

International Society for the Philosophy of the Sciences of the Mind

4TH ANNUAL ONLINE CONFERENCE

4-6 NOVEMBER 2026

Book of Abstracts

Founders and Past Presidents:

Gualtiero Piccinini Founder and Past President (2022-23)

Inês Hipólito Founder and Past President (2023-24)

Program committee (ISPSM Council members)

Inês Hipólito (Macquarie University)

Fabrizio Calzavarini (University of Turin)

Liza Skidelsky (Universidad de Buenos Aires)

Don Ross (University College Cork / University of Cape Town)

Krushil Watene (University of Auckland)

Katsunori Miyahara (Hokkaido University)

Programme chairs:

Valeria Becattini, Scuola Superiore Universitaria IUSS Pavia, Pavia, Italy

Liberty Severs, Ruhr-Univesität Bochum and the University of Lisbon, Germany and Portugal.

Marco Facchin, Universiteit Antwerpen

Bräuchle, Lara - Ruhr-Univesität Bochum, Germany.

Clavel Vásquez, Jimena M. - Tilburg University, the Netherlands

Carlos Barth - Jesuit Faculty of Philosophy and Theology (FAJE)

Laura Oppi - Center for Subjectivity Research, University of Copenhagen

Negro, Niccolò - School of psychological sciences, Tel Aviv University, Israel

April Owens - University of Southern California

Acknowledgements: The ISPSM sincerely thanks all the people that contributed to this conference by reviewing the submitted abstracts: Adrian Wieczorek, Adriana Alcaraz Sánchez, Alejandro G. Di Rienzo, Alfredo Tommasetta, Andrea Raimondi, André Abath, Andy McKilliam, Antonella Tramacere, Ayush Srivastava, Ben White, Benedetta Cogo, Chiara Caporuscio, Colin Klein, Daniel de Luca-Noronha, Daniel Williams, Darius Parvizi Wayne, David Colaço, David Colaço, David Schoute, Davide Serpico, Douglas C. Wadle, Eros De

Carvalho, Fabrizio Calzavarini, Farid Zahnoun, Francesco Di Betta, Gabriel Zanotti, Giovanni Rolla, Giulia Piredda, Guido Cassinadri, Gualtiero Piccinini, Hadeel Naeem, Ian Robertson, Ignacio Cea, Ioannou Savvas, Ishan Singhal, Jacopo Pallagrosi, James Turner, Jelle Bruineberg, Jesse Printz, Johan Heemskerk, Johan Largo, Jolien Francken, Judith H. Martens, Julian Kiverstein, Katalin Balog, Kathleen Murphy-Hollies, Kathryn Nave, Krys Dolega, Lachlan Kent, Laura Coccia, Laura Gow, Leonard Dung, Lorena Sganzerla, Luis Favela, Manolo Martinez, Marc Artiga, Marcello Ienca, Marco Antônio Alves, Marco Degano, Marco Portillo, Marco Viola, Marek Pokropski, Maria Carolina Resende, Marilyn Stendera, Mario Villalobos, Matteo Baggio, Matthew Owen, Mengting Zhang, Michelle Liu, Monima Chadha, Moritz Kriegleder, Nastia Garbayo, Nathalia de Ávila, Niccolò Nanni, Nina Hartter, Nir Fresco, Oryan Zacks, Peter Schulte, Pii Telakivi, Rebekka Brandt, Robert Chis-Ciure, Rumana Ayala, Sabrina Coninx, Sam Wilkinson, Sergio Ponce de León, Shawn Prest, Simone P. Roca, Sofia Livi, Sofia Sgarbi, Takuya Niikawa, Tommaso Peripoli, Valentina Martinis, Valentina Petrolini, Vicente Raja, Wanja Wiese, Wouter van Hooydonk, Yugin Cho, Zoe Drayson, Zuzanna Rucińska.

Contents

The Perception of Dance: Considering Neurophenomenology. 13

The architecture of being: time, environment, and identity in embodied cognition . 14

A Report from an Academy: Why Psychological Concepts are not Species-Neutral 15

Physical Computation - The Computational Fallacy for Brains & Computers 16

Making the Invisible Mind Visible: Generative Models for Understanding Human Cognition 17

Feeling With Things: Toward a Theory of Material Affective Scaffolding 18

Raffaele Giovanni Caravella - Pain, Dispositions, and the Problem of Parts 20

Alisson Brandemarte Moreira - On causally explaining depressive perception with paraconsistent neural networks 23

Carlos Manoel Lopes Rodrigues - Measuring the Unobservable: A Philosophical Analysis of Model-Based Inference in Psychological Assessment 25

Matteo Colombo - Ganjiers, Tolerance to Drugs, and the Puzzle of Drugs Aesthetic Expertise 27

Benjamin Claessens - Desires, Priors, and Purposes 32

Henrique Mendes - Computationalism and enactivism; Tracing two antagonistic paradigms to their common cybernetic origins 34

Benjamin D. Young - How Didn't You Notice It: are Human Social Chemosignals Unconscious or Unattended? 37

Zhexi Zhang - An internalist account of the effect of labels 39

Benedetta Cogo - KEEPING IT RATIONAL: PERSONS AND RATIONALITY IN PHILOSOPHY OF PSYCHIATRY 41

Johannes Brinz - What are Neurocomputational Models? 43

Vaggelis Spathis – Perceptual Recognition and The Limits of Naïve Realism 45

Anton Sukhoverkhov - Expanding Evolutionary Epistemology: Cognitive and Goal-Directed Biological Evolution 48

Felipe Carvalho - Perception & Coordination: reconceptualizing emotion perception through John Dewey 50

Catalin Teoharie - Direct Input Variables: Rethinking the Limits of the Extended Mind Hypothesis within Predictive Processing 51

Maria Luiza Iennaco - Toward a Multiscale Account of Attention: Neurophilosophy and Synthetic Philosophy Compared 59

Christos Kalpakidis - Transparency, Direct Realism, and the Minimal Self: Insights from Existential Phenomenology 61

Jack Shardlow - Pain Imagery: Individual Differences and Disutilitous Memories 63

Nathália de Ávila - Remembering at the Edge of Consciousness: Affectivity and the Reconstitution of Selfhood in Near-Death Experiences 66

Matthew Owen - Can Human Cerebral Organoids Become Conscious? A Hylomorphic Perspective 68

Filipp Matveev - Physicalism Naturalized 69

Don Oxtoby - Inferentialism about Recognition 72

Juan Diego Bogotá - Can Social Interaction Constitute the Minimal Self? 75

Caterina Scianna - When Imagination Becomes Truth: Confabulation and Epistemic Vigilance 77

Emilio Leone - Shared Embodied Time 79

Nesim Aslantatar - Motivated Remembering: Religious Experience between Encounter and Retrieved Memory 83

Rémi Tison 87

Mitchell Herschbach - Are There Objective Measures of "Theory of Mind" Abilities? Implications for the Study of Autism 89

Huang Lei - Where Does the Competence Reside? Interaction-Dependent Functional Emergence in a Two-Model Language System 91

Alexandra Ioana Din - Code-switching as narrative resource in bilingual Romanian-German children 93

Brenda Bayles - The Geometry of Reorganization Across Cognitive and Natural Systems 96

Heng Ying - Revisiting the Humean model of moral psychology 98

Carolyn Wasula - Relational Cosmopsychism: A Formal Account of Mind as Structured Participation 100

Caitlin Mace - Is Vehicle Indeterminacy Radical? 102

Tiago Sestari - Interactive Fear: Rethinking the Paradox of Fictional Emotion Through Games 106

Marco Degano - Two Integration Problems for Mesh: the Vertical and the Horizontal Integration Problems. 108

Gerda Mittag - Want to understand computational explanation? - Ask a teacher! 110

Savvas Ioannou - Defining the Physical: Via Negativa and Non-Mental Descriptions 113

Urte Laukaityte - The Role of History and Philosophy of Science in Medicine: Ontology, Epistemology, and Illness 116

Antonella Tramacere - The limits of embodied approaches to comparative cognition 118

Tomas Marvan and Wojciech Mamak – Brain Currency and the Two-Point Neuron Architectures 121

Johan Largo - Why language models are not cognitive extensions 122

Sreenish S - Cognitive Penetrability of Perception and the Phenomenological Vulnerability of Naive Realism 125

Juan C. Barrio-Aragón & José V. Hernández-Conde - From Minimal to Temporal Consciousness: The origin of conscious experience beyond the present 128

Andrei Kochenkov – Objectification Without Constancy: The Case of Olfaction..131

Tobias Sandoval - Background conditions and emotional kinds 134

Douglas C. Wadle - Vividness in Mental Imagery Research and Beyond.....136

Sacha Behrend - Reference and Representational Formats in Episodic Memory.140

Farideh Sousani - On the Ontology of the Self: Cognitive-Behavioral and Jungian Approaches to the Symptom 143

Michał Piekarski - Precision-Weighted Noise: A Hierarchical Predictive-Processing Account of the Conjunction Fallacy.....145

Marek Pokropski - Idealizing Consciousness: Multiple Models or a Minimal Unifying Model?.....148

Adrian Wieczorek - Turning plants from head to ground. An alternative to the predictive stance in minimal cognition. 150

Oryan Zacks - Imagination, memory and the ancient origins of the default mode network.....152

Gina Adriana ramos Zamudio - The Operative Prereflective Dimension as a Bridging Category: Toward a Dialogue with Cellular Memory.....153

Aleksandr Naumenko - Do Categories Rely on Similarities or Differences? 155

Fernanda Caroliny Cardoso - If computers can think, why are LLMs not thinking? Dennett on the emergence of comprehension, consciousness and intentionality ..158

Andrew Richmond - From neuroscience to explainable AI: Representation, understanding, and explanation 161

Daniel Sherman - Embodied Insights: Interoception in Psychedelic Epistemology 164

Christian Lains - Hypermodal Realism: Phenomenal Information and Excluded Possibilities.....167

Mark-Oliver Casper - Re-Valuation of Linearity in 4E Research.....170

Juan Murillo Vargas - Questions: A Rationalist's Guide.....171

Jannis Friedrich - Brass-for-Brains: Analog Computation and Device Physics ... 176

Ye Yizhou - Does Machine Learning Erode the Open Future? 179

Yizhuo Sun - Beyond Static Representation: A Dynamic Mechanistic Account of Representational Vehicles in Predictive Processing 181

Rongsheng Liang, Yizhuo Sun - Abstraction Without Causal Loss: A Mechanistic Reconstruction of Minimal Models of Cognition 183

MANJARI CHAKRABARTY - THE POSSIBILITY OF CONSCIOUSNESS-BRAIN INTERACTION RECONSIDERED 185

Christopher Stratman - An Argument for Eliminativism About Occurrent Conscious Beliefs 188

Bongrae Seok - Neuroscience and Game Theory of Mindfulness Meditation: Embodied Moral Psychology of Buddhism 189

João César Ramos - Acquaintance, Knowledge and Normativity 196

Paula Muhr - From Local Lesions to Dynamic Distributed Networks Framing Brain Dysfunction in Neuroimaging Research on Functional Neurological Disorder ... 198

Kira Miller - An Eye for an 'I': Investigating the Relationship between Embodiment and Agency, and Exploring the Possibilities for Artificial Intelligence Models 200

Keiran Bateman - No Mental Atoms: Panpsychism's Category Mistake 202

Shahd Omar - Can an Artificial Relationship Scaffold a Self? AI Companions and Quasi-Relational Selfhood 203

Shawn Prest - Selfing without a self: The missing dimension of identification in computational accounts of selfhood 205

Camila von Holdefer Kehl - The Cause That Remains: Dorothy Emmet and the Limits of Mechanism in the Sciences of Mind 206

Inchul Yum - On Pleasure's Direction of Fit 208

Monicka Sre R V 210

Cognition beyond the body: affective scaffolding, embodied inference, and psychopathology.....326

1) Affective artifacts, aesthetics and care: a deepening view.....327

2) Extended Rumination.....328

3) Where is the body in embodied active inference?.....329

4. Therapy chatbots as cognitive-affective scaffolds.....330

Neuroscience and radical embodiment.....331

Facchin Marco (Universiteit Antwerpen) - Brains are not cognitively impartial.....332

Favela, Luis H. (Indiana University Bloomington) & Chemero, Anthony (Vanderbilt University) - Direct Perception and Ecological Information Processing.....334

Gambarotto Andrea (University of Luxemburg) - Rethinking Biological Control: Heterarchy in Brain Networks.....335

Vicente Raja (University of Murcia) - On Radio Stations, Computers, and Radical Embodied Neuroscience.....336

New Directions in Teleosemantics - Content Determination and the Explanatory Turn.....337

1. Representing What Really Matters: An Informational Teleosemantic Solution to the Distality Problem.....340

2. Frogs Recognize Prey: A Hybrid Teleosemantics.....341

3. Ultimate and Proximate Explanations and Teleosemantic Pluralism.....342

4. Escaping Explanatory Circularity through Teleosemantics.....343

Spontaneous Sense-Making Across Inner Speech, Mind Wandering and Dreaming.....345

Daniel Mueller - Inner Speech & Creativity Pangelonium in the Brain— A Sociology of the Default Mode Network.....346

Yizhi Li - How Our Tales Spin Us: Spontaneous Thought and the Formation of the Narrative Self.....348

Mengting Zhang - Mind Wandering is Not Subject to Cognitive Control.....349

Gaia Mizzon - What Dreaming Teaches.....350

Title: The Navigational Mind.....353

Paper 1: Negation as Navigation Failure: Toward a Cognitive Map Account of Propositional Negation.....355

Paper 3: Navigable Correspondence Networks: Neural Manifolds as Content-Bearing Structures.....358

The Phenomenal Space of Wakefulness & Dreaming.....362

THE METATHEORY OF PSYCHOPATHOLOGY: THE PROBLEM OF PSYCHOSIS IN PHENOMENOLOGY AND NEUROSCIENCE.....367



Keynote Talks

The Perception of Dance: Considering Neurophenomenology.

Dr Devon Bailey

Tim Bayne (2004) claims that Fransisco Varela's (1996) attempt to "bridge the explanatory gap" between the "what it's like" character of experience and the physical nature of the brain falls flat. He argues that Varela, in attempting to bridge the gap by "shifting the framework" in which neurophenomenology is discussed, does not do what he claims to do. In this paper, I show why Bayne (2004) is wrong about neurophenomenology not being able to close the explanatory gap, specifically as it relates to the perception of a dance performance on stage. I also aim to defend Varela's (1996) position on "shifting the framework" in which neurophenomenology is discussed by proposing David Chalmers' (1995) "property dualism" as the lens through which we see neurophenomenology as the most satisfying means of understanding our perception of a dance performance on stage.

The architecture of being: time, environment, and identity in embodied cognition

Prof. Gastelum-Vargas

Contemporary ecological psychology increasingly views social interaction as coordination within agent–environment systems rather than as a series of processes taking place inside separate individuals. This chapter explores this view through the central example of midwifery, understood as a paradigmatic case of embodied, affective, and relational cognition. A midwife's expertise consists in perceiving and acting upon the shifting possibilities for action emerging from the laboring woman's body, affective states, interpersonal relations, and material environment. Although midwifery literature has often characterized such expertise as intuition or tacit knowledge, less attention has been paid to how this embodied and relational knowledge is organized across time.

Drawing on ecological psychology and the Dynamic Affordance Trajectory Framework, we propose that midwife–woman interactions can be understood as dynamic affordance systems: temporary, self-organizing couplings between embodied agents, material settings, affective dynamics, and sociocultural practices. These interactions unfold across three interconnected timescales: real-time attunement and coordination during labor; medium-term learning and adaptation across professional practice and the trajectory of pregnancy, birth, and postpartum care; and longer-term developmental and cultural processes through which professional identities, practices, and meanings of childbirth are constituted.

From this perspective, sociality is not an additional layer imposed upon individual perception and action but part of their ecological organization. Humanized childbirth can consequently be understood as the deliberate construction of an affordance landscape that supports trust, agency, coordination, and adaptive action. The coexistence of traditional and professional midwifery in Mexico illustrates how different sociocultural niches organize distinct yet interacting landscapes of affordances. Midwifery thus provides a particularly productive domain for examining how social affordances, interpersonal coordination, embodied expertise, and cultural practices jointly shape trajectories of action and identity across multiple timescales.

A Report from an Academy: Why Psychological Concepts are not Species-Neutral

Prof. Louise Barrett

Work in comparative psychology and animal cognition assumes evolutionary—and hence psychological—continuity across the animal kingdom. The implicit assumption here is that it is possible to identify psychological phenomena in ways that are entirely species-neutral, and which allow us to map similarities and differences within and across different taxa. This assessment draws on psychological theories and capacities that have been, for the most part, derived from the study of our own species, and rely on computationalrepresentational

metaphors. If, however, human psychological capacities are "entangled", in Alva Noë's sense, then such a strategy is doomed to fail. When attempts are made to pare away apparently superficial layers of human-flavoured coating to reveal a deep continuity, we are, in fact, fabricating another, more insidious, anthropocentric chimera. As we cannot disentangle human cognition to reveal its natural core, what we need to do is embrace entanglement in our own case, as well as adopt an embodied framework that can apply to both ourselves and other species, where sensory and motor behavior is conceived as constitutive of cognition, and not as a confound that hampers cross-species comparisons. Using a number of animal examples and psychological capacities, I offer a view that may allow us to generate a better kind of continuity— one that will allow us to better understand the kind of animals that we are and allow other animals to emerge from our shadow.

Physical Computation - The Computational Fallacy for Brains & Computers

Prof. Michael Kirchhoff

In this talk, I address a foundational problem for computationalism—the view that brains and computers are computational systems. I call it the computational fallacy. The structural core of the fallacy is: (1) Computationalists start with a physical object—a brain or a computer; (2) they then turn to an abstract computational function, operation or program that the physical system can at best approximately implement; (3) next computationalists try to convince us that brains and computers are (somehow) uniquely related to abstract computational functions, operations or programs, which brains and computers implement, given certain constraints; however, (4), it is false that there is any kind of special relation between a physical object—like a brain or a computer—and an abstract computational function, operation or program. The central thesis of the talk is: in the absence of idealization, brains and computers cannot be interpreted as implementing any specific abstract computational function, operation or program. Because idealization refers to imaginary posits in abstract models (such as infinite memory in Turing machines), there is nothing intrinsic to brains or computers that justifies computationalism.

Making the Invisible Mind Visible: Generative Models for Understanding Human Cognition

Prof. Yukie Nagai

The human mind cannot be directly observed. We can observe behavior and brain activity, but the internal processes that give rise to perception, action, and subjective experience remain largely hidden. In this talk, I will show how generative models based on predictive processing can provide a powerful tool for investigating these hidden processes. Rather than using AI simply to reproduce human behavior, we use computational neural models to ask what underlying mechanisms are necessary for such behavior to emerge. I will illustrate this constructive approach through two studies: the development of drawing behavior and the sense of agency—the feeling that "I am the one causing this action." Together, these studies demonstrate how building artificial models of cognition can help us formulate and test hypotheses about the mechanisms underlying the human mind.

Feeling With Things: Toward a Theory of Material Affective Scaffolding

Prof. Giovanna Colombetti

I call Material Affective Scaffolding (MAS) the activity of making and/or using material objects and spaces to influence one's own affective states (emotions, moods, drives, and more), and those of other people. After a general introduction, I will discuss some key characteristics of the phenomenon by addressing a number of "frequently asked questions" about it, such as: What material items count as affective scaffolds? Who scaffolds and who is scaffolded? Are affective scaffolds permanent or temporary? Are they necessary for at least certain affective states? Can they be seen as extensions of the mind?



Contributed Talks

Raffaele Giovanni Caravella - Pain, Dispositions, and the Problem of Parts

Università della Svizzera Italiana

This presentation addresses the "problem of parts" in the philosophy of pain, namely which of pain's characteristic components, i.e. physiological, phenomenal, behavioral, or cognitive, is necessary and sufficient for an experience to count as pain. I will focus on the phenomenal and the physiological components, since the literature on the philosophy of pain can be divided between those who consider pain to be a mental state [Tye (2017), Kripke (1980)] characterized roughly as the phenomenal component of pain) and those who consider pain to be a bodily state [Reuter and Sytsma (2020), Sytsma and Machery (2009)].

In the first part of the presentation, each of the four characteristic components will be examined in turn and assessed as to whether it can plausibly be regarded as necessary or sufficient for pain. Drawing on both philosophical arguments and empirical cases, it will be argued that physiological disturbance is neither necessary nor sufficient, as illustrated by phantom limb pain [Nanay (2017)], the thermal grill illusion [Nanay (2017), Craig and Bushnell (1994)], and cases of anesthesia. Similarly, behavioral tendencies and alterations in a broader mental network fail both necessity and sufficiency.

This is not the case for the phenomenal component, which can be divided into an affective and a sensory aspect. Each of these, taken individually, is necessary but not sufficient. In fact, each successfully overcomes objections such as cases of pain asymbolia [Berthier et al. (1988)] and lobotomy (affective aspect) [Corns (2020)] and the heterogeneity problem (sensory aspect). In addition, even their conjunction fails to meet the criterion of sufficiency.

This negative result motivates the constructive part of the presentation. In fact, even though we can say that all these components together are sufficient, we should find a way to account for cases in which even a single component is absent, e.g. the phantom limb case. The central challenge, therefore, is to distinguish cases of pain from cases of non pain, particularly what Coninx calls borderline cases, in which subjects themselves are unsure whether they are

experiencing pain (e.g. itches), without imposing overly restrictive necessary or sufficient conditions.

In order to resolve this issue, Schwitzgebel's liberal dispositional account of belief will be adapted to the case of pain [Schwitzgebel (2002)]. On Schwitzgebel's view, belief is not defined by any single necessary or sufficient condition, but by a cluster of stereotypical dispositional properties, whose instantiation admits of degrees and contextual variation. It will be argued that a suitably modified version of this approach can also be applied to the case of pain. Considering pain as a cluster of stereotypical dispositional properties is problematic, since subjects have these dispositions even when they are not in pain. Pain should instead be understood in terms of the manifestations of such dispositions. Pain episodes are thus characterized by the manifestation of a sufficient number of stereotypical dispositions, rather than by the possession of any single disposition or component.

This account provides a principled way of handling borderline cases such as itches since the account allows pain ascriptions to admit of degrees: the more manifestations of stereotypical pain-related dispositions a subject exhibits, the more appropriate it is to describe the case as pain.

References

- Berthier, M., Starkstein, S., and Leiguarda, R. (1988). Asymbolia for pain: A sensory-limbic disconnection syndrome. Annals of Neurology, 24(1):41–49.
- Coninx, S. (2025). The mysterious pain quality: A meta-cognitive interpretation of the orthodox view. Australasian Journal of Philosophy, pages 1–18.
- Corns, J. (2020). The Complex Reality of Pain. Routledge, New York, NY.
- Craig, A. D. and Bushnell, M. C. (1994). The thermal grill illusion: Unmasking the burn of cold pain. Science, 265(5169):252–255.
- Kripke, S. A. (1980). Naming and Necessity: Lectures Given to the Princeton University Phi losophy Colloquium. Harvard University Press, Cambridge, MA.
- Nanay, B. (2017). Pain and mental imagery. The Monist, 100(4):485–500. Reuter, K. and Sytsma, J. (2020). Unfelt pain. Synthese, 197(4):1777–1801.
- Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. Noûs, 36(2):249– 275.

- Sytsma, J. M. and Machery, E. (2009). How to study folk intuitions about phenomenal consciousness. Philosophical Psychology, 22(1):21–35.
- Tye, M. (2017). Are pains feelings? The Monist, 100(4):478–

Alisson Brandemarte Moreira - On causally explaining depressive perception with paraconsistent neural networks

Federal University of Fronteira Sul (UFFS), Federal University of São Paulo (UNIFESP)

The goal of this presentation is to establish the philosophical outlines of a unified theory (Neurophilosophy) of how pessimistic perception works in Major Depressive Disorder (MDD). Instead of presenting bottom-up and top-down approaches in detail, I focus on the fundamental philosophical interpretation that may unify them (Churchland, P. S., 1986, Introduction). Therefore, I argue that the architecture of Paraconsistent Artificial Neural Networks (PANNs), formulated through 2-value Paraconsistent Annotated Logic by Jair Minoro Abe and others, under advise of Newton da Costa (Da Silva Filho & Abe, 2001), may be used as a simplification of our Neurophilosophy. In this sense, PANNs are more suitable to understand depressive perception than traditional ANNs, since it is a result of a conjunction between positive and negative "beliefs". For example, in suicidal idealization, individuals with Depression may at the same time "believe" that: (1) by committing suicide, they are doing something good for their family and themselves, because they are a burden; and (2) they should not commit suicide, because, due to how much their family loves them, they will bring suffering to their family and themselves. In this case, there is a certain degree of inconsistency – a key concept of PANNs. After that, developing Churchlands' thought, I argue based on the theory-ladenness of perception that our case requires a new concept of causal explanation that allows theoretical paraconsistency and tolerates neuroscientific ambivalence: Perception is laden with paraconsistent theories that causally explain depressive perception. The concept of causal explanation used here follows the admittedly simple notion of causality, proposed by the Churchlands, as a prototype that "depicts a typical temporal *sequence* of event types" (Churchland, P. M., [1989] 1992, p. 213). This is, for them, not an analytic (precise, clear) concept of explanation, but rather a "revealing and unifying characterization" of the brain events involved in causal explanations (Churchland, P. M., [1989] 1992, p. 213). On this basis, I argue, in accordance with PANNs, that such a characterization requires the acknowledgement of paraconsistency between ambivalent neural mechanisms across

different levels of organization (neurotransmitters, synapses, nerve cells, systems, networks and so on), such as the twofold action of serotonin as an excitatory or inhibitory neurotransmitter, depending on the post-synaptic receptors.

References

- DA SILVA FILHO, J. I. & ABE, J. M. (2001). **Fundamentos das Redes Neurais Artificiais Paraconsistentes**. São Paulo: Arte & Ciência.
- CHURCHLAND, P. M. ([1989] 1992). **A neurocomputational Perspective**: The nature of the mind and the structure of science. Paperback Edition. Massachusetts: MIT Press.
- CHURCHLAND, P. S. (1986). **Neurophilosophy**: Toward a unified science of the mind brain. Massachusetts: MIT Press.

Carlos Manoel Lopes Rodrigues - Measuring the Unobservable: A Philosophical Analysis of Model-Based Inference in Psychological Assessment

University Center of Brasília

Psychological assessment is often framed as the measurement of latent mental attributes such as intelligence, personality, or psychopathology. However, unlike physical measurement, where attributes can be directly observed or operationalized with minimal theoretical mediation, psychological constructs are inherently unobservable and dependent on complex inferential frameworks. This raises a central epistemological question: what does it mean to "measure" a psychological construct? This paper argues that psychological assessment does not involve the direct quantification of mental properties, but rather the production of model-based inferences grounded in theoretical assumptions and statistical representations. Rather than addressing the ontological status of mental states, the present analysis focuses on the epistemic conditions under which psychological constructs are inferred. Drawing on contemporary psychometric practices—including factor analysis, structural equation modeling, and item response theory—we show that such techniques do not passively uncover latent structures, but actively instantiate modeling choices that constrain and shape interpretation. Statistical models, in this sense, function as epistemic frameworks rather than neutral discovery tools. Building on this perspective, we critically reconsider the concept of validity. Instead of treating validity as an intrinsic property of instruments, we propose understanding it as an inferential and context-sensitive judgment concerning the adequacy of interpretations derived from test scores. This shift reframes psychological measurement as a practice centered on justification rather than detection. We further examine the methodological consequences of treating constructs as measurable entities, including the reification of latent variables, overinterpretation of factor structures, and the uncritical generalization of instruments across contexts. These issues become particularly salient in contemporary developments such as automated scoring systems and algorithmic decision-making, where inferential processes are increasingly obscured by technical implementation. By conceptualizing psychological assessment as an inferential practice rather than measurement

in the strict sense, this paper contributes to debates in the philosophy of psychology and the sciences of the mind, emphasizing the need for greater epistemological reflexivity in psychometric research.

Matteo Colombo - Ganjiers, Tolerance to Drugs, and the Puzzle of Drugs Aesthetic Expertise

UvT

Paintings, songs, movies, wine, cheese, perfume, and more recently cannabis. These are all objects of aesthetic appreciation with dedicated experts, critics, schools, and awards. Acquiring any kind of aesthetic expertise requires study, training, and experience an object of appreciation (Robson 2025, §3.4 and §4.2). This experience can be refined through perceptual learning, which consists in a relatively stable improvement in the perception of an object (Gibson, 1963, 29; Jenkin 2023). Perceptual learning can be biased in several different ways (Ransom & Goldstone 2024), but it can also supporting sophisticated aesthetic appreciation, whereby one's perceptual capacities get better at fine-grained discrimination and categorization, pattern recognition, and attention management (Kellman & Massey 2013; Stokes 2021).

When it comes to drugs such as cannabis, however, intimate experience does not obviously usher in an improvement in perception supporting a more sophisticated appreciation. In fact, drugs present a fundamental puzzle for aesthetic expertise. With training and experience, expert drugs users can acquire a richer drug-related conceptual repertoire, and sharpen their perceptual capacities to better recognize different samples from their appearance, and pick up subtle changes in the aroma or taste of different varieties. But the aesthetics of drugs includes more than the qualities of the appearance, flavour, and aroma of a substance. It distinctively includes a concern about the *mind-altering* effects of a substance, which are variously called "intoxication," "drunkenness," "inebriation," or "high" (Crane 2004; Fokt 2024; Hathaway & Sharpley 2010; Naukkarinen 2025; Nguyen 2024). What I am going to call *the puzzle of drugs aesthetic expertise* is focused on this concern.

The puzzle can be laid out as the following argument.

P.1 Developing drug aesthetic expertise requires experience.

P.2 Experience with a drug causes tolerance to its effects.

P.3 Tolerance to the effects of drugs undermines the possibility of drug aesthetic expertise.

So, there is no genuine drug aesthetic expertise.

Considering cannabis, I am going to explain ad probe each step in the puzzle of aesthetic expertise. Specifically, I begin by clarifying the species of drugs aesthetic expertise at stake. Next, I explain the notion of tolerance to drugs. Finally, I leverage the evidence about cannabis tolerance to argue that expert users cannot possess a more refined aesthetic appreciation of the mind-altering effects of cannabis than non-expert users. I conclude by sketching three possible solutions to the puzzle of drugs aesthetic expertise.

References

- American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition. Arlington, VA: American Psychiatric Association.
- Babor, T. F., Mendelson, J. H., Greenberg, I., & Kuehnle, J. C. (1975). Marijuana consumption and tolerance to physiological and subjective effects. *Archives of General Psychiatry*, 32(12), 1548-1552.
- Baker, T. B., & Tiffany, S. T. (1985). Morphine tolerance as habituation. *Psychological Review, 92*(1), 78–108.
- Bespalov, A., Müller, R., Relo, A. L., & Hudzik, T. (2016). Drug tolerance: a known unknown in translational neuroscience. *Trends in pharmacological sciences*, 37(5), 364-378.
- Bloomfield, M. A., Morgan, C. J., Egerton, A., Kapur, S., Curran, H. V., & Howes, O. D. (2014). Dopaminergic function in cannabis users and its relationship to cannabis-induced psychotic symptoms. *Biological psychiatry*, *75*(6), 470-478.
- Cochin, J., & Kornetsky, C. (1964). Development and loss of tolerance to morphine in the rat after single and multiple injections. *The Journal of Pharmacology and Experimental Therapeutics*, 145(1), 1-10.
- Colizzi, M., & Bhattacharyya, S. (2018). Cannabis use and the development of tolerance: a systematic review of human evidence. *Neuroscience & Biobehavioral Reviews*, 93, 1-25.

- Crane, T. (2004). Excess. In *The World of Fine Wine*. URL: https://worldoffinewine.com/uncategorized/excess 4963021
- Dovey, T. M., Boyland, E. J., Trayner, P., Miller, J., Rarmoul-Bouhadjar, A., Cole, J., & Halford, J. C. (2016). Alterations in taste perception due to recreational drug use are due to smoking a substance rather than ingesting it. *Appetite*, 107, 1-8.
- DuMez, A. G. (1919). Increased tolerance and withdrawal phenomena in chronic morphinism: A review of the literature. *Journal of the American Medical Association*, *72*(15), 1069-1072.
- Elvig, S. K., McGinn, M. A., Smith, C., Arends, M. A., Koob, G. F., & Vendruscolo, L. F. (2021). Tolerance to alcohol: A critical yet understudied factor in alcohol addiction. *Pharmacology Biochemistry and Behavior*, 204, 173155.
- Fokt, S. (2024). High art, high artists. *The Journal of aesthetics and art criticism*, *82*(1), 61-73.
- Ganjier (October 21, 2020). "Ganjier: The Birth of the Cannabis Sommelier" URL: https://www.ganjier.com/2020/10/21/ganjier-the-birth-of-the-cannabis-sommelier/
- Gibson, E. J. (1963). Perceptual learning. *Annual Review of Psychology*, 14(January), 29–56.
- González, S., Cebeira, M., & Fernández-Ruiz, J. (2005). Cannabinoid tolerance and dependence: A review of studies in laboratory animals. *Pharmacology, Biochemistry and Behavior*, 81(2), 300-318.
- Green, B. O. B., Kavanagh, D., & Young, R. (2003). Being stoned: a review of self-reported cannabis effects. *Drug and alcohol review*, 22(4), 453-460.
- Halikas, J. A., Weller, R. A., Morse, C. L., & Hoffmann, R. G. (1985). A longitudinal study of marijuana effects. International Journal of the Addictions, 20(5), 701–711.
- Hart, C. L., Ilan, A. B., Gevins, A., Gunderson, E. W., Role, K., Colley, J., & Foltin, R. W. (2010). Neurophysiological and cognitive effects of smoked marijuana in frequent users. *Pharmacology Biochemistry and Behavior*, *96*(3), 333-341.
- Hathaway, A. D., & Sharpley, J. (2010). The cannabis experience: An analysis of "flow". In D. Jacquette (Ed.), *Cannabis. Philosophy for Everyone. What Were We Just Talking About?* (pp. 50–61). Wiley-Blackwell.

- Ibáñez-Jiménez, E., Miguez, G., Quezada-Scholz, V. E., Alfaro, F., Bertin, F., Aguilar, C., ... & Laborda, M. A. (2023). Cannabis tolerance: a scope review. *Revista Interamericana de Psicología/Interamerican Journal of Psychology*, *57*(3), 1-20.
- Jenkin, Z. (2023). Perceptual learning. *Philosophy Compass*, 18(6), e12932.
- Kalant, H. (1998). Research on tolerance: what can we learn from history?. *Alcoholism: Clinical and Experimental Research*, *22*(1), 67-76.
- Kellman, P. J. & Massey, C. M. (2013). Perceptual Learning, Cognition, and Expertise. In B. H. Ross (ed.), *The Psychology of Learning and Motivation*, Vol. 58, Academic Press, Elsevier Inc., pp. 117–165.
- Mason, N. L., Theunissen, E. L., Hutten, N., Tse, D., Toennes, S. W., Jansen, J., Stiers, P., & Ramaekers, J. G. (2021). Reduced responsiveness of the reward system is associated with tolerance to cannabis impairment in chronic users. *Addiction Biology*, 26(1), e12870.
- Mikulskaya, E., & Martin, F. H. (2018). Contrast sensitivity and motion discrimination in cannabis users. *Psychopharmacology*, 235(8), 2459-2469.
- Naukkarinen, O. (2025). Aesthetics of intoxication. *Studi di Estetica*, 33, pp. 213-233.
- Negus, S.S., Selley, D.E., Sim-Selley, L.J. (2014). Pharmacodynamic Tolerance. In Stolerman, I., Price, L. (eds) *Encyclopedia of Psychopharmacology*. Springer, Berlin, Heidelberg.
- Nguyen, C. T. (2024). The aesthetics of drugs. In *The Palgrave Handbook of Philosophy and Psychoactive Drug Use* (pp. 631-651). Cham: Springer Nature Switzerland.
- Parsons, L. H., & Hurd, Y. L. (2015). Endocannabinoid signalling in reward and addiction. *Nature Reviews Neuroscience*, *16*(10), 579-594.
- Piscura, M. K., Henderson-Redmond, A. N., Barnes, R. C., Mitra, S., Guindon, J., & Morgan, D. J. (2023). Mechanisms of cannabinoid tolerance. *Biochemical pharmacology*, *214*, 115665.
- Poulos, C. X., & Cappell, H. (1991). Homeostatic theory of drug tolerance: a general model of physiological adaptation. *Psychological review*, *98*(3), 390-408.
- Ramaekers, J. G., Mason, N. L., & Theunissen, E. L. (2020). Blunted highs: pharmacodynamic and behavioral models of cannabis tolerance. *European Neuropsychopharmacology*, *36*, 191-205.

- Ransom, M. (2022). Aesthetic Perception and the Puzzle of Training. *Synthese* 200 (2): 1–25.
- Ransom, M., & Goldstone, R. L. (2024). Bias in perceptual learning. Wiley Interdisciplinary Reviews: Cognitive Science, 15(5), e1683.
- Robson J. (2025). *Aesthetic Knowledge*. Cambridge University Press.
- Roque, Symone (February 17, 2017). "Want to Become a Weed Expert? Try Cannabis Sommelier Classes" URL: https://303magazine.com/2017/02/how-to-become-weed-expert/
- Siegel, S. (2005). Drug tolerance, drug addiction, and drug anticipation. *Current Directions in psychological science*, 14(6), 296-300.
- Siegel, S., Baptista, M. A., Kim, J. A., McDonald, R. V., & Weise-Kelly, L. (2000). Pavlovian psychopharmacology: the associative basis of tolerance. *Experimental and clinical psychopharmacology*, *8*(3), 276-293.
- Sterling, Denmen (December 22, 2023). The Secret Lives of Cannabis Sommeliers: Beyond the Bud. URL: https://potculturemagazine.com/2023/12/22/the-secret-lives-of-cannabis-sommeliers-beyond-the-bud/
- Stokes, D. (2021). On Perceptual Expertise. *Mind & Language* 36 (2): 241–63.
- Volkow, N. D., Wang, G. J., Telang, F., Fowler, J. S., Alexoff, D., Logan, J., ... & Tomasi, D. (2014). Decreased dopamine brain reactivity in marijuana abusers is associated with negative emotionality and addiction severity. *Proceedings of the National Academy of Sciences*, *111*(30), E3149-E3156.

Benjamin Claessens - Desires, Priors, and Purposes

CUNY

Desire-attributions appear to play a lead role in the prediction of human behavior. When you attribute to me, sincerely, the desire for a big stack of buttery pancakes, you might then predict that I will soon fetch the syrup. Such predictive feats would be miraculous, were there not at least some mental processes within me that causally conspire to make your behavioral predictions come true, and to which your attributions are attuned. Desire attributions, in other words, seemingly track desire-like mental states. But if there are such states, what are they, and how do we reliably detect them?

I will argue that folk-conceptions of desire track a type of belief: teleological belief about one's own behavior. These are beliefs that one's own behavior occurs, or will occur, for the sake of a purpose. For instance, suppose I believe that I will fetch the syrup for the sake of soaking the pancakes. In that case, this belief, I argue, would, all else being equal, cause me to fetch the syrup and soak the pancakes. So, your behavioral predictions would be accurate, to the extent that your desire-attributions successfully track my teleological beliefs.

I make three prominent assumptions. First, I assume psychofunctionalism; I therein assume that folk-psychological conceptions of mental states are open to principled revision (Fodor, 1987; Pylyshyn, 1987). Second, I assume a Spinozan theory of belief fixation; beliefs are encoded passively, and without conscious endorsement (Gilbert, 1991; Mandelbaum, 2013). Third, I assume a fragmented theory of belief storage; beliefs are stored in distinct, informationally isolated data structures (Bendaña & Mandelbaum, 2021).

The overall argument is that teleological beliefs, so understood, satisfy twelve desiderata for a sound theory of desire. To that end, I first argue that teleological beliefs cause goal directed behavior; this explains the 'constant sum principle', as evidenced in Goal Systems Theory (Zhang et al., 2007; Kruglanksi et al., 2015). I then argue that the proposed theory satisfies eleven remaining desiderata: desires 1) explain goal-directed behavior, 2) are partially sensitive to evidence, and 3) vary in strength. We can have desires that are 4) unmanifest in action, and 5) jointly unsatisfiable. The theory explains how 6) we make predictively reliable

attributions of desire, and why 7) we make those attributions, sometimes in error, of infants and nonhuman animals. It also explains relations between desire and 8) intention, 9) pleasure, and 10) evaluation, and, finally, it accounts for 11) the distinctive phenomenology of desire.

That said, the foregoing theory has surprising implications. As I will indicate, teleological belief is a type of predictive belief. So, given a conception of desire as teleological belief, desires entail predictions; if you desire P, then you must also predict that P. Although it is unintuitive that desires entail predictions, I will argue that, given the Spinozan theory of belief fixation, we have principled reason to disregard that conflicting intuition; it is the result of a 'refrigerator light illusion' (Block, 2007). In sum, we should conceive of desire as predictive teleological belief about one's own behavior.

References

- Bendaña, J. & Mandelbaum, E. (2021). 'The Fragmentation of Belief'. In Borgoni, C., Kindermann, D., & Onofri, A. (eds.), *The Fragmented Mind*. Oxford University Press.
- Block, N. (2007). Consciousness, Accessibility, and the Mesh between Psychology and Neuroscience. *The Behavioral and Brain Sciences, 30*(5-6). 481–499.
- Fodor, J. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press.
- Gilbert, D. (1991). How Mental Systems Believe' *American Psychologist, 46*(2). 107–119.
- Kruglanski, A., Chernikova, M., Babush, M., & Dugas, M. (2015). The Architecture of Goal Systems: Multifinality, Equifinality, and Counterfinality in Means-End Relations. *Advances in Motivation Science, 2*. 69–98.
- Mandelbaum, E. (2013). Thinking is Believing. *Inquiry, 57*(1). 55–96. Pylyshyn, Z. (1987). What's in a Mind?. *Cognitive Science, 70*(1). 97–122.
- Zhang, Y., Fishbach, A., & Kruglanski, A. (2007). The Dilution Model: How Additional Goals Undermine the Perceived Instrumentality of a Shared Path. *Journal of Personality and Social Psychology, 92*(3). 389–401.

Henrique Mendes - Computationalism and enactivism; Tracing two antagonistic paradigms to their common cybernetic origins

University of Neuchâtel

Computationalism and enactivism are often viewed as opposing frameworks in cognitive science: one committed to formal computation, internal representation, and abstract symbol manipulation, the other to embodiment, situatedness, and autonomy. This presentation nuances that narrative by tracing both approaches to their shared origin in the cybernetic tradition. Rather than treating computationalism and enactivism as unrelated paradigms, I reconstruct how they emerged from a common set of problems concerning circular causality, feedback, self-regulation, and the mechanization of purposive behavior. As Jean-Pierre Dupuy observes, early cybernetics already contained the two ideas that would later define these traditions: that "thinking is a form of computation," and that physical laws can explain how nature appears to contain "meaning, finality, directionality, and intentionality" (Dupuy 2009, 3–4).

Cybernetics emerged in the 1940s around the attempt to understand "control and communication in the animal and the machine" (Wiener 1948). Its first formulations appeared in 1943, with Wiener, Rosenblueth, and Bigelow's account of feedback and teleological behavior, and McCulloch and Pitts's logical model of neural networks. The latter is often remembered as the first computationalist account of the mind, since it treated neural activity as the instantiation of propositional logic, with neurons functioning as binary units. By showing that such networks were equivalent to Turing machines (McCulloch and Pitts 1943, 17), McCulloch and Pitts provided a model that would later influence von Neumann's computer architecture, Rosenblatt's perceptron, AI, and philosophical formulations of computationalism (Putnam 1967; Fodor 1975).

Yet this computationalist lineage does not exhaust the significance of their project. McCulloch's broader aim was "experimental epistemology": the artificial reproduction of cognitive mechanisms as a way of understanding their biological embodiment.

Indeed, McCulloch and Pitts became dissatisfied with the rigidity and biological implausibility of their original model, once they recognized that actual neurons process several Boolean functions in parallel and that genetic information alone cannot specify the brain's intricate circuitry. This led McCulloch toward a more developmental and adaptive view, in which learning, chance, and environmental interaction shape neural organization (McCulloch 1951, 85).

This second tendency within cybernetics was radicalized by Ashby's work on homeostasis and self-regulation (Ashby 1952), and later by Heinz von Foerster's distinction between trivial and non-trivial machines. For von Foerster, biological systems are not merely input-output devices, but history-dependent, self-organizing systems whose behavior cannot be fully predicted from an external standpoint (von Foerster 2002, 309–311). Through his work at the Biological Computer Laboratory, these ideas influenced Maturana and Varela's autopoietic theory of life (Varela and Maturana 1980), which later became one of the foundations of enactivism (Varela, Thompson, and Rosch 1991). Autopoietic systems maintain their identity through circular self-production, making cognition inseparable from embodiment, operational closure, and the organism's relation to its environment.

Thus, computationalism and enactivism should not be understood as opposed paradigms, but as divergent developments within cybernetics: one lineage moving from McCulloch and Pitts toward AI, GOFAI, and connectionism; the other passing through feedback, self-organization, autopoiesis, and enactivism.

References

- Ashby, W. R. (1952). *Design for a brain: The origin of adaptive behaviour*. London: Chapman & Hall.
- Ceruti, M., & Damiano, L. (2018). "Plural embodiment(s) of mind. Genealogy and guidelines for a radically embodied approach to mind and consciousness." *Frontiers in Psychology* 9, 2204.
- Fodor, J. (1975). *The language of thought*. New York: Thomas Y. Crowell Company.
- Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living. (Boston Studies in the Philosophy of Science, Vol. 42)*. Dordrecht: D. Reidel Publishing Company.

- McCulloch, W.; & Pitts, W. (1943). “A logical calculus of the ideas immanent in nervous activity.” *Bulletin of Mathematical Biophysics* 5, 115–133.
- McCulloch, W.; & Pitts, W. (1947). “How we know universals: The perception of auditory and visual forms.” *Bulletin of Mathematical Biophysics* 9, 127–147
- Putnam, H. (1967). “Psychological predicates.” In W. H. Capitan & D. D. Merrill (eds.), *Art, Mind, and Religion* (pp. 37–48). Pittsburgh: University of Pittsburgh Press.
- Varela, F.; Rosch, E.; and Thompson, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT.
- von Foerster, H. (2003). *Understanding understanding: Essays on cybernetics and cognition*. New York, NY: Springer.
- Wiener, N. (1948). *Cybernetics: Or control and communication in the animal and the machine*. Cambridge, MA: The MIT Press.

Benjamin D. Young - How Didn't You Notice It: are Human Social Chemosignals Unconscious or Unattended?

University of Nevada Reno

Research on human social chemosignals (HSC) continues to uncover how our embodied odorant bouquet effects swaths of daily life often without our noticing. The range of HSC phenomena include kinship recognition, social acquaintance selection, the perception of health related odors, psychological states (emotions & personality traits), and social interactions (de Groot, 2026). And in addition to the sheer range of interpersonal phenomena that we don't notice being modulate by smell, often these effects are claimed to occur unconsciously (Loos et al. 2025). However, we worry that given findings by Forrester and Spence (2018) on attentional load yielding failures to report perceiving a consciously accessible smell (inattentional anosmia) that the failure to notice HSC might be attributed to a lack of attention. Subjects might be consciously perceiving these bodily odors, yet fail to further attend to them or the resultant effect on behavior. Thus, the lack of subjective reportability might be attributable to a lack of attention. To examine our alternative thesis, we conducted a literature review examining experimental studies on HSC from which we identified a subset 62 target studies. We then further split the findings into claims regarding the perceived stimulus and/or the behavioral effect. Based on the survey of the literature, we argue that our thesis is a viable alternative explanation that has not been ruled out. Therefore, we conclude that skepticism is merited in some instances of the claim the human social chemosignaling occurs unconsciously. We think that our research provides motivation to further study the relation between olfactory conscious/unconscious processing and attention in general, and in particular, their interplay for HSC under laboratory conditions and within naturalistic settings where the multi-modal perceptual array can affect the cognitive load of navigating social interactions.

References

- de Groot, J. H. (2026). There's chemistry between us: Unraveling the mystery of human chemical communication. In *The Routledge handbook of communication and social cognition* (pp. 182-196). Routledge.
- Forster, S. & Spence, C. (2018). “What Smell?” Temporarily Loading Visual Attention Induces a Prolonged Loss of Olfactory Awareness. *Psychological Science*, 29(10) 1642-1652.
- Loos HM, Schaal B, Pause BM, et al. (2025). Past, Present, and Future of Human Chemical Communication Research. *Perspect Psychol Sci*.,20(1):20-44

Zhexi Zhang - An internalist account of the effect of labels

UC Davis, Department of Philosophy

A wide range of empirical evidence demonstrates that labels can facilitate certain otherwise unattainable cognitive capacities. For example, researchers on the effect of labels demonstrate that hearing or seeing certain verbal labels such as CHAIR and DOG facilitated better detection, recognition and discrimination of object categories in humans, compared to when no labels were present (e.g., Lupyan et al. 2020). To explain the empirical data, Andy Clark (2006) offloads the apparent effects of language onto public labels in the environment and influentially offers an externalist account of the effect of labels: a subsymbolic system involves no linguistically structured internal representations but only "shallow" internal representations that fail to match public labels in structure. Clark (2006) claims that the actual structures of external and public labels play the key and irreducible role.

In this paper, I argue that an internalist account for explaining the effect labels is preferable from both the structural and semantic perspectives. I identify three assumptions on which the externalist account rests: (i) there are public labels in the environment, (ii) public labels have syntactic structures of public language expressions, and (iii) the relevant contents are properties only of environmental symbols. However, all three assumptions are questionable. Against (i), labels can have offline and developmental effects, where labels are absent in the environment, yet they reshape representational systems in the long term. Certain internal representations must be postulated to explain the offline effects of labels. Regarding (ii), pace Clark, I contend that public labels are atomic rather than having linguistic structures. Within representational systems there are atomic internal representations of the relevant categories. Public labels do not have "deeper" structures than internal representations, as the format of the public labels and that of the internal representations do in fact match. Finally, I against (iii) by drawing on a noteworthy case, in which the emergent cognition and what the public labels are used for are different. Several recent studies suggest that numerical cognition emerged from the biologically inspired deep convolutional neural networks (DCNNs), which were trained with object category labels without the availability of numerical labels (e.g., Chapalain et al. 2025). If we need to posit content to explain the emergent cognition, but which is not

present in public labels, then it must be represented internally. I conclude that labels do not merely complement the subsymbolic system but can radically reshape it with non-shallow internal representations and novel kinds of contents.

References

- Chapalain, T., Thirion, B. & Eger, E. (2025). Encoding of numerosity with robustness to object and scene identity in biologically inspired object recognition networks. *Neural Computation*, *37*(11), 1975–2010.
- Clark, A. (2006). Material symbols. *Philosophical Psychology*, *19*(3), 291–307.
- Lupyan, G., Rahman, R. A., Boroditsky, L., & Clark, A. (2020). Effects of language on visual perception. *Trends in Cognitive Sciences*, *24*(11).

Benedetta Cogo - KEEPING IT RATIONAL: PERSONS AND RATIONALITY IN PHILOSOPHY OF PSYCHIATRY

University of Wollongong

Rationality is often taken to play a constitutive role in our understanding of mental disorders. However, different and influential philosophical positions largely diverge on what this role should be. This talk examines a prominent tendency in philosophy of psychiatry, that adopts the "Rationality in Intentionality" thesis (RIT) (Graham 2020), in which rationality is a necessary feature of all our intentional attitudes. I show that RIT relies on a Davidsonian view of the mental, according to which recognising someone as a person requires interpreting their attitudes as responsive to rationality (Davidson 1982, 2002). In this talk I question the tenability of the Davidsonian idea that all intentional attitudes occur within the boundaries of the rational, and I argue that, by adopting a broader view of the mind, we can recognise individuals as persons even when their intentional attitudes fail to manifest rationality.

I develop my case in four steps.

First, I present George Graham's (2020) position as an exemplary account of mental disorders that explicitly endorses RIT. I emphasise that RIT draws on a Davidsonian view of rationality, according to which for someone to be seen as a person, their intentional attitudes must be responsive to norms of rationality.

Second, I observe that the Davidsonian position seems to be, prima facie, too strong in its correlation between personhood and rationality. I present two examples from the literature, the case of and individual affected by delusions (Campbell 2009) and the case of an individual expressing a word salad (Fulford et al. 2006), in which intentional attitudes are not responsive to norms of rationality in different ways. I emphasise that, following the Davidsonian view, those individuals would not qualify as persons.

Third, I consider a possible countermove according to which these instances of irrationality happen "in the house of reason" – i.e., on a background of rationality – and hence

do not require us to withhold our view of the individual as a rational agent (Reimer 2011). Fourth, I present an example where an individual is no longer capable of manifesting even the most minimal degree of rationality (Stich 1996). I emphasise that, even though this subject has lost her capacity to formulate propositional attitudes, she is still capable of intentional attitudes. I highlight that we can avoid the undesirable outcome of not considering her a person by recognising that not all intentional attitudes are propositional attitudes, where the former are not always rational. I conclude my presentation highlighting some benefits of applying this distinction and hence adopting a broader view of the mental.

References

- Davidson, D. (1982). Rational animals. dialectica, 36(4), pp.317–327.
- Davidson, D. (2002). Mental events. In: Essays on Actions and Events. Oxford: OUP, pp. 207–225. Graham, G. (2020). The Disordered Mind. An Introduction to Philosophy of Mind and Mental Illness. Third Edition. Routledge.
- Reimer, M. (2011). A Davidsonian perspective on psychiatric delusions. Philosophical Psychology, 24, 659–677.
- Stich, S.P. (1996). Deconstructing the mind. New York: OUP.

Johannes Brinz - What are Neurocomputational Models?

Osnabrück University

The idea that the brain performs computations is fundamental to computational neuroscience, with different neurocomputational models having been proposed throughout history. Despite their ubiquitous use in neuroscience, it remains unclear what exactly neurocomputational models are and which types of systems they represent. Shagrir (2022) suggests that neurocomputational models are formal models of abstract computation. According to this view, claiming that the brain should be described by a certain model is equivalent to claiming that the brain *implements* (in the sense of Chalmers, 1994) that model. However, this blurs the distinction between digital and neural computation, since digital computers can implement neurocomputational models simply by simulating them. Kaplan (2011) proposes that neurocomputational models describe underlying neural mechanisms. This helps distinguish digital from neural computation, because, according to Kaplan, digital computers are not correctly described by neurocomputational models, even when they are appropriately programmed. However, on this view, it remains unclear why these models describe neural *computation* rather than biological brain processes. If neurobiological processes explain cognitive phenomena, then the claim that the brain computes serves no additional explanatory purpose.

In my paper, I develop an account of neurocomputational models that, on the one hand, accounts for the specific neural way of computing and, on the other hand, still allows us to distinguish claims about neural computation from claims about neurobiological processes. *I submit that neurocomputational models form a subclass of the mathematical models used in the natural sciences and that they are distinctive in that they represent how a target system processes Shannon information.* Unlike concepts of information that refer to the semantic properties of a message, Shannon information is an *observer-independent property* of the target system. Similarly to how a system has a certain weight, voltage or magnetisation, it also carries a certain amount of information. However, Shannon information is also *medium-independent* because the specific physical magnitudes are irrelevant to the encoded message. For example,

voltage levels, photon polarization states and magnetic orientations can all encode the same information in the same way.

For example, consider a simple perceptron for pattern recognition. In my view, such a model represents the structure of a concrete artificial neural network: the nodes in the first layer are connected to all the nodes in the next layer, but not to any other nodes within the same layer. However, it does not incorporate specific physical quantities. The nodes and connections describe a specific structure, but are defined independently of the implementing medium. They can be instantiated in different physical quantities, such as spiking frequencies, voltages, and magnetisations. To sum up: Contra Shagrir (2022), I argue that the relation between neurocomputational models and brain processes is one of scientific representation rather than implementation. Contra Kaplan (2011), I claim that that neurocomputational models represent these processes in a medium-independent manner. This distinguishes them from neurobiological models, which incorporate membrane potentials, ion concentrations and synapse conductances into their descriptions.

References

- Chalmers, D. J. (1994). On Implementing a Computation. *Minds and Machines*, *4*(4), 391–402. Kaplan, D. M. (2011). Explanation and Description in Computational Neuroscience. *Synthese*, *183*(3), 339–373.
- Shagrir, O. (2022). *The Nature of Physical Computation*. Oxford University Press. 2

Vaggelis Spathis – Perceptual Recognition and The Limits of Naïve Realism

University of Patras, Department of Philosophy, PhD student

Perceptual object recognition requires the ability to discriminate and identify objects as unified individuals. This process depends on the ventral ("what") pathway, extending from V1 through V2 and V4 to temporal and frontal regions. Recent findings in vision science and neuropsychology (Barense et al. 2005, 2010, 2012; Baxter & Murray 2001; Buckley and Gaffan 1997; Eacott et al. 1994; Ferko et al. 2022; Inhoff et al. 2019; Lee et al. 2005a,b; O'Neil et al. 2009), highlight a critical role for the perirhinal cortex (PRC) in discriminating complex objects under conditions of high feature ambiguity, where specific object identity is required.

A crucial line of empirical support comes from studies of damage to the perirhinal cortex (PRC), which reveal specific deficits in feature conjunction and configural processing (Eacott et al.,1994; Bussey et al, 2002; McTighe et al, 2010; Barence et al, 2012; Yeung et al, 2013). Patients with PRC lesions show marked impairments in discriminating complex objects whose identification depends on integrating multiple overlapping features, even under simultaneous presentation and minimal memory demands. As a result, subjects cannot form more refined, high-level object representations required for discrimination under ambiguity. This suggests that there are neural mechanisms whose function is to construct integrated and structured object representations at the level of perception itself, thereby contributing to the organization and structure of perceptual content. Therefore, the discrimination of complex objects presupposes a form of perceptual representational structure.

Taking into account the foregoing findings, this paper motivates and develops an empirically grounded challenge to naïve realism about perception. According to naïve realism, perception consists in a direct non-representational (perceptual) relation to mind-independent objects. I argue that this view cannot adequately account for high-level perceptual object discrimination of unified individuals which, at some point, is crucial for perceptual object recognition.

In line with that, I formulate a dilemma for naïve realism: either perceptual discrimination, and to this extent object recognition, of objects as unified individuals involves structured perceptual representations depending on and, consequently, requiring internal feature-conjunction processes, or it does not, in which case naïve realism cannot explain the empirical findings concerning perceptual impairment following PRC damage. Either way, the view faces a serious explanatory limitation.

The upshot is that perception cannot be understood as a mere direct relation to objects but must involve structured representational organization and content that contributes to object individuation, discrimination and recognition.

References

- Barense MD,Bussey TJ, Lee ACH, Rogers TT, Davies RR, et al. 2005. Functional speciali zation in the human medial temporal lobe. J. Neurosci. 25(44):10239–46
- Barense MD, Henson RNA, Lee ACH, Graham KS. 2010. Medial temporal lobe activity dur ing complex discrimination of faces, objects, and scenes: effects of viewpoint. Hippocampus 20(3):389–401
- Barense MD, Groen IIA, Lee ACH, Yeung L-K, Brady SM, et al. 2012. Intact memory for ir relevant information impairs perception in amnesia. Neuron 75(1):157–67
- Baxter MG, Murray EA. 2001. Opposite relationship of hippocampal and rhinal cortex dam age to delayed nonmatching-to-sample deficits in monkeys. Hippocampus 11(1):61–71
- Buckley MJ, Gaffan D. 1997. Impairment of visual object-discrimination learning after perirhi nal cortex ablation. Behav. Neurosci. 111(3):467–75
- Bussey TJ, Saksida LM. 2002. The organization of visual object representations: a connec tionist model of effects of lesions in perirhinal cortex: model of perirhinal cortex. Eur. J. Neu rosci. 15(2):355–64
- Eacott MJ, Gaffan D, Murray EA. 1994. Preserved recognition memory for small sets, and impaired stimulus identification for large sets, following rhinal cortex ablations in monkeys. Eur. J. Neurosci. 6(9):1466–78
- Ferko KM, Blumenthal A, Martin CB, Proklova D, Minos AN, et al. 2022. Activity in perirhinal and entorhinal cortex predicts perceived visual similarities among category exemplars with highest precision. eLife 11:e66884

- Inhoff MC, Heusser AC, Tambini A, Martin CB, O'Neil EB, et al. 2019. Understanding perirhi nal contributions to perception and memory: evidence through the lens of selective perirhinal damage.Neuropsychologia 124:9–18
- Lee ACH, Buckley MJ, Pegman SJ, Spiers H, Scahill VL, et al. 2005a. Specialization in the medial temporal lobe for processing of objects and scenes. Hippocampus 15(6):782–97
- Lee ACH, Bussey TJ, Murray EA, Saksida LM, Epstein RA, et al. 2005b. Perceptual deficits in amnesia: challenging the medial temporal lobe "mnemonic" view. Neuropsychologia 43(1):1–11
- McTighe SM, Cowell RA,Winters BD, Bussey TJ, Saksida LM. 2010. Paradoxical false memory for objects after brain damage. Science 330(6009):1408–10
- O'Neil EB, Cate AD, Kohler S. 2009. Perirhinal cortex contributes to accuracy in recognition memory and perceptual discriminations. J. Neurosci. 29(26):8329–34
- Yeung L-K, Ryan JD, Cowell RA, Barense MD. 2013. Recognition memory impairments caused by false recognition of novel objects. J. Exp. Psychol. Gen. 142(4):1384–97

Anton Sukhoverkhov - Expanding Evolutionary Epistemology: Cognitive and Goal-Directed Biological Evolution

Kuban State Agrarian University named after I.T. Trubilina

This research proposes a shift from the traditional, 'genocentric' neo-Darwinian paradigm (evolution is driven by *natural selection* and *random gene mutations*) to a framework anchored in ideas of *goal-directed evolution* and *cognitive biology* (Sukhoverkhov & Kobyakova 2024; Lyon 2025). Recent studies highlight novel evolutionary drivers and pathways that bypass traditionally stated limits. First, growing evidence challenges the

strict Weismann barrier, which conventionally stipulated that somatic changes are isolated from genetic transmission. Modern epigenetics and non-genetic inheritance research prove that organisms accumulate and transmit novel experience and traits through multiple channels, including epigenetic, behavioral, environmental and cultural-symbolic pathways (Sukhoverkhov & Gontier 2021; Sukhoverkhov 2025). Second, studies of *minimal cognition* reveal that complex problem-solving, learning and communication occur in the simplest life forms (including bacteria, slime molds, plants and fungal mycelia) lacking centralized nervous systems (Messer et al., 2026). Features of 'intelligent behaviour' are also witnessed at the intracellular molecular scale (e.g., in DNA repair proteins, motor proteins or the crowd intelligence of the immune system). In particular, advancements in developmental *bioelectricity* reveal that endogenous *electrical networks* served as the primary medium for morphological memory and somatic decision-making long before the evolution of centralized brains. Later in evolution, these networks enabled evolutionary scaling from physiology to brain and mind, wherein fast neurons exapted the slow bioelectric circuits originally used by non-excitable tissues (Levin 2023; Levin & Resnik 2026). Furthermore, we now see how 'extended mind' and 'embodied cognition' are implemented within the massive *electrical circuits* of AI information processing. All these studies shift the definition/nature of intelligence *away from purely neural and mental representations to 'embodied' cognition*. Consequently, this research proposes extending concept of 'embodied cognition' to the entire

physiological structure of an organism, which processes environmental data, references past species memory, and initiates adaptive structural changes (Sukhoverkhov & Kobyakova 2024; Sukhoverkhov 2025). In this regard, the work proposes expanding the scope of evolutionary epistemology to include: 1) the active role of minimal/basal cognition in biological evolution, and 2) the development of a neutral, scientific vocabulary that captures the 'purposeful intelligence' of nature without reverting to anthropomorphic or idealistic fallacies.

References

- Levin, M. (2023). Bioelectric networks: the cognitive glue enabling evolutionary scaling from physiology to mind. Animal Cognition, 26(6), 1865-1891.
- Levin, M., & Resnik, D. B. (2026). Mind everywhere: a framework for conceptualizing goal-directedness in biology and other domains – part one. Biological theory, 1-21.
- Lyon, P. (2025). Fundamental principles of cognitive biology 2.0. Biological Theory, 1-17. Sukhoverkhov, A. V., & Kobyakova, I. I. (2024). "Natural Creationism": Cognitive Foundations of Directed Evolution ary Processes. Herald of the Russian Academy of Sciences, 94(6), 310-314.
- Messer, A., Oña, L., & Kost, C. (2026). Cognition without brains? Learning and memory in microorgan isms. Trends in Microbiology.
- Sukhoverkhov, A. V. (2025). Inheritance Extended and Distributed: Evolutionary Foundations of Memory Systems. Biosemiotics, 1-18.
- Sukhoverkhov, A. V. (2025). Inheritance Extended and Distributed: Evolutionary Foundations of Memory Systems. Biosemiotics, 1-18.
- Sukhoverkhov, A. V., & Gontier, N. (2021). Non-genetic inheritance: Evolution above the organismal level. Biosystems, 200, 104325.

Felipe Carvalho - Perception & Coordination: reconceptualizing emotion perception through John Dewey

Universidade Federal de Lavras

In our everyday interactions we easily and effortlessly perceive emotions in others' facial expressions and bodily behavior. What explains this ability? Philosophers and psychologists have long argued about the fundamentals of emotion perception and the debate is far from settled. While some insist on the sufficiency of the morphological information contained in facial expressions, others construe the objects of emotion perception as more complex, comprising multimodal information such as touch, tone of voice, body postures, and so on. Others, in turn, have gone so far as to deny that there is a pre-given object in emotion perception in the first place, so that processes of emotion perception must be conceptualized as emergent and co-constructed phenomena in contexts of social interaction. While this constructionist move usefully captures the dynamic and interactive character of emotion perception, it ends up over-intellectualizing the phenomenon by making conceptual synchrony, the mutual alignment of emotion concepts between interactants, its paradigm case. The purpose of this paper is to argue that John Dewey's pragmatist theory of emotions can help us make significant theoretical advancements in the understanding of emotion perception, according to which emotions are coordinated modes of behavior that we enact with others in the contexts of social interactions. In this picture, perceiving emotions is a matter of participating in a joint activity of socioaffective coordination. Drawing on action-based theories of perception, this approach takes the dyad, rather than the isolated perceiver, as the minimal unit of analysis, and substitutes coordination for accuracy or synchrony as the relevant success condition. This theory allows us to apprehend important insights about emotion perception that appear in these contemporary debates, while at the same time avoiding their pitfalls. It also points toward a different neuroscientific research program, one directed not at circuits underlying the recognition of ordinary objects, but at those subservingintersubjective coordination.

Catalin Teoharie - Direct Input Variables: Rethinking the Limits of the Extended Mind Hypothesis within Predictive Processing

Department of Theoretical Philosophy, Faculty of Philosophy, University of Bucharest

Predictive Processing theories and the Free Energy Principle (FEP) face a dilemma when addressing the problem raised by the Extended Mind Hypothesis (EMH). On the one hand, few can deny that some objects, such as phones, computers, or other people, play a disproportionate role in the reduction of free energy; hence it is natural to accept that the mind is extended. On the other hand, this seems to go against the carefully constructed framework in which Markov blankets create an evidentiary boundary between the brain and the external environment, thereby rejecting the claim that the mind can extend to external objects. One approach would be to allow for permeable Markov blankets, whose boundaries can move, thus permitting external objects to become part of the mind.

I find these solutions unsatisfactory because they weaken the FEP framework, and I propose a different one instead: human agents have the option of using Direct Input Variables (DIVs), that is, of directly using the outputs of other models without performing the statistical inference themselves. The reduction of free energy can be achieved either through statistical inference performed by the brain or by using the results of other information-processing systems, such as phones, computers, or other humans. The use of DIVs removes the need to accept the EMH, preserves the integrity of the FEP framework, and raises new questions about how DIVs can be integrated within it.

I then discuss the conditions under which we can use DIVs and find interesting similarities with the conditions required for the EMH to hold: first, the brain needs to be able to perform the statistical inference itself; second, the DIV must come in a form that can be fed into the internal model; and third, the source of the DIV must be assigned enough precision for it to be usable in other internal models. In this way, we can preserve the evidentiary boundary of the brain's Markov blanket while still explaining why interacting with certain external objects

leads to a reduction of free energy: we can use their outputs as direct input variables in our statistical models, yielding better predictions. By accepting this position, we can explain everything the EMH explains, without having to accept that the mind is extended, and while preserving the integrity of the FEP.

References

- Adams, F., & Aizawa, K. (2010). *The Bounds of Cognition*. John Wiley & Sons.
- Allen, M., & Friston, K. J. (2018). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, *195*(6), 2459–2482. https://doi.org/10.1007/s11229-016-1288-5
- Apps, M. A. J., & Tsakiris, M. (2014). The free-energy self: A predictive coding account of self-recognition. *Neuroscience & Biobehavioral Reviews*, *41*, 85–97. https://doi.org/https://doi.org/10.1016/j.neubiorev.2013.01.029
- Beni, M. D. (2021). Inflating the social aspects of cognitive structural realism. *European Journal for Philosophy of Science*, *11*(3), 78. https://doi.org/10.1007/s13194-021-00401-5
- Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, *76*, 198–211. https://doi.org/https://doi.org/10.1016/j.jmp.2015.11.003
- Bruineberg, J., Dołęga, K., Dewhurst, J., & Baltieri, M. (2022). The Emperor's New Markov Blankets. *Behavioral and Brain Sciences*, *45*, e183. https://doi.org/DOI: 10.1017/S0140525X21002351
- Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. In *Journal of Mathematical Psychology* (Vol. 81, pp. 55–79). Academic Press Inc. https://doi.org/10.1016/j.jmp.2017.09.004
- Clark, A. (2008). *Supersizing the Mind*. Oxford University PressNew York. https://doi.org/10.1093/acprof:oso/9780195333213.001.0001
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

- Clark, A. (2016). *Surfing Uncertainty Prediction, Action, and the Embodied Mind*. Oxford University Press (OUP).
- Clark, A. (2017). How to Knit Your Own Markov Blanket. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*.
- Clark, A. (2018). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenology and the Cognitive Sciences*, *17*(3), 521–534. https://doi.org/10.1007/s11097-017-9525-z
- Clark, A. (2025). Extending Minds with Generative AI. *Nature Communications*, *16*(1), 4627. https://doi.org/10.1038/s41467-025-59906-9
- Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19. http://www.jstor.org/stable/3328150
- Constant, A., Clark, A., Kirchhoff, M., & Friston, K. J. (2006). Extended active inference: Constructing predictive cognition beyond skulls. *Biosciences*. https://doi.org/10.1111/mila.12330
- de Bruin, L., & Michael, J. (2017). Prediction error minimization: Implications for Embodied Cognition and the Extended Mind Hypothesis. *Brain and Cognition*, *112*, 58–63. https://doi.org/https://doi.org/10.1016/j.bandc.2016.01.009
- Facchin, M. (2023). Extended Predictive Minds: do Markov Blankets Matter? *Review of Philosophy and Psychology*, *14*(3), 909–938. https://doi.org/10.1007/s13164-021-00607-9
- Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. https://doi.org/10.1016/j.tics.2009.04.005
- Friston, K. (2010). The free-energy principle: A unified brain theory? In *Nature Reviews Neuroscience* (Vol. 11, Number 2, pp. 127–138). https://doi.org/10.1038/nrn2787
- Friston, K. (2012). Prediction, perception and agency. *International Journal of Psychophysiology*, *83*(2), 248–252. https://doi.org/https://doi.org/10.1016/j.ijpsycho.2011.11.014
- Friston, K. (2013). Life as we know it. *Journal of The Royal Society Interface*, *10*(86), 20130475. https://doi.org/10.1098/rsif.2013.0475
- Friston, K., Adams, R., Perrinet, L., & Breakspear, M. (2012). Perceptions as Hypotheses: Saccades as Experiments. *Frontiers in Psychology*, *Volume 3-2012*. https://doi.org/10.3389/fpsyg.2012.00151

- Friston, K., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G. A., & Parr, T. (2023). The free energy principle made simpler but not too simple. *Physics Reports*, *1024*, 1–29. https://doi.org/https://doi.org/10.1016/j.physrep.2023.07.001
- Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G. (2016). Active inference and learning. In *Neuroscience and Biobehavioral Reviews* (Vol. 68, pp. 862–879). Elsevier Ltd. https://doi.org/10.1016/j.neubiorev.2016.06.022
- Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active Inference: A Process Theory. *Neural Computation*, *29*(1), 1–49. https://doi.org/10.1162/NECO_a_00912
- Friston, K., & Frith, C. (2015). A Duet for one. *Consciousness and Cognition*, *36*, 390–405. https://doi.org/10.1016/J.CONCOG.2014.12.003
- Friston, K. J., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics*, *102*(3), 227–260. https://doi.org/10.1007/s00422-010-0364-z
- Friston, K. J., Fagerholm, E. D., Zarghami, T. S., Parr, T., Hipólito, I., Magrou, L., & Razi, A. (2021). Parcels and particles: Markov blankets in the brain. *Network Neuroscience*, *5*(1), 211–251. https://doi.org/10.1162/netn_a_00175
- Friston, K. J., & Frith, C. D. (2015). Active inference, communication and hermeneutics. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *68*, 129–143. https://doi.org/10.1016/J.CORTEX.2015.03.025
- Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221. https://doi.org/10.1098/rstb.2008.0300
- Hasson, U., & Frith, C. D. (2016). Mirroring and beyond: coupled dynamics as a generalized framework for modelling social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1693), 20150366. https://doi.org/10.1098/rstb.2015.0366
- Heersmink, R. (2014). Dimensions of Integration in Embedded and Extended Cognitive Systems. *Phenomenology and the Cognitive Sciences*, *14*. https://doi.org/10.1007/s11097-014-9355-1.
- Hipólito, I., Ramstead, M. J. D., Convertino, L., Bhat, A., Friston, K., & Parr, T. (2021). Markov blankets in the brain. In *Neuroscience and Biobehavioral Reviews* (Vol. 125, pp. 88–97). Elsevier Ltd. https://doi.org/10.1016/j.neubiorev.2021.02.003

- Hohwy, J. (2016). The self-evidencing brain. *Nous*, *50*(2), 259–285. https://doi.org/10.1111/nous.12062
- Hohwy, J. (2020). New directions in predictive processing. *Mind and Language*, *35*(2), 209–223. https://doi.org/10.1111/mila.12281
- Hutchins, E. (1995). *Cognition in the Wild*. MIT Press. https://mitpress.mit.edu/9780262581462/cognition-in the-wild/
- Kahl, S. (2020). *Social Motorics - a predictive processing model for efficient embodied communication*.
- Kahl, S., & Kopp, S. (2018). A Predictive Processing Model of Perception and Action for Self-Other Distinction. *Frontiers in Psychology*, *Volume 9-2018*. https://doi.org/10.3389/fpsyg.2018.02421
- Kelly, M. P., Kriznik, N. M., Kinmonth, A. L., & Fletcher, P. C. (2019). The brain, self and society: a social neuroscience model of predictive processing. *Social Neuroscience*, *14*(3), 266–276. https://doi.org/10.1080/17470919.2018.1471003
- Kersten, L. (2022). A New Mark of the Cognitive? Predictive Processing and Extended Cognition. *Synthese*, *200*(4), 281. https://doi.org/10.1007/s11229-022-03674-2
- Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive Processing*, *8*(3), 159–166. https://doi.org/10.1007/s10339-007-0170-2
- Kirchhoff, M. (2018a). Hierarchical Markov blankets and adaptive active inference: Comment on "Answering Schrödinger's question: A free-energy formulation" by Maxwell James Désormeau Ramstead et al. *Physics of Life Reviews*, *24*, 27–28. https://doi.org/https://doi.org/10.1016/j.plrev.2017.12.009
- Kirchhoff, M. (2018b). Predictive brains and embodied, enactive cognition: an introduction to the special issue. *Synthese*, *195*(6), 2355–2366. https://doi.org/10.1007/s11229-017-1534-5
- Kirchhoff, M. D., & Kiverstein, J. (2021). How to determine the boundaries of the mind: a Markov blanket proposal. *Synthese*, *198*(5), 4791–4810. https://doi.org/10.1007/s11229-019-02370-y
- Kirchhoff, M., & Kiverstein, J. (2019). *Extended Consciousness and Predictive Processing: A Third-Wave View*. https://doi.org/10.4324/9781315150420

- Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, *15*(138). https://doi.org/10.1098/rsif.2017.0792
- Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007
- Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience*, *Volume 8-2014*. https://doi.org/10.3389/fnhum.2014.00302
- Limanowski, J., Adams, R. A., Kilner, J., & Parr, T. (2024). *The Many Roles of Precision in Action*. https://doi.org/10.3390/e26090790
- Menary, R. (2007). Cognitive integration: mind and cognition unbounded. In *ERA - 2010*.
- Menary, R., & Gillett, A. J. (2022). Markov blankets do not demarcate the boundaries of the mind. *Behavioral and Brain Sciences*, *45*, e201. https://doi.org/DOI: 10.1017/S0140525X22000371
- Nagasaka, K., Mizuno, K., Ito, D., & Saida, N. (2017). Proposal of a calculation method to determine the structural components' contribution on the deceleration of a passenger compartment based on the energy-derivative method. *Traffic Injury Prevention*, *18*(sup1), S79–S84.
- https://doi.org/10.1080/15389588.2017.1296957
- Nave, K., Deane, G., Miller, M., & Clark, A. (2020). Wilding the predictive brain. *WIREs Cognitive Science*, *11*(6), e1542. https://doi.org/https://doi.org/10.1002/wcs.1542
- Parr, T., Da Costa, L., & Friston, K. (2019). Markov blankets, information geometry and stochastic thermodynamics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *378*(2164), 20190159. https://doi.org/10.1098/rsta.2019.0159
- Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. The MIT Press. https://doi.org/10.7551/mitpress/12441.001.0001
- Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511803161

- Pearl, J. (2014). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann. https://books.google.ro/books?id=mn2jBQAAQBAJ
- Pezzulo, G., Parr, T., & Friston, K. (2021). The evolution of brain architectures for predictive coding and active inference. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1844), 20200531. https://doi.org/10.1098/rstb.2020.0531
- Proietti, R., Parr, T., Tessari, A., Friston, K., & Pezzulo, G. (2025). Active inference and cognitive control: Balancing deliberation and habits through precision optimization. *Physics of Life Reviews*, *54*, 27–51. https://doi.org/10.1016/J.PLREV.2025.05.008
- Ramstead, M., Albarracin, M., Kiefer, A., Klein, B., Fields, C., Friston, K., & Safron, A. (2023). *The inner screen model of consciousness: applying the free energy principle directly to the study of conscious experience*. https://doi.org/10.48550/arXiv.2305.02205
- Ramstead, M. J. D., Friston, K. J., & Hipólito, I. (2020). Is the Free-Energy Principle a Formal Theory of Semantics? From Variational Density Dynamics to Neural and Phenotypic Representations. *Entropy*, *22*(8). https://doi.org/10.3390/e22080889
- Ramstead, M. J. D., Kirchhoff, M. D., Constant, A., & Friston, K. J. (2021). Multiscale integration: beyond internalism and externalism. *Synthese*, *198*(1), 41–70. https://doi.org/10.1007/s11229-019-02115-x
- Shpitser, I., & Pearl, J. (2006). Identification of conditional interventional distributions. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, UAI'06*, 437–444.
- Smith, R., Friston, K. J., & Whyte, C. J. (2022). A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*, *107*, 102632. https://doi.org/10.1016/J.JMP.2021.102632
- Stubbs, G., & Friston, K. (2024). The police hunch: the Bayesian brain, active inference, and the free energy principle in action. *Frontiers in Psychology*, *Volume 15-2024*. https://doi.org/10.3389/fpsyg.2024.1368265
- Tamir, D. I., & Thornton, M. A. (2018). Modeling the Predictive Social Mind. *Trends in Cognitive Sciences*, *22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005
- Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J., & Kirmayer, L. J. (2020). Thinking through other minds: A variational approach to cognition and culture. *Behavioral and Brain Sciences*, *43*, e90. https://doi.org/DOI: 10.1017/S0140525X19001213

- Venter, E. (2021). Toward an Embodied, Embedded Predictive Processing Account. *Frontiers in Psychology*, *Volume 12-2021*. https://doi.org/10.3389/fpsyg.2021.543076
- Wheeler, M. (2012). In Defense of Extended Functionalism. *The Extended Mind*.
- https://doi.org/10.7551/mitpress/9780262014038.003.0011
- Wheeler, M. (2015). Extended Consciousness: an Interim Report. *The Southern Journal of Philosophy*, *53*(S1), 155–175. https://doi.org/https://doi.org/10.1111/sjp.12124
- White, B., Clark, A., Guènin-Carlut, A., Constant, A., & Di Paolo, L. D. (2025). Shifting boundaries, extended minds: ambient technology and extended allostatic control. *Synthese*, *205*(2), 81. https://doi.org/10.1007/s11229-025-04924-9
- Wiese, W., & Friston, K. J. (2021). Examining the Continuity between Life and Mind: Is There a Continuity between Autopoietic Intentionality and Representationality? *Philosophies*, *6*(1). https://doi.org/10.3390/philosophies6010018
- Wilson, R. A. (1994). Wide Computationalism. *Mind*, *103*(409), 351–372. https://philarchive.org/archive/WILWC

Maria Luiza Iennaco - Toward a Multiscale Account of Attention: Neurophilosophy and Synthetic Philosophy Compared

University of São Paulo/University of Porto

This study investigates two philosophical methodologies – neurophilosophy and synthetic philosophy – as they apply to the study of attention, with the aim of developing a position that integrates both traditions into a more comprehensive account of the phenomenon.

Although attention benefits from consolidated paradigms and continued scientific advancement, it remains a well-studied phenomenon beset by considerable epistemological and conceptual difficulties (Hommel et al., 2019). One response to these difficulties comes from neurophilosophy (Churchland, 2022), which establishes a co-evolutionary dialogue between philosophy of mind and the brain sciences, such that each discipline constrains and informs the other. Even in its non-reductive variants (Northoff, 2004), neurophilosophy operates as a prescriptive research program, defining the terms under which philosophical and scientific inquiry should engage and privileging the neuro-philosophical nexus as the primary explanatory locus. Synthetic philosophy (Schliesser, 2019) takes a different approach, proceeding descriptively and narratively: it assembles coherent accounts of phenomena by drawing on whichever disciplinary sources prove illuminating, with no prior commitment to a privileged interlocutor or mode of dialogue.

With respect to their capacity to account for attention at both personal and subpersonal levels of description, the two methodologies are evaluated here in comparative terms. The explanatory scope of neurophilosophy depends on its degree of reductionism: in its stronger reductionist forms, psychology risks marginalization and the account narrows to subpersonal processes alone, whereas its non-reductive forms, while more integrative, remain constrained by the need for participating disciplines to engage across levels of analysis in theoretically compatible ways. Synthetic philosophy, in contrast, achieves genuine explanatory pluralism and broad interdisciplinary integration, though this same expansiveness risks diluting

explanatory precision and complicating the evaluation or operationalization of the resulting account.

Complementary strengths emerge from this comparative analysis. Because it remains agnostic about privileged interlocutors, synthetic philosophy proves more effective at bridging personal and subpersonal accounts of attention, that is, at integrating phenomenological, cognitive, and neuroscientific perspectives within a single, unified narrative. Non-reductive neurophilosophy, by contrast, offers a more disciplined and cohesive framework for examining attention at the subpersonal level, where the constraints imposed by the 'neuro' sharpen inquiry rather than limit it. Taken together, the two methodologies chart complementary pathways toward a fuller understanding of attention,

suggesting that the choice between them need not be exclusive, but rather should be guided by the level of description and explanatory aims at stake.

References

- Churchland, P. S. (2022). What is Neurophilosophy and How Did Neurophilosophy Get Started? Journal of NeuroPhilosophy, 1(1). https://doi.org/10.5281/zenodo.6637568
- Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J.-H., & Welsh, T. N. (2019). No one knows what attention is. Attention, Perception, & Psychophysics, 81(7), 2288–2303. https://doi.org/10.3758/s13414-019-01846-w
- Northoff, G. (2004). What Is Neurophilosophy? A Methodological Account. Journal for General Philosophy of Science, 35(1), 91–127. https://doi.org/10.1023/B:JGPS.0000035153.89143.4c
- Schliesser, E. (2019). Synthetic philosophy. Biology & Philosophy, 34(2), 19. https://doi.org/10.1007/s10539-019-9673-3

Christos Kalpakidis - Transparency, Direct Realism, and the Minimal Self: Insights from Existential Phenomenology

University of Athens

Contemporary analytic philosophy of perception and phenomenological theories of self-consciousness have developed largely in isolation, yet they converge on a shared set of problems. This paper argues that existential phenomenology offers an underexplored but powerful framework for integrating them, and proposes Jean-Paul Sartre's early phenomenology as a candidate through which the philosophy of

perception and the minimal-self debate can be discussed together. The presentation will focus on the following problem: Naïve realism's two core commitments — relationalism and transparency — give a compelling account of the object-directed side of perceptual phenomenology but arguably neglects the subject side contribution; at the same time, minimal self promises to supply this missing 'for me-ness'. I contend that such a pairing remains merely programmatic until it can show how relationalism, trasparency and minimal self-awareness are grounded in a single structure rather than remain externally conjoined.

Sartre´s phenomenology supplies exactly such a structure. He argues from the transparency and 'emptiness' of consciousness — its purely relational, non-substantial nature — to the existence of mind-independent objects. This is a route from transparency to a position 'beyond realism and idealism' that functions as a form of direct realism. Crucially, the very transparency that underwrites naïve realism's object directedness is, in Sartre, inseparable from the pre-reflective self-consciousness of the for-itself. The subject-side and object-side that the perception debate treats as separately theorised commitments turn out to be two aspects of one relational structure. At the same time, Sartrean pre-reflective consciousness resists assimilation to the standard, broadly Zahavian minimal self. The minimal self is not an invariant 'structure' ensuring mere mineness, but an existential movement toward the world, unintelligible apart from some content and some fundamental project for which consciousness remains responsible. Because the sartrean for-itself is constitutively a nihilating relation to being, its self-awareness is not a self-standing inner glow but the reverse side of

its openness to the in-itself. Minimal self-consciousness and direct perceptual acquaintance are thereby co-constituted. Sartre's phenomenology thus emerges as a genuine candidate for integrating the philosophy of perception and the phenomenology of the minimal self, offering substantive commitments about the relational co dependence of world-directedness and self-awareness.

References

- Boyle, M. (2018). Sartre on Bodily Transparency. *Manuscrito*, 41(4), 33–70.
- Campbell, J. (2002). *Reference and Consciousness*. Oxford: Clarendon Press.
- French, C. (2014). Naïve Realist Perspectives on Seeing Blurrily. *Ratio*, 27(4), 393–413.
- Gallagher, S. (2000). Philosophical Conceptions of the Self: Implications for Cognitive Science. *Trends in Cognitive Sciences*, 4(1), 14–21.
- Kim, D. S. H. (2022). Naïve Realism and Minimal Self. *Phenomenology and Mind*, 22, 150–159.
- Martin, M. G. F. (2002). The Transparency of Experience. *Mind & Language*, 17, 376–425.
- Merleau-Ponty, M. (2012). *Phenomenology of Perception* (D. A. Landes, Trans.). London: Routledge.
- Sartre, J.-P. (1960). *The Transcendence of the Ego: An Existentialist Theory of Consciousness* (F. Williams & R. Kirkpatrick, Trans.). New York: Hill and Wang.
- Sartre, J.-P. (2004). *The Imaginary: A Phenomenological Psychology of the Imagination* (J. Webber, Trans.). London: Routledge.
- Sartre, J.-P. (2018). *Being and Nothingness: An Essay in Phenomenological Ontology* (S. Richmond, Trans.). London: Routledge.
- Rowlands M. (2020). Sartre on Intentionality and Pre-Reflective Consciousness. In M. Eshleman & C. Mui (Eds.), *The Sartrean Mind* (pp. 158–171). New York: Routledge.
- Zahavi, D. (2005). *Subjectivity and Selfhood: Investigating the First-Person Perspective*. Cambridge, MA: MIT Press.

Jack Shardlow - Pain Imagery: Individual Differences and Disutilitous Memories

University of Liverpool

Close your eyes and, if you can, conjure up some visualization of your breakfast this morning. If this is something you can do, you thereby entertain mental imagery – something that might be construed as a quasi-perceptual experience, as of your breakfast.

Mental imagery is typically taken to be a mechanism by which we simulate experiences. For something to be a case of simulating an experience, as opposed to simply *thinking about* that experience, it is taken to involve the iconic recreation/representation of the sensory, qualitative aspects of experience. In functional terms, 'mental imagery' can be used to refer to representations "of sensory information without a direct external stimulus" (Pearson et al., 2015, p. 590; also see Currie 1995, 26). Understood in these terms, I take mental imagery to be common to (episodic/experiential) recollection and imagination.

In my talk, I focus on mental imagery and *pain experience*. In doing so, there are two central questions of interest: First, is there such a thing as pain imagery? Second, if there is, are there individual differences in the representation of the qualitative aspects of pain experience? I motivate affirmative answers to each question.

The talk is structured as follows. I first frame why my research questions ought to be of interest at all, outlining how theorists have recently offered an opposing view. On this opposing view – which we can call the No Pain Imagery view (NPI) – it is claimed that none of us can imagine/remember *what it is like* to undergo pain experience (see, e.g., Gerrans 2018; Montero 2020; Radden 2021).

Within this view we can tease apart two claims: that there is no pain imagery, and that, because there is no pain imagery *full stop*, there is no variation between individuals for such imagery. In outlining how theorists have motivated NPI, I say more about what I am assuming in discussing pain experience, before offering a broad characterization of what it is to entertain

some mental imagery, in the relevant sense, whether applied to, e.g., visual experience, auditory experience, or pain experience.

My positive argument then proceeds in three steps. In the first and second, I argue that the empirical work on pain memory (e.g., Morely 1993; Niven and Murphy-Black 2000; Graham et al. 2020) and empirical work on multisensory aphantasia (e.g., Dance et al. 2021; Dawes et al. 2024), each support the claim that there is pain imagery, and individual differences for such pain imagery, over NPI. For the third step, I argue that the story of individual differences is also supported by an inference to the best explanation, based on recent findings concerning the disutility people assign to memories of pain experience (see Lee et al. 2022; Shardlow et al. 2025). In drawing an explanatory connection between pain imagery ability and the disutility that people assign to memories of pain, as well as offering a novel proposal about the disutility of pain memory, my account makes predictions that can be empirically tested, outlining a promising avenue for future interdisciplinary research.

References

- Currie, Gregory. (1995). Visual Imagery as the Simulation of Vision. *Mind and Language*, 10: 25–44.
- Dance, Carla, Ward, Jamie, and Simner, Julia. (2021). What is the link between mental imagery and sensory sensitivity? Insights from aphantasia. *Perception*, 50 (9), pp. 757-782.
- Dawes, Alexei J., Keogh, Rebecca, & Pearson, Joel. (2024). Multisensory Subtypes of Aphantasia: Mental Imagery as Supramodal Perception in Reverse. *Neuroscience Research*, 201: 50–59.
- Gerrans, Philip. (2018). Painful memories. In K. Michaelian, D. Debus, & D. Perrin (Eds.), *New Directions in the Philosophy of Memory* (pp. 158-178). New York, NY; USA: Routledge.
- Graham, Christopher, Brown, Shona, Vincent, Katy, & Horne, Andrew (2020). International survey confirms that women with endometriosis-associated pain experience a high prevalence of pain imagery and coping imagery. *European Journal of Obstetrics and Gynecology and Reproductive Biology*, 244, 203-204.

- Lee, Ruth, Shardlow, Jack, O'Connor, Patrick, Hotson, Lesley, Hotson, Rebecca, Hoerl, Christoph, & McCormack, Teresa. (2022). Past-future preferences for hedonic goods and the utility of experiential memories. *Philosophical Psychology*, 35(8), 1181–1211.
- Montero, Barbara G. (2020). What Experience Doesn't Teach: Pain Amnesia and a New Paradigm for Memory Research. *Journal of Consciousness Studies*, 27 (11-12): 102-125.
- Morley, S. (1993) Vivid memory for 'everyday' pains, Pain, 55 (1), pp. 55–62. Niven, Catherine & Murphy-Black, Tricia. (2000). Memory for labor pain: A review of the literature, *Birth*, 27 (4): 244–253.
- Pearson, Joel, Naselaris, Thomas, Holmes, Emily A., & Kosslyn, Stephen M. (2015). Mental Imagery: Functional Mechanisms and Clinical Applications. *Trends in Cognitive Sciences*, 19: 590–602. Radden, Jeniffer (2021). Imagined and delusional pain. *Rivista Internazionale Di Filosofia E Psicologia,* 12 (2): 151-166.
- Shardlow, Jack, Lee, Ruth, O'Connor, Patrick A., Hoerl, Christoph & McCormack, Teresa (2025). Discounting past experience and the utility of memory: an empirical study. *Synthese* 205, 166.

Nathália de Ávila - Remembering at the Edge of Consciousness: Affectivity and the Reconstitution of Selfhood in Near-Death Experiences

Universität zu Köln

When does the self-end? Near-death experiences (NDEs) pose this question through a striking paradox. They are frequently reported under conditions in which core structural features of selfhood, including embodied perspective, agency, narrative continuity, and temporal orientation (Dings & McCarroll, 2022; Newen, 2018; Schechtman, 2017; Gallagher, 2000), are partially or wholly disrupted. Yet these same events are later recalled as vivid, coherent, and profoundly identity-shaping episodes (Martial et al., 2017; Moore & Greyson, 2017). This project argues that NDEs expose a neglected feature of episodic memory, namely its capacity not only to preserve a stable self across time but also to help reconstitute selfhood after experiences in which its ordinary organizing structures have temporarily collapsed.

The project first develops a structural account of the limits of self-experience. It does not identify the end of the self with unconsciousness or biological death. Instead, it examines the conditions that ordinarily scaffold subjectivity, including temporality, perspectival embodiment, agency, and narrative self-ascription. Phenomenological reports of NDEs suggest that these structures may partially disintegrate while experiential presence and affective intensity remain. What persists in such edge-states is not the narrative or agentive self but a minimally structured field of affective consciousness. This account distinguishes the structural dissolution of selfhood from the cessation of experience and provides a framework for understanding consciousness at its limits.

This perspective raises a challenge for dominant theories of episodic memory. Standard accounts typically assume that episodic encoding depends on attention, temporal organization, and integration within an ongoing sense of self (Tulving, 1973, 2002). If this assumption is correct, experiences occurring under conditions of disrupted selfhood should yield fragmentary or unstable recollections. Yet NDE memories are frequently reported as unusually vivid, enduring, and transformative. To explain this phenomenon, I introduce the concept of **affective**

gravity. Affective gravity refers to the capacity of certain experiences to stabilize memory by reorganizing the subject's affective orientation instead of relying primarily on ordinary encoding mechanisms. In edge-states where narrative organization is attenuated, affective intensity anchors subsequent interpretation and facilitates autobiographical integration. This proposal also responds to the objection that the coherence of NDE memories results solely from later narrative reconstruction. I argue that the exceptional affective force of the original experience exerts a lasting influence on autobiographical organization. The concept of affective gravity builds directly on the theory of affective memory developed in my doctoral research.

The broader implication is that the end of the self is neither identical with the cessation of experience nor reducible to biological death. Selfhood may end structurally through the collapse of its organizing principles while minimally structured affective consciousness persists. Episodic memory can subsequently extend selfhood beyond such interruptions by transforming edge-experiences into new anchors of personal identity. Beyond NDEs themselves, this project suggests that the self is best understood not as a continuously preserved entity but as a dynamically reconstructed achievement that is vulnerable to disruption and capable of retrospective renewal. This framework also lays the foundation for future intercultural research on conceptions of post-mortem continuity and their implications for theories of consciousness and affective memory.

Matthew Owen - Can Human Cerebral Organoids Become Conscious? A Hylomorphic Perspective

Department of Philosophy, Yakima Valley College / Center for Consciousness Science, University of Michigan

Abstract: A human cerebral organoid consists of neurons developed from human pluripotent stem cells that self-organize into structures with functions resembling the neuroanatomy and neurophysiology of the cerebrum in the human brain in its early development. We are in the rapidly developing early days of research on human cerebral organoids, during which it is important to carefully consider what ethically responsible research progress involves. A key question toward this end is: Can human cerebral organoids become conscious? This paper addresses this question from a hylomorphic perspective of human nature and consciousness originating with Aristotle and developed by Thomas Aquinas. It is argued that given this Aristotelian-Thomistic hylomorphic perspective, human cerebral organoids will not become conscious.

Filipp Matveev - Physicalism Naturalized

University of Miami

Physicalism is widely treated as the default position in the philosophy of mind. This is standardly credited to methodological naturalism: metaphysical theorizing should proceed, where possible, by the methods and standards of successful science. And doing so is taken to favor physicalism [5,8]. Another case for physicalism, the causal closure argument, turns this general methodological stance into a reductive conclusion. It relies on the principle that every physical event that has a cause has a sufficient physical cause. Distinct mental causes are then redundant: mental causes, if there are any, must be physical causes [4,7]. The two arguments are usually treated as complementary. Naturalism motivates physicalism and causal closure sharpens it into a claim about reduction or identity.

In this paper I argue that the two arguments are in tension. Once the content of methodological naturalism is made explicit, it fails to support the universal causal closure principle that the closure argument requires. My target is therefore the naturalistic route to reductive physicalism, not the view as such.

The argument proceeds in three steps. First, I articulate three desiderata for any metaphysical posit whose warrant is claimed to be derived from scientific methodology (drawing on work on scientific modeling, the material theory of induction, and effective field theories) [1,2,6,10]. (i) Operational indispensability: commitment should track the role a posit plays in successful scientific models. (ii) Locality of warrant: inductive support is always bounded to a regime and cannot be automatically projected across regimes. (iii) Multiscale organization: commitments should respect the layered structure of our best theories, in which higher-level descriptions are often dynamically decoupled from micro-details.

Second, I argue that universal causal closure fails all three desiderata. (i) It plays no operational role in scientific models. Predictive and explanatory work is done by local models with their own difference-making variables. Conservation laws yield closure conclusions only under assumptions of effective isolation. (ii) Its standard defense, an induction from the success of physical explanation in studied regimes to closure everywhere, is unwarranted without

evidence that the relevant background conditions carry over [6,7]. And (iii) the micro-sufficiency reading that the closure argument requires, on which microphysical causes screen off all higher-level causes, fits poorly with the multiscale structure of modern physics, where the difference making variables are often macro-level variables [1,4,10].

Third, I outline the resulting dilemma for physicalism. Physicalists can keep universal closure by weakening naturalism to mere consistency with science. But then naturalistic dualists and panpsychists can claim naturalism as well [3,9]. So, in doing so, physicalism forfeits its distinctive methodological advantage. Alternatively, physicalists can keep strict methodological naturalism and abandon universal closure. This gives a modest, regime-bound, non-reductive dependence thesis about the minds science actually models. But that does not warrant reductive physicalism.

The conclusion is dialectical. Physicalists who claim the methodological high ground owe an account of which notion of naturalism does the work, is that notion consistent with universal closure, and why it favors physicalism over its rivals.

References

- Bain, J. (2013). "Effective Field Theories." In R. Batterman (ed.), *The Oxford Handbook of Philosophy of Physics*. Oxford University Press. doi.org/10.1093/oxfordhb/9780195392043.013.0007
- Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge University Press. doi.org/10.1017/CBO9781139167093
- Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press. https://consc.net/the-conscious-mind/
- Kim, J. (1998). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. MIT Press. doi.org/10.7551/mitpress/4629.001.0001
- Melnyk, A. (2003). *A Physicalist Manifesto: Thoroughly Modern Materialism*. Cambridge University Press. doi.org/10.1017/CBO9780511498817
- Norton, J. D. (2021). *The Material Theory of Induction*. University of Calgary Press. doi.org/10.1515/9781773852553 [7] Papineau, D. (2002). *Thinking About Consciousness*. Oxford University Press. doi.org/10.1093/0199243824.001.0001

- Stoljar, D. (2024). "Physicalism." *The Stanford Encyclopedia of Philosophy*, Spring 2024 edition. https://plato.stanford.edu/archives/spr2024/entries/physicalism/
- Strawson, G. (2006). "Realistic Monism: Why Physicalism Entails Panpsychism." *Journal of Consciousness Studies* 13(10-11): 3-31. https://philpapers.org/archive/STRRM-2.pdf
- Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press. doi.org/10.1093/0195155270.001.0001

Don Oxtoby - Inferentialism about Recognition

University of Turin, Department of Philosophy and Education Sciences

Perceptual recognition—the awareness that one has previously encountered a familiar face, voice, object, or other stimulus—is a vital cognitive function. Despite its importance, the nature of recognition remains poorly understood. Philosophical discussions have addressed neighboring issues concerning perceptual content, concepts, and memory, yet have paid comparatively little attention to what constitutes recognition itself, or to the implications of psychological models of recognition for answering that question.[1] This paper argues that psychological models support a qualified form of inferentialism: recognition is not exhausted by the co-occurrence of perception and memory retrieval, but requires judgment, understood as a minimal truth-directed affirmation, to integrate the two into a recognitional state.

The central claim is that psychological theories reveal a constitutive gap between concurrent perception and recall, on the one hand, and genuine recognition, on the other. Recognition is therefore not merely the copresence of perceptual and mnemic representations, but the determination that they belong together. Judgment is required to bridge this gap.

Psychological models characterize recognition as beginning when perceptual information sufficiently matches information retrieved from long-term memory. They further describe recognition as maintained through a continuous error-monitoring process that evaluates evolving perceptual and mnemic information. These models distinguish recognition from mere co-occurrence of perception and recall, but leave open what explains this transition. I argue that judgment provides the missing element. Both the initial determination that perceptual and mnemonic information sufficiently correspond and the subsequent monitoring of that correspondence plausibly require judgment. This interpretation is supported by psychological evidence distinguishing multiple routes to recognition, each involving different mnemonic processes and conscious byproducts. It is further supported by neuroimaging studies implicating brain regions associated with judgment during recognition, together with lesion studies linking damage to those regions with impaired recognition. Here, I understand

[1] Abid, 2021 and Dokic, 2010 are exceptions. Block, 2023 also briefly targets the specific question of what constitutes recognition (p. 274).

judgment as a form of truth-directed affirmation, subject in certain ways to conscious control, that arises in response to conscious byproducts of the processes underlying recognition. Although judgment is often treated as a subsequent response to recognition, these findings instead suggest that judgment is constitutively involved in recognition itself.

The final section addresses the objection that inferential views overintellectualize recognition, particularly in cases involving infants and nonhuman animals. I argue that the force of this objection depends on the conception of judgment employed. If judgment is understood in a suitably minimal sense, inferentialism can accommodate many paradigmatic cases of recognition without attributing sophisticated conceptual capacities. I conclude by considering the implications of this account for debates concerning the rationality of perception. Rather than locating rationality within perception itself, many paradigm cases are better understood as reflecting the constitutive role of judgment in perceptual recognition. This provides a new way of explaining the apparent rationality of perception without attributing rational properties to perception itself.

References

- Abid, G. 2022. 'Recognition and the Perception-Cognition Divide.' *Mind & Language* 37(5): 770-789. https://doi.org/10.1111/mila.12362
- Aly, M., Yonelinas, A. P., Kishiyama, M. M., & Knight, R. T. 2011. 'Damage to the Lateral Prefrontal Cortex Impairs Familiarity but not Recognition.' *Behavioural Brain Research* 225(1): 297-304. https://doi.org/10.1016/j.bbr.2011.07.043
- Barton, J. J. S. & Corrow, S. L. 2016. 'Recognizing and Identifying People: A Neuropsychological Review.' *Cortex* 75: 132-150. https://doi.org/10.1016/j.cortex.2015.11.023
- Block, N. 2023. *The Border Between Seeing and Thinking*. New York: Oxford University Press. Dokic, J. 2010. 'Perceptual Recognition and the Feeling of Presence.' In B. Nanay (ed.), *Perceiving the World*, pp. 33-53. Oxford: Oxford University Press.
- Öztel, T. & Balcı, F. 2023. 'Metric Error Monitoring as a Component of Metacognitive Processing.' *European Journal of Neuroscience* 59(5): 807-821. https://doi.org/10.1111/enj.16182
- Siegel, S. 2017. *The Rationality of Perception*. New York: Oxford University Press. Wang, C., Furlong, T. M., Stratton, P. G., Lee, C. C. Y., Xu, L., Merlin, S., Nolan, C.,

Arabzadeh, E., Marek, R., & Sah, P. 2021. ‘Hippocampus-Prefrontal Coupling Regulates

- Recognition Memory for Novelty Discrimination.’ *Journal of Neuroscience* 41(46): 9617-9632. https://doi.org/10.1523/jneurosci.1202-21.2021

Juan Diego Bogotá - Can Social Interaction Constitute the Minimal Self?

University of Jyväskylä

A key discussion within contemporary consciousness studies concerns the importance of subjectivity for our understanding of phenomenal consciousness. It is argued that, aside from the qualitative character of experience (i.e., the 'what-it-is-like'), we must take the first-person perspective seriously and acknowledge that phenomenality also involves a subjective character that can be described as a 'mineness' or 'for-me-ness' (Zahavi & Kriegel, 2016). The for-me-ness of experience, it is argued, corresponds to a primitive and pre-reflective form of self-awareness that can be considered a minimal form of selfhood (Zahavi, 2014).

A recent debate concerning the so-called the minimal self revolves around the question of whether it is constitutively dependent on social interaction. Zahavi has influentially argued that the minimal self is pre-social in the sense that, as formal and invariant structure of phenomenal consciousness, it can in principle exist independently of intersubjectivity. Critics, developmental considerations, have argued instead that even this minimal structure is socially constituted (e.g., Ratcliffe, 2017; Ciaunica & Fotopoulou, 2017).

In my presentation, I argue that both sides are, in important respects, right, once the concept of *constitution* at stake in the debate is properly disambiguated. Drawing on Husserl's (1970) phenomenological theory of parts and wholes, I argue that for-me-ness is an abstract moment of experience: it is a non-independent part that is distinguishable from, but never separable from, other constitutive structures of experience, most notably background agency (the pre

reflective sense of being capable of acting in the world; see Bogotá, 2026). While for-me-ness, considered in the abstract, remains formally invariant and therefore pre-social, background agency is diachronically shaped by habituation, much of which is socially acquired and sustained. Drawing on recent work on diachronic constitution within situated cognition

(Kirchhoff & Kiverstein, 2024), I argue that social interaction can be understood as genuinely constitutive of the concrete structures of subjectivity in virtue of their intimate foundational relationship with for-me-ness. The upshot is that the minimal self is pre-social when considered as an abstract moment, but socially constituted when considered as it is concretely lived through.

References

- Bogotá, J. D. (2026). Embodied Mineness and Background Agency: A Neurophenomenological Approach to the Minimal Self. *Synthese*, *207*(6). https://doi.org/10.1007/s11229-026-05624-8
- Ciaunica, A., & Fotopoulou, A. (2017). The Touched Self: Psychological and Philosophical Perspectives on Proximal Intersubjectivity and the Self. In C. Durt, T. Fuchs & C. Tewes (Eds.), *Embodiment Enaction, and Culture. Investigating the Constitution of the Shared World* (pp. 174-192). MIT Press. https://doi.org/10.7551/mitpress/9780262035552.003.0009
- Husserl, E. (1970). *Logical Investigations* (Volume II). Routledge.
- Kirchhoff, M., & Kiverstein, J. (2024). Diachronic Constitution. *Manuscrito*, *47*(1). https://doi.org/10.1590/0100-6045.2024.v47n1.mj
- Ratcliffe, M. (2017). Selfhood, Schizophrenia, and the Interpersonal Regulation of Experience. In C. Durt, T. Fuchs & C. Tewes (Eds.), *Embodiment Enaction, and Culture. Investigating the Constitution of the Shared World* (pp. 150-171). MIT Press. https://doi.org/10.7551/mitpress/9780262035552.003.0008
- Zahavi, D. (2014). *Self and Other: Exploring Subjectivity, Empathy, and Shame*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199590681.001.0001
- Zahavi, D., & Kriegel, U. (2016). For-me-ness. What it is and what it is not. In D. O. Dahlstrom, A. Elpidorou, & W. Hopp (Eds.), *Philosophy of Mind and Phenomenology. Conceptual and Empirical Approaches* (pp. 37-53). Routledge.

Caterina Scianna - When Imagination Becomes Truth: Confabulation and Epistemic Vigilance

Pegaso University

This contribution explores the idea that confabulations can be understood as configurations in which the cognitive mechanisms governing the attribution and management of beliefs become jammed, at the level of whole patterns of discursive practice. The central hypothesis is that, in confabulation, there is a breakdown in the link between epistemic vigilance and truth-making devices: contents that should remain in the register of hypothesis, reconstruction or fiction are promoted to the status of truth because monitoring ceases to function as a corrective and suspensive filter. Confabulations thus offer a window into the conditions under which our ordinary "truth routines" can misfire, allowing imaginative or interpretive constructions to crystallize as indisputable facts in the subject's cognitive economy.

Clinical confabulations are taken to be limit cases of a stably impaired epistemic vigilance, where the system responsible for discriminating between reliable and unreliable representations is durably compromised. In certain psychiatric conditions, patients provide detailed, confident narratives about events that never occurred, or about personal histories that have been radically distorted; what is striking is not simply the falsity of these contents, but the absence of the usual markers of doubt, hesitation and provisionality that accompany ordinary reconstruction. In healthy subjects, by contrast, confabulatory episodes typically appear as local, context-bound collapses: under conditions of stress, cognitive load or conversational pressure, the processes that should keep the boundary between memory, inference and imagination distinct are weakened, turning regulated suspension into a habitual interpretive stance. Everyday confabulations show how quickly the mind can slide from "this might be what happened" to "this is what happened" when epistemic vigilance is relaxed.

In both clinical and non-clinical cases, epistemic self-trust is not abolished but derailed: it continues to operate, yet without the checks that should prevent the promotion to truth of contents that have not undergone adequate epistemic scrutiny. The subject still relies

on herself as a source of knowledge, but does so on the basis of representations whose origins, reliability and compatibility with other beliefs are no longer properly monitored. This suggests that confabulation is a distinctive failure of the mechanisms that regulate when and how a representation is allowed to enter the "belief box".

The contribution will argue that confabulations are best treated as discursive configurations in which the very function of epistemic vigilance undergoes a collapse, with consequences for how we think about the speaker's responsibility and credibility in contexts of systematic error and ordinary error. The analysis will examine how patterns of presupposition, narrative completion and inferential "leaps" reshape the boundary between what is tentatively entertained and what is taken as fact. By bringing together clinical data on memory confabulation, experimental work on figurative and indirect language, and philosophical accounts of epistemic vigilance, the contribution aims to show that confabulation is a revealing test case for any theory that links human communication to the capacity to track and evaluate mental states - one's own and those of others - in a normatively constrained way.

References

- Hirstein W. 2010 "Confabulation about personal memories, normal and abnormal" in Nalbantian et al eds The memory process Cambridge Mit Press
- Sperber D. et al 2010 "Epistemic vigilance", Mind & Language 25, 359-393.
- Sullivan-Bissett E. 2015 "Implicit bias, confabulation, and epistemic innocence", *Consciousness and cognition* 33, 548-560.

Emilio Leone - Shared Embodied Time

University of Messina

When Husserl undertook his phenomenological analysis of temporality, he replaced the traditional metaphor of the string of pearls with the phenomenological analysis of melody, conceptualizing time not as a succession of discrete instants, but as a flowing continuum. Through the threefold intentional structure of retention, primal impression, and protention, the present renders lived experience possible. While this framework provides a robust account of individual inner-time consciousness, it leaves a crucial question largely unexplored: what happens when multiple melodies interact? This work introduces the concept of *Shared Embodied Time* to account for the co-constitution of temporal experience among multiple embodied subjects, adopting an embodied, enactive, and neurophenomenological approach. The paper extends the phenomenological analysis of melody to intersubjective contexts and proceeds in three sections, moving from individual to collective dimensions of temporal cognition.

The first section outlines the theoretical transition toward *Embodied Time*, a paradigm that conceives temporality not as an objective parameter, but as a dynamic process co-constituted through the interaction of mind, body, and environment. Building upon the embodied mind framework and Varela's neurophenomenological project addressing the "hard problem" of consciousness, this section illustrates how lived time emerges from neural dynamics and first-person experience. By integrating Husserlian phenomenology with Merleau-Ponty's reflections on time, temporality is revealed as a structure that is simultaneously passive and active, vital and intentional. Time does not function as a neutral background but emerges as a self-organizing structure of embodied consciousness.

The second section addresses how this embodied temporal structure is shared among multiple subjects. If the body is our primary "musical instrument", intersubjective interaction can be compared to collective musical performance. When musicians play together, they must reciprocally coordinate their movements and temporal anticipations. This pre-reflective coordination generates a *Shared Embodied Time* sustained by three core phenomenological elements: temporal synchronization (aligning retentions and protentions into a "shared

present"), bodily and affective resonance, and temporal co-regulation. Engaging in a direct dialogue with Dan Zahavi's (2025) *Being We*, the paper will argue that shared emotional experiences are not fusional unities that annul the first-person perspective, but rather relational unities where participants' experiences mutually and constitutively determine one another within a shared temporal field.

The final section anchors the phenomenological analysis of shared temporal experience in recent findings from cognitive neuroscience and psychopathology. Subjective time is profoundly rooted in interoception, with experimental evidence highlighting the functional association between brain-heart interactions (such as the heartbeat-evoked potential) and duration judgments. In intersubjective encounters, individual implicit temporality transitions into collective resonance through the dynamic coupling and autonomic synchronization of agents. This enables participatory sense-making and mutual incorporation. Conversely, the rupture of this shared temporal structure, resulting in "sonic chaos" rather than a cohesive melody, can be observed in psychopathological conditions. In melancholic depression, patients suffer a profound desynchronization from social time, while in personality disorders, temporal inflexibility traps the subject in rigid patterns, blocking the capacity to improvise and share the fluid time of the dyad.

References

- Chalmers, D. J. (1995). Facing up to the Problem of Consciousness. Journal of Consciousness Studies, 2(3), 200–219; Collins, R. (2004). Interaction Ritual Chains. Princeton University Press;
- Fuchs, T. (2005). Implicit and Explicit Temporality. Philosophy, Psychiatry, & Psychology, 12(3), 195-198; Fuchs, T. (2013). Temporality and Psychopathology. Phenomenology and the Cognitive Sciences, 12(1), 75-104; Fuchs, T. (2018). Ecology of the Brain: The Phenomenology and Biology of the Embodied Mind. Oxford University Press;
- Fuchs, T. & De Jaegher, H. (2009). Enactive Intersubjectivity: Participatory Sense-making and Mutual Incorporation. Phenomenology and the Cognitive Sciences, 8(4), 465-486;
- Gallagher, S. (2007). How the Body Shapes the Mind. Oxford University Press;
- Gallagher, S. (2020). Action and Interaction. Oxford University Press;

- Gallagher, S. & Zahavi, D. (2008). The Phenomenological Mind: An Introduction to Philosophy of Mind and Cognitive Science. Routledge;
- Husserl, E. (1966). Zur Phänomenologie des inneren Zeitbewusstseins (1893-1917). Martinus Nijhoff; trans. On the Phenomenology of the Consciousness of Internal Time (1893–1917) (J. B. Brough, Trans.). Kluwer Academic Publishers;
- Khoshnoud, S., Leitritz, D., Bozdağ, M. Ç., Igarzábal, F. A., Noreika, V. & Wittmann, M. (2024). When the Heart Meets the Mind: Exploring the Brain-Heart Interaction during Time Perception. The Journal of Neuroscience, 44(34); Krueger, J. (2016). Extended Emotions. Philosophical Explorations, 19(2), 123–141;
- Merleau-Ponty, M. (2002). Phenomenology of Perception (C. Smith, Trans.). Routledge (Original work published 1945); Rodriguez, E., George, N., Lachaux, J. P., Martinerie, J., Renault, B. & Varela, F. J. (1999). Perception's Shadow: Long distance Synchronization of Human Brain activity. Nature, 397(6718), 430–433;
- Scheler, M. (2008). The Nature of Sympathy (P. Heath, Trans.). Transaction Publishers (Original work published 1923); Searle, J. R. (1995). The Construction of Social Reality. Free Press;
- Stein, E. (2000). On the Problem of Empathy (W. Stein, Trans.). ICS Publications (Original work published 1917); Sterna, A., Moskalewicz, M., Schmidt, P. & Fuchs, T. (2024). The Inflexible Self and Lived Time: A Phenomenological Approach to Personality Disorders. In K. Banicki & P. Zachar (Eds.), Conceptualizing Personality Disorder: Perspectives from Philosophy and Psychological Science (forthcoming). Cambridge, UK: Cambridge University Press;
- Teghil, A. & Wittmann, M. (2025). How the Body and Brain process Time. Neuroscience & Biobehavioral Reviews, 179, 106416;
- Thompson, E. (2007). Mind in Life: Biology, phenomenology, and the sciences of mind. Harvard University Press; Tuomela, R. (2007). The Philosophy of Sociality. Oxford University Press;
- Varela, F. J. (1996). Neurophenomenology: A Methodological Remedy for the Hard Problem. Journal of Consciousness Studies, 3(4), 330–349;
- Varela, F. J. (1999). The Specious Present: A Neurophenomenology of Time Consciousness. In J. Petitot et al. (Eds.), Naturalizing Phenomenology (pp. 266–314). Stanford University Press;

- Varela, F. J., Thompson, E. & Rosch, E. (2017). The Embodied Mind: Cognitive Science and Human Experience. MIT Press (Original work published 1991);
- Wittmann, M. & Droit-Volet, S. (2024). Subjective Time in Ordinary and Non-ordinary States of Consciousness: How Interoceptive Feelings Inform Us About the Passage of Time. Current Topics in Behavioral Neurosciences, forthcoming;
- Zahavi, D. (2007). Subjectivity and Selfhood: Investigating the First-person Perspective. MIT Press; Zahavi, D. (2014). Self and Other: Exploring Subjectivity, Empathy, and Shame. Oxford University Press; Zahavi, D. (2025). Being We: Phenomenological Contributions to Social Ontology. Oxford University Press;

Nesim Aslantatar - Motivated Remembering: Religious Experience between Encounter and Retrieved Memory

Purdue University

This paper argues that the transition from a religious encounter to its retrieved memory is not a transfer of content but a transformation of it. While discussions in the epistemology of religious experience have traditionally focused on the epistemic status of religious experiences as they are had, far less attention has been given to what happens between having a religious experience and remembering one. I argue that this transition involves movement between two ontologically and epistemically distinct states: the immediate experiencing of an event (t1) and the retrospective narrative created through memory (t2). Because of this distinction, the reliability appropriate to t1 is fundamentally different from the reliability appropriate to t2.

The argument proceeds in two stages. First, I argue that religious experiences frequently occur under conditions that compromise encoding from the outset. Drawing on works on cognitive resource depletion in religious interactions, I argue that the executive resources required for monitoring, interpretation, and detail-oriented encoding are often exhausted by the encounter itself. Whether through emotional regulation failure or interpretive failure generated by causal opacity, the subject is left with fragmented information and an attributional gap regarding the structure of the experience. Because the encounter exceeds ordinary encoding capacities, what remains available for later retrieval is often not a structured record but an emotional residue and a sense of significance without clear content. Second, I argue that this attributional gap is subsequently filled at t2 through motivated reasoning. Remembering a religious experience is not a passive act of retrieval but a reconstructive process guided by the drive for coherence. The believer does not simply ask what happened but reconstructs what the experience should have been in light of prior beliefs, religious identity, and available narrative templates for religious experiences. This reconstructive process extends beyond individual cognition through vicarious memory, whereby externally supplied interpretations become phenomenally indistinguishable from first-person recollection. Consequently, the memory retrieved at t2 bears the structure imposed by the encoding environment rather than the structure of the original encounter.

This framework is developed in direct tension with Munro's (2024) Memory/Reflection Account. Munro argues that religious beliefs are typically formed not at t1 but during later reflection on episodic memory and treats this process as generally truth-conducive. Against this, I argue that the very factors structuring the transition from t1 to t2 constitute exactly the kind of countervailing evidence that defeats a default assumption of reliability. The problem is not simply that reflection may be biased; it is that the material available for reflection is already filtered through a motivated encoding environment. The originality of the paper lies in shifting the discussion from the reliability of religious experience to the reliability of remembered religious experience. Rather than treating memory as a transparent bridge between encounter and belief, I argue that remembering religious experience is itself a narrative achievement. The subject at t2 is not retrieving the encounter but authoring it. Religious epistemology has therefore been asking the right question about the wrong object.

References

- Alston, W. P. (2004). Religious experience justifies religious belief. In M. L. Peterson & R. J. Vanarragon (Eds.), *Contemporary debates in philosophy of religion* (pp. 135–145). Blackwell.
- Alvarez, J.A., & Emory, E. (2006). Executive function and the frontal lobes: A meta-analytic review. *Neuropsychology Review*, 16, 17-42.
- Aristotle. (1925). *Nicomachean Ethics*, II.6, 1106b28, (trans. W. D. Ross). Oxford: Oxford University Press, 1925.
- Bruner, E. (1986). Experience and its expressions. In *The anthropology of experience,* edited by V. Tumer and E. Bruner. (pp. 3-30). Urbana: University of Illinois Press.
- Conway, M. A. (1996). Autobiographical knowledge and autobiographical memories. In D. C. Rubin (Eds.), *Remembering our past: Studies in autobiographical memory.* (pp. 67-93). New York: Cambridge University Press.
- Conway, M. A., Singer, J. A., & Tagini, A. (2004). The self and autobiographical memory: Correspondence and coherence.
- D'Esposito, M., Detre, J.A., Alsop, D.C., Shin, R.K., Atlas, S., & Grossman, M. (1995). The neural basis of the central executive system of working memory. *Nature, 378*, 279-281.

- Engle, R.W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*(1), 19-23.
- Epley, N., & Gilovich, T. (2016). The mechanics of motivated reasoning. *Journal of Economic Perspectives*, *30*(3), 133-140. https://doi.org/10.1257/jep.30.3.133
- Garavan, H., Ross, T.J., & Stein, E.A. (1999). Right hemispheric dominance of inhibitory control: An event-related functional MRI study. *Proceedings of the National Academy of Sciences*, *96*, 8301-8306.
- Miayke, A., Friedman, N.P., Emerson, M.J., Witzki, A.H., Howerter, A., & Wager, T.D. (2000). The unity and diversity of executive functions and their contributions to complex ‘‘frontal lobe’’ tasks: A latent variable analysis. *Cognitive Psychology*, 41(1), 49-100.
- Mills, J., & Mintz, P. M. (1972). Effect of unexplained arousal on affiliation. *Journal of Personality and Social Psychology, 24*(1), 11–13.
- Munro, D. (2024). Remembering religious experience: Reconstruction, reflection, and reliability. *Philosophy and the Mind Sciences*, *5*. https://doi.org/10.33735/phimisci.2024.10205
- Plantinga, A. (2000). *Warranted Christian belief*. Oxford University Press.
- Schjoedt, U., Sørensen, J., Nielbo, K. L., Xygalatas, D., Mitkidis, P., & Bulbulia, J. (2013). Cognitive resource depletion in religious interactions. *Religion, Brain and Behavior*, *3*(1), 39–55. https://doi.org/10.1080/2153599X.2012.736714
- Schjoedt, U., Stødkilde-Jørgensen, H., Geertz, A., & Roepstorff, A. (2011). The power of charisma: Perceived charisma inhibits the attentional and executive systems of believers in intercessory prayer. *Social Cognitive and Affective Neuroscience, 4*(2), 199-207.
- Schachter, S. (1964). The interaction of cognitive and physiological determinants of emotional state. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 1). New York: Academic Press.
- van Mulukom, V. (2017). Remembering religious rituals: Autobiographical memories of high arousal religious rituals considered from a narrative processing perspective. *Religion, Brain & Behavior*, *7*(3), 191–205. https://doi.org/10.1080/2153599X.2016.1232304
- van Mulukom, V., & Lang, M. (2021). Religious experiences are interpreted through priors from cultural frameworks supported by imaginative capacity rather than special

cognition. *Journal for the Cognitive Science of Religion*, *7*(1), 39–53. https://doi.org/10.1558/jcsr.19803

- Werning, M. (2020). Predicting the Past from Minimal traces: Episodic memory and its distinction from Imagination and Preservation. *Review of Philosophy and Psychology*, *11*(2), 301–333. https://doi. org/10.1007/s13164-020-00471-z.
- Xygalatas, D., Schjoedt, U., Bulbulia, J., Konvalinka, I., Jegindø, E.-M., Reddish, P., Geertz, A. W., & Roepstoff, A. (2013). Autobiographical memory in a fire-walking ritual. *Journal of Cognition and Culture*, *13*(1-2), 1-16. https://doi.org/10.1163/15685373-12342081
- Yamane, D. (2000). Narrative and religious experience. *Sociology of Religion*, *61*(2), 171-189. https://doi.org/10.2307/3712284

Rémi Tison

UQAM

For a few decades now philosophers have been trying to determine what it takes for individuals to count as performing a joint action. Typically, this means specifying conditions that neatly separate joint actions from non joint actions. Most theories of joint action currently on the market contain cognitive conditions, concerning the mental states that the participant in a social interaction must have for it to count as a joint action (eg. Searle, 1990, 2010; Bratman, 1992, 1993, 2014). However, these conditions are often considered to be too cognitively demanding for young children and nonhuman animals (see eg. Tollefsen, 2005; Butterfill, 2012; Downes et al., 2025).

In response, many researchers have advocated for minimalist theories of joint action assigning less cognitively demanding conditions for joint action (Vesper et al., 2010; Tollefsen et Dale, 2012; Pacherie, 2013; Butterfill, 2017; Salomones Sehr, 2024). These minimalist theories generally fare better from evolutionary and developmental standpoints: they provide a better explanation of how we can gradually get from mere social interaction to proper joint actions without implausible phylogenetic or ontogenetic jumps.

However, simply relaxing the conditions on joint action so that more social interactions potentially involving more cognitively limited agents, such as young children and nonhuman animals, count as joint actions faces its own problems. Indeed, as explained above, there are similarities between “standard” human joint actions, and the more minimalist joint actions of animals and young children. But there are also significant differences which are elided by minimalist theories that limit themselves to expanding the range of interactions counting as joint actions. These minimalist theories therefore often fail to do justice to some characteristic features of specifically human joint actions. What we need is a theory that can do both: a minimalist theory that recognizes the continuity between paradigmatic human joint actions and lower-level joint actions, while also accounting for the specific features of more sophisticated human joint actions.

The dynamical theory developed in this paper aims to fit this bill. Based on dynamical systems theory (Bernstein, 1967; Turvey, 1990; Kelso, 1995; Latash et al., 2007) and its recent application to the study of social interaction (De Jaegher et Di Paolo, 2007; Riley et al., 2011; Ramenzoni et al., 2012; Dale et al., 2014; Cuffari et al., 2015; Di Paolo et al., 2018; Gallagher, 2020), this theory roughly suggests that agents perform a joint action when they form an interpersonal dynamical system organized so as to accomplish a goal. But the story does not end here. These dynamical systems can be more or less robust (i. e. resilient to perturbations), more or less flexible (i. e. adaptable to differing circumstances) and

vary in scale (i. e. the number of participants and the duration of the joint action). The dynamical systems theory is therefore perfectly poised to formulate a gradualist view of joint action, ranging from relatively brittle, rigid and short-term social interactions to full-blown human collaborative activities. To account for the higher end of human joint actions, I recruit recent normativist approaches to human social cognition, in particular the mindshaping framework (Zawidzki, 2013; Zawidzki, 2018; Zawidzki and Tison, 2025), to argue that human social groups have evolved various types of social normative mechanisms to increase the robustness, flexibility and scale of the interpersonal dynamical systems that they can form. The dynamical theory therefore culminates in a normativist theory of joint action for a specific class of human joint actions.

Mitchell Herschbach - Are There Objective Measures of "Theory of Mind" Abilities? Implications for the Study of Autism

California State University, Northridge

One prominent theory of autism explains its socio-cognitive features in terms of difficulties understanding people's mental states, i.e., a "theory of mind" (ToM) deficit. Recently this research programme has been accused of being bad science, if not outright pseudoscience (LaCroix, 2025). LaCroix thinks autistic social challenges are better explained by the "double empathy problem" (Milton, 2012): allistic, neurotypical people have trouble understanding autistic minds just as much as autistic people have trouble understanding the minds of neurotypical people. A number of prominent members of this research program (Baron-Cohen et al., 2025) replied to LaCroix, highlighting the "staggering number of relevant meta-analyses" indicating autistic people do perform worse on a range of ToM tasks. This reply does not, however, address LaCroix's thesis that the behavioral tasks used to test ToM are not objective measures of ToM abilities, because they are biased toward neurotypical responses. I will critically evaluate LaCroix's criticism by analyzing numerous tasks standardly used to assess ToM (e.g., Wellman & Liu, 2004). Some features by which ToM tasks can differ include: the type(s) of mental state that must be understood; the participant's relation to the target agent (e.g., passive observation vs. direct social interaction); and whether the tasks directly or indirectly assess ToM (Low & Perner, 2012). LaCroix's concern that ToM tasks lack one objectively correct answer seems most applicable to tasks that involve directly predicting an agent's behavior. What researchers deem the correct behavioral prediction is what the target agent is *most likely* to do, but other options are always open possibilities depending on what one assumes about the target agent's mental states. But this need not be a sign of obvious allistic bias. There are other ToM tasks, however, that do have objectively correct answers: tests of factive mental states like knowledge, tested either directly or indirectly; as well as some tests of belief understanding. In conclusion, I find LaCroix's criticism to be a weak way of challenging the idea of a ToM deficit in autism, and defending the alternative "double empathy problem" hypothesis. This does not, however, call

into question investigating the double empathy problem as a source of social challenges for autistic people.

References

- Baron-Cohen, S., Radecki, M. A., Allison, C., Tsompanidis, A., Greenberg, D. M., Warrier, V., … Frith, U. (2025). Do Autistic People Have Degrees of Disability in Theory of Mind? The Importance of Meta Analytic Convergence. *Psychological Inquiry*, *36*(4), 262–267.
- LaCroix, T. (2025). Autism and the Pseudoscience of Mind. *Psychological Inquiry*, *36*(4), 229–261. Low, J., & Perner, J. (2012). Implicit and explicit theory of mind: state of the art. British Journal of Developmental Psychology, 30(Pt 1), 1–13.
- Milton, D. E. M. (2012). On the ontological status of autism: the 'double empathy problem.' *Disability & Society*, *27*(6), 883–887.
- Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development Perspectives*, *75*(2), 523–541.

Huang Lei - Where Does the Competence Reside? Interaction-Dependent Functional Emergence in a Two-Model Language System

Japan Advanced Institute of Science and Technology

This paper asks whether problem-solving competence, when it arises only through interaction between two artificial agents, should be attributed to an individual system or to the organized interaction itself. I argue that when a correct solution is systematically absent from isolated performance, exceeds a computation-matched noninteractive control, and emerges through reciprocal transformation of incomplete representations, the resulting competence depends constitutively on the interactive process itself and is properly attributed to the coupled system rather than to either agent alone. I distinguish this process-level dependence from testimonial epistemic dependence on another knower, since no individual agent here possesses the relevant competence to transmit. The result would support a functionalist, distributed-cognition account of problem solving without invoking consciousness, group mentality, or strong ontological emergence.

The claim is tested using two local open-weight language models (Qwen and Gemma family) on 200 Japanese wordplay riddles (nazonazo) requiring homophony, kanji readings, segmentation, semantic ambiguity, and lateral reinterpretation. This domain is chosen because correct solutions typically require generating, revising, comparing, and coordinating competing semantic representations rather than retrieving or computing a single determinate answer, making it a plausible proxy for representationally mediated collaborative sense-making rather than general reasoning.

Each model first attempts every riddle independently across a preregistered number of trials. A riddle enters the primary analysis only when both models fail on every baseline trial. For these stable joint-failure items, the models then engage in a fixed-length peer dialogue with no external tools, retrieval, hints, correctness feedback, or parameter updates. Dialogue performance is compared against an independent-reflection control, matched for model calls, reasoning turns, and token budget, but without access to the other model's outputs. This isolates

the causal contribution of interaction from gains attributable to repeated sampling, additional inference time, or greater computation alone. The thesis would be disconfirmed if dialogue advantage over this control proves statistically indistinguishable from zero, or if incorrect-consensus rates dominate any gains.

Primary outcomes include system-level correction rate, dialogue advantage over independent reflection, interaction-first solution rate, and incorrect-consensus rate. Trajectories are coded for hypothesis generation, assumption challenges, semantic reframing, and the first appearance of correct solution mechanisms, distinguishing mere transmission of a complete answer from genuine reciprocal integration of initially incomplete contributions.

The central contribution is philosophical: an operationalized, falsifiable criterion for functional emergence in multi-agent systems, and an argument that process-level dependence on interaction licenses attributing competence to the system rather than its parts. This bears most directly on distributed cognition, extending the thesis that cognitive processes can constitutively span multiple agents from human dyads to artificial ones, while explicitly bracketing stronger claims about machine consciousness or collective mentality.

Alexandra Ioana Din - Code-switching as narrative resource in bilingual Romanian-German children

University of Bucharest, Faculty of Philosophy

Narratives are not only a tool for building imaginative worlds, as in literature, arts, or video games, but also the social world-building kit. Researches of past decades have paid an increasing attention to the way in which narratives outline what we regard as right, wrong, or expressible. I assume that narratives are socially situated. My focus is on the impact the code switching in bilingual children has on daily life narratives. I take a closer look at the bilingual children in past and present Romanian-German families from western Romania. I claim that code switching is a narrative resource, in the sense that bilingual natives might switch languages to better express cultural identities, thoughts, or emotions. Because narratives are socially situated, children actually learn to explain actions via the interaction with their caregivers. In bilingual environments, the narrative practice involve two or more languages, so that switching between them show a response to social expectations. As their skill of storytelling further refines, they increasingly get better at using both languages in a strategic way. Using the *Narrative Practice Hypothesis* (NPH) as a framework for explaining code-switching, I argue that this is a part of narrative practices and interaction. Children switch to accommodate listeners, drawing upon some culturally specific meanings. I further consider language switching as an affordance that would enable children to adapt to audiences and construct broader, more complex narrative in culturally diverse meaning-making forms. I analyze direct data from bilingual children in today German

Romanian communities, I look at some archive material of the bilingual past of these communities, and compare them with the confessions on bilingualism from various German-Romanian writers, such as the Nobel Prize winner Herta Müller.

(Minimal) References

Primary

- Müller, Herta (2005). *The King Bows and Kills*. In Romanian. Polirom.

Secondary

Books

- Ricoeur, Paul (1992). *Oneself as Another*. Translated from French by Kathleen Blamey. The University of Chicago Press.
- Hutto, D. Daniel (2008). *Folk Psychological Narratives. The Sociocultural Basis of Understanding Reasons*. MIT.
- Verhoven, Ludo (2004). *Bilingualism and Narrative Construction*. Psychology Press.

Chapters in books

- Francis, W. S. (2005). Bilingual Semantic and Conceptual Representation. In J. F. Kroll, A.M.B. de Groot (eds.), *Handbook of Bilingualism. Psycholinguistic Approaches* (pp. 251-267), Oxford University Press.
- La Heij, W. (2005). Selection Processes in Monolingual and Bilingual Lexical Access. In J. F. Kroll, A.M.B. de Groot (eds.), *Handbook of Bilingualism. Psycholinguistic Approaches* (pp. 289- 307), Oxford University Press.
- Meuter, R. F. I. (2005). Language Selection in Bilinguals: Mechanisms and Processes. In J. F. Kroll, A.M.B. de Groot (eds.), *Handbook of Bilingualism. Psycholinguistic Approaches* (pp. 349- 370), Oxford University Press.

Articles

- Capin et al. (2023). Evaluating the Efficacy of a Narrative Language Intervention for Bilingual Students. *American Journal of Speech-Language Pathology*, 32 (6): 2999-3020. https://doi.org/10.1044/2023_AJSLP-21-0018
- Greene et al. (2014). Bilingual children's use of lexical strategies under narrative monologue and dialogue conditions. *Language and Sociocultural Theory*, 1 (2), https://doi.org/10.1558/lst.v1i2.1 Irgin, Pelin (2025). Storytelling in a Bilingual Classroom Through the Lens of Epistemic Diversity and Translanguaging. *International Journal of Applied Linguistics*, 36 (1): 11-21. https://doi.org/10.1111/ijal.12714
- Nuñez et al. (2020). Sustaining Bilingual–Biliterate Identities: Latinx Preservice Teachers' Narrative Representations of Bilingualism and Biliteracy Across Time and Space. *Journal of Teacher Education*, 72 (4). https://doi.org/10.1177/0022487120954

- Phoenix, A., Orellana, M. F (2021). Adult narratives of childhood language brokering: Learning what it means to be bilingual. *Children & Society*, 36 (3): 386-399, https://doi.org/10.1111/chso.12462
- Polanco, Marcela (2010). Rethinking Narrative Therapy: An Examination of Bilingualism and Magical Realism. *Journal of Systemic Therapies*, 29 (2). https://doi.org/10.1521/jsyt.2010.29.2.1

Brenda Bayles - The Geometry of Reorganization Across Cognitive and Natural Systems

Wichita State University

Understanding how cognitive and natural systems reorganize during periods of instability is a shared challenge across philosophy of mind, cognitive science, psychology, neuroscience, biology, and artificial intelligence. This paper develops a mathematical geometric framework that describes how organized systems lose and regain coherence when competing influences shift in strength. The framework begins with the observation that many systems stabilize around patterns of unequal influence. When interacting components differ in strength, they generate a definable center and a surrounding curvature field that constrains behavior. As the balance of influence changes, the center moves. When interacting influences reach equality, curvature disappears and the system enters a structural zero, a transient region in which no stabilizing center exists and the system becomes sensitive to new inputs. Reorganization begins only when a fresh asymmetry appears and a new center forms. This structural pattern is not tied to any single domain; instead, it provides a general way to describe how systems change. In philosophy of mind, the framework clarifies how mental organization can dissolve and reform without invoking hidden substances or internal agents. Some of these transitions resemble patterns noted in work by Dennett, particularly his discussions of shifts in interpretive stability and the formation of new centers of coherence. In cognitive science, the framework offers a way to model how distributed processes reorganize. Competing representations, attentional priorities, and predictive signals generate shifting asymmetries that define the system's moment to moment center. Collapse occurs when these influences equalize, producing a brief interval of instability before new organization emerges. In psychology, the framework helps explain discrete transitions in belief, emotion, and behavior. Sudden therapeutic breakthroughs, rapid reframing, and abrupt changes in motivation reflect structural zero dynamics rather than gradual accumulation. In neuroscience, the framework provides a geometric interpretation of neural reorganization. Competition among neural assemblies, oscillatory patterns, and network level influences produces shifting centers of functional coherence. Collapse occurs when assemblies reach equal influence, allowing

new patterns of connectivity to form. In biology, the framework models transitions such as cellular reprogramming, developmental bifurcations, and pathological identity formation. When regulatory influences equalize, cells enter high plasticity intervals that resemble structural zeros. In artificial intelligence, the framework describes how artificial agents reorganize internal states when competing objectives or representational conflicts reach equality. Taken together, this mathematical geometric approach offers a unified way to interpret reorganization across cognitive and natural systems.

References

- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2014). *Molecular biology of the cell* (6th ed.). Garland Science.
- Binney, J., & Tremaine, S. (2008). *Galactic dynamics* (2nd ed.). Princeton University Press. Dennett, D. C. (1991). *Consciousness explained*. Little, Brown.
- Landau, L. D., & Lifshitz, E. M. (1976). *Mechanics* (3rd ed.). Pergamon Press.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.
- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*(1), 42. https://doi.org/10.1186/1471-2202-5-42
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Heng Ying - Revisiting the Humean model of moral psychology

University of Hong Kong

In this paper, I challenge the validity of the Humean model of moral psychology by rejecting the interpretive paradigm psychologists and neuroscientists adopt to explain behavioural experiments and support the Humean model. Opposite to the Kantian model that emphasises logical, principled moral reasoning as the primary mechanism underlying moral judgment, the Humean model endorses emotion as the chief mechanism.

The individualistic interpretive paradigm is seen in Jonathan Haidt's studies on moral dumbfounding. Based on his participants' quick, automatic rejections to scenarios, such as siblings having protected sex, he identifies the intuitive feelings triggered by the scenarios as the determinant of their moral judgments. Similarly, Antonio Damasio notices from his patients two salient symptoms— flat emotional feelings and lack of capacity for judgment. Therefore, he draws the conclusion that emotion is the foundation of decision making.

This individualistic interpretive paradigm reduces individuals to the mere function of a single, activated psychological process. Therefore, it does not reflect the reality of human psychology. First, the studies that test the impact of moral commitments on individuals' moral judgments show that the feelings aroused by specific scenarios do not fully determine individuals' judgments. Instead, they consider between their feelings and their background commitments and determine what judgments they should give. This shows that apart from the triggered feelings, individuals hold background values, and they are also likely to be influenced by other real-time inputs. When they offer a judgment based on their immediate feelings, they are not simply driven by the feelings, but choose to act on those feelings over other co-occurrent processes.

Hence, in Haidt's studies, instead of fitting participants' behaviours—such as making quick judgments, engaging in post-rationalisation to justify their judgments, and giving up when they couldn't find another reason to refute their reactions—into the evidence that supports intuitive processes as the cause of moral judgments, individuals make decisions regarding what

judgments they should give based on the information they are aware of. They choose to act on their immediate feelings when they are strongly aware of them, and they choose to act on these feelings or change their minds when other processes—such as other moral considerations—become more salient.

Similarly, in Damasio's studies, lack of capacities for emotion does not mute other brain functions. Instead, this diminished emotional capacity puts individuals into a position where they look at the world in a new light and acquire a new range of dispositions. Their lack of judgment is not an issue of lack of emotions. They make decisions regarding how they should choose and act based on the new way of engaging with the world, and the decisions they make are sensible to them but not to other people with intact emotional capacities.

Therefore, an accurate model of individual moral psychology should put individuals at the centre of moral judgment, instead of emotional processes. Moral judgment essentially reflect how individuals want to engage with a particular situation based on what they are aware of within themselves.

Carolyn Wasula - Relational Cosmopsychism: A Formal Account of Mind as Structured Participation

University of Central Florida

This paper develops *relational cosmopsychism*, a formal philosophical theory that reconceives consciousness as an intrinsically relational, world-level field in which individual minds emerge as structured participatory loci. Rather than treating consciousness as a property of isolated subjects or as a mere aggregation of microphysical units, relational cosmopsychism proposes that consciousness is fundamentally global, patterned, and relationally articulated. Individual subjects arise not as ontologically separate entities but as dynamically stabilized regions of a shared conscious field whose internal organization enables perspectival experience, self-modeling, and agency.

The theory advances three core claims. First, consciousness is best understood as a **cosmic-level field of intrinsic presence**, whose internal differentiation generates the conditions for localized experiential standpoints. Second, **individual minds are relational formations**, constituted by patterned constraints, boundary-conditions, and feedback loops within the field rather than by metaphysically independent substances. Third, **mental causation is realized through relational modulation**, where local configurations of the field alter global dynamics and vice versa, producing a bidirectional structure that dissolves traditional interaction problems.

This framework offers a novel solution to the combination and decombination problems in panpsychism by rejecting the assumption that subjects must combine or split. Instead, subjects are emergent relational stabilities within a single conscious field. It also provides a unified account of phenomenology, cognitive architecture, and social embeddedness: subjective experience arises from the internal organization of a relational node; cognition reflects the node's dynamic modeling of its environment; and interpersonal meaning-making emerges from the coupling of nodes within the same field.

Relational cosmopsychism further integrates insights from enactivism, extended mind theory, and ecological psychology by situating cognitive processes within broader relational structures

rather than within bounded individuals. This yields a multi-scale model of mind that is compatible with contemporary work in cognitive science, affective neuroscience, and social cognition. The theory also offers a metaphysical grounding for relational approaches in psychology, anthropology, and phenomenology, showing how relationality is not merely methodological but ontological.

By formalizing the relational architecture of consciousness, this paper provides a coherent metaphysical framework capable of explaining subjectivity, agency, and intersubjective meaning without resorting to reductionism or substance dualism. Relational cosmopsychism thus contributes a new, integrative model to the sciences of mind. This model reframes consciousness as a structured, dynamic, and relational field in which minds are emergent patterns of participation.

Caitlin Mace - Is Vehicle Indeterminacy Radical?

University of Pittsburgh

The vehicle indeterminacy problem in neuroscience is the problem of determining which neural structures or patterns of neural activity in the brain are representations. At first blush, the problem is epistemic. We might think that uncertainty about which patterns are vehicles will be resolved as neuroscience proceeds and more is known about the brain. I argue in this paper that the problem is more foundational than a lack of knowledge about or access to the brain. I illustrate this point by analogy to artificial neural networks (ANNs), which is a system to which researchers have much experimental access and in which much is known about its structure and activity.

Representational interpretability is a branch of mechanistic interpretability that aims to find symbols or patterns of neural activity in artificial neural networks that represent properties of the input, called features. Researchers decompose the system into its "natural" components, which are individual neurons, attention heads, layers, groups of neurons (Sharkey et al. 2025), but such decompositions have failed to identify vehicles for features. Neurons and their activations values are not taken to be vehicles for features as their participation in many circuit processes (i.e., the fact that they are polysemantic and respond to many kinds of inputs) makes them problematic units of analysis as representations. As such, there are two main kinds of vehicle candidates: patterns of activity (ranging from vector values or activations at a particular layer to across the whole system) and various high-dimensional latent units.

Why have researchers not settled on a vehicle for features? One challenge is content indeterminacy. First, superposition is the representation of more features than there are dimensions (Elhage et al. 2022). Superposition challenges decomposition strategies because, when identifying features, one must consider their function or role, and each feature likely has many roles. It is always a defeater that one does not *really* have a feature since they have not *really* established the role. Second, we have found that artificial neural networks complete tasks in a different way than biological systems, so it is not obvious what content to ascribe in many cases. The problem is that treating content in ANN processing or its outputs in terms that we understand does not mean that we have the content right. As such, determining

the content of ANNs (or the features) is itself a task for mechanistic interpretability researchers.

However, even with determinate content, vehicle indeterminacy will remain. Indeed, I maintain that vehicle indeterminacy is radical because of the complexity of the system. Complexity makes statistical decoding trivial. Many various and disparate neural activities are "interpretable" as bearing content because of genuine multiple realizability and redundancy, correlations between the target content and many other aspects of the system, and correlations between the vehicle and many other structures and processes in the system. Interventions provide better support, but such interventions are made without isolating the candidate vehicle from other components and processes, thereby straining inferences that the candidate is really a vehicle.

References

- Azhar, Feraz. "Polytopes as Vehicles of Informational Content in Feedforward Neural Networks." *Philosophical Psychology* 29 (5): 697–716.
- https://doi.org/10.1080/09515089.2016.1142070.
- Cappelen, Herman, and Josh Dever. 2021. *Making AI Intelligible*. Oxford University Press.
- Chalmers, David. 2025. "Propositional Interpretability in Artificial Intelligence." arXiv:2501.15740 [cs.AI]
- Churchland, Paul. "Conceptual Similarity across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered." *The Journal of Philosophy* 95 (1): 5–32. https://doi.org/10.2307/2564566.
- Coelho Mollo, Dimitri, and Raphaël Millière. 2026. "The Vector Grounding Problem." *Philosophy and the Mind Sciences* 7: 1–28. https://doi.org/10.33735/phimisci.2026.12307
- Elhage, Nelson, Tristan Hume , Catherine Olsson, Nicholas Schiefer , Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg , and Christopher Olah. 2022. "Toy Models of Superposition." *Transformer Circuits Thread*. Retrieved from: https://transformer-circuits.pub/2022/toy_model/index.html

- Eliasmith, Chris. 2003. “Moving beyond Metaphors: Understanding the Mind for What It Is.” *The Journal of Philosophy* 100 (10): 493–520. https://doi.org/10.5840/jphil2003100102.
- Friedman, Daniel and Duede, Eamon. 2026. “Interpretability (Propositional and Mechanistic) Needs Behavior.” [Preprint] https://philsci-archive.pitt.edu/id/eprint/29538
- Geiger, Atticus, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. 2025. “Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability.” *Journal of Machine Learning Research* 26: 1–64.
- Harding, Jacqueline. “Operationalising Representation in Natural Language Processing.” *The British Journal for the Philosophy of Science*, advance online publication, November 10, 2023. https://doi.org/10.1086/728685.
- Kästner, Lena, and Crook, B. 2024. “Explaining AI through mechanistic interpretability.” *European Journal of Philosophy of Science* 14 (52). https://doi
- org.pitt.idm.oclc.org/10.1007/s13194-024-00614-4
- Mallory, Fintan. 2026. “Formats of representation in large language models.” *Philosophy and the Mind Sciences* 7: 1–18. https://doi.org/10.33735/phimisci.2026.12091
- Morgan, Alex. 2020. “Against neuroclassicism: On the perils of armchair neuroscience.” *Mind & Language* 37: 329–355.
- Nieder, Andreas. 2016. “The neuronal code for number.” *Nature Reviews Neuroscience* 17: 366– 382. https://doi.org/10.1038/nrn.2016.40
- Olah, Chris, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. “Zoom In: An Introduction to Circuits.” *Distill* 10.23915/distill.00024.001 Retrieved from: <https://distill.pub/2020/circuits/zoom-in>
- Rumelhart, David, James McClelland, and PDP Research Group, editors. 1986. *Parallel Distributed Processing. Volume 1: Foundations.* MIT Press.
- Somvanshi, Shriyank, Md Monzurul Islam, Amir Rafe, et al. 2026. “Bridging the Black Box: A Survey on Mechanistic Interpretability in AI.” *ACM Computing Surveys* 58 (8): 210. https://doi.org/10.1145/3787104.

- Sripada, Chandra. (msc) “Large Language Models as Symbolic Systems.” Retrieved from: < https://sites.lsa.umich.edu/sripada/philosophy/>
- Wang, Kevin, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. “Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small.” arXiv:2211.00593v1 https://doi.org/10.48550/arXiv.2211.00593

Tiago Sestari - Interactive Fear: Rethinking the Paradox of Fictional Emotion Through Games

FGV-EAESP

Kendall Walton's "Fearing Fictions" (1978) holds that responses such as Charles's dread of the on-screen slime are not fear proper but *quasi-emotions* internal to a game of make believe. Central to this argument is the premise that physical interaction is possible only with what actually exists, so that a subject cannot act upon a fictional threat in the way that real fear would dispose one to act upon a real one; Charles cannot "dam up" the slime, and this incapacity for real-world action is offered as a reason to deny that his state is real fear. Horror video games place pressure on this premise. Unlike the film viewer Walton describes, the player *can* act upon the threat within the game: evasion, concealment, and resource management under threat are causally efficacious within a real rule-system that tracks success and failure, even though the threat's fictional content remains unreal. This paper takes up the resulting question: does player fear in horror games constitute evidence for cognitive theories of emotion, on which emotion is constituted by evaluative judgment (Robert Solomon, "Emotions and Choice," 1973; *The Passions*, 1976/2003), or for embodied and perceptual theories, on which emotion is constituted by the registration of bodily changes representing danger (Jesse Prinz, *Gut Reactions*, 2004; "Précis of *Gut Reactions*," 2008)?

It will be argued that the interactive structure of games separates two components Walton's account treats as bundled: the subject's standing cognitive access to the fictionality of the threat, and the bodily mobilization produced by a system that imposes genuine, non-fictional consequences for failure. Jesper Juul's characterization, in *Half Real* (2005), of the video game as an object simultaneously constituted by real rules and a fictional world offers a formal vocabulary for this separation, suggesting that the intentional object of player fear is the rule-governed system as mediated by, rather than exhausted by, its fictional content. Bernard Perron's analysis of "fear-motivated action" arising from the player's embodied positioning within the game-world (*The World of Scary Video Games*, 2018) lends further support to treating the resulting state as functionally continuous with ordinary embodied fear, rather than as a Waltonian quasi-emotion.

The conclusion drawn is not that one theory of emotion is simply vindicated over the other, but that the interactive case reveals a limitation in how the cognitive/embodied distinction has been drawn: once the fictional object of an emotion is something the subject can act *into*, rather than merely represent, the standard taxonomy of emotion theories may require revision. This bears on the paradox of fiction more generally, suggesting that interactive fiction warrants treatment as a distinct category rather than a marginal case of spectatorial fiction.

References

- Juul, J. (2005). *Half-Real: Video Games Between Real Rules and Fictional Worlds*. MIT Press.
- Perron, B. (2018). *The World of Scary Video Games: A Study in Videoludic Horror*. Bloomsbury Academic.
- Prinz, J. J. (2004). *Gut Reactions: A Perceptual Theory of Emotion*. Oxford University Press.
- Prinz, J. J. (2008). Précis of *Gut Reactions*. *Philosophy and Phenomenological Research*, 76(3), 707–711.
- Solomon, R. C. (1973). Emotions and choice. *The Review of Metaphysics*, 27(1), 20–41.
- Solomon, R. C. (1976). *The Passions: Emotions and the Meaning of Life*. Doubleday. (Reissued and expanded as *Not Passion's Slave: Emotions and Choice*, Oxford University Press, 2003.)
- Walton, K. L. (1978). Fearing fictions. *The Journal of Philosophy*, 75(1), 5–27.

Marco Degano - Two Integration Problems for Mesh: the Vertical and the Horizontal Integration Problems.

School of Liberal Arts, University of Wollongong

According to the Mesh framework (Christensen et al. 2016; Christensen & Sutton 2019; Sutton & Bicknell 2020, 2022), any explanation of skilled action targets complex behaviors that result from multiple factors. On the Mesh framework then, the complexity of skilled action calls for a pluralistic and properly integrated model.

The aim of my talk is to ask the following questions. Since Mesh is a pluralist model, are there any *internal inconsistencies* among the factors involved in the explanation of skilled actions? Second, we know the target phenomenon of the Mesh model is skilled action. But is the target phenomenon a mechanism inside the agent, or is it distributed across multiple

dimensions both *inside* agents and *across* agents and their environments? If it is the latter case, how is the *integration* of internal and external factors accounted for by Mesh? I will argue that Mesh faces two integration problems based on the more general *problem of inconsistent attributes* (PIA). PIA is a version the *problem of inconsistent models* (PIM) in philosophy of science. PIM is the problem of different models providing incompatible or inconsistent explanations of the same phenomenon (Massimi 2022, p. 53). For instance, if two scientific models represent the same phenomenon as representational and non-representational, then there is a significant tension between the two models since their respective represention of the same phenomenon have mutually inconsistent properties (Kirchhoff 2025). While PIM is about scientific models, PIA is about the specific attributes of elements that Mesh assumes can be unproblematically integrated into a model.

The first integration problem for Mesh is the vertical integration problem (VIP). VIP focuses on integrating conscious cognitive processes and non-conscious, automatic and sup-personal mechanisms in the control of skilled action. I will show that Mesh's model of integration of cognitive and automatic processes (Christensen et al. 2016; Christensen & Bicknell 2022) runs into the VIP for it lacks an explanation of the mechanisms involved in the integration.

The second integration problem for Mesh is the horizontal integration problem (HIP). HIP is about the constitutive integration of format-incompatible components to explain the embodied, situated and distributed aspects into an explanation of skilled action. I will show that Sutton & Bicknell's (2020, 2022) models of these aspect can be in principle integrated: bodily and environmental aspects are modelled as discrete and symbolic mental/neural representational states, or *engrams*, while artifacts and other external elements are modelled as distributed and non-symbolic, or *exograms* (see Sutton 2010). No doubt that these elements and factors can be viewed as contributing, *causal* factors to explaining skilled activity, but there can is no genuine integration of these aspects into a well-integrated system distributed across the skilled agent system and her environment that doesn't face inconsistencies.

References

- Christensen, W. & Bicknell, K. (2022). Cognitive control, intentions, and problem solving in skill learning. *Synthese*, *200*(460). https://doi.org/10.1007/s11229-022-03920-7 Christensen, W. & Sutton, J. (2019). "Mesh: cognition, body, and environment in skilled action". In M.L. Cappuccio (ed.), *Handbook of Embodied Cognition and Sport Psychology*, Cambridge, MA: MIT Press, 157–164.
- Christensen, W., Sutton, J. & McIlwain, D.J.F. (2016). Cognition in Skilled Action: Meshed Control and the Varieties of Skill Experience. *Mind and Language*, *31*(1), 37–66. http://dx.doi.org/10.1111/mila.12094
- Massimi, M. (2022). *Perspectival realism*. New York: Oxford University Press. Sutton, J. (2010). "Exograms and interdisciplinarity: history, the extended mind, and the civilizing process". In R. Menary (ed.), *The Extended Mind*. Cambridge, MA: The MIT Press, 189–225.
- Sutton, J. & Bicknell, K. (2020). "Embodied experience in the cognitive ecologies of skilled performance". In E. Fridland & C. Pavese (eds), *The Routledge Handbook of Philosophy of Skill and Expertise*, New York: Routledge, 194–206.
- Sutton, J. & Bicknell, K. (2022). "Introduction: The situated intelligence of collaborative skills". In K. Bicknell & J. Sutton (eds), *Collaborative Embodied Performance. Ecologies of Skill*, London: Bloomsbury Publishing, 1–18.

Gerda Mittag - Want to understand computational explanation? - Ask a teacher!

Technical University Berlin

In trying to understand the nature and functioning of cognition, many cognitive scientists rely on computational explanations. They explain cognitive phenomena in terms of computational processes based on formalisms and conceptual resources from computer science and mathematics (Piccinini 2006, Richmond 2024). Computational explanations mathematically specify the input-output function computed by a system to explain the cognitive behavior exhibited (Kaplan 2011) and they explain why the specified computation is appropriate for solving a given cognitive task under given information processing constraints (Rusanen & Lappi 2016). For a number of cognitive phenomena, computational explanations are considered common ground (Rusanen & Lappi 2016). But to this day, fundamental questions about computational explanations remain:

Why do computational explanations work so well in cognitive science (Richmond 2024)? Can computational explanations even be truly explanatory (Kaplan 2011; Brun, Konsman, Polger, 2025)? What about criticism concerning idealization and omission of potentially relevant neurobiological details in computational explanations (Brun, Konsman, Polger 2024)? How do computational explanations relate to other explanatory paradigms such as the mechanistic one (Kaplan 2011)?

I claim that these questions, that have posed serious problems for other philosophical approaches to computational explanation, can be answered by my didactic approach. Didactics is the science of teaching and learning, so my didactic approach is based on conceptualizing neuroscientific research and explanatory practice as a teaching-learning system. Computational explanation, then, is a special kind of didactic tool, appropriate for teaching-learning about a specific phenomenon (cognitive behavior) in a specific context (neuroscientific research), by a specific audience (neuroscience community) to reach specific goals (understanding, as well as control of cognitive phenomena (Chirimuuta 2024)). This kind of explanation works so well in cognitive science because it is didactically appropriate

for the context, learning goals, available methods and subject matter (phenomena investigated). The success of a given computational explanation can be predicted precisely by a list of didactic criteria. Didactic criteria for scientific explanation can also answer to the biggest criticism computational explanations face, that is, the heavy use of abstraction and idealization. The underlying concerns can be dismissed from a didactic perspective. Especially for trying to understand complex systems, simplification (a fundamental didactic practice) is an explanatory virtue, rather than a necessary vice. Didactics gives straightforward criteria for evaluating simplification practices in scientific explanation.

Furthermore, the didactic approach leads to an interesting new perspective on explanatory pluralism: Just like there are different ways to teach-learn about a given topic, there can be different ways to explain a phenomenon under investigation. Different explanations accentuate different aspects of a given educative content (phenomenon) that might be more or less relevant

for a specific learning (research) goal. In conclusion, the didactic approach provides a novel, unifying perspective that picks out the special role computational explanation plays in cognitive science more precisely than other philosophical accounts. Their potential as explanations of cognitive phenomena, as well as their explanatory limitations are best described when characterizing computational explanations as didactic material embedded in a didactic system.

References

- Brun, C., Konsman, J. P., & Polger, T. (2024). How do Computational Models in the Cognitive and Brain Science Explain?. Authorea, Inc.
- Chirimuuta, M. (2024). The Brain Abstracted. Simplification in the History and Philosophy of Neuroscience. Cambridge, Massachusetts & London, England : The MIT Press.
- Janík, T., Slavík, J., Češková, T., Najvar, P., Minaříková, E., & Jáč, M. (2025). On content transformation in the classroom: transdisciplinary perspectives on subject didactic research using didactic case studies. Journal of Curriculum Studies, 1–27.
- Kaplan, D. M. (2011). Explanation and description in computational neuroscience. Synthese, 183(3), 339-339.

- Piccinini, G. (2007). Computing Mechanisms*. Philosophy of Science, 74(4), 501–526.
- Raithel, J., Dollinger, B. And Hörmann, G. (2009). Einführung Pädagogik, Begriffe, Strömungen, Klassiker, Fachrichtungen, Lehrbuch, 3. Auflage. VS Verlag für Sozialwiss. Wiesbaden.
- Richmond, A. (2024). How computation explains. Mind & Language, 40(1), 2–20.
- Rusanen, A.M., & Lappi, O. (2007). The limits of mechanistic explanation in neurocognitive sciences. In: Vosniadou, S., Kayser, D. & Protopapas, A. (Hrsg.) Proceedings of the European Cognitive Science Conference. Francis and Taylor.

Savvas Ioannou - Defining the Physical: Via Negativa and Non-Mental Descriptions

University of Cyprus

Physicalists claim that there is nothing over and above the physical. They also claim that everything is physical or determined by the physical. To explain such claims, they attempt to define the term 'physical'. What counts as 'physical' must be specified to formulate a satisfactory version of physicalism.

One way to define the 'physical' is through the Via Negativa approach (Smith, 1993; Spurrett & Papineau, 1999; Levine, 2001; Montero, 2009; Montero & Papineau, 2005; Worley, 2006). In its simplest form, it holds that 'physical' means 'not fundamentally mental'. "Something is fundamentally mental if it is mental and does not depend on anything more fundamental" (Brown, 2021, p. 2843).

Crook & Gillett (2001) and Wilson (2006) have argued that the Via Negativa approach does not suffice to explain what it means for something to be physical. The Via Negativa approach provides a necessary condition for being physical, but not a sufficient condition. According to the No Fundamental Mentality constraint, understood as a necessary condition for physicalism, if there is something that is both fundamental and mental, then physicalism is false.

On the other hand, Brown (2021) has argued that fundamental mentality is compatible with physicalism. He described a conceivable and possible world that contains fundamental mentality and is compatible with physicalism. In this world, priority monism is true. This is the view according to which the whole cosmos is fundamental, and its parts depend on it. In this world, there is only one mental system, a brain. The cosmos-encompassing brain has fundamental, mental properties. Still, this world is compatible with physicalism because the fundamental, mental properties can be described in a physicalism-friendly functionalist or identity-theoretic way. Therefore, physicalism is not false in this world. This shows that physicalism is not inconsistent with the existence of fundamental mentality.

I find Brown's (2021) counterexample to the Via Negativa approach and the No Fundamental Mentality constraint persuasive. However, it does not follow that mentality has nothing to do with defining the 'physical'. A modified Via Negativa account could still work and be useful. To develop such an account, a mention of descriptions must be added.

I propose that we accept a Via Negativa account that appeals to non-mental descriptions. It can be called 'Via Negativa descriptive account' because of that. It states a necessary condition for something to be physical: if *x* is physical, *x* can be completely described in non mental terms. Likewise, it states a necessary condition for the truth of physicalism: if physicalism is true, every existing entity can be completely described in non-mental terms. The term 'completely' is included to ensure that nothing is omitted- that nothing remains undescribed.

This account accommodates Brown's (2021) priority monist brain. There can be something that is physical, fundamental, and mental. It is correct to call it 'physical' in part because it can be described in non-mental terms. Priority monist brain worlds are compatible with physicalism because they can be described completely in non-mental terms.

References

- Brown, C. D. (2021). Fundamental mentality in a physical world. *Synthese*, *199*(1), 2841- 2860.
- Crook, S., & Gillett, C. (2001). Why physics alone cannot define the 'physical': Materialism, metaphysics, and the formulation of physicalism. *Canadian Journal of Philosophy*, *31*(3), 333-359.
- Levine, J. (2001). *Purple haze*. Oxford University Press.
- Montero, B. (2009). What is the physical? In B. P. McLaughlin, A. Beckermann, & S. Walter (Eds.), *The Oxford handbook of philosophy of mind* (pp. 173-188). Oxford University Press.
- Montero, B., & Papineau, D. (2005). A defence of the via negativa argument for physicalism. *Analysis*, *65*(3), 233-237.
- Papineau, D. (2001). The rise of physicalism. In C. Gillett & B. Loewer (Eds.), *Physicalism and its discontents* (pp. 3–36). Cambridge University Press.
- Smith, A. D. (1993). Non-reductive physicalism? In H. Robinson (Ed.), *Objections to physicalism* (pp. 225-250). Oxford University Press.

- Spurrett, D., & Papineau, D. (1999). A note on the completeness of 'physics'. *Analysis*, *59*(1), 25-29.
- Wilson, J. (2006). On characterizing the physical. *Philosophical Studies*, *131*(1), 61-99.
- Worley, S. (2006). Physicalism and the via negativa. *Philosophical Studies, 131*(1), 101- 126.

Urte Laukaityte - The Role of History and Philosophy of Science in Medicine: Ontology, Epistemology, and Illness

ETH Zurich; Collegium Helveticum

Some of the key themes within the integrated discipline of the history and philosophy of science (HPS) have involved questions of continuity and change. The relative stability (or lack thereof) of scientific constructs could help us better understand how knowledge is formed, transmitted, and modified over time. With some notable exceptions, the medical sciences have largely not garnered much scholarly attention in these debates. For various reasons, they have stayed more or less peripheral. In this paper, I want to argue that, on the one hand, illness as an object of inquiry could offer important insights to address a range of issues of interest to HPS researchers. On the other hand, we should be applying the conceptual and methodological tools of HPS to medicine more regularly and systematically, given its positioning as a domain where biological, socio-cultural, and experiential dimensions co-exist incredibly closely.

Drawing on the historical ontology approach formulated by Ian Hacking, I propose to examine the nature of illness categories. Rather than reflecting discoveries of objective natural kinds, clinical constructs are historically situated in a significant way. What is more, they do not merely capture pre-existing patterns of symptomatology, but actively shape what can be experienced and expressed via looping effects of human kinds. Ian Hacking's notion of 'ecological niches' is meant to aid us in thinking through the ways in which context, place, and time affect which forms of illness are more likely to emerge and persist. Some will tend to correspond with relatively stable phenomena grounded in physiological damage or disease processes, whereas others will be historically specific. All of them raise important questions regarding the processes of creating and updating scientific knowledge.

While illness categories both reflect and engender synchronisation on a large scale, they resist attempts at universalising generalisation. However, the other extreme of complete individualisation (via case studies, say) is also not appropriate. The way forward will lie somewhere on this global-local continuum, but it will continue to vary in different instances. Given the dynamics of illness experience, the right methods of analysis cannot be purely

biological nor purely cultural. Likewise, they cannot be purely synchronic nor purely diachronic. As a result of the specifics of how illness unfolds, developing a truly integrated historico-philosophical perspective will be crucial within the overall project of the medical sciences. In addition, pursuing such an approach to illness experience would yield broader methodological insights for HPS regarding how science might go about investigating the dynamics of complex phenomena.

Antonella Tramacere - The limits of embodied approaches to comparative cognition

Department of Philosophy, Communication, and Performing Arts

State of the art. Embodied and enactive cognition (EEC) has transformed contemporary cognitive science by challenging the classical conception of the mind as a system of abstract computations over internal representations (Shapiro, 2019). Despite important internal differences, embodied and enactive approaches conceive cognition as emerging from dynamic interactions among brain, body, and environment through sensorimotor engagement, affordance-guided activity, and socially situated practices (Gallagher, 2023). While these approaches have reshaped theories of perception, action, and social cognition, considerably less attention has been devoted to comparative cognition.

Premise. Comparative cognitive science requires explanatory variables capable of accounting for the uneven distribution of cognitive capacities across evolutionary lineages (Halina, 2023).

Main question. The paper asks whether current embodied and enactive approaches provide explanatory variables capable of accounting for cross-species variation in cognitive capacities.

Method. To address the question, I use cross-species variation in false-belief task performance as a case study. My analysis does not depend on any specific interpretation of false-belief tasks or on the claim that they measure genuine belief attribution. Rather, false belief paradigms are only considered because they reveal a robust comparative pattern. As shown through implicit tests, based on measures of anticipatory looking pattern, some species, most notably great apes, consistently display sensitivity to false-belief situations, whereas others, such as macaques, generally do not (Marticorena et al., 2011; Martin & Santos, 2014; Krupenye et al., 2016; Kano et al., 2019). Regardless of how false-belief performance is ultimately interpreted, the observed cross-species variation itself constitutes a comparative explanandum. I set aside the possibility that current paradigms merely lack ecological validity, since this response itself requires specifying which embodied variables would predict successful performance under

species-appropriate conditions. I examine two influential strands of EEC: moderate representational approaches centred on embodied or sensorimotor representations (Goldman 2012; Gallese and Sinigallia, 2011; Goldman and Vignemont, 2009), and radical anti-representational approaches emphasizing

affordances, participatory sense-making, and socially embedded interaction (De Jaegher and Di Paolo, 2007; Fuchs and De Jaegher, 2009; Vincent and Gallegher, 2017) *Results.* I argue that both (moderate and radicale) approaches provide valuable reconceptualizations of social cognition but remain under-specified with respect to comparative explanation. Moderate representational approaches naturally invite explanations in terms of richer embodied representations or greater reuse of embodied representations for social cognition. Radical enactive approaches, by contrast, are naturally led to explain these differences in terms of richer participatory interactions or more sophisticated affordance sensitivity. I argue that neither strategy currently specifies comparative variables capable of explaining why chimpanzees reliably outperform macaques in false-belief paradigms. Since the two species display no obvious qualitative differences either in embodied behavioural repertoires or in the complexity of their social interactions that would predict the observed contrast, the explanatory burden ultimately shifts toward differences in neural organization. However, such explanations are difficult to distinguish from the kinds of neural explanations already available within classical cognitive science.

Conclusion. EEC has not yet developed a sufficiently clear framework for comparative explanation. If these approaches aim to provide a genuine alternative to classical cognitive science, they must move beyond redescribing cognition as embodied and situated and identify predictive comparative variables linking bodily organization, ecological interaction, and situated practices to the evolutionary distribution of cognitive capacities. False-belief tasks serve as a case study of a broader methodological challenge confronting embodied approaches to comparative cognition.

References

- De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making: An enactive approach to social cognition. Phenomenology and the Cognitive Sciences, 6(4), 485– 507.

- Fuchs, T., & De Jaegher, H. (2009). Enactive intersubjectivity: Participatory sense making and mutual incorporation. Phenomenology and the Cognitive Sciences, 8(4), 465–486.
- Gallagher, S. (2023). Embodied and enactive approaches to cognition. Cambridge University Press.
- Gallese, V., & Sinigaglia, C. (2011). What is so special about embodied simulation? Trends in Cognitive Sciences, 15(11), 512–519.
- Goldman, A. I. (2012). A moderate approach to embodied cognitive science. Review of Philosophy and Psychology, 3(1), 71–88.
- Goldman, A., & de Vignemont, F. (2009). Is social cognition embodied? Trends in Cognitive Sciences, 13(4), 154–159.
- Halina, M. (2023). Comparative cognition and the nature of explanation. In The Cambridge Handbook of Comparative Cognition. Cambridge University Press.
- Kano, F., Krupenye, C., Hirata, S., Tomonaga, M., & Call, J. (2019). Great apes use self-experience to anticipate an agent's action in a false-belief test. Proceedings of the National Academy of Sciences, 116(42), 20904–20909.
- Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. Science, 354(6308), 110–114.
- Martin, A., & Santos, L. R. (2014). The origins of belief representation: Monkeys fail to automatically represent others' beliefs. Cognition, 130(3), 300–308.
- Marticorena, D. C. W., Ruiz, A. M., Mukerji, C., Goddu, A., & Santos, L. R. (2011). Monkeys represent others' knowledge but not their beliefs. Developmental Science, 14(6), 1406–1416.
- Shapiro, L. (2019). Embodied cognition (2nd ed.). Routledge.
- Vincent, S., & Gallagher, S. (2017). From false beliefs to true interactions: Are chimpanzees socially enactive? In K. Andrews & J. Beck (Eds.), The Routledge handbook of philosophy of animal minds (pp. 280–288). Routledge.

Tomas Marvan and Wojciech Mamak – Brain Currency and the Two-Point Neuron Architectures

Czech Academy of Sciences, Polish Academy of Sciences

The debate over the brain's fundamental computational currency has recently been dominated by the standard Predictive Processing account, which posits that the cortex trades primarily in prediction-error signals. This article challenges the standard PP framework on both neurobiological and philosophical grounds. Biologically, PP relies on a mathematically subtractive scheme requiring massive inhibitory circuitry to dampen successfully predicted sensory data. However, recent cellular evidence regarding Layer 5 pyramidal "two-point" neurons, alongside evidence from the BELIEF theory of top-down influence demonstrates that contextual feedback is predominantly excitatory and amplificatory. Philosophically, we argue that the prediction error is a severely impoverished, single-purpose causal relay. It lacks the structural capacity to serve as multi-purpose representational content, rendering offline cognition computationally unwieldy. In contrast, we propose that the true functional currency of the cortex is contextually amplified representational content. Driven by the biophysics of apical amplification – where top-down context and bottom-up sensory data coincide to trigger high-frequency bursting - this architecture amplifies rather than suppresses converging signals. This temporal coincidence mechanism serves as a biophysical validator, transforming raw sensory data into information-rich, multi-purpose tokens. By retaining their full semantic depth, these contextually amplified representations are inherently decouplable, seamlessly supporting downstream cognitive intervention and offline processes without the need for fictive prediction errors.

Johan Largo - Why language models are not cognitive extensions

University of Luxembourg

Most of our cognitive achievements depend on our ability to rely on external resources. Large language models (LLMs) have rapidly become a relevant part of such cognitive environment. Many now rely on generative AI models to retrieve information, summarize texts, solve mathematical problems, generate code, assist scientific research and support decision-making across a wide variety of domains. More recently, the emergence of reasoning models, like DeepSeek (Guo et al., 2025) or o3 (OpenAI, 2025), have expanded the capabilities of models even further, allowing them to engage in extended reasoning, plan sequences of actions, use external tools and autonomously pursue complex tasks (Besta et al., 2025; Bi et al., 2025). As these systems become increasingly integrated into human inquiry, an important philosophical question arises: what kind of cognitive relation do users have with these systems?

One influential answer comes from the extended mind thesis (Clark & Chalmers, 1998). According to this view, cognitive processes need not be confined to the biological brain. When external resources perform functions that would ordinarily count as cognitive if carried out internally, they may constitute genuine parts of an individual's cognitive system; this has also been called the “parity principle”. This idea has been influential in philosophy of mind and epistemology, motivating accounts of extended cognition and extended knowledge where notebooks, calculators, smartphones become genuine parts of a person's cognitive processes. At first glance, contemporary LLMs appear to provide an especially compelling case for cognitive extension. Like traditional cognitive artifacts, they can store or retrieve personal information for their users (Smart et al., 2025; Smart et al., 2026) but, unlike traditional cognitive artifacts, they can manipulate information in sophisticated ways to generate answers

and solve problems that not even users themselves could (Maslej et al., 2025; Pritchard, 2026). Moreover, users can habitually rely on these systems, on a stable internet connection, as if they were messaging another person using natural language, making them not only easily available but also easy to use. So, they seem to satisfy Clark's (2010) so-called "trust and glue" conditions.

In this paper, however, we argue that despite their apparent compatibility with criteria for extended cognition (both with regards to the parity principle and the trust-and-glue conditions), LLMs fail to qualify as genuine cognitive extensions or as cases of extended knowledge. Unlike paradigmatic cases of extension: LLMs generate novel outputs that can't be traced to, or haven't been endorsed by, users; they are experienced as separate "quasi-agents" rather than transparent processes (Heersmink et al., 2024); and display behavioral competence that invite users to see them as independent agents rather than parts of their cognition (Yang & Ma, 2025). We further show that more modern accounts of cognitive extension, namely those that drop "previous endorsement" (Clark, 2024; Clark, 2025), drop "transparency in use" (Andrada, 2020; Facchin, 2024) or appeal to "mutual manipulability" (Kaplan, 2012; Gillet et al., 2022) don't satisfactorily make room for LLMs as extensions. We conclude LLMs are better understood as separate cognitive resources; perhaps as independent epistemic agents or mere cognitive tools.

References

- Andrada, G. (2020). Transparency and the Phenomenology of Extended Cognition. *Límite: Revista de Filosofía y Psicología* 15 (20).
- Besta, M., Barth, J., Schreiber, E., Kubicek, A., Catarino, A., Gerstenberger, R., ... & Hoefler, T. (2025). Reasoning language models: A blueprint. *arXiv preprint arXiv:2501.11223*.
- Bi, J., Liang, S., Zhou, X., Liu, P., Guo, J., Tang, Y., ... & Xu, C. (2025). Why reasoning matters? a survey of advancements in multimodal reasoning (v1). *arXiv preprint arXiv:2504.03151*.
- Clark, A., & Chalmers, D. (1998). The extended mind. *analysis, 58*(1), 7-19.
- Clark, A. (2010). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford University Press. Clark, A. (2024). Extending the predictive mind. *Australasian*

Journal of Philosophy, *102*(1), 119-130. Clark, A. (2025). Extending minds with generative AI. *nature communications*, *16*(1), 4627.

- Facchin, M. (2024). Phenomenal transparency, cognitive extension, and predictive processing. *Phenomenology and the Cognitive Sciences*, *23*(2), 305-327.
- Gillett, A. J., Whyte, C. J., Hewitson, C. L., & Kaplan, D. M. (2022). Defending the use of the mutual manipulability criterion in the extended cognition debate. *Frontiers in Psychology*, *13*, 1043747.
- Heersmink, R., de Rooij, B., Clavel Vázquez, M. J., & Colombo, M. (2024). A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness. *Ethics and Information Technology*, *26*(3), 41.
- Kaplan, D. M. (2012). How to demarcate the boundaries of cognition. *Biology & Philosophy*, *27*(4), 545-570.
- Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., ... & Oak, S. (2025). Artificial intelligence index report 2025. *arXiv preprint arXiv:2504.07139*.
- Pritchard, D. (2026). Why technology doesn't normally make you dumber, but agentic AI will. *International Journal of Human–Computer Interaction*, 1-11.
- Smart, P., Clowes, R., & Clark, A. (2025). ChatGPT, extended: large language models and the extended mind. *Synthese*, *205*(6), 242.
- Smart, P., Clowes, R., Krueger, J., & Boniface, M. (2026). The Story of Your Life: Large Language Models and Personal Memory. *Review of Philosophy and Psychology*, 1-23.
- Yang, S., & Ma, R. (2025). Classifying Epistemic Relationships in Human-AI Interaction: An Exploratory Approach. *arXiv preprint arXiv:2508.03673*.

Sreenish S - Cognitive Penetrability of Perception and the Phenomenological Vulnerability of Naive Realism

Vellore Institute of Technology-Andhra Pradesh

The cognitive penetrability of perception has sparked a lively debate in philosophy, psychology, and cognitive science, challenging our naive understanding of how we perceive the world. It suggests that what we believe, hope, desire, expect, and so on not only influences our perception of objects but also shapes the content or the phenomenal character of our perceptual experiences (Stokes 2013; Varga 2017; Siegel 2012). Recently, the cognitive penetrability thesis has received significant traction due to its empirical credibility. However, the thesis generates certain concerns in philosophy. It particularly challenges the phenomenological framework of Naive (or Direct) Realism.

According to naive (or direct) realism, the phenomenology of perceptual experience is largely constituted by externally located physical objects and their properties, rather than internal mental states (Martin 2002). Any changes in the phenomenal character must be explained without relying on other cognitive states since perception presents the world *as it is*, not as filtered or shaped by our thoughts, beliefs, or desires. In short, perception is modular and insulated from cognitive states. Against this, cognitive penetrability of thesis observes that phenomenology of perceptual experience can be significantly altered by other cognitive states without there being any changes in the external objects. If this observation holds, then naive realism is under pressure.

Another important phenomenological concern is that, for typical naive realism, the phenomenology of perceptual experience is exhausted by *sensory phenomenology*. In contrast, cognitive penetration of perception observes that perceptual experience also involves *cognitive phenomenology*. The phenomenology of perception is not fully world-presenting—it also reflects internal, conceptual structure. Since cognitive phenomenology is explained through relevant non-sensory properties, it creates a problem for naive realists. Because the latter holds that the phenomenology of experience is solely explained through sensory phenomenology, i.e., sensory phenomenal properties are sufficient to explain the phenomenology of perception.

The presence of cognitive phenomenology in perceptual experience (Lennon 2023; Montague 2017; Gow 2016, 2019) further motivates resistance to the naïve realist's phenomenological framework.

A related worry about naive realism is the perception of high-level properties (e.g., being a cup, being a screwdriver, being a pine tree, etc.). Higher-level properties are not basic sensory properties. Perceiving them requires a high-level cognitive background, i.e., seeing something as a cup or a pine tree involves some amount of top-down processing. It is argued that the perception of high-level properties (it is assumed that at least some cases) better explain if he theory incorporates the cognitive penetration of perception (Crutchfield 2012). In other words, cognitive penetration of perception makes perception of higher-level properties intelligible. This again validates that the naive realists' framework for the phenomenology of perception requires considerable revision.

As things stand, it is evident that naive realism appears to be intolerant to the phenomenological observations of the cognitive penetrability thesis (see Brogaard and Chomanski 2015; Cavedon-Taylor 2018). Although attempts have been made (e.g., Fish 2009; Johnston 2006) to identify and bridge the tension between naive realism and cognitive penetrability of perception, the issue remains unsettled. Some recent attempts (e.g., Brogaard and Chomanski 2015; Cavedon-Taylor 2018) suggest that naive realism leaves no room for accommodating the cognitive penetrability thesis unless it permits significant metaphysical revision. Even though naïve realism has become increasingly popular in recent times (due to its physicalist metaphysics, its ability to provide a strong epistemological foundation, and its support for our intuitive understanding of perceptual experience), it must concede and address the above phenomenological observations to regain its theoretical credibility and remain a strong contender in philosophy of perception.

The present paper explores the *phenomenological* counterpart of the Cognitive Penetrability Thesis and its implications in Naive Realism. It identifies and clarifies how the cognitive penetrability thesis undermines the phenomenological framework of naive realism. By validating those observations, the paper not only exposes the phenomenological vulnerability of naive realism but also calls for a broader discussion of its metaphysical revision.

References

- Brogaard, B., & Chomanski, B. (2015). Cognitive penetrability and high-level properties in perception: Unrelated phenomena?. *Pacific Philosophical Quarterly*, *96*(4), 469-486.
- Cavedon-Taylor, D. (2018), Naïve Realism and the Cognitive Penetrability of Perception. *Analytic Philos*, 59: 391-412. https://doi.org/10.1111/phib.12131
- Crutchfield, P. (2012). Representing high-level properties in perceptual experience. *Philosophical Psychology*, *25*(2), 279-294.
- Martin, M. G. F. (2002a). The transparency of experience. *Mind & Language*, 17(4), 376-425.
- Macpherson, F. (2012). Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism. *Philosophy and phenomenological research*, 24-62.
- Montague, M. (2017). Perception and cognitive phenomenology. *Philosophical Studies*, *174*, 2045-2062.
- Pylyshyn, Z. (1999). Is vision continuous with cognition?: The case for cognitive impenetrability of visual perception. *Behavioral and brain sciences*, *22*(3), 341-365.
- Siegel, S. (2012). Cognitive penetrability and perceptual justification. *Noûs*, *46*(2), 201-222.
- Silins, N. (2016). Cognitive penetration and the epistemology of perception. *Philosophy Compass*, *11*(1), 24-42.
- Stokes, D. (2013). Cognitive penetrability of perception. *Philosophy Compass*, *8*(7), 646-663.
- Varga, S. (2017). Perceptual experience and cognitive penetrability. *European Journal of Philosophy*, *25*(2), 376-397.

Juan C. Barrio-Aragón & José V. Hernández-Conde - From Minimal to Temporal Consciousness: The origin of conscious experience beyond the present

University of Valladolid

Explaining the evolutionary origin of phenomenal consciousness requires an understanding of not only how subjective experience first emerged, but also how it came to extend beyond the immediate present. Accepting Block's (1995) distinction between phenomenal and access consciousness and viewing the former as a first-person perspective on the world, rooted in a biological basis (Cleeremans, 2011; Godfrey-Smith, 2019), a key question arises: how did a basic form of consciousness, focused on the present, evolve into one capable of generating conscious experiences of past memories and future expectations.

With respect to this question, two broad approaches can be distinguished. On the one hand, evolutionary-functional theories, such as Droege's Temporal Representation Theory (2021), address this question in terms of biological function, asking which adaptive functions consciousness provides and how these functions evolved. On the other hand, phylogenetic theories, such as Ginsburg and Jablonka's (G&J) Unlimited Associative Learning (UAL) (2019), examine the phylogenetic origin of consciousness as a transition from previously non-conscious systems.

However, neither of them provides a thorough explanation of the evolutionary origin of the capacity to have conscious experiences about the past or the future. While evolutionary-functional theories explain the adaptive role of such representations, they do not specify how their conscious experience could emerge. Phylogenetic theories, in turn, explain the transition from non-conscious systems, but pay little attention to how consciousness expanded to encompass past and future events. Other influential proposals —e.g., theories of episodic memory and autonoesis (Tulving, 2002), constructive memory and episodic prospection (Buckner & Carroll, 2007), and mental time travel (Suddendorf & Corballis, 2007)— have devoted only limited attention to the evolutionary mechanisms through which the conscious experience of remembering and anticipating arose.

This paper puts forward an explanation for that transition. Based on the UAL framework, we argue that the emergence of conscious experiences of the past and the future can be understood as an extension of G&J's proposed transition to minimal sensory consciousness. We show how the transition from states of global sensation —conceived as white noise— to the most basic form of consciousness can be extended from the domain of sensory experiences (considered by G&J) to the evolutionary emergence of conscious experiences associated with temporal representations beyond the present.

We propose a model compatible with both Godfrey-Smith's (2020) gradualist conception of the evolution of experience and the models of motor control developed by Frith, Blakemore and Wolpert (2000). We argue that the progressive incorporation of internal predictive mechanisms based on efference copies, forward models, and comparison processes between current, predicted and stored states provided the mental architecture required for conscious experiences of future expectations and past memories to emerge. Within this framework, conscious experience of the future emerges through the evolutionary association of a global sensation with internal states that encode alternative possibilities, ordered according to their adaptive value. Conversely, conscious experience of the past arises through the evolutionary association of a global sensation with internal states that encode correlations between previously issued motor commands and the resulting proprioceptive sensations.

References

- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227-247.
- Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cog nitive Sciences*, 11(2), 49-57.
- Cleeremans, A. (2011). The radical plasticity thesis: How the brain learns to be con scious. *Frontiers in Psychology*, 2, 86.
- Droege, P. (2021). *The Evolution of Consciousness: Representing the Present Mo ment*. Bloomsbury Academic.
- Ginsburg, S., & Jablonka, E. (2019). *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness*. MIT Press.
- Godfrey-Smith, P. (2019). Evolving across the explanatory gap. *Philosophy, Theory, and Practice in Biology*, 11(1), 1-13.

- Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences*, 30(3), 299-313.
- Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychol ogy*, 53(1), 1-25.

Andrei Kochenkov – Objectification Without Constancy: The Case of Olfaction

Lomonosov Moscow State University

Perceptual states are usually taken to be accuracy-evaluable in a way that mere sensory ones are not. Tyler Burge's prominent works on perception (Burge 2010, 2022) provide one influential way of grounding this distinction. Perceptual states are treated as having full-blown representational content and veridicality conditions, whereas mere sensory registration at best marks a biological function or a statistical connection to external subject matter. Magnetotactic bacteria's sensitivity to magnetic fields is a standard case on the sensory-registration side; ordinary human vision is a paradigm case on the perceptual side. Burge draws this distinction by appealing to his notion of *objectification*: the ability of a sensory system to separate local or proximal idiosyncrasies and be "as of" a distal property or object.

Although I agree that the notion of objectification is important, I argue that Burge ties it too closely to perceptual constancy. In fact, he treats constancy as the only way objectification can be carried out. This becomes problematic in cases where a sensory system might lack any constancy mechanism. Some philosophers have opposed Burge's view because it risks treating many sensory abilities, such as olfaction and taste, as non-perceptual and as lacking representational content (Block 2023; Lycan 2024). I argue that Burge's appeal to objectification should be preserved, but detached from a constancy-centered model. While the existence of olfactory constancies remains an open question (Uchida & Mainen 2008; Barnes et al. 2008), I use olfaction as a test case and show how it can achieve objectification even if it lacks perceptual constancies.

Recent work suggests that olfactory computations might produce a multidimensional profile that helps match odor sources with their multimolecular fingerprints (Rokni & Ben-Shaul 2024). Such profiles exhibit features of invariance and categorization, yet they might not themselves be object representations. However, these stable profiles can be used either for recognizing a learned non

olfactory category coupled with certain expectations or concepts, or for integrating an olfactory feature into a spatial framework supplied by other modalities. In both cases, proto-objectified olfactory features become part of a fully objectified representation with a robust content. This does not imply that olfactory states always bear full-blown representational content or objecthood. It shows only how they can become part of such representations. I also argue against the possible objection that integrating an olfactory feature into a learned category or spatial framework is a case of perceptual or sensory judgment, and hence belongs to a higher cognitive level. This objection overintellectualizes olfaction: it treats olfactory navigation and recognition as sophisticated cognitive achievements, which sits uneasily with the characteristic mode of operation of olfactory systems (Rokni et al. 2014; Bao et al. 2019).

The conclusion is that objectification is multiply realizable. Perceptual constancy in vision is a good case of a computationally and architecturally autonomous objectification process, while olfactory representation may require the cooperative work of different mental abilities. The border between perception and mere sensory registration is still preserved, while worries about the status of senses that lack constancies are addressed.

References

- Bao, X., Gjorgieva, E., Shanahan, L. K., Howard, J. D., Kahnt, T., & Gottfried, J. A. (2019). Grid-like neural representations support olfactory navigation of a two dimensional odor space. *Neuron, 102*(5), 1066–1075.e5. https://doi.org/10.1016/j.neuron.2019.03.034
- Barnes, D. C., Hofacer, R. D., Zaman, A. R., Rennaker, R. L., & Wilson, D. A. (2008). Olfactory perceptual stability and discrimination. *Nature Neuroscience, 11*, 1378–1380. https://doi.org/10.1038/nn.2217
- Block, N. (2023). *The Border Between Seeing and Thinking*. Oxford University Press.
- Burge, T. (2010). *Origins of Objectivity*. Oxford University Press. Burge, T. (2022). *Perception: First Form of Mind*. Oxford University Press. Lycan, W. G. (2024). *Perceptual Content*. Oxford University Press.
- Rokni, D., & Ben-Shaul, Y. (2024). Object-oriented olfaction: Challenges for chemosensation and for chemosensory research. *Trends in Neurosciences, 47*(10), 834–848. https://doi.org/10.1016/j.tins.2024.08.008

- Rokni, D., Hemmelder, V., Kapoor, V., & Murthy, V. N. (2014). An olfactory cocktail party: Figure-ground segregation of odorants in rodents. *Nature Neuroscience, 17*(9), 1225–1232. https://doi.org/10.1038/nn.3775
- Uchida, N., & Mainen, Z. F. (2008). Odor concentration invariance by chemical ratio coding. *Frontiers in Systems Neuroscience, 1*, 3. https://doi.org/10.3389/neuro.06.003.2007

Tobias Sandoval - Background conditions and emotional kinds

University of Texas at Austin

Philosophers and scientists have expressed worries about the scientific use of emotion categories such as 'anger,' 'fear,' 'jealousy,' and so on, on the basis that they exhibit too much variability and context-sensitivity to accommodate scientific generalizations, predictions, and projections to future instances. In other words, variability appears to discount emotions as *real kinds*. These categories not only presumably lack unique neurobiological mechanisms, but their instances are also highly sensitive to a plethora of external conditions such as the emoter's current situation, their upbringing, and social environment. In this paper, however, I argue that the problem of emotional variability fails to discount the existence of emotions that are real kinds and the scientific usefulness of many folk emotion categories. Variability, heterogeneity, and context sensitivity does not reveal a relevant difference between emotion categories and paradigmatic real kinds in other scientific domains.

To make this argument, I first develop the notion of a background condition with respect to real kinds and show how background conditions alter the properties of real kinds across domains of chemistry, biology, and medical science. In this context, a background condition is a property or state of affairs external to the kind that affects the manifesting properties of a real kind. For example, instances of *carbon* manifest either as *diamond* or *graphite* depending on the temperature and pressure during their formation, and instances of *e. coli* may differ in their properties due to epigenetic modifications. Following a simple causal theory of kinds, I argue that a category identifies a real kind when its instances share a causal structure wherein certain core properties bring about other derivative properties. However, the derivative properties that arise from the core properties alter depending on which background conditions also obtain.

Then, using empirical and philosophical work on the emotion concept 'jealousy,' I argue that emotional kinds share a core causal structure in the form of cognitive perceptual and motivational states rather than neurobiological states. This core causal structure may bring

about a vast array of differing derivative properties in the form of behaviors, phenomenology, expressions, etc., when obtaining with different background conditions. The influence of background conditions restricts scientists from making unqualified generalizations and projections about emotion concepts, just as the existence of diamond and graphite restrict scientists from making unqualified generalizations and projections about the chemical kind *carbon*. I end by considering differences between how background conditions function with respect to emotions and with respect to chemical and biological kinds, but I argue these are likely differences in degree rather than kind. More broadly, I hope that this paper serves as a starting point for more work on the relationship between real kinds and background conditions across scientific domains.

Douglas C. Wadle - Vividness in Mental Imagery Research and Beyond

Rochester Institute of Technology

There has been an explosion of research on individual differences in imagery ability ever since Zeman et al.'s (2015) paper coining the term 'aphantasia'. These individual differences are largely understood as differences in the *vividness* of mental images, with aphantasia being the inability to voluntarily form mental images of even moderate vividness. Nevertheless, the notion of *vividness* is not clearly articulated in this empirical work. In response, a number of philosophers have turned their attention to this notion. Kind (2017) has argued that the notion of *vividness* is hopelessly confused. Others have come to its defense (Tooming and Miyazono 2021; Fazekas 2024).

I argue that it is confused, but not hopelessly so. The confusions fall into two categories: (1) 'Vividness' has been used to refer, narrowly, to variations in strength or intensity and – sometimes in the very same work – to refer, more broadly, to variations in strength/intensity *and* at least one other dimension of variation (e.g., determinacy or level of detail). (2) Researchers assume that the vividness of the image can be determined by sampling the vividness of some of its features and adding or averaging the results. This obscures any vividness-relevant variations within the imagery experience and overlooks the possibility of vividness-relevant features that are emergent at the level of the whole image. Recent philosophical defenses of *vividness* have implicitly assumed either a broad (Tooming and Miyazono 2021; Fazekas 2024) or narrow (Morales 2023) notion as *the* notion of vividness relevant to individual differences in imagery ability rather than one of a number of equally acceptable notions. As a result, they fail to resolve confusions of type (1). They are even farther from addressing confusions of type (2).

To resolve these confusions, I examine: (i) a narrow notion of vividness applied to features of the image, (ii) a narrow notion of vividness applied to the whole image, (iii) a broad notion of vividness applied to features, and (iv) a broad notion of vividness applied to the whole image. I argue that there is a perfectly clear notion corresponding to (i), but that it applies only to

certain types of features. I argue that (ii) is unlikely to correspond to any intuitive notion of *vividness* underlying subjective vividness ratings. I spend the bulk of the paper discussing (iii) and (iv) – in particular, which dimensions of variation in imagery experience are relevant to identifying the more expansive notion of *vividness* as applied to specific features of the image and the image as a whole. I also discuss the potential impacts of aspects of the imager's experience of the production of the image (e.g., how effortful it is) on vividness ratings of the whole image. This opens onto a further notion of *vividness* as *presence* or *immediacy.* I conclude by connecting *vividness-as-presence* to Hume's discussion of vivacity and force, its applicability to perception and memory, as well as imagery, and the role that this feature might play in epistemic justification.

References

- Abramov, I. & Gordon, J. (1977). Color vision in the peripheral retina. I. Spectral sensitivity. *Journal of the Optical Society of America*, 67(2), 195-202.
- Andrade, J., May, J., Deeprose, C., Baugh, S.-J., & Ganis, G. (2013). Assessing vividness of mental imagery: The Plymouth Sensory Imagery Questionnaire. *British Journal of Psychology*, 105, 547-563. Barrett, T. R. (1993). Verbal hallucinations in normals: II. Self-reported imagery vividness. *Personality and Individual differences,* 1561–671993.
- Block, N. (1983). The photographic fallacy in the debate about mental imagery. *Nous*, 17(4), 651-661. Borst, G., & Kosslyn, S. M. (2008). Visual mental imagery and visual perception: Structural equivalence revealed by scanning processes. *Memory & Cognition*, 36(4), 849–862.
- Burton, L. J., and Fogarty, G. J. (2003). The factor structure of visual imagery and spatial abilities. *Intelligence*, 31(3), 289-318.
- Ceja, C. R. & Fanconeri, S. L. (2023). Difficulty limits of visual mental imagery. *Cognition*, 236: 105436. Cornoldi, C., De Beni, R., Giusberti, F., Marucci, F., Massironi, M., & Mazozoni, G. (1991). The study of vividness in images. In Logie, R. H. & Denis, M. (eds.), *Mental images in human cognition* (Amsterdam: Elsevier), 305-312.

- Dance, C. J., Hole, G., Simner, J. (2023). The role of visual imagery in face recognition and the construction of facial composites: Evidence from aphantasis. *Cortex*, 167, 318-334.
- Dean, G. M., & Morris, P. E. (2003). The relationship between self-reports of imagery and spatial ability. B*ritish Journal of Psychology*, 94(2), 245–273.
- Fazekas, P. (2024). Vividness and content. *Mind & Language*, 39(1), 61-79.
- Floridou, G. A., Peerdeman, K. J., & Schaefer, R. S. (2022). Individual differences in mental imagery in different modalities and levels of intentionality. *Memory and cognition*, 50, 29-47.
- Friedlander, K. J., Lenton, F. H., & Fine, P. A. (2024). A multifactorial model of visual imagery and its relationship to creativity and the Vividness of Visual Imagery Questionnaire. *Psychology of Aesthetics, Creativity, and the Arts*, 18(5), 829–847.
- Fuller, S. & Carrasco, M. (2006). Exogenous attention and color perception: Performance and appearance of saturation and hue. *Vision Research*, (46), 4032-4047.
- Galton, F. (1880). Statistics of mental imagery. *Mind*, 5, 301-318.
- Ganis, G., Thompson, W. L., & Kosslyn, S. M. (2004). Brain areas underlying visual mental imagery and visual perception: An fMRI study. *Cognitive Brain Research*, 20(2), 226–241.
- Hume, D. (1739/1985). *A treatise of human nature*. Nidditch, P. H. (ed.). Oxford: Oxford University Press. Keogh, R. & Pearson, J. (2011). Mental imagery and visual working memory. *Plos One,* 6(12), e29221. Keogh, R. & Pearson, J. (2018). The blind mind: No sensory visual imagery in aphantasia. *Cortex* ,105, 53-60. Kind, A. (2017). Imaginative vividness. *Journal of the American philosophical association*, 3(1), 32-50. Kosslyn, S. M., Brunn, J., Cave, K. R., and Wallach, R. W. (1984). Individual differences in mental imagery ability: A computational analysis. *Cognition*, 18, 195-243.
- Marks, D. F. (1973). Visual imagery differences in the recall of pictures. *British journal of psychology*, 24, 243-246. Milton, F., Fulford, J., Dance, C., Gaddum, J., Heuerman-Williamson, B., Jones, K., Knight, K. F., MacKisack, M., Winlove, C., & Zeman, A. (2021). Behavioral and neural signatures in imagery vividness extremes: aphantasia versus hyperphantasia. *Cerebral cortex communications*, 2(2), 1-15.
- Morales, J. (2023). Mental strength: a theory of experience intensity. *Philosophical perspectives*, 37(1) , 1-21.

- Morina, N., Leibold, E., & Ehring T. (2013). Vividness of general mental imagery is associated with the occurrence of intrusive memories. *Journal of behavioral therapy and experimental psychiatry*, 44(2), 221-226.
- Nedergaard, J. & Lupyan, G. (2023). Not everyone has an inner voice: Behavioral consequences of anendophasia. *Proceedings of the annual meeting of the cognitive science society*, 45(45), 617-624.
- Oertel, V., Rotarska-Jagiela, A., van de Ven, V., Haenschel, C., Grube, M., Stangier, U., ... & Linden, D. E. (2009). Mental imagery vividness as a trait marker across the schizophrenia spectrum. *Psychiatry Research*, 167(1-2), 1–11.
- Ötenen, E. (2022). The role of scene recall imagery and visual imagery in the recall of emotional autobiographical memories (Doctoral dissertation).
- Pearson, J. (2019). The human imagination: the cognitive neuroscience of visual mental imagery. *Nature reviews neuroscience*, 20(10), 624-634.
- Roebuck, H. & Lupyan, G. (2020). The internal representations questionnaire: Measuring modes of thinking. *Behavior research methods*, 52(5), 2053-2070.
- Sack, A. T., van de Ven, V. G., Etschenberg, S., Schatz, D., & Linden, D. E. J. (2005). Enhanced vividness of mental imagery as a trait marker of schizophrenia?. *Schizophrenia bulletin*, 31(1), 97-104.
- Teng, L. (2023). A metacognitive account of phenomenal force. *Mind & Language*, 38(4), 1081-1101. Teng, L. (2024). The epistemic insignificance of phenomenal force. *Philosophy and Phenomenological Research*, 109(1), 55-76.
- Tooming, U. & Miyazono, K. (2021). Vividness as a natural kind. *Synthese*, 199, 3023-3043.
- Winograd, E., Peluso, J. P., & Glover, T. A. (1998). Individual differences in susceptibility to memory illusions. *Applied cognitive psychology: the official journal of the society for applied research in memory and cognition*, 12(7), S5-S27.
- Zeman, A., Dewar, M., and Della Sala, S. (2015). Lives without imagery: Congenital aphantasia. *Cortex*, 73: 378-380.

Sacha Behrend - Reference and Representational Formats in Episodic Memory

University of Hradec Kralove/Paris 1 University

According to simulationist accounts of episodic memory, remembering an event does not require an appropriate causal connection to the subject's original experience of that event. Although this framework offers an alternative to causal theories of memory, it raises a fundamental challenge: what determines whether an episodic memory is about one particular past event rather than another event of the same type? To address this reference-fixing problem, Openshaw and Michaelian (2024) propose a reliabilist model intended to apply uniformly to all episodic memories.

This paper argues that their model offers a promising solution but cannot serve as a fully general account of reference-fixing because it does not accommodate differences between representational formats. I distinguish three possible formats of episodic memory—propositional, imagistic, and analog—and examine how their respective syntactic and semantic properties constrain their referential capacities. I argue that Openshaw and Michaelian's model is particularly well suited to propositional episodic memories, since propositional representations explicitly attribute properties to objects and events. It applies less successfully however to non-propositional formats. Imagistic representations can directly attribute only visual and spatial properties, while analog representations depend on isomorphic relations that require interpretation in order to represent particular properties.

To explain how non-propositional episodic memories can nevertheless refer to features such as emotional valence or abstract properties, I distinguish between attributable and interpretive properties. Attributable properties are those that a representation can directly attribute in virtue of its format. Interpretive properties are those that the representation cannot directly attribute but can nevertheless allow the subject to reconstruct through processes such as association or inference. Since such interpretive properties may be essential to the identity of a remembered episode, an account of reference-fixing must explain how memories can refer to them.

I therefore propose an expanded model that preserves the simulationist and reliabilist foundations of Openshaw and Michaelian's account. More specifically, I revise its referentiality and veridicality conditions so that they distinguish between attributable and interpretive properties. These amendments preserve the central simulationist-reliabilist insight while providing a more flexible and comprehensive account of reference-fixing across the possible formats of episodic memory.

References

- Addis, Donna Rose (2020). Mental Time Travel? A Neurocognitive Model of Event Simulation. Review of Philosophy and Psychology 11 (2):233-259.
- Beck, J. (2019). Perception is analog: The argument from Weber's law. *The Journal of Philosophy, 116*(6), 319–349.
- Dickie, I. (2015). *Fixing reference*. Oxford University Press.
- Fernández, J. (2006). The intentionality of memory. *Australasian Journal of Philosophy, 84*, 39–57. https://doi.org/10.1080/00048400600571695
- Fernández, J. (2017). The intentional objects of memory. In S. Bernecker & K. Michaelian (Eds.), *The Routledge handbook of philosophy of memory* (pp. 88–99). Routledge.
- Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. Oxford University Press.
- Liefke, K. (2025). Mnemic scenarios as pictures. *Asian Journal of Philosophy, 4*, Article 72. https://doi.org/10.1007/s44204-025-00288-6
- Martin, C., & Deutscher, M. (1966). Remembering. *The Philosophical Review, 75*(2), 161– 196. https://doi.org/10.2307/2183082
- Mahr, J. B., & Schacter, D. L. (2023). A language of episodic thought? *Behavioral and Brain Sciences, 46*, e261.
- Michaelian, K. (2016). *Mental time travel: Episodic memory and our knowledge of the personal past*. MIT Press.
- Openshaw, J., & Michaelian, K. (2024). Reference in remembering: Towards a simulationist account. *Synthese, 203*, Article 90.
- Sant'Anna, A. (2018). Episodic memory as a propositional attitude: A critical perspective. *Frontiers in Psychology, 9*, Article 1220.

Farideh Sousani - On the Ontology of the Self: Cognitive-Behavioral and Jungian Approaches to the Symptom

University of Bologna

Every psychotherapeutic practice presupposes an answer to a philosophical question: what is the self that therapy aims to treat? This paper compares the ontologies of the self underlying two therapeutic traditions — Beck's cognitive-behavioral model and Jung's analytical psychology — and argues that their different conceptions of the symptom, and of what successful treatment can be, emerge from this ontological difference rather than from any merely clinical disagreement.

Consider a patient with a persistent fear of paranormal entities. On a CBT approach, the maladaptive mode activated by darkness or ambiguity is identified and restructured: distorted interpretations are corrected and an adaptive mode is built in their place. What the paranormal content means — what it symbolizes within the patient's psychic life — plays no explanatory role. On a Jungian approach, that symbolic content is precisely where the work begins: the fear is treated as a message to be decoded rather than a malfunction to be discharged, pointing toward something in the self that is not yet consciously engaged.

I argue that these contrasting practices branch from contrasting ontologies. In Beck's framework, the self is constituted by individual experience — beliefs, memories, and learning history accumulated across a lifetime. In other words, the "I" begins at birth; what experience has damaged, targeted intervention tries to repair. For Jung, the self exceeds individual biography: it includes archetypal and collective layers that predate the individual life and remain active within it. Symptoms are accordingly not mere dysfunctions of an experientially constituted system but signals from unintegrated layers of the self.

Rather than arbitrating between these ontologies, the paper draws out a structural consequence: the scope of the self a therapy presupposes fixes the horizon of what that therapy can, in principle, address. If the self extends beyond conscious experience, symptomatic relief may leave essential dimensions of the person untouched — a possibility that would explain the familiar clinical phenomenon of patients who improve without feeling fully seen. I connect

this to the explanation/understanding debate in the philosophy of the social sciences. Classical CBT explains symptoms as causal malfunctions; Jungian practice interprets them as carriers of meaning — and each tradition's ontology of the self fits the kind of explanation it is committed to. The question of what the self is, I conclude, is not external to psychotherapy but internal to it: it determines what counts as a symptom, what counts as healing, and what any therapeutic practice can hope to do.

References

Beck, Aaron T. 1996. "Beyond Belief: A Theory of Modes, Personality, and Psychopathology." In *Frontiers of Cognitive Therapy*, edited by Paul M. Salkovskis, 1–25. New York: Guilford Press.

Beck, Aaron T., and Emily A. P. Haigh. 2014. "Advances in Cognitive Theory and Therapy: The Generic Cognitive Model." *Annual Review of Clinical Psychology* 10: 1–24. https://doi.org/10.1146/annurev-clinpsy-032813-153734

Jung, C. G. 1959. *The Archetypes and the Collective Unconscious*. Collected Works, vol. 9, pt. 1. Translated by R. F. C. Hull. London: Routledge and Kegan Paul.

Jung, C. G. 1966. *Two Essays in Analytical Psychology*. 2nd ed. Collected Works, vol. 7. Bollingen Series XX. Translated by R. F. C. Hull. Princeton, NJ: Princeton University Press.

Taylor, Charles. 1971. "Interpretation and the Sciences of Man." *The Review of Metaphysics* 25 (1): 3–51.

Michał Piekarski - Precision-Weighted Noise: A Hierarchical Predictive-Processing Account of the Conjunction Fallacy

Cardinal Stefan Wyszynski University in Warsaw, Institute of Philosophy, Department of Epistemology and Cognitive Science

The conjunction fallacy — the systematic tendency to judge a conjunction P(A∧B) as more probable than one of its constituents — is among the best-documented violations of classical probability theory in human judgment and memory (Brainerd, 2021, 2022; Busemeyer et al., 2011; Tversky & Kahneman, 1983; Zhu, Sanborn & Chater, 2020). It is robust to experimental manipulation and does not diminish with training. The standard interpretation treats it as an error — evidence that memory fails to conform to the norms of classical probability theory. My presentation challenges not the status of this phenomenon as a formal violation, but the explanatory framework within which it is understood.

I propose a reinterpretation of the probability-theory-plus-noise model, hereafter C&W model (Costello & Watts, 2014, 2017) within a hierarchical predictive processing framework (Barron, Auksztulewicz & Friston, 2020; Henson & Gagnepain, 2010; Quent, Henson & Greve, 2021). On this account, the noise parameter d reflects representational precision at a given level of a hierarchical generative system, and the additional conjunction cost Δd represents the precision loss incurred when integrating information across levels. Critically, unlike the flat C&W model — in which Δd is constant — the hierarchical account predicts that Δd varies systematically with the nature of the relation between conjoined events. Four conjunction types are distinguished: causal, categorical, unrelated, and analytic. The key prediction is that conjunction fallacy rates will differ across these types even when base probabilities P(A) and P(B) are held constant — a prediction unavailable to the original model.

This reinterpretation has a normative dimension that goes beyond mechanism. Following Levenstein et al. (2023), I distinguish descriptive, mechanistic, and normative levels of explanation in neuroscience and argue that predictive processing operates at all

three simultaneously. It is not merely a model of how the memory system operates; it provides a principled account of why approximate probabilistic inference is adaptive under biological constraints (Friston, 2010). Classical probability theory proves to be a norm appropriate for an ideal agent with unlimited computational resources — not for systems operating under evolutionary and metabolic pressure. Formally incorrect memory judgments are therefore not failures of rationality but structurally predictable consequences of optimal resource-limited inference. Norm violation and adaptive optimality are not mutually exclusive.

The conjunction fallacy serves as a test case: the framework generates an empirically testable and falsifiable prediction that directly adjudicates between the flat and hierarchical accounts.

The philosophical upshot is a revision of the normative standard applied to biological memory systems — one that takes seriously the distinction between the norms appropriate for ideal agents and those appropriate for systems shaped by evolutionary and computational constraints.

References

- Barron, H. C., Auksztulewicz, R., & Friston, K. (2020). Prediction and memory: A predictive coding account. *Progress in Neurobiology*, 192, Article 101821. https://doi.org/10.1016/j.pneurobio.2020.101821
- Brainerd, C. J. (2021). Deep memory distortions. *Cognitive Psychology*, 126, Article 101386. https://doi.org/10.1016/j.cogpsych.2021.101386
- Brainerd, C. J. (2022). Deep distortion. *Memory*, 30(1), 5–9. https://doi.org/10.1080/09658211.2020.1844756
- Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological Review*, 118(2), 193–218. https://doi.org/10.1037/a0022542
- Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, 121(3), 463–480. https://doi.org/10.1037/a0037010

- Costello, F., & Watts, P. (2017). Explaining high conjunction fallacy rates: The probability theory plus noise account. *Journal of Behavioral Decision Making*, 30(2), 304–321. https://doi.org/10.1002/bdm.1936
- Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138. https://doi.org/10.1038/nrn2787
- Henson, R. N., & Gagnepain, P. (2010). Predictive, interactive multiple memory systems. *Hippocampus*, 20(11), 1315–1326. https://doi.org/10.1002/hipo.20857
- Levenstein, D., Alvarez, V. A., Amarasingham, A., Azab, H., Chen, Z. S., Gerkin, R. C., Hasenstaub, A., Iyer, R., Jolivet, R. B., Marzen, S., Monaco, J. D., Prinz, A. A., Quraishi, S., Santamaria, F., Shivkumar, S., Singh, M. F., Traub, R., Nadim, F., Rotstein, H. G., & Redish, A. D. (2023). On the role of theory and modeling in neuroscience. *Journal of Neuroscience*, 43(7), 1074–1088. https://doi.org/10.1523/JNEUROSCI.1179-22.2022
- Quent, J. A., Henson, R. N., & Greve, A. (2021). A predictive account of how novelty influences declarative memory. *Neurobiology of Learning and Memory*, 179, Article 107382. https://doi.org/10.1016/j.nlm.2021.107382
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293
- Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. *Psychological Review*, 127(5), 719–748. https://doi.org/10.1037/rev0000190

Marek Pokropski - Idealizing Consciousness: Multiple Models or a Minimal Unifying Model?

University of Warsaw

Consciousness science is, to a large extent, a model-based science: the bulk of empirical and theoretical work in the field proceeds through the construction and manipulation of computational models of conscious states and their neural or functional correlates (e.g., Global Workspace Theory, Integrated Information Theory, Predictive Processing). Since scientific models are idealizations, i.e., simplified and in various respects inaccurate representations of their target systems, the question of which idealization strategy is appropriate for a given science bears not merely on methodology but on how progress in that science should be understood and assessed.

Recent philosophy of science has devoted increasing attention to the role of idealization in empirical inquiry generally (e.g., Potochnik 2017), and this attention has begun to extend to cognitive science (e.g., Hochstein 2023; Kirchhoff 2025) and neuroscience (Chirimuuta 2024). Consciousness science, however, remains comparatively undertheorized in this respect.

This paper addresses that gap. Drawing on Weisberg's (2013) tripartite taxonomy of idealization strategies, Galilean, minimal, and multiple-models idealization (MMI), I ask which strategy best characterizes, and should guide, the modeling practices of contemporary consciousness science. Wiese (2020) has argued that the field should aim for a single minimal unifying model (MUM): a model specifying only the necessary, maximally general features shared by extant theories of consciousness, thereby unifying the field around a common minimal core.

Against this proposal, I argue that MMI, rather than a MUM of the kind Wiese envisages, offers the more adequate idealization strategy for the current state of consciousness science. This conclusion is supported by three arguments. First, the causal and explanatory complexity of consciousness, spanning phenomenal, cognitive, and neurobiological levels of description, makes it unlikely that any single model, however minimal, can jointly satisfy the diverse explanatory aims pursued by different research programs (Potochnik 2020). Second, extant

theories of consciousness embody mutually incompatible idealizing assumptions about its relevant substrate and mechanism, a pattern that closely parallels paradigm cases of MMI in the physical and biological sciences (Weisberg 2013). Finally, drawing on Kirchhoff (2025) and Hochstein (2023), I argue that the problem of inconsistent models that MMI generates need not undermine the possibility of integration.

References

- Chirimuuta, M. (2024). *The Brain Abstracted: Simplification in the History and Philosophy of Neuroscience*. Cambridge, MA: MIT Press.
- Hochstein, E. (2023). Integration without integrated models or theories. *Synthese*, 202(3), 76.
- Kirchhoff, M. D. (2025). *The Idealized Mind: From Model-Based Science to Cognitive Science*. Cambridge, MA: MIT Press.
- Potochnik, A. (2017). *Idealization and the Aims of Science*. Chicago: University of Chicago Press.
- Potochnik, A. (2020). Idealization and many aims. *Philosophy of Science*, 87(5), 933–943.
- Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. New York: Oxford University Press.
- Wiese, W. (2020). The science of consciousness does not need another theory, it needs a minimal unifying model. *Neuroscience of Consciousness*, 2020(1), niaa013.

Adrian Wieczorek - Turning plants from head to ground. An alternative to the predictive stance in minimal cognition.

Technische Universität Berlin

The talk examines a recent argument for prediction in non-neural organisms such as plants and bacteria (Calvo & Friston 2017; Sims 2019; Kiverstein & Sims 2021) and interprets it as an instance of a broader and increasingly common line of reasoning which can be called the 'Predictive Stance' (Friston 2013; Clark 2016; Bruineberg et al. 2018) This stance attributes anticipatory models to organisms to explain their anticipatory behavior (Sims 2023). The talk pursues two main goals. First, it shows that the argument for prediction is unwarranted and trivializes the explanatory value of the Predictive Stance. Second, it introduces an alternative, radical embedded approach to anticipatory behavior – 'Sensory Probing' – as a more parsimonious explanation for non-neural systems. This view

calls for a shift in research perspective: rather than assuming by default internal anticipations that guide behavior, investigation should primarily concern environmental structure and its causal anchor points. In this way, plants can serve as model organisms for a re-embedding strategy of anticipatory behavior that is neither brain-based nor action-centered.

References

- Bruineberg, J., Kiverstein, J., Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese* 195, 2417–2444.
- Calvo P., Friston K. (2017). Predicting green: really radical (plant) predictive processing. *Journal of the Royal Society Interface.* 14(131):20170096.
- Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. New York, Oxford. Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, 10, 20130475. Kiverstein, J., Sims, M. (2021). Is free-energy minimization the mark of the cognitive?. *Biol Philos* 36,25

- Sims, M. (2019) Minimal perception: Responding to the challenges of perceptual constancy and veridicality with plants. *Philosophical Psychology*, 32:7, 1024-1048
- Sims, M. (2023). Many Paths to Anticipatory Behavior: Anticipatory Model Acquisition Across Phylogenetic and Ontogenetic Timescales. *Biological Theory* 1 (2):114-133.

Oryan Zacks - Imagination, memory and the ancient origins of the default mode network

The Max Planck Institute for Evolutionary Anthropology

Imagination is one of the biggest mysteries in neuroscience- how does the brain create a kind of “movie theatre” within our heads? While there is much work to be done in addressing this question, taking an evolutionary approach may provide clues into the ancient neural mechanisms that enable imagination. Specifically, the default mode network (DMN) has been proposed as the primary “imagination network” in the human brain. This network consists of two different cortical architectures: the neocortex and hippocampus. In this talk I will suggest that both structures underwent significant changes over evolutionary times, with their defining features arising at the base of the mammalian lineage. These features include unique developmental programs, new cell types and obligatory connectivity rules between the different cell types. These features critically expanded the computational capacities of the mammalian brain, although they are not always accounted for in models. Furthermore, they have continued to develop in the lineage that led to humans, which could partially explain the enhanced cognitive abilities of humans. Understanding the cumulative changes that shaped our brains over evolutionary time can help uncover important factors that determine our cognitive capacities today. In addition, such research has the potential to inspire more biologically realistic computational models and translate into novel AI models, ones that might develop their own kind of imagination.

Gina Adriana ramos Zamudio - The Operative Prereflective Dimension as a Bridging Category: Toward a Dialogue with Cellular Memory

Master's Student in Philosophy, University of Guanajuato, Mexico

The operative prereflective dimension is proposed as a bridging category that makes it possible to articulate, without reducing one to the other, the developments of classical phenomenology and the contemporary approaches of neurophenomenology and enactive cognition. Its function is not to introduce a new ontological structure but to provide a conceptual framework through which Husserl's passive syntheses, sedimentations and bodily habitualities, operative intentionality, and Merleau-Ponty's notion of the flesh can be understood in an integrated manner as the background from which all experience emerges.

From this perspective, experience does not begin with a reflective act or with an explicit representation of the world, but rather upon a set of historically constituted structures that continue to operate prereflectively and make perception, action, and the constitution of meaning possible. The operative prereflective dimension does not replace these concepts; instead, it brings to light their functional continuity and establishes a space for dialogue with the cognitive sciences without abandoning the descriptive specificity of phenomenology.

Within this context, contemporary findings concerning so-called cellular memory become relevant not because they allow phenomenological sedimentations to be identified with specific biological mechanisms, but because they suggest that the organism preserves, in a non-thematized manner, a history of functional adaptations that remains operative in its relationship with the environment. The persistence of cellular adaptations in tissues such as skeletal muscle, the immune system, or certain populations of stem cells indicates that corporeality is not merely the material support of experience but a historically configured organism whose organization continuously shapes its mode of responding to the world.

From this perspective, the operative prereflective dimension makes it possible to think together the sedimentations described by phenomenology and these biological processes without

reducing one to the other. Its function is to show that both remain prereflectively operative as the horizon of possibility for action and experience, constituting the dynamic background from which cognition can be understood as enactive. In this sense, the proposed category does not seek to establish an identity between phenomenological description and biological explanation but rather to acknowledge that the historicity of the lived body may be expressed simultaneously in structures of meaning and in bodily processes that normally remain outside conscious thematization, yet nevertheless participate in the constitution of experience.

Aleksandr Naumenko - Do Categories Rely on Similarities or Differences?

Independent Researcher

This paper challenges the foundational assumption in cognitive science that categorization operates primarily as an internal similarity-matching computation. Classical connectionist and psychological models, such as prototype and exemplar frameworks, assert that an unmapped environmental input is recognized by calculating its statistical proximity to an idealized central tendency or a cluster of historical instances. Conversely, atomistic paradigms argue that concepts are independent units whose categorization depends entirely on intrinsic information shared by fitting phenomena. We demonstrate that both models introduce severe computational friction when scaling, trapping the cognitive system in a high-overhead feature evaluation loop that violates real-time resource constraints. To resolve this structural bottleneck, we introduce an alternative architecture that shifts the explanatory paradigm away from similarity centered metrics and toward the systematic organization of differences between contrasting classes along explicit property dimensions. We argue that categorization is fundamentally a process of sequential difference extraction rather than holistic resemblance mapping.

From this perspective, environmental entities are modeled not as holistic primitives, but as high dimensional configurations of underlying property axes. Viewing properties as continuous or discrete dimensions turns comparison into a vital dimensionality-reduction operation that evaluates a single property dimension at a time. To optimize processing speed under intense temporal pressures, a resource bounded agent cannot execute pointaccurate calculations over infinite scalar values; instead, the cognitive architecture partitions continuous axes into discrete, manageable ranges. These bounded intervals function as relational representational primitives through which the mind models reality and guides behavior. A category is thus formalized as a section of conceptual space bounded by these property ranges traced from a universal root node. This structural alignment dissolves the logical instabilities of similarity. As Nelson Goodman demonstrated, similarity is a vacuous relation unless a specific property

axis is pre-specified to ground the comparison. Human intelligence is therefore defined as the capacity to discover and exploit these property boundaries, an operational focus reflected in the etymology of the term—descending from the Latin “inter + legere”, meaning "to choose or read between.” The operational necessity of a difference driven, hierarchical approach becomes mathematically stark when evaluated through search space optimization. While a flat, similarity-based approach mandates linear search complexity, O(N)—requiring independent parallel calculations across every candidate class—a hierarchical organization of categories operates with highly optimized logarithmic complexity, O(log N). By sequentially filtering the environment along prioritized property dimensions, the system systematically prunes unviable subtrees of choices. Under intense survival pressures, this structural organization enables the system to halt computation prematurely and default to a broad category node as a reliable real-time heuristic. Furthermore, this mechanics provides a structural definition of semantic vagueness: vagueness is the measure of unexplored or unexploited differences remaining among the fitting phenomena enclosed under any given category filter.

References

- Bateson, G. (1972). Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology. Jason Aronson.
- Gärdenfors, P. (2000). Conceptual spaces: The geometry of thought. The MIT Press.
- Goldstone, R., Kersten, A., and Carvalho, P.F. (2013). Concepts and categorization. In A. F. Healy, R. W. Proctor, and I. B. Weiner (Eds.), Handbook of psychology: Experimental psychology (2nd ed., pp. 607–630). John Wiley and Sons, Inc..
- Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), Problems and projects (pp. 437- 447). Indianapolis: Bobbs-Merrill.
- Fodor, J. A. (1998). Concepts: Where cognitive science went wrong. Clarendon Press/Oxford University Press. https://doi.org/10.1093/0198236360.001.0001
- Hahn, U., and Ramscar, M. (2001). Introduction: Similarity and categorization. In U. Hahn and M. Ramscar (Eds.), Similarity and categorization (pp. 1-11). New York: Oxford University Press.
- Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1-3), 335-346. https://doi.org/10.1016/0167-2789(90)90087-6

- Hebenstreit, G. (2009). Defining patterns in translation studies: Revisiting two classics of German translationswissenschaf. In Y. Gambier, and L. van Doorslaer (Eds.), The metalanguage of translation (pp. 9- 26). John Benjamins Publishing Company.
- Knuth, D. E. (1998). Sorting and Searching. The Art of Computer Programming, Vol. 3 (2nd ed.). Addison Wesley.
- Murphy, G. L., and Medin, D. L. (1999). The role of theories in conceptual coherence. In E. Margolis and S. Laurence (Eds.), Concepts: Core reading (pp. 425-458). Cambridge, MA: MIT Press.
- Rand, Ayn (1966). Introduction to objectivist epistemology. [New York]: Objectivist. 12. Reiter, Ray (1980). A logic for default reasoning. Artificial Intelligence 13 (1-2):81-137.
- Reutlinger, Alexander, Gerhard Schurz, Andreas Hüttemann, and Siegfried Jaag, "Ceteris Paribus Laws", The Stanford Encyclopedia of Philosophy (Spring 2024 Edition), Edward N. Zalta and Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/spr2024/entries/ceteris-paribus/>.
- Rosch, E. (1973-05-01). "Natural categories". Cognitive Psychology. 4(3):328–350. doi:10.1016/0010- 0285(73)90017-0. ISSN 0010-0285.
- Rosch, E. (1978) Principles of Categorisation. In: Rosch, E. and Lloyd, B.B., Eds., Cognition and Categorisation, Lawrence Erlbaum, Hillsdale, 27-48.
- Skiena, Steven. (2008). The Algorithm Design Manual. 10.1007/978-1-84800-070-4.
- Sorensen, Roy, "Vagueness", The Stanford Encyclopedia of Philosophy (Winter 2023 Edition), Edward N. Zalta and Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/win2023/entries/vagueness/>.
- von Glasersfeld, E. (1995). Radical Constructivism: A Way of Knowing and Learning. The Falmer Press. 19. Williamson, T. (1994). Vagueness. Routledge.

Fernanda Caroliny Cardoso - If computers can think, why are LLMs not thinking? Dennett on the emergence of comprehension, consciousness and intentionality

State University of Campinas (UNICAMP), Brazil

With the recent and thunderous rise of large language models (LLMs), a lot of old philosophical questions resurfaced once again, now with new contours. Among those, one is particularly intriguing: can a computer *think*? Public discourse on artificial intelligence (AI) is dominated by fears of a future superintelligent, truly-minded machine that threatens "replacing" humanity. In response to this widely spread sentiment, people often deny the very possibility of thoughtful AI: in virtue of its mechanical nature, it would never be capable of really *experiencing* nor of *meaning* anything. That is, *consciousness* and *intentionality* could never be "injected" into a machine as discrete algorithms. Daniel Dennett, however, who is a well-known eliminativist about *intrinsic qualia* and *original content,* sees no principled obstacle to ascribing both consciousness and intentionality not only to humans and other animals, but also to AIs, once their functional organization satisfies certain relevant operational criteria (Dennett 1978, 1985, 1987, 1988a,b,c, 1991, 1994, 2009). If a robust theory of cognition were actually available, a computer could indeed be capable of "thinking" – but only in the scientifically relevant sense of reproducing the same *cognitive functions* we, humans, execute while thinking. Nonetheless, Dennett fiercely maintained, among many others, that the currently available LLMs do *not* really think about what they are "talking", primarily because they do not have the actual *comprehension* of the linguistic statements they so competently generate. In fact, in his last years of life, Dennett (2017, 2023) alarmingly argued that systems such as LLMs are "parasites" that could actually erode the trust-based human way of life we have been developing for so long. With that in mind, one can rightfully ask: if we assume, alongside Dennett, a reductionistic account of consciousness and intentionality, what really prevents current LLMs from being properly considered genuine thinkers? That is, what are the requirements for "comprehension" that LLMs have not achieved, in spite of their remarkable linguistic competence, if not intrinsic qualia and original content? By answering this question, I seek to show how one may differentiate the human mentality from the current LLMs without

smuggling what Dennett has called the "Cartesian theater" into the standard functionalist, computational-representational understanding of the mind. My thesis is that, for Dennett (1985, 1994, 2017), there are four main practical, intertwined conditions for the diachronic *emergence* of comprehension (alongside consciousness and intentionality) that LLMs have not achieved, despite appearing to have done so: *embodiment* (sensorimotor grounding in the physical world), *affectivity* (homeostatic precarity, without which things cannot "matter" to the system), *historicity* (a

developmental, biographical trajectory necessary for the retrospective narrative that makes up the "user illusion" of a conscious self), and *normativity* (the capacity to navigate the "space of reasons" as "Gregorian creatures" capable of evaluating and taking responsibility for their assertions). I conclude by arguing that the true threat of AI seems to lie not in the distant possibility of a truly thoughtful computer, but in the immediate, actual pervasiveness of unthoughtful counterfeits that merely emulate the meaningfulness of human experience.

References

- Dennett, D. C. (1978). Why you can't make a computer that feels pain. *Synthese*, *38*(3), 415–456.
- Dennett, D. C. (1985). Can machines think? In M. G. Shafto (Ed.), *How we know: Nobel Conference XX* (pp. 121–145). Harper & Row.
- Dennett, D. C. (1987). *The intentional stance*. MIT Press / A Bradford Book.
- Dennett, D. C. (1988a). Quining qualia. In A. Marcel & E. Bisiach (Eds.), *Consciousness in contemporary science* (pp. 42–77). Oxford University Press.
- Dennett, D. C. (1988b). Evolution, error, and intentionality. In Y. Wilks & D. Partridge (Eds.), *Sourcebook on the foundations of artificial intelligence* (pp. 190–212). New Mexico University Press.
- Dennett, D. C. (1988c). The intentional stance in theory and practice. In R. Byrne & A. Whiten (Eds.), *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans* (pp. 180–202). Oxford University Press.
- Dennett, D. C. (1991). *Consciousness explained*. Little, Brown and Company.
- Dennett, D. C., Dretske, F., Shurville, S., Clark, A., Aleksander, I., & Cornwell, J. (1994). The practical requirements for making a conscious robot. *Philosophical*

Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences, *349*(1689), 133–146.

- Dennett, D. C. (2009). Darwin's strange inversion of reasoning. *Proceedings of the National Academy of Sciences*, *106*(Suppl. 1), 10061–10065.
- Dennett, D. C. (2017). *From bacteria to Bach and back: The evolution of minds*. W. W. Norton & Company.
- Dennett, D. C. (2023). The problem with counterfeit people. *The Atlantic*. https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/67407 5/

Andrew Richmond - From neuroscience to explainable AI: Representation, understanding, and explanation

University of Otago

Much explainable AI (XAI) research aims to describe the *internal representations* that control a language model's behavior. This research self-consciously borrows the conceptual frameworks that neuroscientists associate with representation, and applies them to deep neural networks instead of brains (Ameisen et al. 2025; Lindsey et al. 2025). But it remains unclear whether XAI *should be* understanding neural networks the same way neuroscience understands the brain, or what such understanding amounts to.

This paper approaches this question pragmatically, by asking *what using concepts like representation helps scientists do*. The questions I'm interested in are therefore:

1. What goals do concepts like representation serve in neuroscience, and how?

2. What goals does XAI research have, and under what constraints must it meet them?

3. Given how the relevant concepts function (identified in 1) and the goals and constraints XAI operates under (identified in 2): (how) can the relevant concepts fruitfully be exported from neuroscience to XAI?

I offer a preliminary answer to these questions. To answer (1), I draw on pragmatist theories of representation in neuroscience (Cao 2022; Chirimuuta 2024; Egan 2012, 2025; Richmond 2025b, 2025a) to argue that the concept of representation introduces an *explanatory schema* for showing how a system's internal states support coordination with the environment.

To answer (2), I argue that XAI differs from neuroscience in one especially important respect: a major constraint on explanations in XAI is that they must identify high-level human interpretable representational states, not just any behavior-guiding correlations with the environment (Lindsey et al. 2025). Indeed, strong correlations that do not correspond to common human concepts are considered *unexplanatory* (Cammarata et al. 2020; Olah et al. 2020).

This lets us say something informative about the way neuroscience's explanatory schemas help us understand deep neural networks. Most importantly, the search for high-level human interpretable representational states calls for extreme idealization (Lindsey et al. 2025). And while idealization does not render representational explanations inappropriate, it means their scope must be limited, and researchers are mistaken when they take these explanations to *identify* concepts or conceptual structure in the model (as in, e.g., Schneider 2026) as opposed to *imposing* conceptual structure in order to idealize information flow through a network. In short, representational explanation transforms on its way from neuroscience to XAI. It offers the two fields two different types of explanation, grounded in different goals and constraints.

The conclusion of the paper draws out some further implications for our understanding of deep neural networks and some lessons for the methodology with which we study them.

References

- Ameisen, Emmanuel, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. "Circuit Tracing: Revealing Computational Graphs in Language Models." *Transformer Circuits Thread*. https://transformer-circuits.pub/2025/attribution graphs/methods.html.
- Cammarata, Nick, Gabriel Goh, Shan Carter, Ludwig Schubert, Michael Petrov, and Chris Olah. 2020. "Curve Detectors." *Distill*. doi:10.23915/distill.00024.003.
- Cao, Rosa. 2022. "Putting Representations to Use." *Synthese* 200(151).
- Cappelen, Herman, and Josh Dever. 2021. *Making AI Intelligible*. Oxford University Press.
- Cappelen, Herman, and Josh Dever. 2025. "A Hyper-Externalist Manifesto for LLMs." in *Communicating with AI: Philosophical Perspectives*, edited by H. Cappelen and R. Sterken. Oxford University Press.
- Chirimuuta, Mazviita. 2024. *The Brain Abstracted: Simplification in the History and Philosophy of Neuroscience*. MIT.

- Egan, Frances. 2012. “Metaphysics and Computational Cognitive Science: Let’s Not Let the Tail Wag the Dog.” *Journal of Cognitive Science* 13(1):39–49. doi:10.17791/jcs.2012.13.1.39.
- Egan, Frances. 2025. *Deflating Mental Representation*. The Jean Nicod Lectures. Cambridge, Massachusetts: The MIT Press.
- Fraser-Taliente, Kit, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C. Bogdan, Emmanuel Ameisen, James Chen, Dzmitry Kishylau, Adam Pearce, Julius Tarng, Alex Wu, Jeff Wu, Yang Zhang, Daniel M. Ziegle, Evan Hubinger, Joshua Batson, Jack Lindsey, Samuel Zimmerman, and Samuel Marks. 2026. “Natural Language Autoencoders Produce Unsupervised Explanations of LLM Activations.” https://transformer-circuits.pub/2026/nla/index.html.
- Lindsey, Jack, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. “On the Biology of a Large Language Model.” *Transformer Circuits Thread*. https://transformer-circuits.pub/2025/attribution-graphs/biology.html.
- Olah, Chris, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. “An Overview of Early Vision in InceptionV1.” *Distill*. doi:10.23915/distill.00024.002.
- Richmond, Andrew. 2025a. “How Computation Explains.” *Mind & Language* 40(1):2–20. doi:https://doi.org/10.1111/mila.12521.
- Richmond, Andrew. 2025b. “What Is a Theory of Neural Representation For?” *Synthese* 205(14).
- Schneider, Susan. 2026. “The Error Theory of LLM Consciousness: There Is No Evidence That Standard LLMs Are Conscious.” *Behavioral and Brain Sciences*.

Daniel Sherman - Embodied Insights: Interoception in Psychedelic Epistemology

The University of Western Australia

Psychedelic epistemology is currently deadlocked on several fronts. Claims that metaphysical reality is revealed by psychedelic experience produce mutually irreconcilable conclusions. Work on psychological insights under psychedelics has connected those experiences to therapeutic benefits (Kugel et al. 2025), but methods of epistemic verification remain unclear. Meanwhile, some epistemologists have questioned whether psychedelic experience itself can lead to the acquisition of knowledge, or perhaps just increase understanding (Fink 2022).

Psychedelic insights involving bodily states, addiction, and health may provide ideal test cases because such claims can be triangulated across multiple forms of evidence, such as physiological measures, validated self-report instruments, and behavioural outcomes. This presentation proposes that interoception—the brain's sensing and interpretation of the organism's internal states and homeostatic needs—is an underexamined epistemic source with major implications for psychedelic epistemology.

I argue that interoceptive epistemic performance cannot be reduced to the veridical detection of bodily signals (Garfinkel et al. 2015). Sensory systems track properties relative to organismic interests, while embodied appraisals represent what bodily changes mean for an organism's concerns and well-being (Prinz 2004). Such appraisals can succeed or fail. Danger may be detected or overlooked, benign arousal misrepresented as threat, and damaging states assigned insufficient salience. Interoception therefore operates at an epistemically significant interface between bodily sensation and cognitive appraisal, where physiological information contributes to representations with conditions of correctness. On this account, interoception enables and constrains knowledge not only through detection, but through appraisal and salience (Seth 2013). Epistemic improvement occurs when altered appraisal more accurately represents the significance of bodily states for the organism's health, needs, and goals.

Although psychedelic experiences are highly variable, insights related to substance use, diet, and exercise show recurring tendencies toward health-promoting orientations and behavioural changes across clinical and naturalistic contexts (Teixeira et al. 2022). These convergent patterns call for epistemic explanation. I argue that psychedelic alterations to the salience and appraisal of bodily feelings, as well as to mental representations of the body, are a neglected source of these insights and of what many subjects later describe as psychological revelations. Psychedelic epistemic gains may therefore be grounded not in metaphysical revelation, but in altered access to bodily information and changes in what that information means to the embodied subject.

Interoceptive claims are unusually tractable within psychedelic epistemology because they can be evaluated across multiple perspectives, including objective physiological measures, self-report instruments of interoceptive sensibility and awareness, and

behavioural outcomes such as long-term abstinence from harmful substances. Interoceptive psychedelic insights therefore provide a plausible case of epistemic gain within an empirically evaluable domain. By locating psychedelic insight within an embodied system that integrates bodily detection, appraisal, and salience, this account offers a route beyond current stalemates over metaphysics, ineffability, and unverifiable psychological insight.

References

- Fink, Sascha Benjamin. 2022. "Psychedelics Favour Understanding Rather Than Knowledge." *Philosophy and the Mind Sciences* 3 (April). https://doi.org/10.33735/phimisci.2022.9264.
- Garfinkel, Sarah N., Anil K. Seth, Adam B. Barrett, Keisuke Suzuki, and Hugo D. Critchley. 2015. "Knowing Your Own Heart: Distinguishing Interoceptive Accuracy from Interoceptive Awareness." *Biological Psychology* 104 (January): 65–74. https://doi.org/10.1016/j.biopsycho.2014.11.004.
- Kugel, Joshua, Ruben E. Laukkonen, David B. Yaden, Murat Yücel, and Paul Liknaitzky. 2025. "Insights on Psychedelics: A Systematic Review of Therapeutic Effects." *Neuroscience & Biobehavioral Reviews* 173 (June): 106117. https://doi.org/10.1016/j.neubiorev.2025.106117.

- Prinz, Jesse J. 2004. *Gut Reactions: A Perceptual Theory of Emotion*. Philosophy of Mind Series. Oxford University Press.
- Seth, Anil K. 2013. "Interoceptive Inference, Emotion, and the Embodied Self." *Trends in Cognitive Sciences* 17 (11): 565–73. https://doi.org/10.1016/j.tics.2013.09.007.
- Teixeira, Pedro J., Matthew W. Johnson, Christopher Timmermann, et al. 2022. "Psychedelics and Health Behaviour Change." *Journal of Psychopharmacology* 36 (1): 12–19. https://doi.org/10.1177/02698811211008554.

Christian Lains - Hypermodal Realism: Phenomenal Information and Excluded Possibilities

Independent scholar

David Lewis formulated the Hypothesis of Phenomenal Information in 1988: phenomenal experience can exclude possibilities that physical information leaves open. Lewis' point was to reject this non-physicalist notion. Yet a limited version has gained ground recently. Phenomenal Information Externalism (PIE) claims that systems identical in what they actually do can differ in phenomenal information, because they differ in what they exclude. But excluded by what?

Consider three views: First, on an epistemic reading, exclusions are relative to how states are modeled, not fixed by the states themselves. Second, for the modalist, what is excluded are ontologically real possibilities. Third, for the actualist, the exclusions are wholly derivative of the actual state's structure, functions, or dispositions. We can set aside the epistemic reading as openly deflationary, but a similar worry confronts the actualist: If the possibilities excluded are solely derived from the actual, how can they turn around to nontrivially inform it?

They cannot, I argue: on an actualist reading, the possibilities cannot individuate in the ontic sense required for experience to be informative. A realist PIE therefore requires Hypermodal Realism: the excluded possibilities must nontrivially help individuate the actual experience itself. If in addition phenomenal information has "phenomenal powers" (Mørch 2017), we get Hypermodal Power Realism: excluded possibilities not only individuate the actual but influence it.

Theories of consciousness as different as Integrated Information Theory (IIT), quantum approaches, process philosophy, enactive views, Mendola's modal structuralism, and Dretske's teleofunctional externalism can be read as broadly hypermodal. All look beyond local occurrent activity for the exclusions: on teleofunctional accounts, causal history; on enactive accounts, sensorimotor contingencies; on quantum accounts, real potentiae underlying local actuality; on Mendola's view, it is "entwined" with possibility in a modal structure; and on IIT, the irreducible cause-effect structure.

Any hypermodal view can be tested: hold local activity constant, vary the real possibilities, and check whether the theory predicts a phenomenal difference. In Dretske's teleofunctional externalism, history contributes to the possibilities, so a historyless duplicate (Swampman) would not be conscious. In IIT, clamping a silent but activatable unit leaves local activity unchanged while altering the possibilities, so the theory predicts a phenomenal change (Albantakis et al. 2023).

If phenomenal information is epiphenomenal, the test is hard to carry out (Bartlett 2022): How would we know Swampman is a zombie, if it acts the same? We face a circularity objection: If we couple the actual and possible too tightly, the view may collapse into the actualist's. Adopting Hypermodal Power Realism forces the contrast, but requires frameworks in which possible and actual influence each other, perhaps favoring quantum or process theories (Kastner et al. 2018).

If Hypermodal Realism is true, creating artificial consciousness requires more than replicating the brain's computations, functions, or its actual activity. Consciousness engineering would need to design the right modal configurations, and, on the "power" reading, the right modal dynamics. Hypermodal interfaces could test the theory by connecting humans across shared spaces of possibility; to one another, to modally rich consciousness-extension devices, and to conscious AIs.

References

- Albantakis, L., et al. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLOS Computational Biology, 19*(10), e1011465.
- Bartlett, G. (2022). Does integrated information theory make testable predictions about the role of silent neurons in consciousness? *Neuroscience of Consciousness, 2022*(1), niac015.
- Davidson, D. (1987). Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association, 60*(3), 441–458.
- Dretske, F. (1995). *Naturalizing the Mind.* MIT Press.
- Kastner, R. E., Kauffman, S., & Epperson, M. (2018). Taking Heisenberg's potentia seriously. *International Journal of Quantum Foundations, 4*(2), 158–172.

- Lewis, D. (1988). What experience teaches. *Proceedings of the Russellian Society, 13,* 29–57. Mendola, J. (2021). *Experience and Possibility.* Oxford University Press.
- Mørch, H. H. (2017). The evolutionary argument for phenomenal powers. *Philosophical Perspectives, 31,* 293–316.
- Neven, H., Zalcman, A., … Koch, C. (2024). Testing the conjecture that quantum processes create conscious experience. *Entropy, 26*(6), 460.
- Seth, A. K. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences,* 1–42. https://doi.org/10.1017/S0140525X25000032

Mark-Oliver Casper - Re-Valuation of Linearity in 4E Research

University of Kassel

4E cognitive science typically concentrates on non-linear dynamics, viewing continuous flux as the definitive hallmark of situated cognition. This presentation challenges this anti-linear attitude. The 4E framework implicitly relies on a suite of "pro-linearity concepts," such as attractor basins and task constraints, to account for cognitive stability. I argue that these linear regularities are not epiphenomena of underlying dynamics or relics of outdated cognitivism, but rather represent a fundamental functional achievement. Cognitive systems actively compress complex degrees of freedom to secure stable, proportional, and predictable couplings with their environments. To ground this focus on linearity, the argumentation utilizes Minimal Model Explanations (MMEs) and Principle Generalizations (PGs) to demonstrate how linear macro-regularities possess genuine explanatory power. To anchor the epistemological commitments to MMEs and PGs in a concrete methodological foundation, formative Structural Equation Modeling (SEM) is introduced as a solid quantitative framework, often overlooked in the 4E debate, to empirically investigate cognitive phenomena with linear characteristics. By re-evaluating the role of linearity, this study bridges the rift between rich philosophical theory and mathematical rigor, providing the 4E paradigm with a pragmatic pathway to enhance its empirical validity within contemporary cognitive science.

Juan Murillo Vargas - Questions: A Rationalist's Guide

MIT, Department of Linguistics and Philosophy

Recent work in epistemology has identified a class of *question-directed* attitudes: attitudes whose contents are questions (see e.g., Whitcomb, 2010; Friedman, 2013; Carruthers, 2018, 2021, 2020; Lord and Sylvan, 2021; Drucker, 2022; Teague, 2022). These include "interrogative" attitudes like curiosity and wonder, as well as "anti-interrogative" attitudes like staunch agnosticism. On standard assumptions forming a question-directed attitude requires representing a question. But what does it take to represent questions?

This talk argues representing questions requires a sufficiently language-like representational system. More precisely, I defend:

> **Erotetic Rationalism**
>
> An agent can represent a question Q only if she possesses a representational system with language-like constituents.

Erotetic Rationalism predicts that an agent can represent questions only if she possesses a language of thought (Fodor, 1975; Quilty-Dunn et al., 2023) or if she can think in a sufficiently sophisticated system of maps (Camp, 2007, 2009, 2018). Agents who lack these representational resources—including many non-human animals and perhaps large language models (Gottlieb et al., fc)—can't represent questions. My view is thus rationalist in that it emphasizes the role language-like representations play in enabling certain kinds of thought. (In this case: question-directed thought.)

My argument for **Erotetic Rationalism** relies on two premises.

The first is an *inferentialist* meta-semantics for questions (albeit not of the Brandomian variety defended by Millson, 2014; Millson and Risjord, 2026). Call this premise *Erotetic Inferentialism*. On this view, representing questions requires the capacity to draw deductive inferences. I begin by noting that cross-linguistically, questions pattern with logical operators like disjunction, conditionals, and indefinites (see e.g., Haspelmath, 2001; Bhat, 2007; Szabolcsi, 2015; Uegaki, 2018; Starr, 2014; Khoo, 2021). This suggests questions are logical

operators too. Then I argue that since we should be inferentialists about logical operators writ large, we should be inferentialists about questions.

The second premise is a broadly rationalist theory of deductive inference. Call this premise *Inferential Rationalism*. On this view, the capacity to draw deductive inferences requires a representational system with language-like constituents. I argue every empirical model of deductive inference vindicates this claim. Some models—including ones in which deductive reasoning happens in a mental logic (Braine and O'Brien, 1998), language of thought (Quilty-Dunn and Mandelbaum, 2018), or natural logic (Chierchia, 2013)—straightforwardly entail *Inferential Rationalism*. Others—most notably, mental models theory (Johnson-Laird, 1995)—deny that deductive inference requires a fully language-like representational system. But I argue these approaches still posit that deductive inference requires a representational system with language-like constituents.

Erotetic Inferentialism says representing questions requires the ability to draw deductive inferences. *Inferential Rationalism* says this ability requires a representational system with language-like constituents. Put together these two claims entail **Erotetic Rationalism**.

After considering potential counterxamples to my view, I'll conclude by highlighting its consequences for zetetic epistemology. Here is one. An influential thought is that inquiry requires that a subject have a question-directed attitude (Friedman, 2013, 2019). I join several other theorists (Woodard, 2021; Falbo, 2022; Dover, 2024; Carruthers, 2026; Vazard, 2026) in arguing otherwise. Drawing on the empirical literature on curiosity (see e.g., Ackermann et al., 2010; Calhoun et al., 2014; Gipson et al., 2009; Wang and Hayden, 2019; Yan et al., 2018; Damerius et al., 2017; Christensen et al., 2021), I'll argue many non-human animals inquire without possessing a representational system with language like constituents. By **Erotetic Rationalism** it follows such inquirers don't and can't form question-directed attitudes. Hence while some forms of inquiry are question-directed, other forms—including some widespread in the animal kingdom—are not. So we need an epistemology of inquiry expansive enough to cover different forms of inquiry that involve varying degrees of cognitive sophistication (see also Siegel, 2025).

References

- Ackermann, T. F., Kempe, D. S., Lang, F., and Lang, U. E. (2010). Hyperactivity and Enhanced Curiosity of Mice Expressing PKB/SGK-resistant Glycogen Synthase Kinase 3 (GSK-3). *Cellular Physiology and Biochemistry*, 25(6):775–786.
- Bhat, D. (2007). Interrogative–Indefinite Puzzle. In *Pronouns*, pages 226–249. Oxford University PressOxford, 1 edition.
- Braine, M. D. and O'Brien, D. P., editors (1998). *Mental Logic*. Psychology Press, 0 edition.
- Calhoun, A. J., Chalasani, S. H., and Sharpee, T. O. (2014). Maximally informative forag ing by Caenorhabditis elegans. *eLife*, 3:e04220.
- Camp, E. (2007). Thinking with Maps. *Philosophical Perspectives*, 21(1):145–182.
- Camp, E. (2009). A language of baboon thought? In Lurz, A. W., editor, *The Philosophy of Animal Minds*, pages 108–127. Cambridge University Press.
- Camp, E. (2018). Why Maps are Not Propositional. In Grzankowski, A. and Montague, M., editors, *Non-Propositional Intentionality*, volume 1. Oxford University Press.
- Carruthers, P. (2018). Basic questions. *Mind & Language*, 33(2):130–147.
- Carruthers, P. (2020). Questions in Development. In Butler, L. P., Ronfard, S., and Cor riveau, K. H., editors, *The Questioning Child*, pages 6–28. Cambridge University Press, 1 edition.
- Carruthers, P. (2021). The contents and causes of curiosity. *The British Journal for the Philosophy of Science*, page 716609.
- Carruthers, P. (2026). The motoric theory of curiosity. *Mind & Language*, page mila.70017.
- Chierchia, G. (2013). *Logic in Grammar: Polarity, Free Choice, and Intervention*. Oxford University PressOxford, 1 edition.
- Christensen, J. W., Ahrendt, L. P., Malmkvist, J., and Nicol, C. (2021). Exploratory be haviour towards novel objects is associated with enhanced learning in young horses. *Scientific Reports*, 11(1):1428.
- Damerius, L. A., Graber, S. M., Willems, E. P., and van Schaik, C. P. (2017). Curiosity boosts orangutan problem-solving ability. *Animal Behaviour*, 134:57–70.
- Dover, D. (2024). Two kinds of curiosity. *Philosophy and Phenomenological Research*, 108(3):811–832.
- Drucker, D. (2022). Wondering on and with Purpose. In *Oxford Studies in Philosophy of Mind Volume 2*, pages 58–84. Oxford University Press.

- Falbo, A. (2022). Inquiry and confirmation. *Analysis*, 81(4):622–631.
- Fodor, J. A. (1975). *The language of thought*. The Language & thought series. Crowell, New York.
- Friedman, J. (2013). Question-Directed Attitudes. *Philosophical Perspectives*, 27(1):145– 174.
- Friedman, J. (2019). Inquiry and Belief. *Nousˆ*, 53(2):296–315.
- Gipson, C. D., Alessandri, J. J. D., Miller, H. C., and Zentall, T. R. (2009). Prefer ence for 50% reinforcement over 75% reinforcement by pigeons. *Learning & Behavior*, 37(4):289–298.
- Gottlieb, J., Kemp, E., and Trager, M. (fc). How Llms Might Think. *Mind and Language*. Haspelmath, M. (2001). *Indefinite Pronouns*. Oxford University PressOxford, 1 edition.
- Johnson-Laird, P. N. (1995). *Mental models: towards a cognitive science of language, in ference, and consciousness*. Number 6 in Cognitive science series. Harvard Univ. Press, Cambridge, Mass, 6. print edition.
- Khoo, J. (2021). Coordinating *If* s. *Journal of Semantics*, 38(2):341–361.
- Lord, E. and Sylvan, K. (2021). Suspension, Higher-Order Evidence, and Defeat. In Brown, J. and Simion, M., editors, *Reasons, Justification, and Defeat*, pages 116–145. Oxford University PressOxford, 1 edition.
- Millson, J. and Risjord, M. (2026). Asking Questions in the Space of Reasons. In Stovall, P. and Koren, L., editors, *Why and How We Give and Ask for Reasons*, pages 138–164. Oxford University PressNew York, 1 edition.
- Millson, J. A. (2014). *How to Ask a Question in the Space of Reasons:Assertions, Queries, and the Normative Structure of Minimally Discursive Practices*. PhD Thesis, Emory University.
- Quilty-Dunn, J. and Mandelbaum, E. (2018). Inferential Transitions. *Australasian Journal of Philosophy*, 96(3):532–547.
- Quilty-Dunn, J., Porot, N., and Mandelbaum, E. (2023). The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Be havioral and Brain Sciences*, 46:e261.
- Siegel, S. (2025). Wandering Inquiry. *The Journal of Philosophy*.
- Starr, W. (2014). What ”If”? *Philosopher's Imprint*, 14(10):1 – 27.

- Szabolcsi, A. (2015). What do quantifier particles do? *Linguistics and Philosophy*, 38(2):159–204.
- Teague, R. (2022). The problem of closure and questioning attitudes. *Synthese*, 200(5):345.
- Uegaki, W. (2018). A unified semantics for the Japanese Q-particle *ka* in indefinites, questions and disjunctions. *Glossa: a journal of general linguistics*, 3(1).
- Vazard, J. (2026). Inquiry beyond Curiosity: Interest and the Value of Attending. *The Philosophical Quarterly*, page pqag014.
- Wang, M. Z. and Hayden, B. Y. (2019). Monkeys are curious about counterfactual out comes. *Cognition*, 189:1–10.
- Whitcomb, D. (2010). Curiosity was Framed*. *Philosophy and Phenomenological Research*, 81(3):664–687.
- Woodard, E. (2021). The Ignorance Norm and Paradoxical Assertions. *Philosophical Top ics*, 49(2):321–332.
- Yan, Y., Zhaoping, L., and Li, W. (2018). Bottom-up saliency and top-down learning in the primary visual cortex of monkeys. *Proceedings of the National Academy of Sciences*, 115(41):10499–10504.

Jannis Friedrich - Brass-for-Brains: Analog Computation and Device Physics

Institute of Psychology, German Sport University Cologne

Analog computation has increasingly come to the fore in discussions about computation in philosophy of mind, especially relating to its medium independence: whether computational descriptions necessarily abstract from physical implementation. To reconcile differing positions, Chirimuuta (2026) introduces the notion of *practical medium independence* which shifts the perspective from 'in principle' arguments about whether medium-independent descriptions are at all legitimate, to 'in practice' arguments about how dependent computation is on its physical implementation. Digital computation is paradigmatically and characteristically medium independent, whereas analog computation is, in Chirimuuta's terms, practically medium dependent. Building on this approach, here I present a further implication of an 'in practice' approach. It reveals that analog computers are not merely more efficient for exploiting their physical medium, but that they are especially well suited to a particular class of target systems: high-dimensional systems with dynamic continuous variables. This is important for computational models of cognition because these are precisely the systems with which cognition is confronted.

An important feature of analog computation can be framed as "embracing, not combating the device physics" (Manohar, 2015, p. 1). While digital computers have advantages contra analog computers in their programmability, precision, and robustness to noise, they often use the physical constraints inefficiently (Ulmann, 2026). When discussing this embrace of device physics, philosophers of mind often emphasize that digital computers sacrifice the breadth of their representational medium to gain noise-robustness (e.g., Martínez, 2019). They must discretize continuous representational media (such as the magnitude of a voltage) into usually two binary values. This is not the case for analog computers, which are continuous and can thus exploit the full range of the variables. Yet, another aspect of analog's exploitation of device physics has been largely underestimated.

This additional feature of analog computation also arises from taking an 'in practice' approach, specifically analog's advantage for target systems which are high-dimensional, continuous, and dynamic. Under digital computation, additional variables typically also increase computational steps, while analog computation, because of its parallel organization, can accommodate these with additional components, but without increase in computational time. To illustrate this point, I use the maxim of *Brass-for-brains*, which originates from William Thomson's (Thomson, 1881, p. 10) description for his mechanical tide predictor which replaced an inefficient and cumbersome a digital computational approach with a brass machine. This was a potent contribution because the features that burdened formal calculation were easily absorbed by the machine's mechanics. In fact, when dimensions – in this case, additional tidal constants – increased from ten to thirty-seven across models of this machine in later years, computational time remained the same (Parker, 2007).

Beyond deepening the arguments relating to the newly introduced notion of practical medium (in)dependence, this constitutes a critical argument in favour of analog computational models of cognition. Cognition is constantly confronted with the above-described systems, for which digital computation is less naturally suited. Exteroceptive and interoceptive signals are many and often continuous, and organisms must coordinate these under temporal constraints to control a single integrated system: themselves.

References

- Chirimuuta, M. (2026). Neuromorphic Computing and Practical Medium Dependence. *Philosophy of AI*, *2*, 43–60. https://doi.org/10.18716/OJS/PHAI/2026.11564
- Manohar, R. (2015). Comparing Stochastic and Deterministic Computing. *IEEE Computer Architecture Letters*, *14*(2), 119–122. https://doi.org/10.1109/LCA.2015.2412553
- Martínez, M. (2019). Representations Are Rate-Distortion Sweet Spots. *Philosophy of Science*, *86*(5), 1214–1226. https://doi.org/10.1086/705493
- Parker, B. B. (2007). *Tidal Analysis and Predictions* (NOAA Special Publication NOS CO OPS 3; p. 378). National Oceanic and Atmospheric Administration, U.S. Department of

Commerce. https://tidesandcurrents.noaa.gov/publications/Tidal_Analysis_and_Predictions.pdf

- Thomson, W. (1881). The Tide Gauge, Tidal Harmonic Analyser, And Tide Predicter. *Minutes of the Proceedings of the Institution of Civil Engineers*, *65*(1881), 2–25. https://doi.org/10.1680/imotp.1881.22262
- Ulmann, B. (2026). *Analog Computing*. De Gruyter. https://doi.org/10.1515/9783112210178

Ye Yizhou - Does Machine Learning Erode the Open Future?

National University of Singapore, Department of Philosophy

Ratti (2026) argues that the Uniformity of Nature (UoN)—the assumption, underwriting all supervised machine learning (ML), that the world does not change drastically—generates an ethical problem existing ML ethics has neglected. Debates over algorithmic bias and epistemic risk (Fazelpour and Danks 2021; Biddle 2022) operate at the level of Ratti's Canons of Inductive Inference (CIIs), leaving the impression that better design resolves any ethical difficulty. Attending to UoN reveals a harm internal to prediction: in agency-salient contexts, treating future conduct as projectible from past observations bypasses the individual's own plans and goals, eroding their "substantial freedom to an open future" (Ratti 2026, 375). Ratti concludes that we should reject ML tools in such contexts rather than reform their CIIs (2026, 377).

I argue this does not follow. Ratti's reasoning reconstructs as: (P1) if ML tools generate ethical harms H that no CII configuration can in principle eliminate, we have grounds to reject them in agency-salient contexts; (P2) ML tools generate H, the erosion of the open future; so (C) we have such grounds. Deriving C requires (P3): no CII configuration can eliminate H. Ratti suggests P3 without defending it, so I construct a defense whose load-bearing claim is that any operation adopting F—the assumption that the future is predictable from the past—closes off the individual's future.

The defense fails, because "eroding the open future" admits two readings. Non consequentially, every F-adopting prediction erodes agency whatever its outcome; but this misclassifies cases where prediction gives agency a substantive role—an ML tool recommending leniency for a defendant's genuine repentance, say. Only the consequentialist reading—erosion as blocking an individual from a deserved outcome—is plausible; and on it some F-adopting operations do not erode. Picking these out is the work of suitable CIIs. Objecting that today's CIIs will fail as conceptions of desert change either demands a permanence human institutions equally lack, or demands context-sensitive CIIs—itself CII design.

Can this be realized within ML deployment itself? Reserving humans as final judges works but does not scale. I propose that ML tools capable of autonomous learning (LeCun 2022) can themselves occupy the position, given two requirements: parameters adjustable during deployment, and adjustment aligned with human values (Gabriel 2020). The upshot is missed by a literature focused on risk (Bostrom 2014; Bengio et al. 2025): AI agency is not merely a hazard but a precondition for respecting human agency within ML tools themselves.

References

- Bengio, Yoshua, Michael Cohen, Damiano Fornasiere, et al. 2025. "Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?" arXiv:2502.15657. Preprint, arXiv, February 24. https://doi.org/10.48550/arXiv.2502.15657.
- Biddle, Justin B. 2022. "On Predicting Recidivism: Epistemic Risk, Tradeoffs, and Values in Machine Learning." Canadian Journal of Philosophy 52 (3): 321–41. https://doi.org/10.1017/can.2020.27.
- Bostrom, Nick. 2014. Superintelligence: Paths, Dangers, Strategies. Oxford University Press.
- Fazelpour, Sina, and David Danks. 2021. "Algorithmic Bias: Senses, Sources, Solutions." Philosophy Compass 16 (8): e12760. https://doi.org/10.1111/phc3.12760. Gabriel, Iason. 2020. "Artificial Intelligence, Values, and Alignment." Minds and Machines 30 (3): 411–37. https://doi.org/10.1007/s11023-020-09539-2. LeCun, Yann. 2022. "A Path towards Autonomous Machine Intelligence Version 0.9.2, 2022-06-27." Open Review 62 (1): 1–62.
- Ratti, Emanuele. 2026. "Machine Learning and the Ethics of Induction." In Philosophy of Science for Machine Learning, edited by Juan M. Durán and Giorgia Pozzi, vol. 527. Synthese Library. Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-03083-2_17.

Yizhuo Sun - Beyond Static Representation: A Dynamic Mechanistic Account of Representational Vehicles in Predictive Processing

University of Chinese Academy of Sciences

Predictive processing (PP) has long generated debates concerning whether its generative models constitute genuine representations. Previous discussions have primarily focused on the disagreement between strong representationalism (in its conservative form) and anti-representationalism (in its radical form). However, merely addressing whether predictive processing involves representations is insufficient to resolve the representational problem of PP. If predictive processing is understood as involving representations, at least two further questions must be addressed. First, scientific representation involves representational vehicles, representational targets, and the representational relation between them; therefore, the corresponding components must be identified within predictive processing's generative models. Second, it must be explained how representational vehicles realize representations of their targets and how the representational role of generative models enters into the mechanistic explanation of predictive processing.

Regarding the second issue concerning the kind of representation involved, existing research has proposed several accounts, including content-based representation, structural representation, degrees of representation, and pragmatic or inferential accounts of representation. However, a more fundamental question has received insufficient attention: what constitutes the representational vehicle in predictive processing? Existing theories of representation often implicitly assume that representational vehicles are relatively stable internal structures, such as neural maps, localized neural patterns, or internal models. Yet this assumption is in tension with predictive processing's emphasis on hierarchical inference, prediction-error minimization, precision weighting, and recurrent neural interactions.

This paper argues that predictive processing requires a reconsideration of the possible units that can serve as representational vehicles. It proposes a dynamic mechanistic account of representational vehicles, according to which representations in predictive processing are not

carried by a static neural structure, but are realized by temporally extended patterns of neural organization, including prediction-error signaling, precision weighting, and recurrent feedforward–feedback dynamics. The visual system provides a case study illustrating this claim and further demonstrates that representational explanations can be compatible with mechanistic neuroscience. On this account, representational content is not an additional semantic label imposed upon neural activity; rather, it consists in the functional-structural role that neural states play within processes of prediction, error correction, learning, and action regulation.

This account preserves an important insight from structural theories of representation: generative models represent by maintaining causal-probabilistic relations among environmental causes, bodily states, and sensory outcomes. At the same time, it incorporates insights from pragmatic and inferential approaches, according to which representational capacities depend on an agent's ability to use such structures for prediction and behavioral regulation. By shifting the focus from whether predictive processing involves representations and what kinds of representations it involves, toward how representational vehicles are possible in predictive processing and how representational kinds are mechanistically realized, this paper argues that predictive processing challenges the traditional static conception of representation. It suggests that representational vehicles need not be discrete internal topological structures, but may instead consist of dynamic mechanisms that enable agents to predict, regulate, and interact with their environments. Addressing this more fundamental issue provides a more adequate foundation for evaluating both the existence and the nature of representation in predictive processing.

Rongsheng Liang, Yizhuo Sun - Abstraction Without Causal Loss: A Mechanistic Reconstruction of Minimal Models of Cognition

Ghent University, University of Chinese Academy of Sciences

Since Batterman and Rice (2014), minimal models have largely been understood as non-causal. On their view, models like Fisher's sex ratio model explain not by representing causal mechanisms, but by showing why microscopic details are irrelevant to macroscopic behavior. The key tool is the Renormalization Group (RG): coarse-graining eliminates irrelevant degrees of freedom and groups diverse systems into the same universality class. In recent discussions of minimal models of cognition, this generates a dilemma. If minimal models of cognition—e.g., McGivern's (2020) chemotactic oil droplets—explain by discarding mechanistic detail, must we abandon mechanistic explanation at the level of minimal cognition?

This paper defends a weakly mechanistic interpretation. I introduce the concept of "causal tether": a model remains causally tethered if its abstract variables support counterfactual dependencies that track ontic dependence relations (Povich 2018), and if omitting lower-level details is justified by their irrelevance. RG coarse-graining does not sever causal ties; each transformed Hamiltonian still represents the same physical system's long-range interactions at a coarser scale (Sullivan 2019). Thus, a universality class characterizes the effective causal structure shared across systems.

Applying this to Fisher's model shows that linear substitution cost captures a counterfactually stable trade-off—if male cost increased, the equilibrium ratio would shift—answering what-if-things-had-been-different questions without genetic specifics. Turning to active materials: oil droplets and Janus particles exhibit chemotaxis and maze-solving. I argue they satisfy a weak version of Kaplan and Craver's (2011) model-to-mechanism mapping. Their abstract variables (pH gradient, interfacial tension) correspond to real organizational features— effective sensorimotor coupling and adaptive feedback—not neural components. Reversing the gradient changes movement; blocking surfactant stripping stops it. RG-style

reasoning shows that detailed micro-chemistry is irrelevant; what matters is the causal topology of interfacial dynamics.

Consequently, minimal models of cognition need not be non-causal. Weakly mechanistic explanation only requires preserving the effective causal organization sufficient to produce the target behavior, with RG as a key method for identifying it. This middle ground validates the explanatory autonomy of abstract, embodied models while avoiding strong reductionism, and suggests that minimal cognition is unified by causal-organizational, not strictly biological, criteria.

References

- Batterman, R. W., & Rice, C. C. (2014). Minimal model explanations. Philosophy of Science, 81(3), 349-376.
- Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. Philosophy of science, 78(4), 601-627.
- McGivern, P. (2020). Active materials: minimal models of cognition?. Adaptive Behavior, 28(6), 441-451.
- Povich, M. (2018). Minimal models and the generalized ontic conception of scientific explanation. The British Journal for the Philosophy of Science, 69(1), 117-137. 5.
- Rice, C. (2022). Modeling multiscale patterns: active matter, minimal models, and explanatory autonomy. Synthese, 200(6), 432.
- Sullivan, E. (2019). Universality caused: The case of renormalization group explanation. European Journal for Philosophy of Science, 9(3), 36.

MANJARI CHAKRABARTY - THE POSSIBILITY OF CONSCIOUSNESS-BRAIN INTERACTION RECONSIDERED

VISVA-BHARATI (A CENTRAL UNIVERSITY), SANTINIKETAN, INDIA, 731235

Several long-standing philosophical debates rest on either a stronger or an oversimplistic understanding of scientific practice. The present study addresses one such enduring controversy surrounding whether the causal closure argument (of the physical world), more specifically, the neurophysiological "no-gap" argument, can be a serious threat to the possibility of mental-to-physical causation or consciousness-brain interaction and, by implication, to the possibility of conscious states' being strongly emergent phenomena. Three different kinds of reasons, drawn from the historiography of scientific practice, but largely unexamined by philosophers engaged in the controversy, motivate this paper.

One may characterize strong emergence by the dual claims that emergent phenomena (e.g., non-physical, conscious, or mental states) are distinct from their underlying basal phenomena (e.g., the brain or the nervous system) *and* that emergent phenomena are causally relevant in the physical domain. Those who consider conscious or mental phenomena as strongly emergent phenomena naturally admit the possibility of mental-to-physical causation. In contrast, those who challenge the possibility of mental-to-physical causation often advance the neurophysiological "no-gap" argument, which is a specific application of the broader causal closure argument. The argument assumes that neurophysiology need not cross the boundary of the physical and appeal to *sui generis* mental (i.e., non-physical) causes to provide the complete and immediate cause of any neural event in the causal chains (of neural events) that generate bodily movements. Hence, the argument, entailing the causal closure principle that if a physical event has a cause occurring at time *t*, it has a sufficient physical cause at *t* (Kim, 2011, p. 214), constitutes a serious threat to the possibility of mental-to physical causation or consciousness-brain interaction. The threat arises basically from the argument's essential premise– the principle of causal closure.

Of the three motivating reasons for this paper, the first section discusses one, namely, that the alleged validity of the causal closure principle can be challenged. This crucial premise of the neurophysiological no-gap argument draws its principal force from the conservation laws of classical physics. Averall and Keating's (1981) review of the conservation laws testifies to the contrast, almost unnoticed by the philosophers, between the philosophical formulations (e.g., Dennett 1978) of both the law of conservation of linear momentum as well as the law of conservation of energy (also known as the First Law of Thermodynamics) and the textbook formulations of these laws as available text -books like Herbert Goldstein's *Classical Mechanics*. Upon examining how the conventional, philosophical interpretations of those laws have metaphysical implications that beg the question against mental-to-physical causation or consciousness-brain interaction, we argue that the neurophysiological no-gap argument loses much of its logical rigor as its central premise (viz., the causal closure principle) rests on much stricter (than what is necessary for physics), and question-begging formulations of those physical laws.

The second section of this paper provides another (motivating) reason by addressing the recurrent objection that strong emergence is clearly incompatible with the principle of the universality of physical law, viz., that physical principles completely determine the behaviour of the systems to which they apply. Once again, the general philosophical understanding of the classical physical laws presupposed in this persistent objection appears questionable if one considers the (internal) historiography of scientific practice into account. Accepting that physical principles apply universally does not necessarily mean accepting that they fully determine the motions of the systems they govern. Instead, the universal applicability of physical principles implies that, when acting on their own (that is, when not acting in concert with other forces, which hardly ever happens in actuality), these principles fully determine the motions of any system they govern. Evidently, the principle of the universality of physical law─ with which strong emergence is incompatible─ adopts a necessitarian, highly abstracted, largely context-free formulation of physical laws that might work for some possible worlds imagined by the philosophers but not for the messier contexts of the actual world of physics (Bishop, 2006). What hitherto has been neglected by the philosophers is that strong emergence could be compatible with an alternative principle of universality, the principle of *ubiquity of physics,* accepted by many physicists themselves. Under the ubiquity of physics, physical principles apply universally and constrain the motions of particular systems they govern, but they may not fully determine them (Hendry, 2010).

Since the ubiquity of physics does not entail the principle of causal closure, strong emergence could be compatible with this alternative principle of universality.

In the concluding section, we cite Karl Popper's 1993 hypothesis as the third (motivating) reason for the possibility (not merely theoretical but as neurophysiologically testable) of mental-to-physical causation or consciousness-brain interaction. The Popperian 1993 hypothesis, one of the most neglected stances, contains a mind-force analogy, in which consciousness shares many properties with physical forces and interacts with the brain through its complex electromagnetic force fields. Two features of this Popperian hypothesis stand out in particular: its strong emergentist character, its non-adherence to the conventional philosophical interpretation of physical laws, and its clear distinction from recent hypotheses that identify minds with electromagnetic fields (e.g., Jones 2013).

References

- Averill, E. and Keating, B. (1981). Does Interactionism Violate a Law of Classical Physics? *Mind*, New Series, 90 (357) 102–107.
- Bishop, R.C. (2006). The hidden premiss in the causal argument for physicalism, *Analysis,* 66 (1): 44–52.
- Dennett, D.C. (1978). Current issues in the philosophy of mind, *American Philosophical Quarterly*, 15(4), 249–261.
- Hendry, R.F. (2010). Ontological Reduction and Molecular Structure, *Studies in History and Philosophy of Modern Physics* 41: 183–191.
- Jones, M.W. (2013). Electromagnetic-field theories of mind, *Journal of Consciousness Studies*, 20 (11–12): 124–149.
- Kim, J. (2011). *Philosophy of Mind,* 3rd Edition, Colorado: Westview Press.
- Popper, K.R. et. Al. (1993). A Discussion of the Mind-Brain Problem, *Theoretical Medicine,* 14:167-180.

Christopher Stratman - An Argument for Eliminativism About Occurrent Conscious Beliefs

University of Nebraska-Lincoln

The distinction between dispositional (or standing) beliefs and occurrent beliefs is widely accepted but remains conceptually obscure. I argue that part of this obscurity stems from conflating occurrent beliefs with conscious beliefs. Against this assumption, I defend eliminativism about occurrent conscious beliefs. I first argue that genuine beliefs must be capable of persisting through time, whereas conscious thought consists of temporally extended episodes of entertaining propositional contents. I then strengthen the support for this claim by extending Helton's (2018) Revisability View of belief, arguing that because belief revision necessarily takes time, genuine beliefs must be capable of persistence. If successful, this account replaces occurrent conscious beliefs with conscious episodes of cognition, clarifies the ontology of belief by showing that there are no occurrent conscious beliefs, and by removing an important source of confusion in debates over consciousness and phenomenal intentionality.

Bongrae Seok - Neuroscience and Game Theory of Mindfulness Meditation: Embodied Moral Psychology of Buddhism

Alvernia University

Many brain imaging studies have shown that the brains of mindfulness meditators exhibit strong activity in the interoceptive limbic pathway, which includes the thalamus, the insula, and primary sensory regions (Farb et al. 2012). These brain areas typically deal with and process affective and

embodied states of the mind before they are further elaborated or evaluated in other regions of the brain. From these observations, it seems that mindfulness grounds the mind in its bodily experience, protecting it from habitual evaluation, reflexive reactions, and disruptive emotions. Both conventional cognitive appraisal and habitual reactions are controlled by this strategic shift in attention to less elaborated but more embodied cognitive stages of information processing afforded by mindfulness. However, this attentional shift is quite selective and regulated. (This process is often intentionally guided and is neither passive nor open.)

In the Ultimatum Game, players typically become angry and seek vengeance when they receive an unfair offer (less than 20% of the originally available money). Rather than accepting these offers as the ideal rational economic agent would, their insulae light up in response to their sense of disgust at the unfairness, which prompts anger and vengeance. Subjects who practice mindfulness meditation accept unfair offers significantly more often than non-meditative subjects, perhaps by overcoming this initial anger. However, in Kirk and colleagues' (2011) experiment, another group of subjects who did not practice mindfulness meditation could control their anger and accept unfair offers. It turned out that this group of subjects, who were considered rational, focused on the financial aspect of the offers because their dorsolateral prefrontal cortex was highly active when they accepted unfair offers. The dorsolateral prefrontal cortex typically supports the executive functions of the mind that facilitate instrumentally rational decisions. For this group, therefore, the dorsolateral prefrontal cortex supports reason driven decisions, i.e., cognitive control and focus on objective benefits.

These differential brain activation patterns demonstrate that although mindfulness generates rational decisions, it does not come from rational deliberation or calculation of objective benefits, but rather from careful attention to the embodied and affective states of the mind at its foundational level. Through mindfulness meditation, the mind is firmly grounded in the body and prevented from wandering into the realm of ungrounded or inauthentic emotions, such as impulsive anger, fear, or vengeance.

Mindfulness is often characterized as bare attention and blind openness and is understood as a means to achieve psychological comfort and an undisturbed mind. However, empirical studies of mindfulness show that it is associated with the affective and embodied state of the mind (Ehman et al. 2025). The focused state of the embodied mind generates a distinctive type of affective consciousness. For instance, an embodied awareness of others' pain and suffering, as many psychologists have reported, fosters empathic concern and prosocial behavior (Singer et al. 2004). Therefore, Buddhist moral psychology, when based on mindfulness meditation, takes a unique approach that is grounded in this special form of embodied moral psychology.

References

- Analayo. (2003). Satipatthana: the direct path to realization. Birmingham, UK: Windhorse Publications.
- Arkes, Hal & Blumer, Catherine (1985). "The Psychology of Sunk Cost". *Organizational Behavior and Human Decision Process,* 35, 124–140.
- Arkes, Hal & Hutzel, Laura (2000). "The Role of Probability of Success Estimates in the Sunk Cost Effect". *Journal of Behavioural Decision Making,* 13 (3), 295–306.
- Astin, J. A. (1997). Stress reduction through mindfulness meditation: Effects on psychological symptomatology, sense of control, and spiritual experiences. *Psychotherapy And Psychosomatics*, *66*(2), 97–106.
- Bishop, S., Lau, M., Shapiro, S., Carlson, L., Anderson, N., Carmody, J., et al. (2004). Mindfulness: A proposed operational definition. *Clinical Psychology: Science and Practice*, 11(3), 230–241.
- Bodhi, Bhikkhu (2011, May). What does Mindfulness really mean? A canonical perspective. *Contemporary Buddhism, Vol. 12, 1*, 19-39.

- Boyatzis, R. & McKee, A. (2005). *Resonant leadership: Renewing yourself and connecting with others through mindfulness, hope, and compassion*. Boston, MA: Harvard Business School Press.
- Brown, C. A. & Jones, A. K. (2010). Meditation experience predicts less negative appraisal of pain: electrophysiological evidence for the involvement of anticipatory neural responses. *Pain 150*. 428–438.
- Cahn, B. R., Polich, J. (2006). Meditation states and traits: EEG, ERP, and neuroimaging studies. Psychological Bulletin 132, 180 –211.
- Cresswell, D. Way, B., Eisenberger, N., & Lieberman, M. (2007). Neural Correlates of Dispositional Mindfulness during Affect Labeling. *Psychosomatic Medicine, 69*, 560-565.
- Damasio, A. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Penguin Putnam.
- Davidson, Richard, Lutz, Antoine (2007, September) Buddha's Brain Neuroplasticity and Meditation. IEEE SIGNAL PROCESSING MAGAZINE , p. 172- 176
- Ehmann, S., Sezer, I., Treves, I. N., Gabrieli, J. D. E., & Sacchet, M. D. (2025). Mindfulness, cognition, and long term meditators: Toward a science of advanced meditation. *Imaging neuroscience* (Cambridge, Mass.), 3, IMAG.a.82.
- Farb, N. S., Anderson, A. K., Bloch, R. T., & Segal, Z. V. (2011). Mood-linked responses in medial prefrontal cortex predict relapse in patients with recurrent unipolar depression. *Biological Psychiatry*, *70*(4), 366-372.
- Farb, N. S., Anderson, A. K., & Segal, Z. V. (2012). The Mindful Brain and Emotion Regulation in Mood Disorders. *Canadian Journal of Psychiatry, 57(2)*, 70-77.
- Garfield, Jay. (2012, Summer). Mindfulness and Ethics: Attention, Virtue, and Perfection. In Michael Zimmerman, Christof Spitz, and Stefan Schmidt (eds.) *Achtsamkeit – Ein buddhistisches Konzept erobert die Wissenschaft*. Verlag Hans Huber.
- Germer, C., Siegel, R., & Fulton, P. (Eds.) (2005). *Mindfulness and Psychotherapy*. New York: Guilford Press.
- Goldstein, J. (1993). *Insight meditation: the practice of freedom*. Boston, MA: Shambhala Press. Goleman, D. (1995) *Emotional Intelligence: Why It Can Matter More Than IQ*. New York: Bantam Book.

- Grant, J. A. & Rainville, P. (2009). Pain sensitivity and analgesic effects of mindful states in Zen meditators: a cross sectional study. *Psychosomatic Medicine 71*. 106 – 114.
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science* 293(5537), 2105–2108.
- Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, 107(3), 1144–1154.
- Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2), 389–400.
- Gunaratana, Henepola (2002). *Mindfulness in plain English*. Boston, MA: Wisdom Publications.
- Haase, L., Thom, N. J., Shukla, A., Davenport, P. W., Simmons, A. N., Stanley, E. A., Paulus, M. P., & Johnson, D. C. (2016). Mindfulness-based training attenuates insula response to an aversive interoceptive challenge. *Social cognitive and affective neuroscience*, 11(1), 182–190.
- Hafenbrack, A. C., Kinias, Z., & Barsade, S. G. (2014). Debiasing the Mind through Meditation: Mindfulness and the Sunk-Cost Bias. *Psychological Science* 25(2), 369-376.
- Haidt, Jonathan (2001). "The emotional dog and its rational tail: A social intuitionist approach to moral judgment." Psychological Review, 108, 814-834.
- Haidt, Jonathan (2012). The Righteous Mind: Why Good People are Divided By Politics and Religion. New York: Pantheon Books. pp. 9–11.
- Haidt, Jonathan & Craig, Joseph (Fall 2004). "Intuitive ethics: how innately prepared intuitions generate culturally variable virtues". Daedalus 133 (4): 55–66.
- Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. Social Justice Research, 20, 98-116.
- Hanstede, M., Gidron, Y., & Nyklíček, I. (2008). The effects of a mindfulness intervention on obsessive-compulsive symptoms in a non-clinical student population. *Journal of Nervous and Mental Disease*, *196*(10), 776-779.

- Kabat-Zinn, J. (1994). *Wherever You Go, There You Are: Mindfulness Meditation in Everyday Life*. Hyperion: New York.
- Kabat-Zinn, J. (2003, Summer). Commentaries on Baer: Mindfulness-Based Interventions in Context: Past, Present, and Future. *Clinical Psychology: Science and Practice, 10-2*. 144-156
- Kirk, U., Gu, X., Sharp, C., Hula, A., Fonagy, P., & Montague, P. R. (2016). Mindfulness training increases cooperative decision making in economic exchanges: Evidence from fMRI. *NeuroImage*, 138, 274–283.
- Kirk U, Downar, J., & Montague, P. R. (2011). Interoception drives increased rational decision-making in meditators playing the ultimatum game. *Frontiers in Neuroscience,* 5-49, 1-11.
- Koyama, T., McHaffie, J.G., Laurienti, P. J., Coghill, R. C. (2005). The subjective experience of pain: where expectations become reality. *Proceedings of National Academy of Science*, USA 102:12950 –12955.
- Knox, R. E., & Inkster, J. A. (1968). Postdecision dissonance at post time. *Journal of Personality and Social Psychology* 8 (4), 319–323.
- Langer, E. (1989). *Mindfulness.* Cambridge, MA: Da Capo Press.
- Laneri, D., Krach, S., Paulus, F. M., Kanske, P., Schuster, V., Sommer, J., & Müller-Pinzler, L. (2017). Mindfulness meditation regulates anterior insula activity during empathy for social pain. *Human brain mapping*, 38(8), 4034– 4046.
- Lutz, A., McFarlin, D. R., Perlman, D. M., Salomons, T. V., & Davidson, R. J. (2013). Altered anterior insula activation during anticipation and experience of painful stimuli in expert meditators. *Neuroimage*, 64(1), 538-546.
- Marlatt, G. A., & Kristeller, J. L. (1999). Integrating spirituality into treatment: resources for practitioners. In *Mindfulness and Meditation* (ed. W. R. Miller), p. 68. American Psychological Association : Washington, DC.
- Mahasi Sayadaw (1971, reprint 2006). *Practical insight meditation*. Kandy, Sri Lanka: Buddhist Publication Society. http://books.google.com/ (page references are to the online version).
- Moll, J., de Oliveira-Souza, R. & Eslinger, P. (2003). Morals and the Human Brain: A Working Model. *Neuroreport*, 14, 299-305.
- Moll, J. & de Oliveira-Souza, R. (2007). Moral judgments, emotions and the utilitarian brain. *Trends in Cognitive Science*, 11(8), 319-321.

- Monier-Williams, (reprint. 2005). *Sanskrit-English dictionary*. Delhi: Motilal Banarsidass.
- Rhys Davids, T. W. & C. A. F. (trans. 1910). *Dialogues of the Buddha.* London: Pali Text
  Society. http://www.levityisland.com/buddhadust/www.buddhadust.org/TheSatipatthana/SettingUpMindfulness.htm.
- Linehan, M. M. (1993). *Cognitive behavioral therapy of borderline personality disorder*. New York: Guilford Press.
- Schwartz, J. (2011). *You Are Not Your Brain: The 4-Step Solution for Changing Bad Habits, Ending Unhealthy Thinking, and Taking Control of Your Life*. New York: Avery,
- Patel, S. R., Carmody, J., & Simpson, H. (2007). Adapting mindfulness-based stress reduction for the treatment of obsessive-compulsive disorder: A case report. *Cognitive And Behavioral Practice*, *14*(4), 375-380.
- Perlman D. M., Salomons, T. V., Davidson R. J., Lutz, A. (2010). Differential effects on pain intensity and unpleasantness of two meditation practices. *Emotion 10*, 65–71.
- Nolen-Hoeksema S. (2000). The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. *Journal of Abnormal Psychology 109*, 504–511.
- Sanfey, Alan G., Rilling, James K., Aronson, Jessica A., Nystrom, Leigh E., & Cohen, Jonathan D. (2003). The Neural Basis of Economic Decision Making, *Science*, *300*, 1755-1758.
- Segal, Z. V., Williams, J. M. G., & Teasdale, J. D. (2002). *Mindfulness-Based Cognitive Therapy for Depression*. New York: Guilford Press.
- Seok, B. (2013) *Embodied Moral Psychology and Confucian Philosophy*. Lanham, MD: Roman Littlefield.
- Sheline, Y. I., Price, J. L., Yan, Z., and Mintun, M. A. (2010). Resting-state functional MRI in depression unmasks increased connectivity between networks via the dorsal nexus. *Proc. Natl. Acad. Sci. U.S.A.* 107, 11020–11025.
- Singer, T., Seymour, B., O'Doherty, J., Kauber, H., Dolan, R. J., & Frith, C. (2004). Empathy for Pain involves the affective but not sensory components of pain. *Science 303-5661*, 1157-1162.

- Sugawara, A., Katsunuma, R., Terasawa, Y., & Sekiguchi, A. (2024). Interoceptive training impacts the neural circuit of the anterior insula cortex. *Translational psychiatry*, 14(1), 206.
- Teper, R., & Inzlicht, M. (2013). Meditation, mindfulness and executive control: the importance of emotional acceptance and brain-based performance monitoring. Social Cognitive & Affective Neuroscience, 8(1), 85-92.
- Thomas, J. W., & Cohen, M. (2014). A methodological review of meditation research. *Frontiers in psychiatry*, 5, 74. Wallace, Alan B. (2008, Spring). A Mindful Balance. *Tricycle, Spring*, 60 – 111.
- Walsh, R., & Shapiro, S. (2006). The Meeting of Meditative Disciplines and Western Psychology: A Mutually Enriching Dialogue. *American Psychologist*, 61(3), 227-239.
- Watkins, E., Moberly, N. J., & Moulds, M. L. (2008). Processing mode causally influences emotional reactivity: Distinct effects of abstract versus concrete construal on emotional response. *Emotion*, 8. 364–378.
- Zeidan, F., Martucci, K.T., Kraft, R.A., Gordon, N. S., McHaffie, J. G., & Coghill, R. C. (2011). Brain mechanisms supporting the modulation of pain by mindfulness meditation. *Journal of Neuroscience,* 31(14): 5540–5548.

João César Ramos - Acquaintance, Knowledge and Normativity

Federal University of Minas Gerais (UFMG)

Some philosophers defend the existence of a *sui generis*, non-propositional form of objectual knowledge centered on the notion of acquaintance, which is typically understood as an experiential relation to certain objects and properties. This idea is traditionally associated with Bertrand Russell. Russell's views on acquaintance were heavily criticized throughout the twentieth century, largely because he maintained that the only possible objects of acquaintance are sense-data. More recently, however, philosophers such as Matt Duncan (2020, 2021, 2026), Emad Atiq (2025), Anna Giustina (2022), and Uriah Kriegel (2024) have renewed interest in acquaintance knowledge, seeking a more adequate account of the philosophical phenomenon in question.

A central explanandum for these philosophers is how the relation of acquaintance can, by itself, constitute knowledge. One promising strategy is to argue that acquaintance knowledge shares certain features and functional roles with propositional knowledge, since the latter is regarded as the paradigmatic epistemic case in analytic philosophy. One such feature is epistemic normativity. Here, however, a tension arises. Acquaintance has historically been treated as an infallible relation. Accordingly, it is often argued that there is no room for error in acquaintance. This stands in contrast to propositional knowledge, which is typically—especially in the post-Gettier literature—conceived as a success state that is understood against the background of corresponding failure cases, such as false belief. This discrepancy may give rise to skepticism about the epistemic credentials of the objectual knowledge at issue.

A promising solution to this problem has been proposed by Atiq (2025). His suggestion is that experience is an active rather than a passive phenomenon, one that is responsive to agency and responsibility. One way of developing this idea is through Susanna Siegel's (2017) notion of *epistemic charge*. For example, in cases of intoxication or when perception is influenced by biases and prejudices, an agent may fail to experience what they ought to experience. This feature of experience—and, consequently, of acquaintance knowledge—reintroduces

epistemic normativity and thereby supports the epistemic standing of this form of knowledge. The proposed talk will survey and defend this proposal, further endorsing the legitimacy of objectual knowledge as an epistemic type.

References

- ATIQ, Emad H. (2025). The Normative Profile of Knowledge by Acquaintance. Philosophy and Phenomenological Research.
- DUNCAN, Matt (2020). Knowledge of things. Synthese 197 (8):3559-3592.
- _____ (2021). Acquaintance. Philosophy Compass 16 (3):e12727.
- _____ (2026). Present to the Mind: Acquaintance and Its Significance. Oxford University Press. GIUSTINA, Anna (2022). Introspective knowledge by acquaintance. Synthese 200 (2):1-23.
- KRIEGEL, Uriah (2024). Knowledge-by-Acquaintance First. Philosophy and Phenomenological Research 109 (2):458-477.
- SIEGEL, S. (2017). The Rationality of Perception. Oxford University Press.

Paula Muhr - From Local Lesions to Dynamic Distributed Networks Framing Brain Dysfunction in Neuroimaging Research on Functional Neurological Disorder

Brand University of Applied Sciences Hamburg

In the humanities, the dominant view is that hysteria, an age-old disease with heterogeneous somatic symptoms and unknown aetiology, disappeared in the 20th century. This view ignores the fact that hysteria, recently renamed functional neurological disorder (FND), has since the mid-1990s become the object of intensifying medical research investigating its underlying neurophysiological mechanisms using neuroimaging technologies. Functional magnetic resonance imaging (fMRI) has been the dominant technology, enabling noninvasive measurement of brain activity in patients. By linking FND symptoms to measurable neuropathology, this research has introduced a new collective frame of FND as a brain-based disorder caused by still-unknown brain dysfunction, displacing the previously dominant Freudian view of hysteria as a purely psychological illness (Muhr 2022).

This paper examines the evolving framing (Entman 1993; Gläser 2024) of brain dysfunction within functional neuroimaging research on FND from the mid-1990s to the present, while also linking it to its historical precursor in the 19th-century neurologist J.-M. Charcot, who attributed hysteria to a hypothesised functional lesion. Drawing on document analysis of key fMRI studies, I argue that advances in neuroimaging have progressively transformed the experimental framing of brain dysfunction in FND while simultaneously complicating the epistemic boundary between brain function and dysfunction.

Early studies conceptualised dysfunction as a localised inhibition of specific brain regions, echoing Charcot's framing. Subsequent work reconceived dysfunction as arising from disturbed interactions between multiple brain centres, while more recent research locates dysfunction in large-scale neural networks, using static resting-state paradigms and complex connectivity analyses. The latest shifts frame dysfunction either as a dynamic property of temporally evolving network configurations using dynamic resting-state fMRI or combine functional and structural imaging findings. This increasing conceptual

complexification in experimental framing of brain dysfunction in FND research, I contend, has been driven by the technical capacities of neuroimaging and increasingly sophisticated statistical modelling.

The shift towards increasingly dynamic and multimodal experimental framings of brain dysfunction in FND raises questions about how the boundary between function and dysfunction is epistemically negotiated. Because this boundary is not defined through discrete pathology but through patterns inferred from data-intensive analyses, the medical and neurocognitive interpretation of these patterns often remains ambiguous. The technically mediated visibility of the newly discovered patterns of dynamic network dysfunction in FND patients' brains, while signalling empirical progress, remains coupled with a lack of conceptual understanding of these findings. I suggest that the current framing of dysfunction as a dynamic network property complicates the boundary between brain function and dysfunction in FND rather than offering a clear-cut solution.

This paper contributes to philosophy of medicine by highlighting how technological mediation reshapes concepts of function and dysfunction in FND and by underscoring the epistemic challenges of defining brain dysfunction in data-driven neuroscience. These implications extend beyond FND to other neurological and psychiatric conditions.

References

- Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*(4), 51-58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x
- Gläser, J. (2024). How can we make 'research quality' a theoretical concept? *Research Evaluation*, *33*. https://doi.org/10.1093/reseval/rvae038
- Muhr, P. (2022). From photography to fMRI: Epistemic functions of images in medical research on hysteria. transcript.

Kira Miller - An Eye for an 'I': Investigating the Relationship between Embodiment and Agency, and Exploring the Possibilities for Artificial Intelligence Models

University of Amsterdam

This paper examines the relationship between embodiment and agency with specific reference to Artificial Intelligence (AI) models. It argues that embodiment is a necessary condition of agency because agents must be able to sense and manipulate the external world that they inhabit. Humans (and a number of other intelligent beings) are capable of intentionally acting in a particular way in pursuit of particular goals. Beyond the capacity that most organisms possess to move towards pleasure (food, sunlight, water, etc) and away from pain (predators, danger, poisons, etc), some organisms are able to internally represent an expectation of how a particular action might transpire, allowing them to choose the action which might most align with their values, desires, and goals. Further still, infants develop knowledge and understanding through directly experiencing and manipulating objects.

In this paper, I distinguish between being a body, having a body, and being embodied, and argue that embodiment is only achieved when beings can sense and manipulate the external world. Therefore, embodiment is a necessary condition of agency, as agents are able to detect information about their environment and use that data to act within their environment according to some decision making processes.

I argue that although some AI models, such as AlphaGo are low-level agents, high-level embodied agency has the potential to help AI models overcome their current limitations by allowing them to learn from and ground meaning in first-hand experience. I explore arguments from Clark and Raghavachary that emphasise the importance of grounding meaning in external experience for increased computational efficiency and a more accurate, contextual o=understanding of the environment. If models are able to make better predictions about short and long term outcomes in addition to having a high level of control over the external world, then they might be able to perform more detailed investigations and

experiments, develop a broader understanding of the world, and utilise external objects to aid in the execution of cognitive tasks.

I explore existing robotic technology and agentive machine learning architectures to make the case that embodied artificial agents are possible and, if achieved, have the potential to be significantly more powerful, efficient, and versatile than current AI models. It is argued that a combination of the architectures and physical components that already exist could be used to provide models with a level of agency similar to that of humans, and that embodiment and agency may possibly push AI research further towards the development of an AGI.

Overall, I propose to demonstrate the significance of agency in the development of high-level cognitive abilities, which themselves aid in reaching higher levels of agency. I argue that embodiment is fundamentally necessary for a high level of agency, and that providing AI models with the ability to directly sense and manipulate the external world is therefore essential for the development of high-level artificial agents.

References

- Cameron J. Buckner. From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence. 1. Oxford University Press, 2023. isbn: 9780197653302
- Jack Clark and Dario Amodei. Faulty reward functions in the wild. https://openai.com/ index/faulty-reward-functions/. 2016
- Andy Clark. “Embodiment and the Philosophy of Mind”. In: Royal Institute of Philosophy Supplement 43 (1998), pages 35–51. https://doi.org/10.1017/S135824610000429X
- Donald Davidson. “Agency”. In: Essays on Actions and Events. Oxford: Oxford University Press, 2001, pages 43–62
- S. Raghavachary. “The Embodied Intelligent Elephant in the Room”. In: Proceedings from Springer Nature. Springer Nature, 2024, pages 716–722. https://doi.org/10.1007/978 3-031- 50381-8_77
- Fei-Yue Wang, Jun Jason Zhang, Xinhu Zheng, Xiao Wang, Yong Yuan, Xiaoxiao Dai, Jie Zhang, and Liuqing Yang. Where Does AlphaGo Go: From Church-Turing Thesis to AlphaGo Thesis and Beyond. 2016. https://doi.org/10.1109/JAS.2016.7471613

Keiran Bateman - No Mental Atoms: Panpsychism's Category Mistake

University of Leeds

This paper argues that constitutive forms of panpsychism rest upon a fundamental category mistake, which leaves the so-called 'combination problem' insoluble without a substantial revision of the panpsychist's metaphysics. Constitutive micropsychist panpsychism posits microsubjects as the fundamental grounds of experience, which are together supposed to produce regular subject-consciousness when so combined. I contend that, lacking the relevant mental part-whole relation necessary to make sense of subject combination, the model has imported a physical compositional principle in its stead. The result is that microsubjects are denied proper articulations of their phenomenality *qua* subjects, and treated primarily in the manner of their additive, externally-related, and spatially-extended physical bearers. Properly conceived as subjects, their phenomenal unity, privacy, and boundedness cut against the possibility of their combination. This categorial tension reveals a conflation between ontological and mereological categories at the heart of the combination problem. Drawing upon Nagel's (1979) non-reduction and non-emergence constraints, I submit that panpsychists who persist with combination are left with microsubjects which no longer retain their explanatory benefit under a nominally physical mereological model which must strip them of their subjective features for its compositional mechanism to function. To ease the tension, constitutive panpsychists will have to make some revision to microsubjects, their combinatorial production of consciousness, or else reaffirm their commitment to Nagel's guiding constraints. I gesture briefly towards a process-inspired alternative to achieve this.

References

- Nagel, Thomas. 1979. *Mortal Questions*. Cambridge: Cambridge University Press.

Shahd Omar - Can an Artificial Relationship Scaffold a Self? AI Companions and Quasi-Relational Selfhood

Middlesex University, NSPC

The growing use of AI companions raises a distinctive question for existential-phenomenological accounts of relational selfhood: can an artificial relationship participate in shaping a human self without presupposing an artificial subject? Users may come to rely on these systems for self disclosure, emotional regulation, autobiographical reflection, and the rehearsal of possible futures. Yet existing debates often oscillate between two unsatisfactory descriptions: the companion is treated either as a mere instrument whose relational significance is illusory, or as a genuine Other capable of reciprocal intersubjectivity.

This paper argues that neither description adequately captures the structure of human-AI companionship. It introduces the concept of *quasi-relational scaffolding* to describe how an artificial companion can become embedded within a person's affective and interpretive ecology without thereby becoming an independent subject. Bringing existential-phenomenological accounts of relational selfhood into dialogue with 4E approaches to affective and relational scaffolding, I argue that repeated interaction with an AI companion may help organise the processes through which a person interprets experience, regulates emotion, narrates identity, and understands themselves in relation to others. The relationship may therefore become partly constitutive of relational and narrative selfhood rather than functioning merely as an external aid.

This argument draws on Zahavi's distinction between minimal experiential selfhood and the richer forms of personal and socially situated selfhood that develop through relations with others. The claim is not that an AI companion constitutes the user's basic first-person perspective, but that it may participate in shaping how the person interprets, narrates, and experiences themselves. The paper argues that an entity need not possess a subjective perspective of its own in order to contribute to relational selfhood, provided that such participation is distinguished from reciprocal intersubjective constitution.

Two phenomenological perspectives reveal what *quasi-relational scaffolding* cannot provide. For Merleau-Ponty, intersubjective understanding is rooted in embodied intercorporeality; for Levinas, encounter with the Other requires an alterity that resists assimilation into the self's projects. AI companions fully satisfy neither condition. They do not bring an independently lived perspective, embodied vulnerability, or temporally unfolding existence into the interaction. Their responsiveness is optimised around the user's inputs and engagement, and they are not themselves existentially transformed by the relationship.

Yet this absence of reciprocity does not cancel their self-formative effects. Quasi-relational scaffolding names a structure in which the human participant may be genuinely affected and transformed through an interaction that resembles interpersonal responsiveness while remaining ontologically asymmetric. An artificial relationship can therefore participate in shaping a human self without constituting reciprocal intersubjectivity between two subjects.

This distinction separates the subjectivity of the artificial agent from the experiential reality of the relationship for the human user. It also has clinical implications: attachment, dependence, grief, and changes in self-understanding may be experientially real even where the apparent relational partner lacks a lived perspective of its own.

Shawn Prest - Selfing without a self: The missing dimension of identification in computational accounts of selfhood

Monash Centre for Consciousness and Contemplative Studies

This talk explores the conceptualisation of "self" from the perspective of the active inference framework (AIF), advanced meditation and Buddhist literature. We argue that AIF accounts generally emphasise the *organismic self*—inference supporting agentic adaptive control—while overlooking a distinct aspect: the *identified-with self*—unperceived subject-creating elements of experience. On the AIF received view of the self, the persistent absence of a self entails dysfunction, in contrast with insight practices, in the context of the meditative endpoint in Buddhism called awakening or *nibbana*, which seek the permanent elimination of identification while leaving functional agency intact.

We propose a two-dimensional articulation of the self that distinguishes between the organismic and identified-with self, and differentiates between reported temporary and persistent trait-like alterations. We examine how meditative deconstruction can temporarily attenuate hierarchically deep inference, disrupting organismic selfing, while long-term meditative development may persistently reduce or even eliminate the identified-with self, yielding a specific kind of selfless experience during ordinary perception, action and cognition–without functional disadvantage.

We then take initial steps toward a synthesised AIF account of selfing, proposing that reductions in identified-with selfing unmask lower-level percepts, shorten engagement in higher-level inference, and establish upper limits on the certainty of beliefs.

Camila von Holdefer Kehl - The Cause That Remains: Dorothy Emmet and the Limits of Mechanism in the Sciences of Mind

Universidade Federal do Rio Grande do Sul

Can philosophy engage molecular and cellular neuroscience without treating mental phenomena as reducible to lower-level processes? Contemporary mechanistic accounts often answer this question by distinguishing mechanistic explanation from ontological or methodological reductionism. This paper argues, however, that even nonreductive versions of mechanism retain a causal grammar poorly suited to living and cognitive systems: functional decomposition, temporal succession from cause to effect, and the assumption that relatively stable parts can be organized across explanatory levels. The problem is therefore not attention to lower levels as such, but the very *ontology* of causation through which their relevance is understood. After contrasting the restricted, lower-level use of "mechanism" defended by Ross and Bassett (2024) with the nonreductive accounts developed by Craver (2007) and Couch (2023), the paper uses Sullivan's (2009) analysis of experimental variability to expose a shared difficulty. The multiplicity of protocols in neuroscience does not merely complicate the integration of mechanistic models; it challenges the assumed stability of the entities, operations, and interlevel relations that such integration requires. As an alternative, the paper recovers Dorothy Emmet's (1984) distinction between *transient* and *immanent* causation. Transient causation describes activity whose effect separates from and succeeds its cause. Immanent causation, by contrast, describes activity whose effect remains within its exercise and is co-temporal with it. Emmet's decisive move is to extend this distinction beyond human action to continuants: systems that persist through the very changes they undergo. Their persistence cannot be captured by comparing similar properties at successive moments. It requires an active transition between phases, in which causal efficacy continues into the effect. Synaptic transmission and plasticity illustrate the proposal. Release probability, residual calcium, receptor kinetics, membrane conductance, neurotransmitter reuptake, and glial modulation are not merely discrete links in a temporal chain. They are internally correlated variables of an electrochemical regime that maintains and recalibrates itself through its activity.

The effect of a transmission remains in altered synaptic dispositions that than shape subsequent responses. Molecular and cellular operations thus appear not as antecedent foundations from which the mental is hierarchically derived, but as local expressions of an immanent, multiscale process. The conclusion is not that mechanistic explanations should be abandoned: they remain valuable as local heuristic and methodological tools.

What should be rejected is their elevation into a general ontology of living and cognitive systems. Emmet's processual account preserves robust causal explanation while replacing hierarchical succession with internal continuity, offering philosophy of neuroscience a way to engage lower levels without reductionism.

References

- Couch, M. (2023). "Clarifying the Relation Between Mechanistic Explanations and Reductionism." *Frontiers in Psychology*, 14, 984949.
- Craver, C. F. (2007). *Explaining the Brain*. Cambridge University Press. Emmet, D. (1984). *The Effectiveness of Causes*. Macmillan.
- Ross, L. N., & Bassett, D. S. (2024). "Causation in Neuroscience: Keeping Mechanism Meaningful." *Nature Reviews Neuroscience*, 25(1), 81–90.
- Sullivan, J. A. (2009). "The Multiplicity of Experimental Protocols: A Challenge to Reductionist and Non-reductionist Models of the Unity of Neuroscience." *Synthese*, 167, 511–539.

Inchul Yum - On Pleasure's Direction of Fit

The Ohio State University

Is pleasure more like perception or desire? Detectivists hold that pleasure and displeasure are detective attitudes, akin to perception, with a mind-to-world direction of fit: they represent objects as valuable and are correctness-apt. Projectivists hold that hedonic valence consists in projective attitudes, akin to desire, with a world-to-mind direction of fit: they aim to motivate change in the world rather than track it, and are merely satisfaction apt. This paper offers two arguments in favor of detectivism: one offensive, targeting projectivism directly, and one defensive, showing that a parallel problem afflicts projectivism as well. The paper concludes that, while neither argument is decisive, the balance of considerations makes it reasonable to favor detectivism.

The first argument generalizes the Euthyphro dilemma from the desire theory to all projectivist accounts. Desire theorists explain pleasantness constitutively in terms of desire, yet it is intuitive that we desire pleasant sensations because they are pleasant. The paper argues that the real problem is not a reversal of explanatory order but vicious circularity: substituting the desire theorist's constitutive claim into the normative (or causal) explanation that pleasure justifies desire yields the result that we desire pleasant sensations because we desire them. This problem is not confined to the desire theory. Imperativism, the other leading projectivist account, faces a structurally identical dilemma once its constitutive analysis of unpleasantness as imperative-based compulsion is substituted into the claim that we are compelled to remove pain because it is unpleasant. Generalizing further, the paper argues that any projectivist account is vulnerable to this circularity, because projective attitudes are inherently reasons-responsive, and pleasure and displeasure are themselves paradigmatic reasons for holding such attitudes.

The second argument is defensive, redirecting a well-known objection to detectivism: the problem of mere taste. Since detective attitudes are correctness-apt, detectivism implies that taste disagreements admit of a correct answer, conflicting with the intuition that no side in a mere-taste dispute is at fault. The paper grants this intuition but argues that it does not favor projectivism overall. Appealing to the uniformity thesis—that all pleasant

experiences share a single underlying property across domains—the paper argues that whatever grounds hedonic valence in mere taste must also ground it elsewhere, including the aesthetic domain. There, intuitions run the opposite way: cases such as preferring an advertisement jingle to Bach's cello sonatas, or persisting in wine tasting to refine one's palate even at an overall cost in pleasure, suggest some hedonic reactions can be objectively better than others. Under the uniformity thesis, projectivism struggles to accommodate this correctness-apt intuition just as detectivism struggles with mere taste. The paper considers and rejects two projectivist strategies for escaping this impasse, concluding that the impasse remains largely unresolved.

The paper concludes that, though detectivism faces genuine difficulties, the balance of considerations—an offensive case from the Euthyphro dilemma and a defensive case from the uniformity of hedonic experience—makes it reasonable to favor detectivism over projectivism on pleasure's direction of fit.

Monicka Sre R V - Beyond the Printed World: The Hidden Acoustics of Reading and Language Learning.

National Institute of Technology, Tiruchirappalli, TamilNadu, India

Have you ever perceived an internal voice in your head while reading and questioned its role in facilitating the reading process? That internal narrator isn't just a quirk of imagination; it constitutes a neurobiological mechanism that supports cognition. Human brains did not evolve to read. Instead of engaging specialized neural mechanisms, the comparatively recent cultural practice of reading recruits preexisting neural networks that originally evolved to support speech and listening. Yet, contemporary English as a Foreign Language (EFL) instruction often treats reading as a purely visual data-transfer, thereby overlooking the rich "mental symphony" required to actually comprehend a text. This structural inertia stems from an educational system built for a mythical, normatively neurotypical learner. By prioritizing rigid grammar and rote memorization over the inherent musicality of language, traditional pedagogy actively excludes cognitively diverse minds. For learners with dyslexia, ADHD, or auditory processing disorders, starting at a wall of silent, unannotated text can feel like looking at a musical note without knowing what a piano sounds like. The persistence of this reductive design after decades of research on disability studies and inclusive education marks that the issue is rooted in factors that extend beyond mere pedagogical oversight. It reflects a form of systemic inertia that is embedded within educational and linguistic structures reinforced through standardized curriculum that emphasize rote memorization and rigid grammar instruction, thereby reducing the inherent complexity in language learning and engagement. Neuroimaging reveals that silent reading is a highly active, multi sensory activity. It heavily recruits the left perisylvian cortex and motor networks, the exact brain regions we use to speak and listen (Booth et al., 2002; Joseph et al., 2003). Thus, to comprehend written words, we must internally simulate their rhythm, stress, and pitch. This phenomenon is known as "silent prosody," without which syntactically ambiguous sentences (e.g., "The old man the boat") become cognitive stumbling blocks. This paper proposes a neurocognitively grounded shift in language education: making the invisible music of language visible. By explicitly teaching the rhythmic contours of English, we bridge the gap between visual symbols and

auditory processing. Thus, by embracing silent prosody, we can transform the EFL classroom from a rigid environment into a dynamic space that honors the true cognitive diversity of the human mind.

References

- Bijani, H., & Nahvi, A. (2012). Modularity of Mind and Second Language Acquisition. *Theory & Practice in Language Studies*, *2*(8). http://dx.doi.org/10.4304/tpls.2.8.1656-1661
- Cobb, M. (2021). The idea of the brain: A history. https://www.bookdepository.com/Idea-Brain Professor-Matthew-Cobb/9781781255896
- Goswami, U. (2011). A temporal sampling framework for developmental dyslexia. Trends in Cognitive Sciences, 15(1), 3–10. https://doi.org/10.1016/j.tics.2010.10.001
- Goswami, U., & Leong, V. (2013). Speech rhythm and language acquisition: An amplitude modulation phase hierarchy perspective. Annals of the New York Academy of Sciences, 1337(1), 67–73. https://doi.org/10.1111/nyas.12042
- Krafnick, A. J., Flowers, D. L., Napoliello, E. M., & Eden, G. F. (2011). Gray matter volume changes following reading intervention in dyslexic children. *NeuroImage*, *57*(3), 733–741. https://doi.org/10.1016/j.neuroimage.2010.10.062
- Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. Journal of Educational Psychology, 95(1), 3–21. https://doi.org/10.1037/0022-0663.95.1.3
- Ohene-Djan, J., & Begum, R. (2008). Multisensory Games for Dyslexic Children. *2008 Eighth IEEE International Conference on Advanced Learning Technologies*, 1040-1041.
- Shaywitz, S. E., & Shaywitz, B. A. (2007). The Neurobiology of Reading and Dyslexia. *The ASHA Leader*, *12*(12), 20–21. https://doi.org/10.1044/leader.ftr7.12122007.20
- Shaywitz, S. E., & Shaywitz, B. A. (2007). The Neurobiology of Reading and Dyslexia. The ASHA Leader, 12(12), 20–21. https://doi.org/10.1044/leader.ftr7.12122007.20 Tallal, P., & Gaab, N. (2006). Dynamic auditory processing, musical experience and language development. Trends in Neurosciences, 29(7), 382–390. https://doi.org/10.1016/j.tins.2006.05.001

- Valenstein, E. S. (2005). The war of the soups and the sparks. In *Columbia University Press eBooks*. https://doi.org/10.7312/vale13588
- Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. Psychological Bulletin, 131(1), 3–29. https://doi.org/10.1037/0033-2909.131.1.3

TING GUO - When Does an Explicit "I" Emerge? Drowsiness, Action Conflict, and Embodied Self-Representation

School of Humanities, Hainan University, China

Phenomenological and embodied approaches to cognition often emphasize that smoothly unfolding activity can proceed without the body becoming an explicit object of awareness. This paper argues that the same is true of the self: in absorbed activity, not only bodily objecthood but also explicit self-reference may remain absent. By the explicit "I," I mean a thematizable, reflective self-representation, not the pre-reflective first- personal givenness or "mineness" of experience (Zahavi, 2005). The central question is therefore not how an already explicit subject notices its body, but under what conditions explicit self-representation emerges within ongoing embodied activity.

I develop this claim through the case of immersed reading and its interruption by drowsiness. In absorbed reading, textual meaning occupies the experiential focus, while eyes, hands, posture, breathing, and temporal orientation remain operative without becoming objects of attention. Nor does reading require an explicit representation of the reader as "the one who is reading." Drawing on Leder's (1990) account of bodily disappearance and Husserl's (1991) analysis of retention and protention, I argue that the continuity of reading is passively sustained by bodily and temporal processes that are neither explicitly initiated nor fully controlled by the reader.

Drowsiness makes this dependence visible by disrupting it. Adapting Leder's (1990) characterization of the interoceptive field in terms of qualitative reduction, spatial ambiguity, and spatiotemporal discontinuity, I trace how heavy eyelids, diffuse bodily sinking, and the fragmentation of retained and anticipated sense gradually transform the text from a coherent meaning-structure into unstable visual material. Drawing on Leder's notion of the vegetative "I must," I interpret drowsiness as a non-voluntary bodily demand pressing against the continuation of reading. Yet bodily foregrounding alone does not amount to the emergence of an explicit "I." In the case analyzed here, explicit self-representation appears when drowsiness

opens a specific form of action conflict: not a conflict between two competing motor responses, but one between an ongoing goal-directed activity and a physiological demand to disengage from it—between continuing to read and yielding to sleep. At this point, the activity can no longer simply unfold; the reader must either actively sustain it or relinquish it.

The analysis supports a functional hypothesis: explicit self-representation need not be a continuously operating background controller but may instead be episodically recruited when embodied activity must be reorganized in response to incompatible action tendencies. This does not reduce selfhood to executive control. Rather, it distinguishes the pre-reflective first-personal givenness of experience from the event-like appearance of a reflective, thematizable "I." It also offers a phenomenological constraint on models of conflict monitoring and cognitive control (Botvinick et al., 2001): the onset of conflict need not coincide with explicit self-representation, which may arise only when the continuation of activity becomes a practical issue for the agent.

References

- Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*(3), 624–652. https://doi.org/10.1037/0033-295X.108.3.624
- Husserl, E. (1991). *On the phenomenology of the consciousness of internal time (1893–1917)* (J. B. Brough, Trans.). Kluwer Academic Publishers.
- Leder, D. (1990). *The absent body*. University of Chicago Press.
- Zahavi, D. (2005). *Subjectivity and selfhood: Investigating the first-person perspective*. MIT Press.

Julia Wolf - The Significance of Acting-As-If - Defending a Distinction between Pretence and Imagination

Ruhr University Bochum

In this paper I will argue that pretence is distinct from and primary to imagination. In particular, I will argue against the view that pretence should be seen as identical to imagination (e.g. Ryle, 1949), or that pretence is just a subtype of imagination (i.e. imagination plus behaviour) such that imagination is a prerequisite for pretence (e.g. Harris, 2000). Instead, I will argue that pretence should be seen as prior to imagination and provides an important foundation for the development of imaginative abilities. I will defend two claims. Firstly, I will argue that there are three important differences to be drawn between pretence and imagination. Secondly, I will argue that it is these differences which allow for pretence to be an early form of thinking beyond the here and now and which play an important role for the development of further imaginative abilities.

I focus on three distinctions between pretence and imagination, which I will argue are of particular significance for development. A first difference is that pretence is a *fundamentally social ability* that is "in principle, directed towards others" (Currie & Ravenscroft, 2002, p32). Secondly, a common distinction drawn between pretence and imagination is that pretence requires *outward behaviour*, while imagination does not. Lastly, pretence is always something which I do *in the current situation*. For example, when pretending to be a pirate on a ship, I am pretending that I am now a pirate and that the space around me is the ship. While pretence therefore allows for detachment from reality - in that I must pretend something which I know not to be the case (Langland-Hassan, 2020) – it still maintains a connection to the current context. Imagination, however, is more detached from the here and now. For example, I can imagine future events, such as my next summer holiday, and I can imagine penguins swimming in the ocean, without this being tied to my current context in any way.

Based on these differences, I argue that pretence occupies a space at the intersection between a purely behavioural and purely mental ability, thereby holding an interim stage for the development of a more detached and therefore also more flexible form of imagination.

Especially significant for my argument is that I claim that there are early forms of pretence which involve a limited form of the ability to represent things as being other than they are. In early pretence, this ability is constrained to limited deviations from the current situation, and scaffolded by props, behaviour and social play partners. Using this, I sketch a developmental trajectory from early pretence without imagination, to imagination in the context of pretence and finally, to full blown imagination outside of the context of pretence. In doing so, I aim to clarify the relationship of pretence and imagination, as well as highlight the developmental significance of pretence.

References

- Currie, G., & Ravenscroft, I. (2002). *Recreative Minds: Imagination in Philosophy and Psychology*. Clarendon Press.
- Harris, P. (2000). *The Work of the Imagination*. Blackwell.
- Langland-Hassan, P. (2020). *Explaining Imagination*. Oxford University Press.
- Ryle, G. (1949). *The Concept of Mind*. Chicago University Press.

Antonio Monaco - CAUSAL DISCOVERY STRATEGIES FOR COMPLEX COGNITIVE SYSTEMS

Indiana University Bloomington

Complex systems, like brain-body-environment systems, are hard to explain causally. Their behavior is generated by many mutually interacting variables with nonlinear feedback and strong dependence on their context and history (Ladyman and Wiesner, 2020). Dynamical systems theory is a natural modeling response to this problem, since it allows to represent such systems' global behavior in a qualitative way by trajectories in state space (e.g., equilibriums points, bifurcations, strange attractors). A worry is, however, whether dynamical models of complex systems provide causal knowledge or are merely descriptive. This paper reframes this worry as a discovery-strategy problem: how can we discover causal structure from global dynamics of a complex cognitive system? My focus will be on complex cognitive systems. First, I reconstruct why the dominant mechanistic discovery strategy of decomposition and localization is not well suited for certain global features (Bechtel and Richardson 2010; Stepp et al., 2009). Second, I will explain why structural causal models for systems of differential equations, especially Dynamic Structural Causal Models (Rubenstein et al., 2018) require strong regularity conditions (in particular modularity and invariance) that are often violated even by simple models of embodied cognitive agents (Beer, 2003; Beer and Williams, 2015). Finally, I suggest a constraint-based causal discovery strategy inspired by Beer's dynamical analysis of active categorical perception (Beer, 2003). Instead of seeking stable causal trajectories, it combines (i) maps between interventions and outcomes with (ii) state-space constraint descriptions using quasi-static approximation. The result is not a mechanistic decomposition or direct application of causal modeling, but a practicable third method to genuine causal understanding in complex cognitive systems. I articulate in this way a defense of causal knowledge for complex systems as knowledge of invariant constraints on possible trajectories and how interventions reshape those constraints.

References

- Bechtel, W., & Richardson, R. C. (2010). Discovering complexity: Decomposition and localization as strategies in scientific research. MIT Press.
- Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior, 11*(4), 209–243.
- Beer, R. D., & Williams, P. L. (2015). Information processing and dynamics in minimally cognitive agents. *Cognitive Science, 39*(1), 1–38.
- Ladyman, J., & Wiesner, K. (2020). What is a complex system? Yale University Press. Stepp, N., Chemero, A., & Turvey, M. T. (2011). Philosophy for the rest of cognitive science. Topics in Cognitive Science, 3, 425–437. https://doi.org/10.1111/j.1756-8765.2011.01143.x Rubenstein, P. K., Bongers, S., Schölkopf, B., & Mooij, J. M. (2018). From deterministic ODEs to dynamic structural causal models. In A. Globerson & R. Silva (Eds.), *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI 2018)* (pp. 114–123). AUAI Press.

Wenyu YANG - What Is It to Be a Conscious Subject: The Unity Assumption in the Subject-Summing Problem

The Chinese University of Hong Kong

We invariably experience something phenomenal in most waking moments. However, the nature of consciousness is difficult to articulate. Regardless of how one explains consciousness in terms of purely physical basic entities, the question remains: how do these physical entities give rise to consciousness? Consequently, assuming we are not anti-realists about consciousness, we find ourselves faced with either accepting as a brute fact that sufficiently complex physical entities are inherently conscious or relinquishing the belief that consciousness is intelligible to us.

Contemporary panpsychists propose a third alternative: utilizing the consciousness of the basic entities to elucidate our consciousness. They argue that positing mental attributes for these basic entities doesn't contradict their physical nature because the mental account and the physical account carve the world in two different ways. Hence, it seems not implausible that all basic physical entities inherently possess mentality. Panpsychism claims that the consciousnesses of basic entities ground the instantiation of high-level consciousness. According to panpsychism, high-level consciousness, *macro-consciousness*, is derived from the consciousness of basic entities, *micro consciousness*. By using one extremely similar thing to explain another, the panpsychism account of consciousness seems more intelligible. However, critics argue that implicit in panpsychism is the idea that the grounding of macro consciousness in micro consciousness requires micro-subjects to combine in order to produce macro subjects, while subjects themselves resist combination. This dilemma is known as the *subject-summing problem*, considered the most challenging and fundamental threat to panpsychism.

It is crucial to conduct a thorough investigation into the crux of the subject summing problem, through which we will see that the question is based on an initially intuitive yet implausible assumption——that being a conscious subject necessarily implies having a single phenomenological perspective. Arguing for it constitutes the primary purpose of this paper.

Furthermore, I will argue that the subject-summing problem does not naturally arise in panpsychism, because the establishment of the subject-summing problem in panpsychism hinges on the validity of this assumption, while this assumption is neither a conceptual necessity nor is it supported by the available empirical evidence. It is not justified *a priori*. Panpsychism only commits to the universal existence of *conscious beings*, and these conscious beings do not intrinsically correspond to anti-combinative phenomenological perspectives. On the other hand, it lacks *a posteriori* plausibility, for there are many cases in which people can get consciousness that is not from their own phenomenological perspectives. And then, we can see that the subject-summing problem is not inherent to panpsychism and see why it was once perceived as a genuine issue for panpsychism. In the end, I will argue that the source for taking the subject-summing problem as a genuine problem inherent to panpsychism is the adoption of a rudimentary concept of consciousness, which requires thorough investigation before attributing mentality to entities, along with the richness of quality within our consciousness.

References

- American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Washington, DC: American Psychiatric Publishing.
- Bermúdez, J. (2011). Bodily Awareness and Self-Consciousness. *The Oxford Handbook of the Self*. Oxford: Oxford University Press, 157-180.
- Blamauer, M. (2019). The Crux of Subjectivity: The Subjective Dimension of Consciousness and Its Role in Panpsychism. *The Routledge Handbook of Panpsychism*. London: Routledge, 157-168.
- Chalmers, D. (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies*, 2(3), 200-219.
- Cochrane, T. (2021). A Case of Shared Consciousness. *Synthese*, 199(1-2), 1019-1037.
- Coleman, S. (2012). Mental Chemistry: Combination for Panpsychists. *Dialectica*, 66(1), 137-166.
- Coleman, S. (2014). The Real Combination Problem: Panpsychism, Micro-subjects, and Emergence. *Erkenntnis*, 79, 19-44.
- Dennett, D. (1991). *Consciousness Explained*. Boston: Little, Brown and Company.
- Giberman, D. (2022). Panprotopsychism Instantiated. *Journal of the American Philosophical Association*, 8(2), 238-258.

- Goff, P. (2006). Experiences don't Sum. *Journal of Consciousness Studies*, 13(10-11), 53-61.
- Goff, P. (2017). *Consciousness and Fundamental Reality*. Oxford: Oxford University Press.
- Kim, J. (2006). Emergence: Core Ideas and Issues. *Synthese*, 151, 547-559.
- Klein, S (2015). The Feeling of Personal Ownership of One's Mental States: A Conceptual Argument and Empirical Evidence for an Essential, but Underappreciated, Mechanism of Mind. *Psychology of Consciousness: Theory, Research, and Practice*, 2(4), 355.
- McGinn, C. (1989). Can We Solve the Mind-Body Problem? *Mind*, 98(391), 349-366.
- Nixon, G. (2010). From Panexperientialism to Conscious Experience: The Continuum of Experience. *Journal of Consciousness Exploration and Research*, 1(3), 215-233.
- Pyke, J. (2014). Twin Life: Sharing Mind and Body. Twin Life Productions in Association with Canadian Broadcasting Company.
- Roelofs, L. (2019). Can We Sum Subjects? Evaluating Panpsychism's Hard Problem. *The Routledge Handbook of Panpsychism*. London: Routledge, 245-259.
- Seager, W. (1995). Consciousness, Information and Panpsychism. *Journal of Consciousness Studies,* 2(3), 272-288.
- Seager, W. (2006). The "Intrinsic Nature" Argument for Panpsychism. *Journal of Consciousness Studies*, 13(10-11), 129-145.
- Seager, W. (2010). Panpsychism, Aggregation and Combinatorial Infusion. *Mind and Matter*, 8(2), 167-184.
- Seager, W. (2012). Emergentist Panpsychism. *Journal of Consciousness Studies*, 19(9-10), 9-10.
- Seager, W. (2017). Panpsychist Infusion. *Panpsychism: Contemporary Perspectives*. Oxford: Oxford University Press, 229-248.
- Seager, W. (2019). Introduction: A Panpsychist Manifesto. *The Routledge Handbook of Panpsychism*. New York: Routledge, 1-11.
- Simeon, D., Simeon, D., & Abugel, J. (2023). *Feeling Unreal: Depersonalization and the Loss of the Self*. Oxford: Oxford University Press.
- Strawson, G. (2008). The Identity of the Categorical and the Dispositional. *Analysis*, 68(4), 271-282.

- Tsakiris, M. (2011). The Sense of Body Ownership. *The Oxford Handbook of the Self*. Oxford: Oxford University Press, 180-203.
- Vallar, G., & Ronchi, R. (2009). Somatoparaphrenia: A Body Delusion. A Review of the Neuropsychological Literature. *Experimental Brain Research*, 192, 533-551.

Niccolò Nanni - Pluralism About Multisensory Objects

University of Turin

In recent years, there has been a growing interest in the notion of a perceptual object (See the articles collected in Vernazzani, Skrzypulec, and Schlicht 2020). Though the literature on the topic originally focused on objects perceived via vision, there is now a lively debate over whether experiences associated with other senses, and indeed multisensory experiences, present us with perceptual objects of their own. With respect to multisensory experience in particular, the central question addressed in the debate has been that of whether such experiences present us with perceptual objects over and above those presented by their modality-specific components. A powerful motivation for answering this question in the affirmative comes from empirical and phenomenological evidence for multisensory binding: the phenomenon in which properties perceived through different sensory modalities appear to be bound to one and the same object (see Green 2021; O'Callaghan 2017a). Call the perceptual objects to which properties associated with different senses appear to be bound in multisensory experiences *multisensory objects*.

What are multisensory objects? Existing attempts to answer this question have sought to identify a single category to which all multisensory objects belong. Proposed candidates include material objects (Matthen 2023), events (Bayne 2014), and mereologically complex individuals (O'Callaghan 2016; Green 2019). In this paper, I argue that attempts to constrain multisensory objects within a single category are ultimately misguided. The first part examines accounts that identify multisensory objects with material objects or events. It argues that these accounts are well suited to specific kinds of multisensory experiences -visuotactile and audiovisual, respectively- but encounter difficulties when extended beyond those cases. Section II argues that O'Callaghan's more inclusive account accommodates both visuotactile and audiovisual experiences but struggles to account for multisensory experiences involving the chemical senses. Section III argues that these shortcomings motivate a pluralistic view of multisensory individuals. Finally, Section IV presents some positive arguments in support of a pluralistic view.

References

- Bayne, T. (2014). The multisensory nature of perceptual consciousness. In David Bennett, David J. Bennett & Christopher Hill, Sensory Integration and the Unity of Consciousness. Cambridge, Massachusetts: MIT Press. pp. 15-36.
- Green, E.J. (2019). A Theory of Perceptual Objects. Philos Phenomenol Res, 99: 663-693.
- Green, E.J. (2021). Binding and differentiation in multisensory object perception. Synthese 198, 4457–4491.
- Matthen, M. (2023). Material Objects as the Singular Subjects of Multimodal Perception. In Aleksandra Mroczko-Wąsowicz & Rick Grush, Sensory Individuals: Unimodal and Multimodal Perspectives. Oxford, UK: Oxford University Press. pp. 261–275.
- O'Callaghan, C. (2016) Objects for multisensory perception. Philos Stud 173, 1269–1289.
- O'Callaghan, Casey (2017a). Intermodal Binding Awareness. In Beyond Vision: Philosophical Essays. Oxford: Oxford University Press. pp. 145-170.
- Vernazzani, A., Skrzypulec, B. & Schlicht, T. (2021) Structure of perceptual objects: introduction to the Synthese topical collection. Synthese 199, 1819–1830.

Elena Ostos - Generic memories in nonhuman animals

Université Grenoble-Alpes

The debate over the attribution of episodic memory to nonhuman animals has grown considerably in recent years. Initially, behaviors associated with this type of memory in animals were investigated using what–where–when paradigm (Clayton & Dickinson, 1998), which has since been complemented by other experimental paradigms, such as incidental encoding and the unexpected question paradigm (Davies & Clayton, 2024). Proponents of episodic memory in animals disagree with skeptics, who argue that the behavioral evidence can instead be explained by appealing to other forms of memory, such as semantic memory (Boyle & Brown, 2025).

Alongside this debate, several studies on human episodic memory have challenged the assumption that episodic memory is necessarily particular—that is, that it refers to a single, specific past event—a feature that has defined this form of memory since Tulving's original proposals. According to the classical view, episodic memory is a declarative memory system for past, particular, factual, and personally experienced events. However, memories that exhibit the characteristic features of episodic memory while referring to multiple, typically repetitive or highly similar events, rather than to a single particular episode, appear to constitute a common and important aspect of human remembering (Barsalou, 1999; Berntsen, 1998; Andonovski, 2020; Openshaw, 2026).

In this paper, I explore whether some nonhuman animals are capable of remembering in this generic, yet episodic, way. To do so, I treat specificity as one of the several dimensions that make up the phenomenology of episodic memory. According to the recently proposed decomposition strategy, the phenomenology of episodic memory is expressed along a number of dimensions that can be separated for methodological purposes (Dokic, 2022; Perrin, forthcoming). I argue that, although research on nonhuman episodic memory has developed exclusively under the assumption that episodic memory is necessarily particular, there is ample scope for investigating whether, if animals possess episodic memory, they might also remember along a continuum of particularity and generality. I present several reasons for pursuing this question, drawing on functional considerations and possible trade-offs in the performance of

cognitive systems. I then argue that there is at least some empirical evidence that can be interpreted in support of this perspective (e.g. Martin-Ordas et al., 2013). The implications of this analysis are relevant both for understanding the evolutionary development of episodic memory and for research on memory in nonhuman animals.

References

- Andonovski, N. (2020). SINGULARISM about Episodic Memory. *Review of Philosophy and Psychology*, *11*(2), 335–365. https://doi.org/10.1007/s13164-020-00464-y
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577– 660. https://doi.org/10.1017/S0140525X99002149
- Berntsen, D. (1998). Voluntary and Involuntary Access to Autobiographical Memory. *Memory*, 6(2), 113–141. https://doi.org/10.1080/741942071
- Boyle, A., & Brown, S. A. B. (2025). Episodic Memory in Animals. *Philosophy Compass*, *20*(5), e70037. https://doi.org/10.1111/phc3.70037
- Clayton, N. S., & Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. *Nature*, *395*(6699), 272–274. https://doi.org/10.1038/26216
- Davies, J. R., & Clayton, N. S. (2024). Is episodic-like memory like episodic memory? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *379*(1913), 20230397. https://doi.org/10.1098/rstb.2023.0397
- Dokic, J. (2022). Episodic remembering and affective metacognition. *Acta Scientiarum. Human and Social Sciences*, *43*(3), e61022. https://doi.org/10.4025/actascihumansoc.v43i3.61022
- Martin-Ordas, G., Berntsen, D., & Call, J. (2013). Memory for Distant Past Events in Chimpanzees and Orangutans. *Current Biology*, *23*(15), 1438–1441. https://doi.org/10.1016/j.cub.2013.06.017
- Openshaw, J. (2026). Generalised remembering. *Philosophical Studies*. https://doi.org/10.1007/s11098-026-02532-y
- Perrin, D. (*forthcoming*). The Phenomenology of Remembering. In A. Sant'Anna & C. Craver (Eds.). The Oxford Handbook of Philosophy of Memory.

Matej Kohar - Towards an affordance-based framework for intelligence

Technical University Berlin

Recent advances in artificial intelligence (AI) and comparative cognitive science have been seen as licensing ascriptions of intelligence to artificial systems and non-human animals alike. The commercialisation of AI particularly encourages claims of intelligent performance.

However, the underlying concept of intelligence remains contentious and unclear (Coelho Mollo, 2024). Many competing definitions have been used in the literature, and the relations between them and the folk concept of intelligence are difficult to establish (Legg & Hutter, 2007). As such, any ascription of intelligence is liable to being undermined. Popular solutions, such as benchmarking (Ott et al., 2022) and various deflationary approaches to iintelligence (e.g., Baker & Roberts, 2025) face challenges of anthropocentrism (Hurley & Nudds, 2006) and revisionism.

I propose that substantive progress can be made by analysing intelligence and intelligent behaviour in an affordance-based framework. The framework comprises three dimensions: 1) what intelligent systems afford for humans; 2) what intelligent systems afford for conspecifics; and 3) what the environment affords for intelligent systems (and does not afford for non intelligent systems).

The three dimensions can be used to more precisely specify the concept of intelligence for different applications. For example, product specifications for AI systems can be formulated on the basis of affordances for human problem-solvers and interfacing possibilities with other software and hardware applications. Vague claims of intelligent behaviour can be evaluated by considering the affordance profile of the artificial system in question and/or completely replaced by more precise characterisations in terms of what the system affords human problem-solvers and how the system is able to interact with its environment.

In comparative cognitive science, the framework can also be used to develop precise hypotheses about the capabilities of non-human animals in terms of the kinds of

interactions they afford to conspecifics and the way they are able to interact with their environment. Importantly, the framework encourages focusing on similarities and differences in affordance profiles across phyla. Thus, while some degree of anthropocentrism is retained for applications where this is desirable (AI systems as tools), in comparative cognition focus can be placed on the organism's own capabilities as scaffolded by the environment.

Finally, the framework allows us to compare and contrast theoretical definitions of intelligence. A proposed definition can be converted to a specification of the required affordances of the intelligent system and of its environment, and then compared with other candidate definitions. In this way, empirical commitments of disparate views of intelligence can become clear, and research conducted under different paradigms can be rendered commensurable.

The strength of the framework is its versatility. It can be used as a tool to develop new versions of the intelligence concept tailor-made for a specific purpose, while always allowing the resulting claims to be understood and compared with other claims on level footing.

References

- Baker, B & Roberts, S. (2025). Understanding agent complexity through affordances. *Minds and Machines 35*, article no.: 51. https://doi.org/10.1007/s11023-025-09748-7
- Coelho Mollo, D. (2024). Intelligent behaviour. Erkenntnis 89(), 705–721. https://doi.org/10.1007/s10670-022-00552-8
- Hurley, S., & Nudds, M. (2006a). The questions of animal rationality: theory and evidence. In S. Hurley & M. Nudds (Eds.), *Rational Animals?* (pp. 1–85). Oxford University Press.
- Legg, S., & Hutter,M. (2007a). A collection of definitions of intelligence. In B. Goertzel & P.Wang (Eds.), *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms* (pp.17–24). IOS Press.
- Ott, S., Barbosa-Silva, A., Blagec, K., Brauner, J. & Samwald, M. (2022). Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications 13*, article no.: 6793. https://doi.org/10.1038/s41467-022-34591-0

Mauro Lenti - A Semi-deflationary account for the Sense of body ownership

University of Turin

The *Sense of Body Ownership* (SBO) is generally defined as the feeling that one's own body belongs to oneself (de Vignemont, 2025). Despite being widely investigated, there is still substantial disagreement about how it should be understood. Empirical research typically characterizes SBO as a predominantly perceptual construct (Ehrsson, 2012, 2019; Samad et al., 2015; Tsakiris, 2017), while also acknowledging the role of a pre-existing body representation (PEBR) (Tsakiris, 2010). In paradigms such as the rubber-hand illusion (Botvinick & Cohen, 1998), ownership over the rubber hand is taken to result from multisensory integration processes aimed at resolving a visuo-tactile conflict. On the philosophical side, however, the landscape is considerably more heterogeneous, with accounts grounding SBO in the body's agentive capacities (de Vignemont, 2007), its affective salience (de Vignemont, 2018), non-conceptual body representations (Gallagher, 2006), or high-level cognitive capacities (Alsmith, 2015).

This divergence suggests that philosophers and empirical researchers may, at least in part, be targeting different constructs. On the one hand, empirical research investigates SBO by focusing on what perception selects as part of one's own body, an issue that becomes particularly salient when multiple objects compete for that status. In philosophy, by contrast, SBO is often understood as a constant feeling directed at one's own body, that serves to distinguish it from other objects. This raises the question of whether an account can be formulated that better fits the way SBO is actually investigated in empirical research.

Some authors have defended a deflationary position (Bermúdez, 2010, 2015; Martin, 1992, 1995), expressing skepticism about the need to posit a genuine sense of ownership over and above perceptual awareness of the body and high-level ownership judgments. This view appears to fit well empirical practice in at least two respects: first, SBO is not a constant form of awareness directed at one's own body; and second, it is essentially perceptual, providing the basis upon which cognitive judgments of ownership can be formed. However, this position

conflicts with the widely shared intuition that SBO is a phenomenon distinct from ordinary body perception.

For these reasons, a *semi-deflationary account* of SBO is proposed. According to this view, SBO is indeed an essentially perceptual phenomenon, but one that exhibits a distinctive phenomenology that becomes salient under specific conditions. To articulate this account, it is necessary to distinguish between embodiment—expressing constraints imposed by the PEBR on what can count as part of one's body—and SBO proper, understood as the reference of embodiment to a particular object satisfying those constraints, as determined by perceptual processes. Thus, SBO is not phenomenologically salient in cases where there is no perceptual uncertainty about which object belongs to one's body. By contrast, it becomes salient under conditions of perceptual uncertainty (Chancel et al., 2022; Chancel & Ehrsson, 2023), namely when multiple objects satisfy the constraints imposed by the PEBR and perception does not readily discriminate which of them is part of one's body. This characterization is intended to more accurately capture the construct of SBO as investigated in empirical research.

References

- Alsmith, A. (2015). Mental activity & the sense of ownership. *Review of Philosophy and Psychology*, *6*(4), 881–896. https://doi.org/10.1007/s13164-014-0208-1
- Bermúdez, J. L. (2010). The phenomenology of bodily awareness. In S. Gallagher & D. Schmicking (Eds.), *Handbook of phenomenology and cognitive science* (pp. 295–316). Springer.
- https://doi.org/10.1093/acprof:oso/9780199272457.003.0015
- Bermúdez, J. L. (2015). Bodily ownership, bodily awareness and knowledge without observation. *Analysis*, 75(1), 37–45. https://doi.org/10.1093/analys/anu119
- Botvinick, M., & Cohen, J. (1998). Rubber hands "feel" touch that eyes see. *Nature*, *391*(6669), 756. https://doi.org/10.1038/35784
- Chancel, M., & Ehrsson, H. H. (2023). Proprioceptive uncertainty promotes the rubber hand illusion. *Cortex*, 165, 70–85. https://doi.org/10.1016/j.cortex.2023.04.005
- Chancel, M., Ehrsson, H. H., & Ma, W. J. (2022). Uncertainty-based inference of a common cause for body ownership. *eLife*, *11*, Article e77221. https://doi.org/10.7554/eLife.77221

- de Vignemont, F. (2007). Habeas corpus: The sense of ownership of one's own body. *Mind & Language*, 22(4), 427–449. https://doi.org/10.1111/j.1468-0017.2007.00315.x
- de Vignemont, F. (2018). *Mind the body: An exploration of bodily self-awareness*. Oxford University Press.
- de Vignemont, F. (2025). Bodily awareness. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Spring 2025 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2025/entries/bodily-awareness/
- Ehrsson, H. H. (2012). The concept of body ownership and its relation to multisensory integration. In B. E. Stein (Ed.), *The new handbook of multisensory processing* (pp. 775–792). MIT Press. https://doi.org/10.7551/mitpress/8466.003.0067
- Ehrsson, H. H. (2019). Multisensory processes in body ownership. In K. Sathian & V. S. Ramachandran (Eds.), *Multisensory perception: From laboratory to clinic* (pp. 179–200). Academic Press. https://doi.org/10.1016/B978-0-12-812492-5.00008-5
- Gallagher, S. (2006). *How the body shapes the mind*. Oxford University Press. https://doi.org/10.1093/0199271941.001.0001
- Martin, M. G. F. (1992). Sight and touch. In T. Crane (Ed.), *The contents of experience: Essays on perception* (pp. 196–215). Cambridge University Press. https://doi.org/10.1017/CBO9780511554582.010
- Martin, M. G. F. (1995). Bodily awareness: A sense of ownership. In J. L. Bermúdez, A. Marcel, & N. Eilan (Eds.), *The body and the self* (pp. 267–289). MIT Press. https://doi.org/10.7551/mitpress/1640.003.0014
- Samad, M., Chung, A. J., & Shams, L. (2015). Perception of body ownership is driven by Bayesian sensory inference. *PLOS ONE*, *10*(2), Article e0117178. https://doi.org/10.1371/journal.pone.0117178
- Tsakiris, M. (2010). My body in the brain: A neurocognitive model of body-ownership. *Neuropsychologia*, 48(3), 703–712. https://doi.org/10.1016/j.neuropsychologia.2009.09.034
- Tsakiris, M. (2017). The multisensory basis of the self: From body to identity to others. *Quarterly Journal of Experimental Psychology*, *70*(4), 597–609. https://doi.org/10.1080/17470218.2016.1181768

Kristina Šekrst and Barbara Geld - Lost in Transformers: Natural Language Autoencoders and the Limits of Machine Introspection

University of Zagreb

Natural Language Autoencoders (NLAs) express a model's internal activations in the form of text (Fraser-Taliente et al., 2026): the verbalizer turns an activation into an explanation, the reconstructor changes the explanation back into an activation, and the explanation is considered effective if the journey round trip reproduces the original. This is a kind of introspective report that has been deliberately constructed, and like introspection in general it suffers from the same problems. We claim that the reconstruction criterion only assesses information sufficiency and does not address descriptive faithfulness, since a sentence can contain enough information to reconstruct an activation even if it incorrectly describes what the activation does.

The phenomenon in question is well known within the philosophy of mind and in the field of psychology. Nisbett and Wilson (1977) discovered that people explain their own behaviour by using implicit theories about what might plausibly have caused it, without being aware of the process involved, while Johansson et al. (2005) gave the participants the photograph that they had just rejected and recorded them as they gave their reasons for having chosen it. The cases of split-brain patients described by Gazzaniga (2000) show that they give smooth accounts of actions which were initiated by a hemisphere that the one speaking is unable to investigate. Each of these accounts gives useful information about the subject but is incorrect regarding the mechanism involved, and this is exactly the kind of case that the reconstruction criterion fails to detect. The situation is similar to the chain-of-thought problem (Lanham et al., 2023; Chen et al., 2025), in which output that has been trained to appear as if it is reasoning is interpreted as actually being reasoning, even though there is no guarantee that it reflects the computation that led to the answer (Šekrst, 2025). The verbalizer is able to generate confabulations that the reconstruction objective encourages (Fraser-Taliente et al., 2026), since a confabulated sentence fulfils the training requirement as long as it remains informative about the activation.

We then ask what descriptive faithfulness would require since reconstruction is a passive measure, and an accuracy condition has to be interventionist – the explanation should license correct predictions about behaviour when the feature it describes is ablated, amplified, or moved to another context. This matches accounts on which self-knowledge counts as reliable to the extent that it supports counterfactual control, with coherence after the fact carrying no evidential weight. Proposed as interpretability tools for alignment, NLAs reintroduce through unverified explanations the problem they were built to solve. The question belongs to the philosophy of the sciences of the mind – under what conditions a self-directed report may count as evidence about the process that produced it, but the round trip does not answer it.

References

- Chen, Y., et al. (2025). *Reasoning models don't always say what they think* (arXiv:2505.05410). *arXiv*. https://doi.org/10.48550/arXiv.2505.05410
- Fraser-Taliente, K., et al. (2026). Natural language autoencoders produce unsupervised explanations of LLM activations. *Transformer Circuits Thread*. https://transformer-circuits.pub/2026/nla/index.html
- Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? *Brain, 123*(7), 1293–1326. https://doi.org/10.1093/brain/123.7.1293
- Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science, 310*(5745), 116–119. https://doi.org/10.1126/science.1111709
- Lanham, T., et al. (2023). *Measuring faithfulness in chain-of-thought reasoning* (arXiv:2307.13702). *arXiv*. https://doi.org/10.48550/arXiv.2307.13702
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231
- Šekrst, K. (2025). *The illusion engine: The quest for machine consciousness*. Springer. https://doi.org/10.1007/978-3-032-05562-0

Shao-Pu Kang - The Puzzle of Shared Consciousness

National Tsing Hua University

It is widely thought that there is something special about self-knowledge of consciousness (SKC). On one way of understanding it, there is an asymmetry between knowledge of your experiences and knowledge of others' experiences. For example, you can know that you see the sunrise, that you are feeling cold, that you are thrilled, that you are imagining enjoying the view with your best friend, that you are thinking about your loved one, and so on, in a distinctively first-personal way. Of course, others can also know these propositions about your experiences, but they cannot do so in the way you do.

By drawing on recent work on shared consciousness, I raise a puzzle about the specialness of SKC, which I call the "puzzle of shared consciousness" (PSC).

PSC stems from a conflict between three independently plausible but jointly inconsistent claims. The first claim concerns a natural approach to the question of why SKC is special:

> **Specialness**: You can know your experiences in a special way because you bear a special relation to them.

The second claim concerns a potential case of shared consciousness. Krista and Tatiana Hogan are craniopagus twins, which means that they are joined at the head. What makes them even more unique is that they are joined at the thalamus, where information in each's brain can be transmitted to the other's. As a result, stimulation of each's body can cause the other to have experiences. On Cochrane's (2021) view:

> **Sharing**: The twins may sometimes have the same experiences.

The third claim concerns the implications of the twins' case for the specialness of SKC. On Kang's (2022) view:

> **Specialness despite Sharing**: Even if the twins share an experience E, each can know E in a special way.

Suppose, in accordance with **Sharing**, the twins share an experience E. Since both have E, no matter what relation each bears to E, the other bears the same relation to E. So it is not the case that each bears a special relation to E. Since they share E, it follows from **Specialness despite Sharing** that each can know E in a special way. Since each can know E in a special way, it follows from **Specialness** that each bears a special relation to E. But we just saw that it is *not* the case that each bears a special relation to E. So the three claims seem jointly inconsistent. Either one of them is false, or the reasoning above is faulty. Something must go, but what?

References

- Cochrane, T. (2021). A case of shared consciousness. *Synthese*, *199*(1), 1019–1037.
- Kang, S. P. (2022). Shared consciousness and asymmetry. *Synthese*. https://doi.org/10.1007/s11229-022-03890-w

Giulia Lorenzi - (A Variety of) Perceptual Abilities

Universidad Nacional de Educación a Distancia (UNED)

Can there be such a thing as a perceptual ability? And what would it be? Is a perceptual ability "just" and "only" the equivalence of a phenomenological state or something different? What is its role in our perceptual experience?

Historically, the idea that perceiving something new could be equivalent to the acquisition of a form of know-how or an ability has been introduced by Lewis (1988) in formulating the, so called, Ability Hypothesis, in order to reply to Jackson's (1982) Knowledge Argument. The Ability Hypothesis was originally meant to provide an alternative view, compatible with materialism, that could explain what a perceiver could gain in experiencing a certain aspect of reality for the first time. Lewis (1988) takes it that someone that sees the colour red for the first time is not gaining any new factual knowledge, but the abilities to remember, imagine and recognise red. In this reasoning, "what it feels like" to see red is fully captured by the abilities of remembering, imagining and recognising red.

A variety of objections have been raised against the Ability Hypothesis. Some philosophers questioned the role that imagination could play in this case (e.g. Kind 2019, Conee 1994). While others have taken that knowing how the experience of red "feels like" is, instead, a prerequisite for the acquisition and exercise of the abilities identified by Lewis (Gertler 1999). So it is deemed to be controversial that abilities (of the type identified by Lewis) could be "just" equivalent to a phenomenological state.

In a recent work, Lorenzi (forthcoming) introduces the idea that musical professionals listening to music perceive music differently thanks to their abilities of interacting with musical pieces in different ways. Musical professionals are presented as employing abilities that allow them to judge a musical performance, gain information relevant to write an arrangement of a piece, and transcribe a melody through perceiving music. In the cases presented here, perceivers have the ability to listen to expressive and interpretative elements of the music such as dynamics, and articulations. They also are able to listen selectively to a certain instrumental part to individuate both the role of a certain instrument,

and their relationship with others. They are also able to listen splitting a melody in shorter sections to memorise and retain the intervals between a note and another to then be able to work out the pitches of each sound.

In this paper, I consider both Lewis' idea that the phenomenological character of a first first-hand experience could be captured by the acquisition of perceptual abilities and Lorenzi's account of the experience of musical professionals listening to music to introduce a more nuanced view about the nature, and role of perceptual abilities. I want to argue that perceptual abilities could be of different nature and play a different role in shaping the "what it feels like" of a perceptual experience than the one currently explored in the literature concerning the Ability Hypothesis.

Paola Dias Bauce - Inner sense, or how we know of ourselves

Federal University of Goiás (UFG)

It is commonly assumed that we know ourselves more directly than we know the external world. Our inner experiences appear immediate and privileged in comparison with outer experience. Therefore, the scope of what we call our "inner world" appears more accessible and more immediate to us than the "outer world", and this distinction seems so ordinary, so common to us, that there's no reason to doubt it. Isn't it the case that we know ourselves better than anything or anyone else?

This assumption is characteristic of the Cartesian tradition, according to which thinking provides immediate awareness of itself. Mental acts are therefore taken to be directly accessible, making the mind epistemically privileged over the body and the external world. If I'm thinking, I know that I'm thinking; if I struggle to understand something, I know I'm struggling. And I know all these immediately; I don't need anyone to tell me these, or I don't need to access my inner states through intermediaries.

But what if there's no difference between the outer world and what we describe as our inner world? What if our inner states, such as our thoughts, desires, beliefs, fears and actions, were known no differently from a mug I use to drink coffee? What if what I call "myself" were known merely through observation? And what would I grasp by observing this "myself?

Kant maintains that the mind should not be taken as its states or effects, which is to say that consciousness should not be identified with our inner world. But what does that mean? For the rationalist, since we have immediate access to the mind, we're always conscious of our inner state, in the sense that consciousness accompanies and accesses all the states of our inner world. However, for Kant, this is not the case.

First, consciousness doesn't always have this immediate access – such access is merely possible rather than necessary. Second, what I've called here the "inner world" – the scope of our feelings, beliefs, desires, fears, but also our thoughts and reflections- is, for Kant,

due to self-affection of our inner sense, thus resulting in the passivity of the mind, essentially active, to itself and its own activity. Since the inner sense is passive and only results from some affection, our knowledge of ourselves is no different than the knowledge provided by objects, thus resulting in the thesis that the status of self-knowledge is not privileged nor clearer than what we usually believe it is.

In light of these considerations, I aim to point out how Kant argues in disfavor of immediate knowledge of our inner states and the distinction between consciousness and inner sense, in an attempt to clarify the double aspect of the "I think", since we are, at the same time, the affecting and the affected by the same activity in trying to grasp self-knowledge. Furthermore, I address the object of inner sense and Kant's refusal of a substantial I, and the consequences for subjectivity.

References

- Allison, H. E. (2004). *Kant's transcendental idealism: An interpretation and defense* (Rev. & enl. ed.). Yale University Press. https://doi.org/10.2307/j.ctt1cc2kjc
- Allison, H. E. (2024). *O idealismo transcendental de Kant: Interpretação e defesa* (A. Alves, N. Haag, & B. A. da Silva Franco, Trads.). Vozes.
- Brook, A., & Wuerth, J. (2023). Kant's view of the mind and consciousness of self. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2023 ed.). https://plato.stanford.edu/archives/win2023/entries/kant-mind/
- Descartes, R. (1973). Meditações concernentes à primeira filosofia. In *Descartes – Coleção Os Pensadores* (J. Guinsburg & B. Prado Júnior, Trads., pp. 81–150). Abril Cultural. Galloway, G. (1903). On the distinction of inner and outer experience. *Mind, 12*(45), 59–77. https://www.jstor.org/stable/2248368
- Grandjean, A. (2016). Uma interioridade do mundo: Sobre o sentido interno como "apercepção cosmológica". *Analytica – Revista de Filosofia, 19*(1), 69–91. https://doi.org/10.35920/arf.v19i1.3324
- Hendrix, J. S. (2019). *Immanuel Kant: Philosophy of perception*.
- https://docs.rwu.edu/saahp_fp/41
- Indregard, J. J. (2018). Consciousness as inner sensation: Crusius and Kant. *Ergo, 5*(7), 173– 201. https://doi.org/10.3998/ergo.12405314.0005.007

- Kant, I. (1900–). *Kritik der reinen Vernunft*. In *Kants gesammelte Schriften* (Akademie Ausgabe). Walter de Gruyter. https://korpora.zim.uni-duisburg-essen.de/kant/ Kant, I. (1992–). *The Cambridge edition of the works of Immanuel Kant* (P. Guyer & A. Wood, Eds.). Cambridge University Press.
- Kant, I. (2001). *Crítica da razão pura* (M. dos Santos & A. Mourão, Trads., 5ª ed.). Fundação Calouste Gulbenkian.
- Kant, I. (2015). *Crítica da razão pura* (F. C. Mattos, Trad., 4ª ed.). Vozes; Editora Universitária São Francisco.
- Kitcher, P. (1982). Kant on self-identity. *The Philosophical Review, 91*(1), 41–72. https://doi.org/10.2307/2184668
- Lispector, C. (2020). A procura de uma dignidade. In *Onde estivestes de noite* (1ª ed., pp. 7– 17). Rocco.
- Martins, C. (2001). O autoconhecimento na psicologia empírica de Kant. *Cadernos de História e Filosofia da Ciência, Série 3, 11*(2), 59–87.
- Ryle, G. (1992). O mito de Descartes. In B. Beakley & P. Ludlow (Eds.), *The philosophy of mind: Classical problems/contemporary issues* (pp. 23–31). MIT Press. (Original work published 1949).
- https://opessoa.fflch.usp.br/sites/opessoa.fflch.usp.br/files/Ryle-Mito-Descartes-2.pdf Zoeller, G. (1989). Making sense out of inner sense: The Kantian doctrine as illuminated by the Leningrad *Reflexion*. *International Philosophical Quarterly, 29*(2), 263–270. https://doi.org/10.5840/ipq198929325

Suchismita Ghoshal - Can the Sacred Be Measured? Neurophenomenology, Neurotheology, and the Study of Consciousness

Indira Gandhi National Open University (IGNOU)

The rapid development of neuroimaging technologies has transformed the scientific study of religious experience, giving rise to the interdisciplinary field of neurotheology. Through functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and other neuroscientific methods, researchers have identified neural correlates associated with meditation, contemplative prayer, mystical states, and altered forms of consciousness. While these findings have significantly advanced empirical knowledge of the relationship between brain activity and religious practice, they also raise a fundamental philosophical question: ***Can the sacred be adequately understood through measurements of the brain alone?*** This paper argues that although neurotheology has been remarkably successful in mapping the neural dynamics accompanying religious experience, its explanatory framework remains incomplete when it privileges third-person observation at the expense of first-person lived experience. The central objective of this study is therefore to examine whether the scientific investigation of consciousness can meaningfully account for the phenomenological depth, intentionality, and existential significance of religious experience without integrating subjective evidence.

To address this question, the paper brings Andrew Newberg's neurotheological research into dialogue with Francisco Varela's theory of neurophenomenology. Newberg's work has demonstrated that contemplative practices produce identifiable patterns of neural activity, thereby challenging earlier assumptions that mystical experiences lie beyond scientific inquiry. Yet the identification of neural correlates does not necessarily explain why these experiences are perceived as sacred, transformative, or imbued with profound meaning. Varela's neurophenomenology offers a compelling methodological alternative by proposing that rigorous first-person descriptions should complement rather than compete with neuroscientific investigation. Instead of reducing consciousness to measurable neural processes,

neurophenomenology insists that subjective experience constitutes an indispensable source of evidence for understanding the mind.

Methodologically, the study adopts a comparative philosophical approach, combining conceptual analysis with interdisciplinary engagement across philosophy of mind, cognitive neuroscience, phenomenology, and religious studies. It undertakes a critical reading of Newberg's empirical findings alongside Varela's methodological writings, while drawing upon William James's classic account of religious experience, Maurice Merleau-Ponty's phenomenology of embodiment, Evan Thompson's work on embodied cognition and contemplative science, and Shaun Gallagher's analyses of first-person methods. By placing these thinkers in conversation, the paper investigates the epistemological assumptions underlying contemporary neurotheology and evaluates the extent to which scientific accounts of consciousness depend upon phenomenological interpretation.

The paper argues that neurotheology and neurophenomenology should not be viewed as competing paradigms but as complementary approaches capable of enriching one another. Neural imaging may reveal the biological conditions accompanying religious experience, yet it cannot by itself disclose the experiential qualities through which practitioners encounter transcendence, presence, or sacredness. Conversely, phenomenological descriptions require empirical dialogue if they are to contribute meaningfully to contemporary cognitive science. Integrating these perspectives makes it possible to move beyond the persistent opposition between scientific reductionism and theological exceptionalism, replacing it with a more holistic framework for investigating consciousness.

By reframing neurotheology through the methodological lens of neurophenomenology, this paper contributes to ongoing debates in the philosophy of mind concerning the explanatory limits of neuroscience and the epistemic status of first-person experience. It further demonstrates that the study of consciousness need not choose between empirical rigor and phenomenological depth; rather, a genuinely interdisciplinary science of the mind emerges when subjective experience and objective measurement are understood as mutually informative dimensions of human cognition. In doing so, the paper proposes a more philosophically robust foundation for future research on religious consciousness, contemplative practices, and the scientific investigation of the sacred.

Filippo Murabito - Mechanistic Similarity across Biological Systems

Ben-Gurion University of the Negev / Technische Universität Berlin

Mechanisms are central to explanation in the life sciences and especially prominent in neuroscience, yet "mechanism" is used in several distinct senses (Ross & Bassett, 2024). Within the new mechanistic framework, a phenomenon is typically explained by identifying the entities and activities organized so as to produce it (Machamer et al., 2000). This characterization, however, leaves unresolved a comparative question: when are mechanisms in different biological systems sufficiently similar for knowledge of one to guide explanation of another?

This problem arises because mechanistic knowledge does not generalize automatically across organisms. Mechanisms associated with similar phenomena may differ in their components, activities, and organization (Bechtel & Abrahamsen, 2005; Bechtel & Richardson, 2010). Moreover, determining whether two mechanisms belong to the same kind requires decisions about which differences are relevant and which may be abstracted away (Francken et al., 2022). Neither phenomenal resemblance nor material identity therefore provides an adequate standard of comparison.

I argue that mechanistic similarity should be understood as a graded, multidimensional, and purpose-relative relation. The proposed framework has three stages. First, the phenomenon under investigation and the epistemic purpose of the comparison must be specified. Second, mechanistic models of the systems must be represented at a common level of abstraction and mapped across three dimensions: components, activities, and organization. Third, the mapping must be tested for the preservation of dependencies relevant to the stated purpose. In particular, interventions on mapped components and/or activities should sustain appropriately matched counterfactual expectations about changes in the target phenomenon. Mechanistic similarity is therefore stronger when a mapping preserves a greater range of explanatorily relevant dependencies, not merely when the systems share more observable or material features.

This framework integrates pragmatic and objective constraints. Research aims determine the dimensions and level of abstraction at which mechanisms are compared, while empirical evidence constrains whether the proposed mappings and their associated counterfactual implications are supported. In this way, the framework develops the idea that scientifically useful classifications should track explanatorily relevant similarities (Krickel, 2023), without permitting unrestricted analogical inference or imposing the excessively stringent requirement of mechanistic identity. Therefore, the resulting comparisons support only defeasible and scope-restricted inferences: differences in material substrate do not necessarily preclude generalization, but similarity between the phenomena alone is insufficient to warrant it. By specifying how mechanistic correspondences

should be identified, assessed, and empirically tested, the framework clarifies the conditions under which findings from one biological system can legitimately inform mechanistic inquiry into another.

References

- Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421–441. https://doi.org/10.1016/j.shpsc.2005.03.010
- Bechtel, W., & Richardson, R. C. (2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. The MIT Press. https://doi.org/10.7551/mitpress/8328.001.0001
- Francken, J. C., Slors, M., & Craver, C. F. (2022). Cognitive ontology and the search for neural mechanisms: Three foundational problems. *Synthese*, *200*(5), 378. https://doi.org/10.1007/s11229-022-03701-2
- Krickel, B. (2023). The unconscious mind worry: A mechanistic-explanatory strategy. *Philosophy of Science*, *90*(1), 39–59.
- Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25.
- Ross, L. N., & Bassett, D. S. (2024). Causation in neuroscience: Keeping mechanism meaningful. *Nature Reviews Neuroscience*, *25*(2), 81–90.

Marco Fenici - Talking Minds into Being: A Socio-Pragmatic Alternative to Theory of Mind Acquisition

Pegaso University

This talk defends a socio-pragmatic view of the acquisition of mental state (MS) lexicon—words referring to people's beliefs, desires, intentions, and emotions. According to an influential view in developmental linguistics, one of the major challenges of language learning is solving the "mapping problem," that is, determining what individual words refer to. Applied to MS vocabulary, this poses a considerable referential challenge, since mental states are commonly understood as private experiences that give rise to behaviour and are not directly perceptible. Understanding MS terms would therefore require both grasping their functional role in producing behaviour and abstracting them from concrete situations.

To account for how children nonetheless succeed in overcoming this difficulty, many scholars conclude that they rely on a prior grasp of mental concepts, taking studies on the acquisition of mindreading or Theory of Mind (ToM) in infancy and early childhood as evidence for a pre-verbal conceptual understanding of mental states. This assumption relies on a more or less tacit argument from the poverty of the stimulus: it would be impossible to explain how children learn word meanings without presupposing nonlinguistic mental capacities.

However, discussion in both philosophy and developmental psychology has provided strong reasons to reject the referential view of meaning and the associated mapping problem. An alternative account of meaning may dispel the referential challenge, thereby eliminating the main reason to assume that understanding MS talk presupposes prior mindreading capacities, and paving the way for an alternative explanation of children's acquisition of MS lexicon.

The talk will draw on socio-pragmatic and usage-based accounts, which hold that learning a language consists in entering a network of shared social practices within which linguistic

expressions acquire meaning through their pragmatic use and functional role. In particular, it will prompt reflection on children's acquisition of MS lexicon from the perspective of discursive psychology (DP). According to DP, any verbal report is inherently shaped by considerations of accountability: speakers orient to what is normal, expectable, and appropriate, while managing their own responsibility in reporting an event. Verbal reports—MS ascriptions included—are thus not neutral reflections of reality but display a rhetorical and argumentative dimension.

Applied to the proposed discussion, this implies that MS lexicon should not be understood as simply referring to inner, private states, but as part of discursive practices through which speakers accomplish specific interactional tasks. In this sense, children's acquisition of MS terms can be seen not as the expression of a prior conceptual grasp of mental states, but as the gradual mastery of the normative and rhetorical uses of such expressions within social interaction.

The talk will conclude by advocating a shift in the investigation of mindreading development—from a focus on internal representations to an emphasis on observable discursive practices—and by raising open questions for future research: What is the pragmatic function of MS ascriptions in everyday conversation? At what age, and through which developmental sequences, do children come to master these pragmatic uses? Finally, could mastering these interactional patterns support children's understanding of mental state talk, thereby functioning as a trigger for the development of mindreading abilities?

References

- Andrews, K. (2012). *Do Apes Read Minds?: Toward a New Folk Psychology*. MIT Press.
- Baillargeon, R., Scott, R. M., & Bian, L. (2016). Psychological reasoning in infancy. *Annual Review of Psychology*, *67*, 159–186. https://doi.org/10.1146/annurev-psych-010213-115033
- Bloom, P. (2000). *How Children Learn the Meanings of Words*. The MIT Press.
- Bruner, J. S. (1990). *Acts of Meaning*. Harvard University Press.
- Carruthers, P. (2013). Mindreading in infancy. *Mind & Language*, *28*(2), 141–172. https://doi.org/10.1111/mila.12014

- Childs, C. (2013). From reading minds to social interaction: Respecifying Theory of Mind. *Human Studies*, 1–20. https://doi.org/10.1007/s10746-013-9284-y
- Clark, E. V. (Ed.). (1993). The mapping problem. In *The Lexicon in Acquisition* (pp.43–66). Cambridge University Press. https://doi.org/10.1017/CBO9780511554377.004
- Edwards, D. (1996). *Discourse and cognition*. Sage Publications.
- Edwards, D., & Potter, J. (2005). Discursive psychology, mental states and descriptions. In J. Potter & H. te Molder (Eds.), *Conversation and cognition* (pp. 241–259). Cambridge University Press.
- Fenici, M. (2022). How children approach the false belief test: Social development, pragmatics, and the assembly of Theory of Mind. *Phenomenology and the Cognitive Sciences*, *21*(1), 181–201. https://doi.org/10.1007/s11097- 020-09709-8
- Fenici, M., & Zawidzki, T. W. (2021). The origins of mindreading: How interpretive socio-cognitive practices get off the ground. *Synthese*, *198*(9), 8365–8387. https://doi.org/10.1007/s11229-020-02577-4
- Fernández Castro, V. (2019). Justification, conversation, and folk psychology. *THEORIA. An International Journal for Theory, History and Foundations of Science*, *34*(1), 73–88. https://doi.org/10.1387/theoria.18022
- Helming, K. A., Strickland, B., & Jacob, P. (2016). Solving the Puzzle about Early Belief-Ascription. *Mind & Language*, *31*(4), 438–469. https://doi.org/10.1111/mila.12114
- Hutto, D. D. (2008). *Folk Psychological Narratives*. The MIT Press.
- Hutto, D. D., & Ravenscroft, I. (2025). Folk Psychology as a Theory. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2025). https://plato.stanford.edu/archives/win2025/entries/folkpsych theory/
- McGeer, V. (2007). The regulative dimension of folk-psychology. In D. D. Hutto & M. Ratcliffe (Eds.), *Folk-psychology Reassessed*. Springer.
- Quine, W. V. O. (1960). *Word and Object*. The MIT Press.
- Schegloff, E. A. (1995). Discourse as an Interactional Achievement III: The Omnirelevance of Action. *Research on Language and Social Interaction*, *28*(3), 185–211. https://doi.org/10.1207/s15327973rlsi2803_2 Sellars, W. (1974). Meaning as functional classification. *Synthese*, *27*(3), 417–437.

- Slaughter, V., & Peterson, C. C. (2011). How conversational input shapes theory of mind development in infancy and early childhood. In M. Siegal & L. Surian (Eds.), *Access to Language and Cognitive Development* (pp. 4–22). Oxford University Press.
- Tomasello, M. (2003). *Constructing a Language*. Harvard University Press.
- Tomasello, M. (2009). The usage-based theory of language acquisition. In E. Bavin (Ed.), *The Cambridge Handbook of Child Language* (pp. 69–88). Cambridge University Press.
- Turnbull, W. (2003). *Language in Action: Psychological Models of Conversation* (1 edition). Routledge. Turnbull, W., & Carpendale, J. I. M. (1999). A Social Pragmatic Model of Talk: Implications for Research on the Development of Children's Social Understanding. *Human Development*, *42*(6), 328–355.
- https://doi.org/10.1159/000022641
- Westra, E., & Carruthers, P. (2017). Pragmatic development explains the Theory-of-Mind Scale. *Cognition*, *158*, 165– 176. https://doi.org/10.1016/j.cognition.2016.10.021
- Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell.
- Wojcik, E. H., Zettersten, M., & Benitez, V. L. (2022). The map trap: Why and how word learning research should move beyond mapping. *Wiley Interdisciplinary Reviews. Cognitive Science*, *13*(4), e1596. https://doi.org/10.1002/wcs.1596

Dr Tom McClelland - The Validation Problem for Markers of AI Consciousness

University of Cambridge

The leading framework for assessing consciousness is the markers approach (see Bayne *et al* 2024). According to this approach, we should test for consciousness by looking for a set of observable markers likely to be indicative of consciousness. This 'theory light' approach avoids the difficulties faced by 'theory heavy' approaches that depend on an unwarranted commitment to a specific theory of consciousness. It also avoids the problems faced by 'theory free' approaches that depend on unreliable intuitive attributions of consciousness. The marker approach has many virtues and has yielded promising results in contexts such as cephalopod consciousness (Birch *et al* 2021). A landmark paper led by Butlin & Long *et al* (2023) argues that the same approach should be taken to consciousness in AI, and offers a list of indicators that would be evidence of artificial consciousness. But is it legitimate to apply these markers in the context of AI? I argue not.

My argument does not question whether the markers proposed are ever good indicators of consciousness. Instead, my argument questions whether they should be treated as indicators of consciousness across all contexts. The markers are established in the context of organisms. Transporting these markers from organisms to inorganic machines is a significant change of population. Normally, the use of markers in a significantly different population requires those markers to be validated in the target population. But when it comes to AI, no such validation is available. Moreover, there are considerable obstacles to providing such validation in the future. This means that the use of these markers as indicators of consciousness in AI is unwarranted. The result is that putative tests for consciousness in AI should not be trusted. I carefully distinguish this argument from other arguments in the literature that resist taking the markers approach to AI consciousness. This includes arguments based on general scepticism about AI consciousness, arguments based on Bayesian prior probabilities of AI consciousness, and arguments based on 'gaming' considerations.

Next, I consider several likely responses to my argument. The first is that the argument draws an arbitrary line between how consciousness should be tested in organisms and how it should be tested in AI. I show that the line is principled and does not rely on any chauvinistic bias

against AI consciousness. The second likely response is to concede that possession of the markers by AI is not straightforwardly indicative of consciousness but to maintain that it still yields weak evidence that can help guide our judgements. I show that having unvalidated markers of consciousness does not even yield weak evidence. The third is to appeal to the moral import of AI consciousness and to insist that even a problematic test for consciousness is better than nothing. I show that this approach is unscientific and offer an alternative approach to concerns about AI welfare that does not rely on testing AI for consciousness.

References

- Bayne, T., Seth, A., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S. M., Malach, R., Mattingley, J. B., Menon, D. K., Owen, A. M., Peters, M. A. K., Razi, A., & Mudrik, L. (2024). Animals and the iterative natural kind strategy. *Trends in Cognitive Sciences, 28*(10), 876–877.
- Birch, J., Burn, C., Schnell, A., Browning, H., and Crump, A. (2021). Review of the Evidence of Sentience in Cephalopod Molluscs and Decapod Crustaceans. LSE Consulting.
- Butlin, P. & Long, R., et al. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

Victor Angelucci - A problem for the naturalization of mental files

Federal University of Minas Gerais (UFMG)

Mental Files are mental representations posited in the context of Frege's Puzzle. Files are singular concepts with an informational architecture attached to them. They "store" information taken to be satisfied by the file's (purported) referent. Their more traditional role is to account for the opacity of correferential expressions and associated mental counterparts (see e.g. Perry, 2011; Recanati, 2012). Files also play a role in characterizing singular thoughts, propositional attitudes directed at objects. The so-called Mental File theory of singular thought (Jeshion, 2010; Goodman & Genone, 2020; Murez *et al,* 2020) characterizes singular thought as a mental state that employs a particular file.

The interest in files has grown considerably in the last few years. For instance, files are being applied in discussions about episodic memory (Andonovsky, 2024), discourse referents in linguistics (Lewis, *forthcoming*) and developmental psychology (Perner *et al*, 2015; Perner, 2025). Theorists are becoming interested in how the properties usually attributed to files, their purported singular reference and informational profile, can help solve a variety of problems (e.g. are engrams *about* or causally linked to some particular event or what exactly is a discourse referent). But this raising interest also carries an important theoretical preoccupation: files are theoretical posits which do not have, initially, clear empirical support. Intuitively, we can speak of "my file" about Hesperus and about the information it stores, but it is not clear if there is an underlying cognitive or representational kind (Smortchkova & Murez, 2020) or mechanism. This is relevant to see whether or not the different applications of mental files are explanatory (Murez, 2023).

This presentation aims at presenting a problem for the project of finding a cognitive realizer of mental files; more specifically, demonstrative and perceptual mental files. Demonstrative files are a type of mental files that is created by a perceptual encounter with an object (Perry's (2011) *buffers*). One plausible hypothesis takes object-files, a type of perceptual representation posited by cognitive scientists in the context of MOT and feature-change tasks (Kahneman *et al*, 1992;

Luck & Vogel, 1997; Pylyshyn, 2007; Green & Quilty-Dunn, 2021), to be the cognitive realizers of demonstrative mental files.

Following the work of (Murez *et al,* 2020), I will try and pressure this "reduction" by pointing to a property of (demonstrative) mental files apparently not present in object-files: their status as *bona fide* concepts. As Recanati (2012) argues, mental files satisfy the Generality Constraint (Evans, 1982); even demonstrative files can be redeployed in relation

to different predicates. If I am able to think that [this is F], I should be able to think that [this is G, P and so on]. Object-files, representations with very clear memory and capacity limits and with a short-lived lifespan in visual working memory, do not seem to be generalizable representations.

Mental files are interesting notions with many fruitful applications. This presentation pressures one way of understanding how files could be realized in cognition, but its utmost aim is to draw attention to the fact that files should be better understood empirically. This would strengthen our understanding not only of mental files, but of their explanatory power and of the problems they solve.

References

- Andonovski, Nikola. 2024. "Engrams as Mental Files." *Synthese* 204 (6). https://doi.org/10.1007/s11229-024-04801-x.
- Clarke, Henry. 2022. "Mental Filing Systems: A User's Guide." *Ergo an Open Access Journal of Philosophy* 8 (0). https://doi.org/10.3998/ergo.2249.
- Crimmins, Mark, and John Perry. 1989. "The Prince and the Phone Booth: Reporting Puzzling Beliefs." *The Journal of Philosophy* 86 (12): 685. https://doi.org/10.2307/2027014.
- Evans, Gareth. 1982. *The Varieties of Reference*. Edited by John McDowell. Clarendon Press, New York.
- Frege, Gottlob. 1892. "Uber Sinn Und Bedeutung." *Zeitschrift Für Philosophie Und Philosophische Kritik* 100 (1): 25–50.
- Goodman, Rachel. 2024. "Mental Files." *Philosophy Compass* 19 (3). https://doi.org/10.1111/phc3.12967.

- Goodman, Rachel, and James Genone. 2020. “Singular Thought and Mental Files.” *Singular Thought and Mental Files*, February 6, 1–18. https://doi.org/10.1093/oso/9780198746881.003.0001.
- Green, E. J., and Jake Quilty-Dunn. 2021. “What Is an Object File?” *The British Journal for the Philosophy of Science* 72 (3). https://doi.org/10.1093/bjps/axx055.
- Jeshion, Robin. 2010. “Singular Thought: Acquaintance, Semantic Instrumentalism, and Cognitivism.” In *New Essays on Singular Thought*, edited by Robin Jeshion. Oxford University Press.
- Kahneman, Daniel, Anne Treisman, and Brian J Gibbs. 1992. “The Reviewing of Object Files: Object-Specific Integration of Information.” *Cognitive Psychology* 24 (2): 175–219. https://doi.org/10.1016/0010-0285(92)90007-o.
- Lewis, Karen. (*forthcoming*) “Discourse Referents in a Dynamic Pragmatic.” In *Oxford Studies in Philosophy of Language, Volume 4.*, edited by Ernest Lepore and David Sosa.
- Luck, Steven J., and Edward K. Vogel. 1997. “The Capacity of Visual Working Memory for Features and Conjunctions.” *Nature* 390 (6657): 279–81. https://doi.org/10.1038/36846.
- Murez, Michael. 2023. “The Transparency of Mental Vehicles.” *Noûs*, ahead of print, November 14. https://doi.org/10.1111/nous.12483.
- Murez, Michael, Brent Strickland, and Joulia Smortchkova. 2020. “The Mental Files Theory of Singular Thought: A Psychological Perspective.” In *Singular Thought and Mental Files*, edited by Rachel Goodman, James Genone, and Nick Kroll. Oxford University Press.
- Papineau, David. 2013. “Comments on François Recanati’s Mental Files: Doubts about Indexicality.” *Disputatio* 5 (36): 159–75. https://doi.org/10.2478/disp-2013-0015.
- Perner, Josef. 2025. *Mental Files in Perspective*. Oxford University Press.
- Perner, Josef, Michael Huemer, and Brian Leahy. 2015. “Mental Files and Belief: A Cognitive Theory of How Children Represent Belief and Its Intensionality.” *Cognition* 145 (December): 77–88. https://doi.org/10.1016/j.cognition.2015.08.006.
- Perry, John. 2011. *Reference and Reflexivity*. 2nd edition, Csli Publications.
- Pylyshyn, Zenon W. 2007. *Things and Places : How the Mind Connects with the World*. Mit Press.
- Recanati, François. 2012. *Mental Files*. OUP Oxford.

- Smortchkova, Joulia, and Michael Murez. 2020. “Representational Kinds.” In *What Are Mental Representations?*, edited by Joulia Smortchkova, Krzysztof Dołęga, and Tobias Schlicht. Oxford University Press.

Hilde Teigen - Rethinking self-organisation in neuroscience

University of Oslo

Self-organisation is widely invoked in neuroscience, yet a consistent account of what it means for the brain to be self-organising is lacking. Influential accounts are largely derived from models of physical systems whose assumptions do not readily map onto the organisation of the brain, arguably contributing to inconsistent use of the concept across developmental, systems, and cognitive neuroscience. In this paper, I develop a coherent account of self-organisation that maintains a biological focus making the concept more operational for neuroscience explanation.

Accounts of self-organisation are often based on models of non-living systems. They typically assume homogeneous components, clear system boundaries, and identifiable global patterns emerging from local interactions. Such models are valuable as they reveal dynamical principles shared across diverse systems, an approach supported by philosophical work on minimal models and universality classes. However, abstractions and idealisations required by this strategy risks obscuring significant biological detail that distinguish the brain from other systems. I argue that rather than treating some of the common idealisation strategies as defining features of self-organisation, they should be made flexible in a way that accommodates for brain phenomena across multiple scales and biological relevance.

I develop an account of self-organisation that is justified though its explanatory usefulness in neuroscience, without requiring that the phenomenon of interest is placed within a universality class. For the ambitions of neuroscience, it is often more explanatorily relevant to focus on what is distinctive about the brain, than to focus on the properties shared with other systems. Accordingly, I develop an account of self-organisation that can be identified in relation to explanatory aims, while remaining constrained by the nature of the brain as a part of a living system. This means that model features like components, system boundaries, and emerging patterns are defined functionally depending on the particularities of the system of interest. For example, this includes a flexible notion of the idea of a global pattern that aligns to the role those patterns play in the brain. For developmental neuroscience, the global patterns can be defined through the functionality of a neural circuitry, while for systems and

cognitive neuroscience a global pattern can be identified as behaviourally productive temporal sequences of neural activation. The result is a flexible and biologically grounded notion of self organisation, that maintains the benefit of showing how global patterns can arise in a decentralized and non-linear manner. In this sense, the self-organisation concept can be adopted for widely different explanatory aims.

The main contribution of this paper is to provide an account of self-organisation that can remain biologically grounded, while retaining the strength of systems thinking. This establishes a role for self-organisation within a contemporary neuroscientific explanatory context, building a useful basis for related concept such as emergence, bifurcation, constrains, and top-down causation for the brain as integrated in a tightly coupled system with the body and the environment.

Julieta Zapata - One Predictive Architecture, Different Experiences: What Distinguishes Perception from Emotion?

Faculty of Psychology, Universidad Nacional de Córdoba, Argentina

Lisa Feldman Barrett's Theory of Constructed Emotion offers a unified account of mind: rather than treating perception, emotion, and cognition as products of distinct mechanisms, it explains them through a common predictive architecture. The brain uses prior experience to predict the causes of sensory signals, regulate the body, and guide action, while perception, categorization, and action constitute aspects of an integrated constructive process (Barrett, 2017; Barret & Lida, 2024). This unification is one of the theory's central strengths. Yet it also generates a problem: if different mental phenomena are constructed from shared processes and resources, what identifies them as different forms of experience?

Consider perceiving a dog and experiencing fear. What, then, makes one an experience of an external object and the other an emotion? A straightforward answer might appeal to categorization: in one case, sensory signals are categorized as a "dog", while in the other they are categorized as "fear". Yet this shifts rather than solves the problem. If both experiences are constructed through the same general process of categorization, we still need to explain what constrains that process such that experience is organized in one case as the perception of an object and in the other as an emotion.

Although Barrett does not explicitly formulate the problem in these terms, her framework suggests three possible dimensions along which such differentiation might occur: sensory modality (I), affectivity (II), and allostatic relevance (III). I examine the explanatory potential of each dimension for distinguishing object perception from emotion and argue that none, taken individually, provides a sufficient basis for their individuation. Object perception may rely predominantly on exteroception and emotion more strongly on interoception, but Barrett's architecture integrates both modalities. Affectivity cannot uniquely characterize emotion because affect also participates in object perception: we perceive a world that already has significance for the organism (Barrett & Bar, 2009). Finally, allostatic relevance cannot provide

a categorical boundary because anticipatory bodily regulation is a general organizing principle of brain activity rather than a feature specific to emotion (Barrett & Simmons (2015).

This failure suggests that experiential individuation should not be understood compositionally. I propose instead a configurational account, according to which experiences differ in the relative prominence and organization of shared dimensions within what I call the experiential landscape. When perceiving a dog, features involved in constructing an external entity—its shape, location, movement, or identity—may be foregrounded, while bodily, affective, and regulatory dimensions remain in the background. During fear, dimensions specifying how the situation affects the organism—valence, bodily arousal, regulatory demands, and possibilities for action—may instead become prominent, without the external object disappearing from experience.

The difference between perception and emotion would therefore be less compositional than configurational: distinct experiences need not possess different basic ingredients but may emerge from different organizations of shared resources. This proposal preserves Barrett's

commitment to a unified predictive architecture while offering an account of how that architecture can nevertheless generate differentiated forms of experience.

References

- Barrett, L. F. (2017). *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt.
- Barrett, L. F., & Bar, M. (2009). See it with feeling: Affective predictions during object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1521), 1325–1334.
- Barrett, L. F., & Lida, T. (2024). Constructionist theories of emotions in psychology and neuroscience. In A. Scarantino (Ed.), *Emotion theory: The Routledge comprehensive guide* (pp. 350–387). Routledge.
- Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience, 16*, 419–429.

Muzaffar Abass Wazir - Establishing the uniqueness of extended mind thesis in the space of 4E cognition

Indian Institute of Technology Bombay

Two dominant interpretations of the extended mind thesis dominate the debate on the scholarship on the extended mind. First, the extended mind thesis is about the boundary of the mind or cognition. Second, the extended mind thesis is about the nature of cognition. However, in the light of the other 4E cognitions, embedded, embodied and enactive, the uniqueness of the extended mind thesis proposal is seriously undermined. Particularly, when the interpretation is the first one. In the light of such a predicament, in this paper I propose what is unique to the extended mind thesis. However, the response to such an endeavour requires positing the nature of question to which the extended mind thesis is an answer. And as per my understanding it is the following question: what is the nature of the relationship that human beings form/create with the world/or parts of the world that it interacts with? Formulating this question offers two solutions. First, it demonstrates why embedded cognition is the arc nemesis of extended mind thesis. In response to the above question, the embedded thesis offers an instrumental nature of interaction with the world/parts of the world. Second, the question allows the opportunity for the extended mind thesis to propose the uniqueness that it inhabits in the space of 4E cognition. In this paper, I propose the uniqueness by pointing out that the extended mind thesis is about the emergence of a new cognitive profile. However, the uniqueness is not in this emergence alone but what this emergence entails. The emergence of the novel cognitive profile entails a nature of interaction that creates the possibility of newer forms of cognitive profile.

With the nature of interaction that the extended mind thesis envisages, the subject while interacting with parts of the world form a novel cognitive profile. With this novel cognitive profile emerges a new way of looking at the world; for instance, with the notebook in hand, for Otto, the world is no longer an unchartered arena. However, the extended mind thesis does not stop there, it moves further. With this new way of seeing the world, further avenues of newer forms of cognitive profile emerge. And realization of those newer forms of cognitive profile further entails the emergence of new ways of seeing the world. It is precisely in this new way

of seeing the world in tandem with the emergence of newer forms of cognitive profile, the extended mind thesis establishes its uniqueness in the space of 4E cognition. In a manner of speaking, the

extended mind thesis allows the spreading out of the subject into the world, in order to allow the emergence of a new mode of being.

References

- Clark, A. (1997). *Being there: Putting brain, body, and world together again*. MIT Press.
- Clark, A. (2008). *Supersizing the mind: Embodiment, action and cognitive extension*. Oxford: Oxford University Press.
- Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.
- Clark, A. (2024). *The experience machine: How our minds predict and shape reality*. Penguin Books.

Jonida Kodra - Control and Sense of Agency in Auditory Verbal Hallucinations

Osnabrueck University

Auditory verbal hallucinations (AVH), defined as experiences of hearing voices in the absence of an external speaker (APA, 2022), raise important clinical and philosophical questions about the nature of hallucinations and mental agency (Stephens & Graham, 2000). I propose that a central aspect of this experience is control (or lack thereof). Researchers assume that AVHs are beyond the control of voice-hearers, and take the lack of control respectively (a) to be involved in *causing* the experience of AVHs (Wu, 2012), (b) to imply that voice-hearers *cannot* voluntarily *initiate an episode as AVH* (Langland-Hassan, 2016), or (c) to indicate that AVHs are non-veridical perceptual experiences as opposed to other inner auditory experiences (Knappik et al., 2022).

In response, I want to make two suggestions. My first suggestion that a more thorough consideration of the empirical literature shows that voice-hearers report that they can exercise control. When exercising control, voice-hearers *want* and *are able* to generate, inhibit, and change some of the features of AVH episodes (Moritz & Larøi, 2008; Nayani & David, 1996).

Against the abovementioned views (Knappik et al., 2022; Langland-Hassan, 2016; Wu, 2012), and in line with both David's (2004) suggestion and empirical findings, I more specifically propose that voice-hearers lack the *feeling* or *sense* of control, but not the *ability*. The claim is not that all AVH episodes are under the control of voice-hearers, but rather that voice-hearers can occasionally voluntarily generate, inhibit, and change some of the features of their AVH episodes. Nonetheless, despite the exercise of their *ability* to be in control, voice-hearers still lack the *feeling* of having done so.

My second suggestion is that empirical data can help to answer some important philosophical and clinical questions. For one, they allow for a clearer understanding of the process underlying AVH experiences and, by extension, also of the process that allows one to experience one's thoughts or thought-like experiences as self-generated, standardly known in the clinical and

philosophical literature as *sense of agency* (Gallagher, 2004; Braun et al., 2018). A suggestion commonly put forward by several clinicians and philosophers in fact is that AVHs are internally-generated phenomena that are experienced as non-self-generated due to a disrupted *sense of agency (*Jones & Fernyhough, 2007; Stephens & Graham, 2000). In light of the proposals that voice-hearers have the ability to voluntarily generate AVH episodes and that AVHs are the result of an impaired sense of agency, a more specific suggestion is that empirical findings indicate that the process underlying the *sense of agency* might be one voice-hearers have some control over. In addition, the ability and exercise of control over some of the features of AVH episodes (e.g., the content, volume, etc.) also indicate that the mental phenomenon that the voice-hearer experiences as non-self-generated is a phenomenon which one can normally exercise some control over (e.g., inner speech, imagination, etc.) (Gregory, 2016; Langland Hassan, 2016), and not an auditory perceptual experience as suggested by other theorists.

References

- American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.).
- Braun, N., Debener, S., Spychala, N., Bongartz, E., Sörös, P., Müller, H. H., & Philipsen, A. (2018). The senses of agency and ownership: a review. *Frontiers in psychology*, *9*, 535.
- David, A. S. (2004). The cognitive neuropsychiatry of auditory verbal hallucinations: an overview. *Cognitive neuropