$A(t+1) = g(Q(t), S(t), A(t))$$

This simple formulation is a primitive form of an instrument. Currently, we do not have enough data to constrain the form of the functions f and g . However the equations generally formalize how changes of sensory inputs⁸ affect both what we experience and how we attend. They also capture how attending to uncertain aspects of qualia (e.g., a spatial boundary) can change qualia. For specific and empirical applications of instruments in quantum cognition, see Ozawa and Khrennikov (2021).

3.6 A common mathematical and philosophical structure between quantum phenomena and qualia

QQ proposes an application of some aspects of the mathematical structure from quantum theory (e.g., separation of observables, states and averaged measured outcomes, and instruments). In parallel with the mathematical structure, we surmise that there is a common

⁷ Consider all possible temperatures of water as observables. The temperature of water is a complex physical concept, which depends not only on the average kinetic energy of water molecules but also on the measuring probe device’s temperature, surface areas, and many other factors. We treat all of these factors that relate to measurement as “states.” In the case of measuring water temperature, depending on how invasive the measurement probe is (with a probe from either a very cold or very hot environment), the measured outcome of the temperature of water can change.

⁸ While some theories consider a possible role of conscious agents on the control of $S(t + 1)$ through motor control and intention, we consider that they are better left out from the formalism of this update rule of instrument for qualia. Consider the sensory input while you are looking at an ever-changing shape and colors of a burning fireplace. Also, in an experimental situation, experimenters can change sensory input $S(t + 1)$ to a participant in any way they want.

philosophical stance covering both quantum phenomena and qualia. Through such a philosophical connection, QQ naturally situates some of the perplexing psychological findings in qualia and attention as detailed below.

3.6.1 Noncommutativity, complementarity, uncertainty relations in quantum theory, quantum cognition, and QQ

One of the foundational ideas behind quantum theory is “complementarity.” In the context of qualia, two qualia are complementary when they cannot be experienced simultaneously, as we consider in more detail below (Bruza et al., 2023).⁹ Complementarity is a philosophical concept that one of the founders of quantum theory, Niels Bohr, introduced in physics, indirectly inspired by one of the founders of modern experimental psychology, William James, through Edgar Rubin (Holton, 1988).

The idea of complementarity can be mathematically expressed via the concept of noncommutativity (Streater, 2007; Atmanspacher and Filk, 2018). Noncommutativity implies sensitivity to the order of an operation. In general, the effect of processing A then B may not be the same as B then A. Noncommutativity is the default for many processes, from cooking to chemical reactions.¹⁰ In the brain, this could correspond to the effect of processing A leaving some trace, in terms of synaptic plasticity or neuronal activity, which impacts on processing B. If this is the case, processes A and B are expected to be noncommutative and likewise for the corresponding qualia.

If observables A and B are noncommutative, measuring A after B typically yields a different outcome to B after A. It is generally accepted that many aspects of human cognition are noncommutative. Even in arithmetic, subtraction and division are noncommutative. While multiplication is commutative for numbers, it is not for matrices. Note that matrix operations are fundamental to quantum theory (Busemeyer and Bruza, 2024). Noncommutative observables can be used to formalize important features of qualia, such as the aforementioned indeterminacy. Starting with the well established noncommutative formalization of quantum theory as a guiding framework, it should be possible to appropriately extend this formalism for QQ. Then, as we explain later, it should be possible to empirically demonstrate its necessity.

Regarding qualia, in general, when we consider “processes,” whereby the order of the processes matters. In an example drawn from masking, presenting target T briefly before mask M at a particular interval can make T completely invisible. But swapping the order into M then T, both of them can become highly visible. This is an example of noncommutativity. Quantitative and coherent explanations of order effects, fallacies in decision making, conceptual combination, evidence

accumulation, over/under distribution effects in memory and other cognitive phenomena is one of the hallmarks of the quantum cognition framework (Busemeyer and Bruza, 2024; Busemeyer and Wang, 2017; Pothos and Busemeyer, 2022). Complementarity as noncommutativity is experimentally demonstrated as uncertainty relations (Atmanspacher and Filk, 2018).

Complementarity, noncommutativity and uncertainty relations are the basis of quantum theory, from which the field of quantum cognition arose. Quantum cognition started from explaining enigmatic phenomena in decision making (Aerts et al., 2018; Mistry et al., 2018; Basieva et al., 2019; Busemeyer et al., 2019; Broekaert et al., 2020), concept combination (Bruza et al., 2015; Wang et al., 2021; Aerts and Arguëlles, 2022), and judgment (Wang and Busemeyer, 2013; White et al., 2020; Ozawa and Khrennikov, 2021). It has recently expanded into modeling for language (Surov et al., 2021), emotion (Khrennikov, 2021; Huang et al., 2022), music (beim Graben and Blutner, 2019), and social judgments (Tesař, 2020). It is beginning to be applied to solve real-world problems (Arguëlles, 2018; Song et al., 2022; Wojciechowski et al., 2022) and it has been influencing the design of artificial intelligence and robots that aim to interact with the world (Ho and Hoorn, 2022).

To the extent that cognition is continuous with perception (Barsalou, 2010), quantum cognition is a relevant framework to consider quality of perceptual consciousness, or qualia. Indeed, certain applications of quantum cognition to perceptual judgments are already emerging (Conte et al., 2009; Atmanspacher and Filk, 2010; Asano et al., 2014; Yearsley et al., 2022; Bruza et al., 2023; Epping et al., 2023) as we will discuss below.

3.6.2 A common philosophical structure between quantum phenomena and qualia

On the philosophical side, both quantum phenomena and qualia arise from “interactions.” In the above, we introduced “a state as an interface,” which is an idea almost equivalent to the philosophical concept of “mediation” (Taguchi, 2019). Quantum phenomena arise from interactions between quantum objects, such as photons, and measurement devices (Plotnitsky, 2021).

Notably, Niels Bohr stated that the “reality” responsible for quantum phenomena is indeterminate and beyond representation (Plotnitsky, 2021). By “reality,” we mean a definite single event before any measurement. Such a concept is not problematic in the classical view, which assumes that anything can exist before measurement and it is in principle not affected by measurement. In quantum theory, a property of an observable is not defined without a state and there is no meaning to a single measurement outcome. In this sense, we adopt a view analogous to Bohr’s that “reality” is “indeterminate” and “beyond representation” before any measurement.

Likewise, QQ proposes that the reality of qualia defies concrete representation in a similar way, such as points in a high dimensional space in classical models. Note that classical models can consider a distribution of points rather than a single point. However, this still assumes the existence of “reality” of qualia before measurement. Moreover, measurement is assumed to introduce noise so that a probability distribution is needed to model it. In this view, the underlying uncertainty is epistemic due to the limitation of our measurement technique or lack of knowledge. However, QQ proposes that measurement outcomes statistically arise from interrelationship between qualia

⁹ Note that we are not saying that all qualia are complementary to each other. At least some combinations of qualia are likely to be complementary and cannot be experienced at the same time. Indeed, at each moment, we are experiencing multiple qualia at the same time. This is consistent with our introduction of a concept of “broad-sense” qualia. A broad-sense quale is composed of qualia in narrow sense in a unified way.

¹⁰ Note that non-commutativity includes commutativity as a special case. This is similar to the statement that quantum probability theory includes classical probability theory as a special case.

observables and states of sensory inputs and attention. In other words, the underlying uncertainty of qualia is ontic due to the nature of the very “being-ness” of qualia phenomena. If qualia are ontologically uncertain, we would be unable to establish what property each qualia observable corresponds to, for at least some states at a single event, even if we had all relevant information available.¹¹ For such qualia, the act of measurement does not reveal pre-existing properties of qualia observables. Rather the measured property emerges as part of the interrelationship between qualia observables and a state where a measurement takes place.

In classical philosophy literature, representationalism states that the phenomenal character of experience is reducible to representational content (Block, 1998). These views typically conceive of a definitive single event, regardless of a state, which is reduced to a cognitive representation. By contrast, anti-representational views of consciousness propose that such a definitive representation does not exist (Koenderink, 2010; Gibson, 2014; Varela et al., 2017; Schlicht and Starzak, 2021). While the precise reasoning behind the latter views is not the same, the QQ hypothesis shares the same conclusion.

The point of quantum theory, as argued by Bohr, is to abandon the assumption that “reality” must be definitive and to argue that, due to indeterminacy, the underlying “reality” cannot be represented in a classical way. Instead, quantum theory offers a suitable predictive and explanatory framework.

The analogy with qualia is that, due to their indeterminacy, some qualia cannot be “represented” as points in the dimensional space, as is usually assumed. Specifically, QQ points out that at least some qualia are indeterminate when they are in an unattended state. In many cases, when attention is directed to a particular qualia observable, measurement outcomes about the attended property would become more determinate. This corresponds to an intentional, content-bearing phenomenal object with an associated cognitive representation as proposed by the orthodox cognitive science. However, in an unattended state, these qualia observables have properties, which do not have well established values or qualities. Classical representationalism does not consider such a possibility. Further, as we elaborate later, QQ predicts that the measurement outcomes are not only statistical but they additionally violate some statistical laws that must be satisfied if qualia properties are always determinate.

¹¹ As “ontologically” indeterminate qualia, we consider several cases where measurements of qualia have non-ignorable impacts (periphery, similarity judgements, attention related experiments). In Section 4, we provided empirical experiments to address this issue. In classical physics objects exist independent of measurement. Similarly, classical qualia models tend to assume existence of qualia independent of measurement. For example, in encountering an unfamiliar painting, classical models tend to assume that you have some preference even if you do not articulate it or even if it is uncertain. Our QQ is more explicit about this. Some qualia are affected by measurement and measurement instrument theory (in the future) should specify how a particular type of measurement should affect qualia in what way. This also means that QQ also anticipates some qualia are not affected by measurements as well (say, the color of apple in front of you).

3.7 Interim summary: what is the Quantum-like Qualia hypothesis?

In summary, QQ hypothesizes the following. First, observables correspond to all possible aspects of experience that a system can have, including experiences from all sensory modalities, as well as thoughts, concepts, memories and feelings, that is, anything, as long as it is part of an experience (i.e., qualia in the broad sense). States are a particular arrangement of the system. When the system is in a given state, averaged measurement outcomes from qualia observables can be lawfully specified. States represent sensory inputs and any internal condition of the system, including how the system attends to or accesses observables. Second, averaged measurement outcomes are results of interactions between observables and states and they can be reported outside the system. Third, observables and states change dynamically and interact with each other, as formalized by the instrument theory. From mathematical and philosophical perspectives, qualia have an analogical correspondence with quantum phenomena. Table 1 summarizes these basic concepts and how they are used in quantum theory, quantum cognition, and QQ.

4 What are the benefits of QQ and how can we test QQ predictions?

As explained above, QQ accords with fundamental intuitions about qualia, such as their indeterminacy, dynamics, and interaction with internal processes. Furthermore, QQ offers some important insights concerning our empirical knowledge about qualia and provides novel perspectives about the nature of qualia. Here we provide some details of three lines of investigation comprising order effects, violation of the Bell inequality, and relationships between qualia and attention, thereby showcasing how to empirically test various predictions from QQ.

4.1 Order effects in similarity judgments among color qualia

The QQ hypothesis is empirically testable in surprisingly simple ways. One way is to ask if the order of questions or stimuli matters for the resulting reports. Epping et al. (2023) presented a pair of color patches to participants, then asked if the reported similarities are symmetric with respect to the order of color patch presentation.

Since seminal work by Rosch (1975) and Tversky (1977), perceptual similarity judgments about colors, faces, and objects have been repeatedly shown to be asymmetric (Polk et al., 2002; Roberson et al., 2007; Hodgetts and Hahn, 2012; Best and Goldstone, 2019). These studies challenge standard points-in-space type models, requiring arguably *ad hoc* modifications (Krumhansl, 1978; Ashby and Perrin, 1988; Nosofsky, 1991).

The extremely high citation rate of Tversky’s paper attests to the fact that researchers are aware of this asymmetry. Yet, it is not common to empirically take asymmetries into account in similarity studies, as this doubles the numbers of trials. Even when different orders are included, researchers often remove them by symmetrizing the originally asymmetric similarity matrix, so that they can use popular, existing analytic algorithms, such as multidimensional scaling.

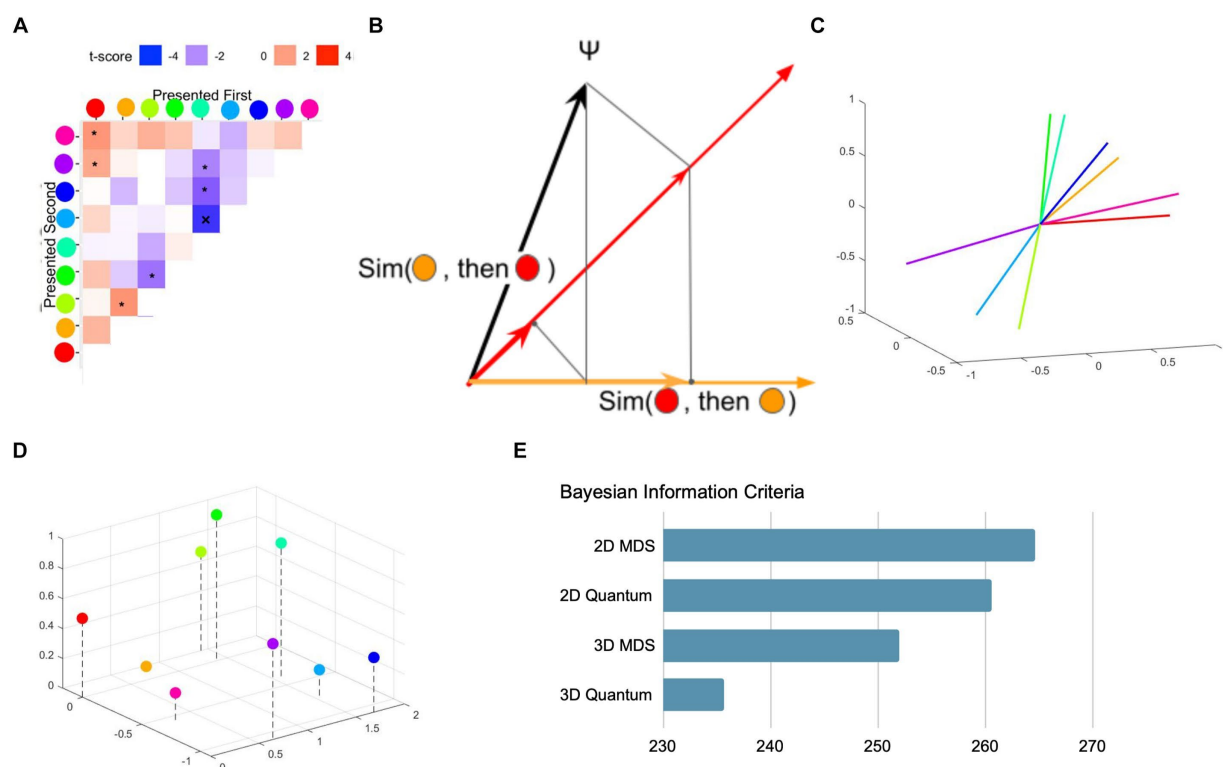


FIGURE 3

Quantum model of color similarity. **(A)** Empirical asymmetry matrix. The raw similarity matrix is subtracted from its transpose to reveal the degree of asymmetry in similarity judgments. Taken from Epping et al. (2023). **(B)** How quantum operations (projections) give rise to perceived similarity (Pothos et al., 2013; Yearsley et al., 2022; Epping et al., 2023). Assume an initial (mental) state as a unit vector Ψ (the black line). Color qualia observables (red and orange) are represented as two “subspaces” in a space (the red and orange axes). The vector is projected onto a subspace representing the color that is first experienced. From there, it is further projected onto the subspace corresponding to the second color. The resulting length of the final projection can be related to the perceived similarity between the two colors. Importantly, the resulting length can depend on the order with which the colors are experienced. **(C)** The best fit quantum similarity model for the data in (A) (Epping et al., 2023). In the quantum model, each of 9 color qualia observables is modeled as a subspace in 3D space. Experienced similarity between the two subspaces is related to the square value of the cosine angle between them (e.g., the red and the pink subspaces have a narrow angle, but the red and the green subspaces have a near 90 deg. angle). **(D)** Traditional 3D MDS representation of 9 colors based on their pairwise similarity. **(E)** Bayesian information criteria (BIC) for best fit 2D and 3D MDS and quantum models. Note that MDS models needed additional free parameters to account for asymmetries in similarity judgments (Nosofsky, 1991), resulting in more complex models. The 3D quantum model offered the best fit to the empirical data.

While an isolated instance of asymmetry [e.g., “Is China similar to North Korea” vs. “Is North Korea similar to China,” (Tversky, 1977)] can be explained in many possible models, a collection of perceptual reports for many stimuli, such as color patches, and a particular pattern of asymmetries across many stimuli represent a more substantial challenge (Figure 3A). Epping et al.’s quantum models, which consider a state as a density matrix (this is a generalization of the idea that a state can be a vector), and similarity as arising from sequential projections (Figure 3B), offered a better fit to the empirical data (Figure 3C), compared to points-in-space models of qualia (Figures 3D,E), with flexibility to accommodate asymmetry when mapping distance between points to similarity.

As noted previously, most similarity experiments tend to ignore the effect of order of presentation, using a simultaneous presentation paradigm, or paradigms that allow longer and uncontrolled inspection of the items. This is understandable due to the increased cost of experiments that manipulate order, because the number of the trials increases quadratically with the number of items to examine. Distributing pairs of items across many participants in online samples may solve this issue (Kawakita et al., 2023).

4.2 Violation of the Bell inequality in the domain of qualia

Quantum theory was developed in the 1920s by Bohr, Heisenberg, Shrodinger, Born and others. This theory challenged the predominant realist view of nature. In 1935, Einstein, Podolsky and Rosen (Einstein et al., 1935) (EPR) challenged this view, claiming that quantum theory is incomplete. In 1962, Bell discovered one fundamental inequality (Bell, 1964) must be satisfied assuming EPR’s view is correct. Subsequently, the violation of the Bell inequality was empirically demonstrated (Freedman and Clauser, 1972; Aspect et al., 1982). The Nobel Prize for Physics in 2022 was awarded for the demonstration of violations of the Bell inequality.

Since the initial EPR experiments, there has been debate about loopholes in the experiments that were being conducted. Over the years these loopholes have been successively closed. Nowadays, it is generally accepted that the EPR experiments do empirically verify that microscopic particles can violate the Bell inequalities and are therefore entangled. What this implies about the underlying nature of these particles has been debated (Zeilinger, 2010). In

parallel, a classical realist view has been questioned in relation to cognitive phenomena when these violate the Bell inequalities (Bruza et al., 2023).

Bell's inequality can be represented as follows:

$$S = E(a, b) - E(a, b') + E(a', b) + E(a', b'),$$

where a and a' are two measurement settings for system A, b and b' for B, and $E(\cdot)$ is the expected value of the corresponding measurements. These expected values have to be measured in separate experimental conditions. In classical systems, $|S| \leq 2$, unless there are direct influences or signaling, between measurements of system A and system B. Contextuality-by-Default (CbD) is a generalization of the Bell inequalities. CbD allows a determination of contextuality in the presence of direct influences [For its application, see Basieva et al. (2019) and Cervantes and Dzhamfarov (2019)]. The Bell inequality can be violated by quantum phenomena. A generally accepted explanation for the violation is that the properties of the phenomena do not have definite values at all times, that is, they are indeterminate.

For the QQ hypothesis, demonstrating that qualia violate the Bell inequality will play a similarly fundamental role. If these types of inequalities are violated, qualia can be assumed to be quantum-like (which implies additional properties, such as noncommutativity). There are many ways to psychophysically test the Bell inequalities (Basieva et al., 2019; Cervantes and Dzhamfarov, 2019; Bruza et al., 2023).

4.2.1 Establishing violations of the temporal Bell inequality in multistable perception

Multistable perception (Maier et al., 2012; Brascamp et al., 2018) can be used to demonstrate violations of a type of Bell inequality. Atmanspacher and Filk (2010) focused on the number of reversals between three time points of an ambiguous figure. They proposed empirical tests involving the temporal version of the Bell inequality (Yearsley and Pothos, 2014). Specifically, Atmanspacher and Filk's proposal was to measure perceptual switches between times t_1 , t_2 , and t_3 , where $t_1 < t_2 < t_3$, selecting two time points per condition and for all three possible combinations. The probability of the perceptual state being different at time i vs. time j is denoted by p_{ij} . If qualia are determinate at all time (as hypothesized Figure 5 and Table 1 of Atmanspacher and Filk, 2010), then it has to be the case that $p_{12} + p_{23} \geq p_{13}$. If violations of this inequality are found under some conditions, it gives reason to believe that the qualia are generally indeterminate, which is fundamental to the QQ hypothesis. (Note that qualia can be in a determinate state under some conditions under the QQ. Indeterminacy includes determinacy as a special case).

On the other hand, if qualia are generally determinate and can never be indeterminate, $p_{12} + p_{23} \geq p_{13}$ have to always apply. Without doubt, there will be many instances of qualia which indeed behave in such a classical way (as we noted above, the classical probability theory is a special case of the quantum probability theory). What is of interest is whether we can identify cases of qualia for which $p_{12} + p_{23} \geq p_{13}$ is violated. When this happens, then we can conclude that the qualia should

be considered quantum-like in general (even if they might be classical-like, in many cases).¹² The research effort for identifying such violations is still in its infancy, but there are already some promising results (Waddup et al., 2023) that showed violations of the temporal Bell inequality within a decision paradigm.

A closely related phenomenon concerns quantum Zeno effects (Atmanspacher et al., 2004; Yearsley and Pothos, 2016). Quantum Zeno effects are the surprising prediction that, everything else being equal, an increased frequency of measurements can slow down change in the relevant state. Yearsley and Pothos (2016) demonstrated the Zeno effect at the cognitive level (i.e., the switch of opinion about someone to be judged from guilty to not guilty over the accumulated evidence). If "measurements" do not affect qualia, any kind of gradual changes in qualia should not be affected by measurements. While multistable percepts change spontaneously, other types of qualia changes, such as morph-induced categorical perception and gradual change blindness, can be used to test if the effects of measurement can be precisely predicted from the quantum formulation of the Zeno effects (Atmanspacher et al., 2004; Yearsley and Pothos, 2016).

4.2.2 Establishing violations of Bell inequality in multiple qualia about an object

Another way to test the Bell inequality is to set up a task with at least three qualia observables, measuring two observables at a time, but against three different states. If qualia can be modeled classically and if measurements do not change qualia, then we expect the logical constraints, as exemplified by a Venn diagram (Figure 4A) to be satisfied by the set of probabilities. A simple diagrammatic analysis reveals various inequalities, described by George Boole as "conditions of possible experience" (Pitowsky, 1994). Pitowsky convincingly argues that quantum phenomena violate Boole's "conditions of possible experience" as these are predicated on an assumption of realism. As quantum phenomena do not always have definite properties at all times, like marbles being pulled from an urn, they can violate probabilistic relationships expressed in these inequalities.

Figure 4A demonstrates probability relationships among the three averaged measurement outcomes about three qualia observables, Color = {red, purple, orange, ...}, Position = {up, down, center, left, right}, and Shape = {circle, octagon, hexagon, ...}. Let us say, you are briefly presented with an object and you experience it with associated (narrow-sense) qualia. In classical theory, these qualia should stay the same regardless of which of two observables you report. Let $\text{Prob}(C = \text{'red'}) = p(R)$, $\text{Prob}(S = \text{'circle'}) = p(C)$, and $\text{Prob}(P = \text{'left'}) = p(L)$ represent the probability that the averaged measurement outcomes of your

¹² It is worth repeating here that even if we were to find violations of temporal Bell inequality, it does not mean that brains that support qualia are operating in non-classical mechanisms. Instead, it would exclude mathematical structures for qualia that are purely based on classical notions (e.g., determinacy). Rather more broader mathematical structures, such as quantum-like, need to be considered.

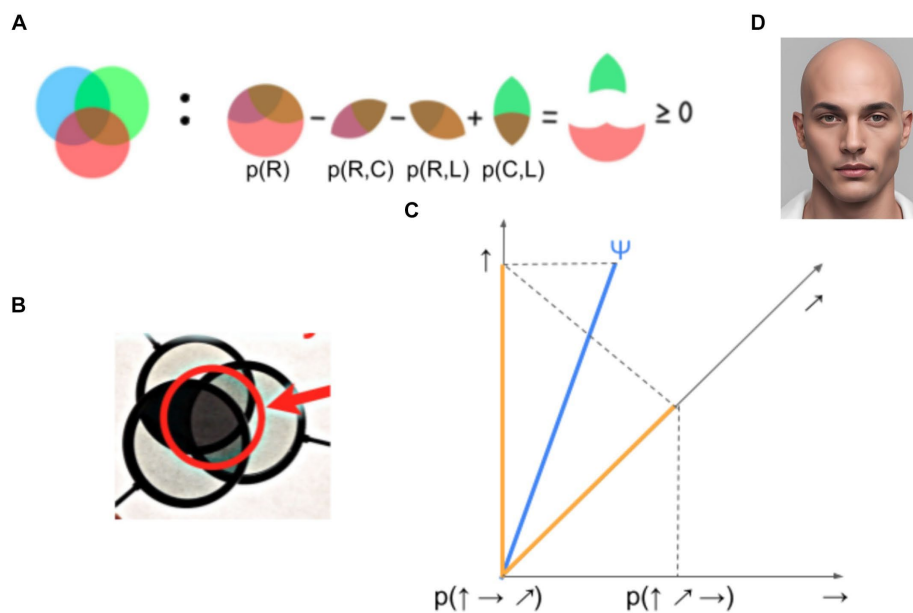


FIGURE 4

Classical probability predictions and their violations in perceptual and quantum phenomena. (A) Venn diagram of Boole's idea of possible experience. (B) Intuitive physical demonstration of the violation of the Venn diagram constraints using polarizers. See <https://www.youtube.com/watch?v=zcqZHYo7ONs>. The main idea is this: prepare 3 polarizers. By arranging two of them, you can completely block any light through them. That is, the probability of passing photons across two polarizers can be set to 0. Then, insert a third polarizer between the two. Depending on the angle of the third, the three filters can pass more photons, and thus the output beam would be brighter at the intersection of the three polarizers. (C) An explanation of (B) with a quantum projection scheme. Assume the state can be influenced by measurement. After we project the initial state Ψ to the \uparrow axis, further projection to the \rightarrow gives 0 length, which corresponds to a perfect block of photons. However, if we project to the \nearrow axis, after the \uparrow one, then third projection to the \rightarrow gives a non-zero length, explaining why more photons pass through three filters than just the two original ones. (D) An artificial face (generated by AI Canva), similar to the one used in Bruza et al. (2023), where the relationship in (A) does not hold for three aspects of the face (dominance, trustworthiness, and intelligence). Consequently, there is reason to believe that some of these facial traits were indeterminate prior to judgment.

qualia observables of the object is red, circular, and on the left, respectively. Then, we obtain that $p(R) - p(R,C) - p(R,L) + p(C,L)$ has to be always non-negative. This is easily confirmed from a Venn diagram (Figure 4A).

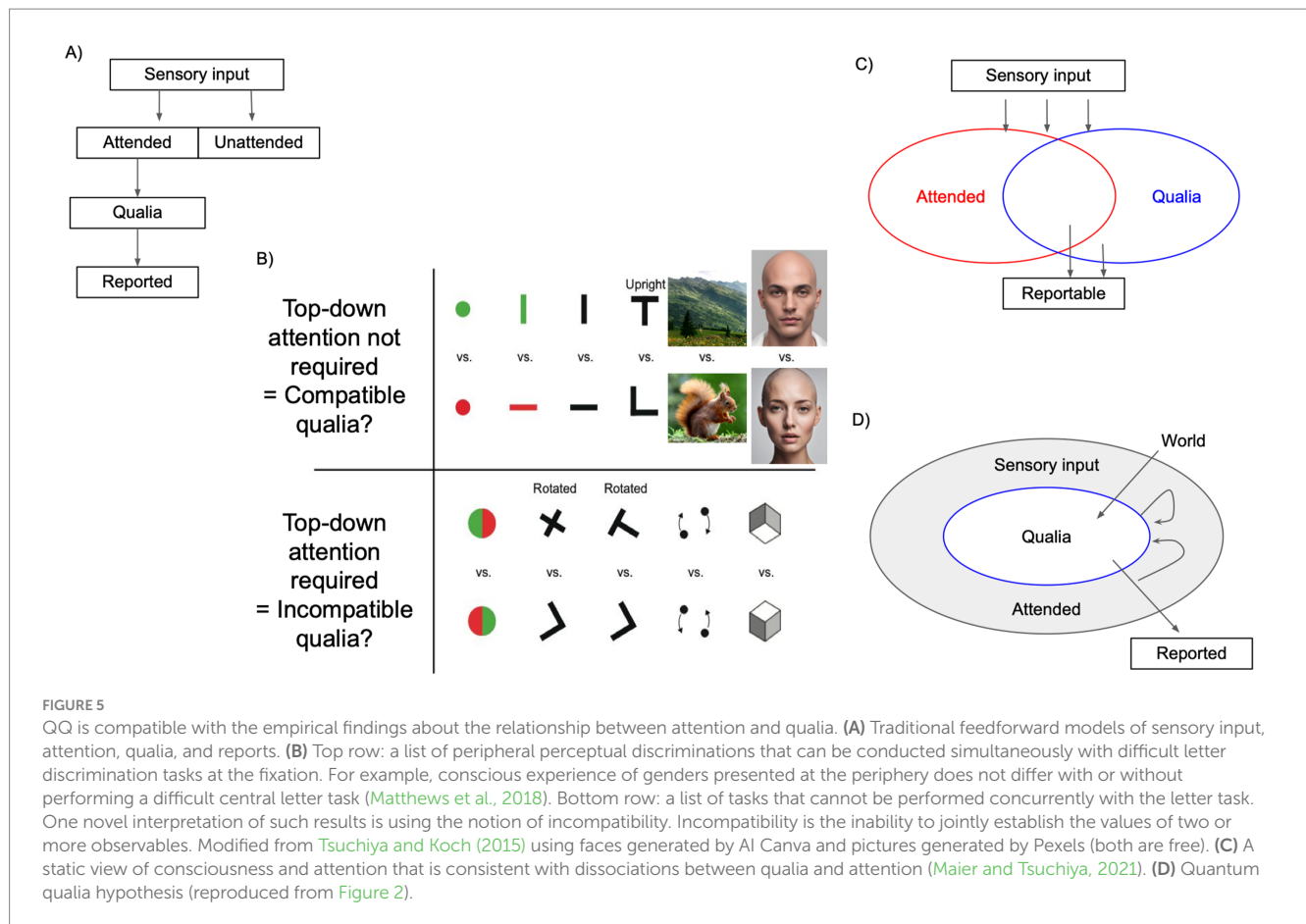
Now, imagine the object was “masked” to reduce its visibility or two such objects are simultaneously tested. The three properties can be randomly changed from trial to trial. In such a situation, your answers are likely to become probabilistic, that is, $\text{Prob}(C = \text{'red'})$, $\text{Prob}(S = \text{'circle'})$, $\text{Prob}(P = \text{'left'})$ are all smaller than 1. But, answers will still have to satisfy various probabilistic constraints. For example, $p(R) - p(R,C) - p(R,L) + p(C,L)$ has to be greater than or equal to 0, if these qualia properties follow the common sense assumptions regarding the objects being observed. Boole termed such probabilistic constraints “conditions of possible experience.” It is worth noting that classical intuitions regarding the averaged measurement outcomes are so entrenched, it is hard to imagine how things could be otherwise. Violations of such Venn diagram constraints can physically arise and are even easy to demonstrate in a classroom using just 3 polarizers (Figures 4B,C).¹³ This is an excellent demonstration to become familiar with the interesting reality of quantum phenomena, directly observable at the macro level.

Bruza and colleagues (Bruza et al., 2023) examined this constraint for qualia of a face. They considered three qualia observables. Whether faces appear trustworthy = {yes, no}, dominant = {yes, no}, and intelligent = {yes, no} (Figure 4D). It turned out that the Boole's “possibility of experience” can be violated (i.e., $p(A) - p(A,B) - p(B,C) + p(C,A) < 0$), implying that the simple classic probabilistic picture in Figure 4A is inappropriate.¹⁴

Several extensions to the above task are possible. For example, it is plausible that the degree of violation of the Bell inequality may depend on the characteristics of the qualia. If this were the case, performing the same face experiment but with reduced visibility might induce greater violations of the Bell inequality. Visual psychophysics offer a multitude of techniques to reduce visibility of an object (Kim and Blake, 2005; Stein and Peelen, 2021). As mentioned in the opening section, one of the fundamental visibility manipulations is masking. It is interesting to note that masking among three objects (Dember and Purcell, 1967; Breitmeyer et al., 1981) has been reported to be quite

¹³ <https://www.youtube.com/watch?v=zcqZHYo7ONs>

¹⁴ Note that this does not mean that the quantum-like explanation is unique and the only way to explain this result. Rather, quantum theory is able to bring together a body of insights and mechanisms, in a coherent, axiomatic framework.



complex and might reveal a promising alternative demonstration of Bell inequality violations.

One might argue that properties of faces, such as trustworthiness, dominance, and intelligence are not directly experienced qualia, but rather they are cognitively inferred constructs or concepts (Kemmerer, 2015; McClelland and Bayne, 2016). It would be a fruitful future experiment to examine if similar conclusions can be obtained when using more perceptual aspects of qualia of an object, such as color, orientation, size, location, and so on.

To sum up, one explanation for a violation of a Bell inequality is that the underlying phenomena do not have well-defined properties that exist prior to observation and distributed in a certain manner (Pitowsky, 1994). Consequently, when the inequality is violated, there is reason to believe that the phenomena are indeterminate prior to measurement. While superficially simple, definitive tests of such inequalities are subject to several checks and assumptions (Blasiak et al., 2021), and this makes it hard to definitely establish the inference from violations to indeterminacy.

While the fundamental ideas are fairly simple, almost no research on qualia has adopted a task design, where three qualia observables are measured under three states. This is understandable given that it would be difficult to motivate such a task or interpret the results, in the absence of a quantum-like theoretical framework. We believe there is a huge opportunity to test novel ideas about consciousness with the QQ formulation involving three or more observables.

4.3 Dual-task interference and non-interference between qualia in terms of incompatible and compatible observables

The relationship between consciousness and attention is one of the most debated topics in psychology, neuroscience and philosophy (Iwasaki, 1993; Hardcastle, 1997; Lamme, 2003; Dehaene et al., 2006; Block, 2007; Koch and Tsuchiya, 2007; Mole, 2008; van Boxtel et al., 2010a; Tallon-Baudry, 2011; Bor and Seth, 2012; Cohen et al., 2012; Pitts et al., 2018; Bronfman et al., 2019; Maier and Tsuchiya, 2021). QQ is mostly consistent with the known empirical findings. Moreover, QQ makes further testable predictions which are critical to empirical research in this area.

Traditionally, sensory inputs are considered to be filtered by attention first (Figure 5A), implying that attention is necessary for consciousness. Information selected with attention is experienced as qualia and subsequently reported in a feedforward manner. Only some aspects of sensory input are attended, which ostensibly give rise to particular qualia. Behavioral reports reflect the experienced qualia. In this model, typically, attention is considered as a single limited resource and any task consumes some amount of attention.

This view goes against empirical findings concerning reports of sensory inputs outside of attention. Among many empirical findings, a particularly intriguing one is a pattern of tasks that consume almost all attention and those that do not consume any

attention, as shown in Figure 5B. These properties of task combinations have been documented over the years within the “dual task” research program (Braun and Sagi, 1990; Braun and Julesz, 1998; Reddy et al., 2004; Fei-Fei et al., 2005; Pastukhov et al., 2009; Matthews et al., 2018; Bronfman et al., 2019). For example, conscious experience of genders presented at the periphery do not differ with or without performing a difficult central letter task. Meanwhile, the experience of red/green bisected disks becomes totally unclear under a dual-task with the same central task (Reddy et al., 2004, 2006). Notably, this is even the case when the disk and the face are superposed transparently at the same location (Matthews et al., 2018). One possible explanation of this pattern is the existence of attention-free specialized modules in the cortex, possibly due to biological significance or extended training (VanRullen et al., 2004).

There are many alternatives to the traditional view of attention and consciousness. One view considers consciousness and attention to operate independently (Figure 5C; Lamme, 2004; Koch and Tsuchiya, 2007). In this scheme, unattended conscious and attended unconscious processes are both possible. Attention and consciousness do not proceed in a feedforward manner. While this view is consistent with empirical findings, it does not explain how consciousness and attention interact dynamically.

The QQ hypothesis (Figures 2, 5D) explicitly considers how qualia can be affected by attention through the formalism of instruments. This does not mean that all qualia are equally affected by attention, as demonstrated by the dual task. In fact, QQ provides two novel explanations about why a given pair of tasks may not interfere with one another.

One explanation has to do with the existence of “commutative” qualia. While any process is generally noncommutative (see 3.6.1), in quantum theory, some observables, called “centers,” are always commutative with any other observables. Centers do not show any order effects. Such observables include mass. It is plausible that some types of qualia (e.g., extreme pain, bright light, loud sound) may also behave like centers and be commutative with other types of qualia. These would also be predicted to be less affected by states of measurement including attention. This is an empirical question for future research, which can be addressed by testing the presence of order effects in similarity experiments, for example.

Another explanation relates to the idea of “incompatibility.” In quantum theory, when the properties of two or more observables cannot not be generally established together, these observables are called “incompatible.” According to QQ, pairs of qualia observables that cannot be simultaneously established are deemed “incompatible.”

From the QQ perspective, it is important to point out that, in many dual tasks, a letter discrimination task is used as the primary difficult fixation task (Tsuchiya and Koch, 2015; Matthews et al., 2018). Thus, the conclusions from these studies may be revealing “incompatibility” between qualia observables of letters and others. In other words, some qualia observables, such as face gender (Matthews et al., 2018) and the presence of animals in a natural scene (Li et al., 2002; Figure 5B top row), may just be “compatible” with a letter qualia observable. These qualia observables may be “incompatible” with others. If the attentional interference happens only at the task level, we should not expect systematic

patterns in interference and order effects. However, if interference is a result of the incompatibility between specific qualia combinations, then interference would result in specific order effects with a quantitative explanation based on a quantum-like model (Epping et al., 2023).

Reconsidering the patterns of attentional limits in terms of incompatibilities between observables might allow novel insights into the qualia-attention research. With traditional psychological theories, we consider attention as a fixed resource (Joseph et al., 1997), which can amplify aspects of qualia, it is hard to explain why in some visual illusions stronger attention leads to poorer visibility of the target (Schölvinck and Rees, 2009; van Boxtel et al., 2010b). Further, it is also hard to understand why distracting participants sometimes leads to better psychological performance in various paradigms (Koch and Tsuchiya, 2007; Tsuchiya and Koch, 2015). Attention can change the neuronal circuitry momentarily (Harris and Thiele, 2011; Gilbert and Li, 2013), thus it might be possible to understand such effects as a change, for a pair of observables, from incompatible into compatible. This change can be formalized as an instrument where attention as a state affects qualia observables. This explanation offers a coherent explanation of these seemingly odd relationships between qualia and attention.

Unlike the limited resource model, QQ predicts an existence of pairs of “compatible” qualia observables, even though each one consumes a significant amount of a presumed attentional “resource.” QQ also predicts pairs of “incompatible” qualia observables, which cannot be simultaneously established, even if each does not consume much attentional resource. Discoveries of such pairs of qualia observables would further support QQ.

5 Conclusion

We proposed a Quantum-like Qualia (QQ) hypothesis based on a quantum theoretical framework (e.g., noncommutative observables, states, and instruments; Figure 2; Table 1). QQ proposes qualia as observables, not the “things” or results of “cognitive processes” as traditionally assumed. QQ explains intuitive and known properties of qualia, such as their inherent indeterminacy, dynamics, and interaction with attention. Predictions from QQ can be empirically tested with demonstrations of asymmetry in perceptual similarity judgments, violations of the Bell inequality, and apparent incompatibilities between particular qualia. Among these, particularly powerful are demonstrations of Bell inequality violations. In order to test them, we minimally need to measure three observables, two at a time across three different states (Figure 4). Such experiments have been rarely conducted systematically, due to the lack of theoretical background and motivation. Additionally, there are subtle loopholes that need to be considered, before compelling empirical evidence is provided that substantiates our claim that qualia are indeterminate (Emery, 2017; Atmanspacher and Filk, 2019; Basieva et al., 2019). In physics, it took more than twenty years from the theoretical proposal by Bell through to the initial experiment by Clauser and then to the compelling demonstration by Aspect (Section 4.2.1). Will a similar pathway await the Quantum-like Qualia hypothesis in the future? Only time will tell. With increasing evidence that QQ provides a coherent explanation on the mathematical structure of qualia, QQ

may well emerge as a promising mathematical and philosophical framework to link qualia and the brain.

Author contributions

NT: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. PB: Conceptualization, Investigation, Methodology, Supervision, Writing – review & editing. MY: Investigation, Methodology, Writing – review & editing. HS: Supervision, Writing – review & editing, Methodology, Investigation, Conceptualization. EP: Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. NT, MY, and HS were supported by the Japan Society for the Promotion of Science Grant-in-Aid for Transformative Research Areas (B) (20H05709, 20H05710, and 20H05711) and (A) (23H04829, 23H04830, 23H04833), JST CREST (Grant No. JPMJCR23P4), Japan, and KAKENHI (Grant No. 22K18265), Japan. NT and HS were supported by Foundational Question Institute (NT, FQXi-RFP-CPW-2017). MY was supported by JST Moonshot R&D Grant (JPMJMS2295-01), JSPS KAKENHI Grant (22H01108), and MEXT Quantum Leap Flagship Program (MEXT QLEAP) (Grant No. JPMXS0120330644). EP was supported by AFOSR (Grant No.

FA8655-23-1-7220). NT was funded by Australian Research Council Discovery Projects DP180104128, DP180100396, and DP240102680 from the National Health and Medical Research Council (APP1183280). PB was supported by the Air Force Office of Scientific Research under award number: FA9550-23-1-0258. The funders had no role in study design, decision to publish, or preparation of the manuscript.

Acknowledgments

Thanks to Angus Leung, Alex Maier, and Steven Phillips for comments on the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Aerts, D., and Arguëlles, J. A. (2022). Human perception as a phenomenon of quantization. *Entropy* 24:1207. doi: 10.3390/e24091207
- Aerts, D., Arguëlles, J. A., Beltran, L., Geriente, S., Sassoli de Bianchi, M., Sozzo, S., et al. (2018). Spin and wind directions I: identifying entanglement in nature and cognition. *Found. Sci.* 23, 323–335. doi: 10.1007/s10699-017-9528-9
- Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., et al. (2022). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLOS Computational Biology*. 19, e1011465. doi: 10.1371/journal.pcbi.1011465
- Arguëlles, J. A. (2018). The heart of an image: quantum superposition and entanglement in visual perception. *Found. Sci.* 23, 757–778. doi: 10.1007/s10699-018-9547-1
- Asano, M., Hashimoto, T., Khrennikov, A., Ohya, M., and Tanaka, Y. (2014). Violation of contextual generalization of the Leggett–Garg inequality for recognition of ambiguous figures. *Phys. Scr.* T163:014006. doi: 10.1088/0031-8949/2014/T163/014006
- Ashby, F. G., and Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychol. Rev.* 95, 124–150. doi: 10.1037/0033-295X.95.1.124
- Aspect, A., Dalibard, J., and Roger, G. (1982). Experimental test of Bell's inequalities using time-varying analyzers. *Phys. Rev. Lett.* 49, 1804–1807. doi: 10.1103/PhysRevLett.49.1804
- Atmanspacher, H. (2017). "Quantum approaches to brain and mind: an overview with representative examples" in *The Blackwell companion to consciousness*. eds. S. Schneider and M. Velmans. 1st ed (Hoboken, NJ: Wiley), 298–313.
- Atmanspacher, H., and Filk, T. (2010). A proposed test of temporal nonlocality in bistable perception. *J. Math. Psychol.* 54, 314–321. doi: 10.1016/j.jmp.2009.12.001
- Atmanspacher, H., and Filk, T. (2018). Non-commutativity and its implications in physics and beyond. Orpheus' Glance. Selected Papers on Process Psychology: The Fontarèches Meetings, 2002–2017, 45–60.
- Atmanspacher, H., and Filk, T. (2019). Contextuality revisited: signaling may differ from communicating. In Barros, A. de and C. Montemayor (Eds.), *Quanta and mind* (Vol. 414, pp. 117–127). New York: Springer International Publishing.
- Atmanspacher, H., Filk, T., and Römer, H. (2004). Quantum Zeno features of bistable perception. *Biol. Cybern.* 90, 33–40. doi: 10.1007/s00422-003-0436-4
- Atmanspacher, H., and Müller-Herold, U. (2016). *From chemistry to consciousness*. New York: Springer International Publishing.
- Bachmann, T. (2000). *Microgenetic approach to the conscious mind*, vol. 25. Amsterdam: John Benjamins Publishing.
- Balduzzi, D., and Tononi, G. (2009). Qualia: the geometry of integrated information. *PLoS Comput. Biol.* 5:e1000462. doi: 10.1371/journal.pcbi.1000462
- Barsalou, L. W. (2010). Grounded cognition: past, present, and future: topics in cognitive science. *Top. Cogn. Sci.* 2, 716–724. doi: 10.1111/j.1756-8765.2010.01115.x
- Basieva, I., Cervantes, V. H., Dzhaferov, E. N., and Khrennikov, A. (2019). True contextuality beats direct influences in human decision making. *J. Exp. Psychol. Gen.* 148, 1925–1937. doi: 10.1037/xge0000585
- beim Graben, P., and Blutner, R. (2019). Quantum approaches to music cognition. *J. Math. Psychol.* 91, 38–50. doi: 10.1016/j.jmp.2019.03.002
- Bell, J. S. (1964). On the einstein podolsky rosen paradox. *Physics Physique Fizika* 1, 195–200. doi: 10.1103/PhysicsPhysiqueFizika.1.195
- Best, R. M., and Goldstone, R. L. (2019). Bias to (and away from) the extreme: comparing two models of categorical perception effects. *J. Exp. Psychol. Learn. Mem. Cogn.* 45, 1166–1176. doi: 10.1037/xlm0000609
- Blasiak, P., Pothos, E. M., Yearsley, J. M., Gallus, C., and Borsuk, E. (2021). Violations of locality and free choice are equivalent resources in Bell experiments. *Proc. Natl. Acad. Sci.* 118:e2020569118. doi: 10.1073/pnas.2020569118
- Block, N. (1998). Is experiencing just representing? *Philos. Phenomenol. Res.* 58, 663–670. doi: 10.2307/2653766
- Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behav. Brain Sci.* 30, 481–499. doi: 10.1017/S0140525X07002786
- Bor, D., and Seth, A. K. (2012). Consciousness and the prefrontal parietal network: insights from attention, working memory, and chunking. *Front. Psychol.* 3:63. doi: 10.3389/fpsyg.2012.00063
- Borg, I., and Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications*. 2nd Edn. Berlin: Springer.

- Brascamp, J., Sterzer, P., Blake, R., and Knapen, T. (2018). Multistable perception and the role of the Frontoparietal cortex in perceptual inference. *Annu. Rev. Psychol.* 69, 77–103. doi: 10.1146/annurev-psych-010417-085944
- Braun, J., and Julesz, B. (1998). Withdrawing attention at little or no cost: detection and discrimination tasks. *Percept. Psychophys.* 60, 1–23. doi: 10.3758/BF03211915
- Braun, J., and Sagi, D. (1990). Vision outside the focus of attention. *Percept. Psychophys.* 48, 45–58. doi: 10.3758/BF03205010
- Breitmeyer, B. G., and Ogmen, H. (2007). Visual masking. *Scholarpedia* 2:3330. doi: 10.4249/scholarpedia.3330
- Breitmeyer, B. G., Rudd, M., and Dunn, K. (1981). Metacontrast investigations of sustained-transient channel inhibitory interactions. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 770–779. doi: 10.1037/0096-1523.7.4.770
- Broekaert, J. B., Busemeyer, J. R., and Pothos, E. M. (2020). The disjunction effect in two-stage simulated gambles. An experimental study and comparison of a heuristic logistic, Markov and quantum-like model. *Cogn. Psychol.* 117:101262. doi: 10.1016/j.cogpsych.2019.101262
- Bronfman, Z. Z., Jacobson, H., and Usher, M. (2019). Impoverished or rich consciousness outside attentional focus: recent data tip the balance for overflow. *Mind Lang.* 34, 423–444. doi: 10.1111/mila.12217
- Bruza, P. D., Fell, L., Hoyte, P., Dehdashti, S., Obeid, A., Gibson, A., et al. (2023). Contextuality and context-sensitivity in probabilistic models of cognition. *Cogn. Psychol.* 140:101529. doi: 10.1016/j.cogpsych.2022.101529
- Bruza, P. D., Kitto, K., Ramm, B. J., and Sitbon, L. (2015). A probabilistic framework for analysing the compositionality of conceptual combinations. *J. Math. Psychol.* 67, 26–38. doi: 10.1016/j.jmp.2015.06.002
- Bujack, R., Teti, E., Miller, J., Caffrey, E., and Turton, T. L. (2022). The non-Riemannian nature of perceptual color space. *Proc. Natl. Acad. Sci.* 119:e2119753119. doi: 10.1073/pnas.2119753119
- Busemeyer, J. R., and Bruza, P. D. (2024). *Quantum models of cognition and decision*. Cambridge: Cambridge University Press.
- Busemeyer, J. R., Kvam, P. D., and Pleskac, T. J. (2019). Markov versus quantum dynamic models of belief change during evidence monitoring. *Sci. Rep.* 9:18025. doi: 10.1038/s41598-019-54383-9
- Busemeyer, J. R., and Wang, Z. (2017). Is there a problem with quantum models of psychological measurements? *PLoS One* 12:e0187733. doi: 10.1371/journal.pone.0187733
- Carrasco, M., and Barbot, A. (2019). Spatial attention alters visual appearance. *Curr. Opin. Psychol.* 29, 56–64. doi: 10.1016/j.copsyc.2018.10.010
- Casarotto, S., Comanducci, A., Rosanova, M., Sarasso, S., Fedchio, M., Napolitano, M., et al. (2016). Stratification of unresponsive patients by an independently validated index of brain complexity. *Ann. Neurol.* 80, 718–729. doi: 10.1002/ana.24779
- Cervantes, V. H., and Dzhabarov, E. N. (2019). True contextuality in a psychophysical experiment. *J. Math. Psychol.* 91, 119–127. doi: 10.1016/j.jmp.2019.04.006
- Chalmers, D. J., and McQueen, K. J. (2021). Consciousness and the collapse of the wave function. *ArXiv*. doi: 10.48550/arXiv.2105.02314
- Cheng, E. (2022). *The joy of abstraction an exploration of math, category theory, and life*. Cambridge: Cambridge University Press.
- Churchland, P. (2005). Chimerical colors: some phenomenological predictions from cognitive neuroscience. *Philos. Psychol.* 18, 527–560. doi: 10.1080/09515080500264115
- Clark, A. (2000). *A theory of sentience*. Oxford: Oxford University Press.
- Cohen, M. A., Cavanagh, P., Chun, M. M., and Nakayama, K. (2012). The attentional requirements of consciousness. *Trends Cogn. Sci.* 16, 411–417. doi: 10.1016/j.tics.2012.06.013
- Conte, E., Khrennikov, A. Y., Todarello, O., Federici, A., Mendolicchio, L., and Zbilut, J. P. (2009). Mental states follow quantum mechanics during perception and cognition of ambiguous figures. *Open Syst. Inform. Dyn.* 16, 85–100. doi: 10.1142/S1230161209000074
- Cowen, A. S., Fang, X., Sauter, D., and Keltner, D. (2020). What music makes us feel: at least 13 dimensions organize subjective experiences associated with music across different cultures. *Proc. Natl. Acad. Sci.* 117, 1924–1934. doi: 10.1073/pnas.1910704117
- Cowen, A. S., and Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Natl. Acad. Sci.* 114, E7900–E7909. doi: 10.1073/pnas.1702247114
- Davies, E. B., and Lewis, J. T. (1970). An operational approach to quantum probability. *Commun. Math. Phys.* 17, 239–260. doi: 10.1007/BF01647093
- Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J., and Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *TICS* 10, 204–211. doi: 10.1016/j.tics.2006.03.007
- Dember, W. N., and Purcell, D. G. (1967). Recovery of masked visual targets by inhibition of the masking stimulus. *Science* 157, 1335–1336. doi: 10.1126/science.157.3794.1335
- Einstein, A., Podolsky, B., and Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.* 47, 777–780. doi: 10.1103/PhysRev.47.777
- Emary, C. (2017). Ambiguous measurements, signalling and violations of Leggett-Garg inequalities. *Phys. Rev. A* 96:042102. doi: 10.1103/PhysRevA.96.042102
- Epping, G. P., Fisher, E. L., Zeleznikow-Johnston, A. M., Pothos, E. M., and Tsuchiya, N. (2023). A quantum geometric framework for modeling color similarity judgments. *Cogn. Sci.* 47:e13231. doi: 10.1111/cogs.13231
- Esteban, F. J., Galadí, J. A., Langa, J. A., Portillo, J. R., and Soler-Toscano, F. (2018). Informational structures: a dynamical system approach for integrated information. *PLoS Comput. Biol.* 14:e1006154. doi: 10.1371/journal.pcbi.1006154
- Exner, S. (1868). Über die zu einer Gesichtswahrnehmung nöthige Zeit. *Wiener Sitzungsber Math-Naturwissensch Cl Kaiser Akad Wissensch* 58, 601–632.
- Fei-Fei, L., VanRullen, R., Koch, C., and Perona, P. (2005). Why does natural scene categorization require little attention? Exploring attentional requirements for natural and synthetic stimuli. *Vis. Cogn.* 12, 893–924. doi: 10.1080/13506280444000571
- Fekete, T., and Edelman, S. (2011). Towards a computational theory of experience. *Conscious. Cogn.* 20, 807–827. doi: 10.1016/j.concog.2011.02.010
- Filk, T. (2009). Quantum physics and consciousness: the quest for a common, modified conceptual foundation. *Mind Matter* 7, 59–79.
- Filk, T. (2013). Temporal non-locality. *Found. Phys.* 43, 533–547. doi: 10.1007/s10701-012-9671-7
- Freedman, S. J., and Clauser, J. F. (1972). Experimental test of local hidden-variable theories. *Phys. Rev. Lett.* 28, 938–941. doi: 10.1103/PhysRevLett.28.938
- Gärdenfors, P. (2000). “Conceptual spaces: The geometry of thought” in *A Bradford Book* (Cambridge, MA: MIT Press).
- Gibson, J. J. (2014). *The ecological approach to visual perception*. Classic Edn. London: Psychology press.
- Gilbert, C. D., and Li, W. (2013). Top-down influences on visual processing. *Nat. Rev. Neurosci.* 14, 350–363. doi: 10.1038/nrn3476
- Graham, D. J., Friedenberg, J. D., Rockmore, D. N., and Field, D. J. (2010). Mapping the similarity space of paintings: image statistics and visual perception. *Vis. Cogn.* 18, 559–573. doi: 10.1080/13506280902934454
- Hameroff, S., and Penrose, R. (2014). Consciousness in the universe: a review of the ‘Orch OR’ theory. *Phys Life Rev* 11, 39–78. doi: 10.1016/j.plrev.2013.08.002
- Hardcastle, V. G. (1997). Attention versus consciousness: a distinction with a difference. *Cogn. Stud. Bull. Jpn. Cogn. Sci. Soc.* 4, 56–66.
- Harris, K. D., and Thiele, A. (2011). Cortical state and attention. *Nat. Rev. Neurosci.* 12, 509–523. doi: 10.1038/nrn3084
- Haven, E., and Khrennikov, A. (2013). *Quantum Social Science*. 1st Edn. Cambridge: Cambridge University Press.
- Hebart, M. N., Zheng, C. Y., Pereira, F., and Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nat. Hum. Behav.* 4, 1173–1185. doi: 10.1038/s41562-020-00951-3
- Ho, J. K. W., and Hoorn, J. F. (2022). Quantum affective processes for multidimensional decision-making. *Sci. Rep.* 12:20468. doi: 10.1038/s41598-022-22855-0
- Hodgetts, C. J., and Hahn, U. (2012). Similarity-based asymmetries in perceptual matching. *Acta Psychol.* 139, 291–299. doi: 10.1016/j.actpsy.2011.12.003
- Hoffman, D. D., Prakash, C., and Prentner, R. (2023). Fusions of consciousness. *Entropy* 25:129. doi: 10.3390/e25010129
- Hoffman, D. D., Singh, M., and Prakash, C. (2015). The Interface theory of perception. *Psychon. Bull. Rev.* 22, 1480–1506. doi: 10.3758/s13423-015-0890-8
- Holton, G. (1988). “The Roots of Complementarity.” In *Quantum Mechanics*, Routledge. 253–93.
- Huang, J. A., Zhang, Q., Busemeyer, J. R., and Breithaupt, F. (2022). A quantum walk model for emotion transmission in serial reproduction of narratives. *Proceedings of the Annual Meeting of the Cognitive Science Society*. 44. Available at: <https://escholarship.org/uc/item/2tj1w6hw>
- Indow, T. (1988). Multidimensional studies of Munsell color solid. *Psychol. Rev.* 95, 456–470. doi: 10.1037/0033-295X.95.4.456
- Iwasaki, S. (1993). Spatial attention and two modes of visual consciousness. *Cognition* 49, 211–233. doi: 10.1016/0010-0277(93)90005-G
- Joseph, J. S., Chun, M. M., and Nakayama, K. (1997). Attentional requirements in a “preattentive” feature search task. *Nature* 387, 805–807. doi: 10.1038/42940
- Kahneman, D. (2003). Experiences of collaborative research. *Am. Psychol.* 58, 723–730. doi: 10.1037/0003-066X.58.9.723
- Kanai, R., and Tsuchiya, N. (2012). Qualia. *Curr. Biol.* 22, R392–R396. doi: 10.1016/j.cub.2012.03.033
- Kawakita, G., Zeleznikow-Johnston, A., Takeda, K., Tsuchiya, N., and Oizumi, M. (2023). Is my “red” your “red”? unsupervised alignment of qualia structures via optimal transport. *PsyArXiv*. doi: 10.31234/osf.io/h3pqm
- Kemmerer, D. (2015). Are we ever aware of concepts? A critical question for the global neuronal workspace, integrated information, and attended intermediate-level representation theories of consciousness. *Neurosci. Consciousness* 2015:6. doi: 10.1093/nc/niv006

- Khrennikov, A. Y. (2010). *Ubiquitous Quantum Structure*. Berlin: Springer Berlin Heidelberg.
- Khrennikov, A. (2015). Quantum-like model of unconscious-conscious dynamics. *Front. Psychol.* 6:997. doi: 10.3389/fpsyg.2015.00997
- Khrennikov, A. (2021). Quantum-like model for unconscious-conscious interaction and emotional coloring of perceptions and other conscious experiences. *Biosystems* 208:104471. doi: 10.1016/j.biosystems.2021.104471
- Kim, C. Y., and Blake, R. (2005). Psychophysical magic: rendering the visible 'invisible'. *TICS* 9, 381–388. doi: 10.1016/j.tics.2005.06.012
- Kirchhoff, M., Parr, T., Palacios, E., Friston, K., and Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *J. R. Soc. Interface* 15:20170792. doi: 10.1098/rsif.2017.0792
- Kleiner, J. (2024). Towards a structural turn in consciousness science. *Conscious. Cogn.* 119:103653. doi: 10.1016/j.concog.2024.103653
- Kleiner, J., and Ludwig, T. (2024). What is a mathematical structure of conscious experience? *Synthese* 203:89. doi: 10.1007/s11229-024-04503-4
- Klincewicz, M. (2011). "Quality space model of temporal perception" in *Multidisciplinary aspects of time and time perception*. eds. A. Vatakis, A. Esposito, M. Giagkou, F. Cummins and G. Papadelis, vol. 6789 (Berlin: Springer Berlin Heidelberg), 230–245.
- Koch, C., Massimini, M., Boly, M., and Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nat. Rev. Neuro.* 17, 307–321. doi: 10.1038/nrn.2016.22
- Koch, C., and Tsuchiya, N. (2007). Attention and consciousness: two distinct brain processes. *Trends Cogn. Sci.* 11, 16–22. doi: 10.1016/j.tics.2006.10.012
- Koenderink, J. J. (2010). "Vision and information" in *Perception beyond inference: The information content of visual processes* (Cambridge, MA: MIT Press), 27–57.
- Kostic, D. (2012). The vagueness constraint and the quality space for pain. *Philos. Psychol.* 25, 929–939. doi: 10.1080/09515089.2011.633696
- Krumhansl, C. L. (1978). Concerning the Applicability of Geometric Models to Similarity Data: The Interrelationship between Similarity and Spatial Density. 85, 445–163.
- Lamme, V. A. (2003). Why visual attention and awareness are different. *Trends Cogn. Sci.* 7, 12–18. doi: 10.1016/S1364-6613(02)00013-X
- Lamme, V. A. (2004). Separate neural definitions of visual consciousness and visual attention; a case for phenomenal awareness. *Neural Netw.* 17, 861–872. doi: 10.1016/j.neunet.2004.02.005
- Lee, A. Y. (2021). Modeling Mental Qualities. *Philos. Rev.* 130, 263–298. doi: 10.1215/00318108-8809919
- Li, F. F., VanRullen, R., Koch, C., and Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proc. Natl. Acad. Sci. USA* 99, 9596–9601. doi: 10.1073/pnas.092277599
- Lin, L. (2023). The Relational Structure of Emotional Experience: A Novel Paradigm for Characterizing Differences Between Alexithymic and General Online Participants. *Zenodo*. doi: 10.5281/zenodo.10252326
- Lyre, H. (2022). Neuropsychological structuralism. A philosophical agenda for a structuralist neuroscience of consciousness. *Neurosci. Consciousness* 2022:niac012. doi: 10.1093/nc/niac012
- Maier, A., Panagiotaropoulos, T. I., Tsuchiya, N., and Keliris, G. A. (2012). Introduction to research topic—binocular rivalry: a gateway to studying consciousness. *Front. Hum. Neurosci.* 6:263. doi: 10.3389/fnhum.2012.00263
- Maier, A., and Tsuchiya, N. (2021). Growing evidence for separate neural mechanisms for attention and consciousness. *Atten. Percept. Psychophys.* 83, 558–576. doi: 10.3758/s13414-020-02146-4
- Mashour, G. A., Roelfsema, P., Changeux, J.-P., and Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105, 776–798. doi: 10.1016/j.neuron.2020.01.026
- Matthews, J., Schroder, P., Kaunitz, L., van Bostel, J. J. A., and Tsuchiya, N. (2018). Conscious access in the near absence of attention: critical extensions on the dual-task paradigm. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 373:352. doi: 10.1098/rstb.2017.0352
- McClelland, T., and Bayne, T. (2016). Concepts, contents, and consciousness. *Neurosci. Conscious.* 2016:12. doi: 10.1093/nc/niv012
- Melloni, L., Mudrik, L., Pitts, M., and Koch, C. (2021). Making the hard problem of consciousness easier. *Science* 372, 911–912. doi: 10.1126/science.abj3259
- Melloni, L., Schwiedrzik, C. M., Muller, N., Rodriguez, E., and Singer, W. (2011). Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness. *J. Neurosci.* 31, 1386–1396. doi: 10.1523/JNEUROSCI.4570-10.2011
- Mistry, P. K., Pothos, E. M., Vandekerckhove, J., and Trueblood, J. S. (2018). A quantum probability account of individual differences in causal reasoning. *J. Math. Psychol.* 87, 76–97. doi: 10.1016/j.jmp.2018.09.003
- Mole, C. (2008). Attention in the absence of consciousness? *Trends Cogn. Sci.* 12:44. doi: 10.1016/j.tics.2007.11.001
- Moyal, R., Fekete, T., and Edelman, S. (2020). Dynamical emergence theory (DET): a computational account of phenomenal consciousness. *Mind. Mach.* 30, 1–21. doi: 10.1007/s11023-020-09516-9
- Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cogn. Psychol.* 23, 94–140. doi: 10.1016/0010-0285(91)90004-8
- Nummenmaa, L., Hari, R., Hietanen, J. K., and Glerean, E. (2018). Maps of subjective feelings. *Proc. Natl. Acad. Sci.* 115, 9198–9203. doi: 10.1073/pnas.1807390115
- Ojima, I. (2005). "Micro-Macro Duality in Quantum Physics." in *Stochastic Analysis: Classical and Quantum*, WORLD SCIENTIFIC, 143–161. doi: 10.1142/9789812701541_0012
- Ozawa, M., and Khrennikov, A. (2021). Modeling combination of question order effect, response replicability effect, and QQ-equality with quantum instruments. *J. Math. Psychol.* 100:102491. doi: 10.1016/j.jmp.2020.102491
- Pastukhov, A., Fischer, L., and Braun, J. (2009). Visual attention is a single, integrated resource. *Vis. Res.* 49, 1166–1173. doi: 10.1016/j.visres.2008.04.011
- Pitowsky, I. (1994). George Boole's 'conditions of possible experience' and the quantum puzzle. *Br. J. Philos. Sci.* 45, 95–125. doi: 10.1093/bjps/45.1.95
- Pitts, M. A., Lutsyshyna, L. A., and Hillyard, S. A. (2018). The relationship between attention and consciousness: an expanded taxonomy and implications for 'no-report' paradigms. *Philos. Trans. Royal Society B* 373:20170348. doi: 10.1098/rstb.2017.0348
- Plotnitsky, A. (2021). *Reality without realism: Matter, thought, and Technology in Quantum Physics*. New York: Springer International Publishing.
- Polk, T. A., Behensky, C., Gonzalez, R., and Smith, E. E. (2002). Rating the similarity of simple perceptual stimuli: asymmetries induced by manipulating exposure frequency. *Cognition* 82, B75–B88. doi: 10.1016/S0010-0277(01)00151-2
- Pothos, E. M., and Busemeyer, J. R. (2022). Quantum Cognition. *Annu. Rev. Psychol.* 73, 749–778. doi: 10.1146/annurev-psych-033020-123501
- Pothos, E. M., Busemeyer, J. R., and Trueblood, J. S. (2013). A quantum geometric model of similarity. *Psychol. Rev.* 120, 679–696. doi: 10.1037/a0033142
- Prakash, C., Fields, C., Hoffman, D. D., Prentner, R., and Singh, M. (2020). Fact, fiction, and fitness. *Entropy* 22:514. doi: 10.3390/e22050514
- Prentner, R. (2021). Dr Goff, tear down this wall! The Interface theory of perception and the science of Consciousness. *J. Conscious. Stud.* 28, 91–103. doi: 10.5376/20512201.28.9.091
- Reddy, L., Reddy, L., and Koch, C. (2006). Face identification in the near-absence of focal attention. *Vis. Res.* 46, 2336–2343. doi: 10.1016/j.visres.2006.01.020
- Reddy, L., Wilken, P., and Koch, C. (2004). Face-gender discrimination is possible in the near-absence of attention. *J. Vis.* 4, 106–117. doi: 10.1167/4.2.4
- Renner, A. (2014). Consciousness and mental qualities for auditory sensations. *J. Conscious. Stud.* 21, 179–204. doi: 10.5376/20512201.21.9.179
- Roberson, D., Damjanovic, L., and Pilling, M. (2007). Categorical perception of facial expressions: evidence for a "category adjustment" model. *Mem. Cogn.* 35, 1814–1829. doi: 10.3758/BF03193512
- Rosch, E. (1975). Cognitive reference points. *Cogn. Psychol.* 7, 532–547. doi: 10.1016/0010-0285(75)90021-3
- Rosenthal, D. (2015). "Quality spaces and sensory modalities" in *Phenomenal qualities*. eds. P. Coates and S. Coleman (Oxford: Oxford University Press), 33–65.
- Saigo, H. (2021). Quantum Fields as category algebras. *Symmetry* 13:9. doi: 10.3390/sym13091727
- Saigo, H., Naruse, M., Okamura, K., Hori, H., and Ojima, I. (2019). Analysis of soft robotics based on the concept of category of mobility. *Complexity* 2019, 1–12. doi: 10.1155/2019/1490541
- Sakurai, J., and Napolitano, J. (2020). *Modern Quantum Mechanics* (3rd ed.). Cambridge: Cambridge University Press.
- Schlicht, T., and Starzak, T. (2021). Prospects of enactivist approaches to intentionality and cognition. *Synthese* 198, 89–113. doi: 10.1007/s11229-019-02361-z
- Schölvinc, M. L., and Rees, G. (2009). Attentional influences on the dynamics of motion-induced blindness. *J. Vis.* 9, 38–38.9. doi: 10.1167/9.1.38
- Seth, A. K., and Bayne, T. (2022). Theories of consciousness. *Nat. Rev. Neurosci.* 23, 439–452. doi: 10.1038/s41583-022-00587-4
- Shepard, R. N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika* 27, 125–140. doi: 10.1007/BF02289630
- Shepard, R. N. (1962b). The analysis of proximities: multidimensional scaling with an unknown distance function. II. *Psychometrika* 27, 219–246. doi: 10.1007/BF02289621
- Shepard, R. N., and Susan, C. (1970). "Second-Order Isomorphism of Internal Representations: Shapes of States." *Cognitive Psychology* 1, 1–17.
- Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science* 210, 390–398. doi: 10.1126/science.210.4468.390

- Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychol. Rev.* 89, 305–333. doi: 10.1037/0033-295X.89.4.305
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science* 237, 1317–1323. doi: 10.1126/science.3629243
- Shepard, R. N., and Cooper, L. A. (1992). Representation of colors in the blind, color-blind, and normally sighted. *Psychol. Sci.* 3, 97–104. doi: 10.1111/j.1467-9280.1992.tb00006.x
- Simons, D. J., and Rensink, R. A. (2005). Change blindness: past, present, and future. *Trends Cogn. Sci.* 9, 16–20. doi: 10.1016/j.tics.2004.11.006
- Smolin, L. (2022). “On the place of qualia in a relational universe” in *Consciousness and quantum mechanics*. ed. S. Gao (Oxford: Oxford University Press), 0.
- Song, C., Haun, A. M., and Tononi, G. (2017). Plasticity in the structure of visual space. *Eneuro* 4:ENEURO.0080-17.2017. doi: 10.1523/ENEURO.0080-17.2017
- Song, Q., Wang, W., Fu, W., Sun, Y., Wang, D., and Gao, Z. (2022). Research on quantum cognition in autonomous driving. *Sci. Rep.* 12:300. doi: 10.1038/s41598-021-04239-y
- Stein, T., and Peelen, M. V. (2021). Dissociating conscious and unconscious influences on visual detection effects. *Nat. Hum. Behav.* 5, 612–624. doi: 10.1038/s41562-020-01004-5
- Streater (2007). *Lost causes in and beyond physics*. Berlin: Springer Berlin Heidelberg.
- Surov, I. A., Semenenko, E., Platonov, A. V., Bessmertny, I. A., Galofaro, F., Toffano, Z., et al. (2021). Quantum semantics of text perception. *Sci. Rep.* 11:4193. doi: 10.1038/s41598-021-83490-9
- Taguchi, S. (2019). Mediation based phenomenology. *Metodo* 7, 17–44. doi: 10.19079/metodo.7.2.17
- Tallon-Baudry, C. (2011). On the neural mechanisms subserving consciousness and attention. *Front. Psychol.* 2:397. doi: 10.3389/fpsyg.2011.00397
- Tesař, J. (2020). A quantum model of strategic decision-making explains the disjunction effect in the Prisoner's dilemma game. *Decision* 7, 43–54. doi: 10.1037/dec0000110
- Tsuchiya, N., and Koch, C. (2015). “The relationship between consciousness and top-down attention” in *The neurology of consciousness*. eds. S. Laureys, G. Tononi and O. Gosseries. 2nd ed (Cambridge, MA: Academic Press), 69–89.
- Tsuchiya, N., Phillips, S., and Saigo, H. (2022). Enriched category as a model of qualia structure based on similarity judgements. *Conscious. Cogn.* 101:103319. doi: 10.1016/j.concog.2022.103319
- Tsuchiya, N., and Saigo, H. (2021). A relational approach to consciousness: categories of level and contents of consciousness. *Neurosci. Consciousness* 2021:34. doi: 10.1093/nc/niab034
- Tsuchiya, N., Saigo, H., and Phillips, S. (2023). An adjunction hypothesis between qualia and reports. *Front. Psychol.* 13:1053977. doi: 10.3389/fpsyg.2022.1053977
- Tsuchiya, N., Taguchi, S., and Saigo, H. (2016). Using category theory to assess the relationship between consciousness and integrated information theory. *Neurosci. Res.* 107, 1–7. doi: 10.1016/j.neures.2015.12.007
- Tversky, A. (1977). Features of similarity. *Psychol. Rev.* 84, 327–352. doi: 10.1037/0033-295X.84.4.327
- Tye, M. (2021). “Qualia” in *The Stanford encyclopedia of philosophy (fall 2021)*. ed. E. N. Zalta (Stanford, CA: Stanford University).
- Tyler, C. W. (2015). Peripheral color demo. *I-Perception* 6:204166951561367. doi: 10.1177/2041669515613671
- van Boxtel, J. J., Tsuchiya, N., and Koch, C. (2010a). Consciousness and attention: on sufficiency and necessity. *Front. Psychol.* 1:217. doi: 10.3389/fpsyg.2010.00217
- van Boxtel, J. J., Tsuchiya, N., and Koch, C. (2010b). Opposing effects of attention and consciousness on afterimages. *PNAS* 107, 8883–8888. doi: 10.1073/pnas.0913292107
- VanRullen, R., Reddy, L., and Koch, C. (2004). Visual search and dual tasks reveal two distinct attentional resources. *J. Cogn. Neurosci.* 16, 4–14. doi: 10.1162/089892904322755502
- Varela, F., Evan Thompson, J., and Rosch, E. (2017). *The embodied Mind_ cognitive science and human experience*. 2nd Edn. Cambridge, MA: The MIT Press.
- Waddup, O. J., Yearsley, J. M., Blasiak, P., and Pothos, E. M. (2023). Temporal Bell inequalities in cognition. *Psychon. Bull. Rev.* 30, 1946–1953. doi: 10.3758/s13423-023-02275-5
- Wang, Z., and Busemeyer, J. R. (2013). A quantum question order model supported by empirical tests of an a priori and precise prediction. *Top. Cogn. Sci.* 5, 689–710. doi: 10.1111/tops.12040
- Wang, D., Sadzadeh, M., Abramsky, S., and Cervantes, V. H. (2021). On the quantum-like Contextuality of ambiguous phrases (arXiv:2107.14589). *arXiv*. doi: 10.48550/arXiv.2107.14589
- Wang, Z., Solloway, T., Shiffrin, R. M., and Busemeyer, J. R. (2014). Context effects produced by question orders reveal quantum nature of human judgments. *Proc. Natl. Acad. Sci.* 111, 9431–9436. doi: 10.1073/pnas.1407756111
- White, L. C., Pothos, E. M., and Jarrett, M. (2020). The cost of asking: how evaluations Bias subsequent judgments. *Decision* 7, 259–286. doi: 10.1037/dec0000136
- Wojciechowski, B. W., Izydorczyk, B., Blasiak, P., Yearsley, J. M., White, L. C., and Pothos, E. M. (2022). Constructive biases in clinical judgment. *Top. Cogn. Sci.* 14, 508–527. doi: 10.1111/tops.12547
- Yearsley, J. M., and Pothos, E. M. (2014). Challenging the classical notion of time in cognition: a quantum perspective. *Proc. R. Soc. B Biol. Sci.* 281:20133056. doi: 10.1098/rspb.2013.3056
- Yearsley, J. M., and Pothos, E. M. (2016). Zeno's paradox in decision-making. *Proc. R. Soc. B Biol. Sci.* 283:20160291. doi: 10.1098/rspb.2016.0291
- Yearsley, J. M., Pothos, E. M., Barque-Duran, A., Trueblood, J. S., and Hampton, J. A. (2022). Context effects in similarity judgments. *J. Exp. Psychol. Gen.* 151, 711–717. doi: 10.1037/xge0001097
- Young, B., Keller, A., and Rosenthal, D. (2014). Quality-space theory in olfaction. *Front. Psychol.* 5:1. doi: 10.3389/fpsyg.2014.00001
- Zeilinger, A. (2010). *Dance of the photons: From Einstein to quantum teleportation*. New York: Farrar, Straus and Giroux.
- Zeleznikow-Johnston, A., Aizawa, Y., Yamada, M., and Tsuchiya, N. (2023). Are color experiences the same across the visual Field? *J. Cogn. Neurosci.* 35, 509–542. doi: 10.1162/jocn_a_01962

Frontiers in Psychology

Paving the way for a greater understanding of human behavior

The most cited journal in its field, exploring psychological sciences - from clinical research to cognitive science, from imaging studies to human factors, and from animal cognition to social psychology.

Discover the latest Research Topics

[See more →](#)

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact

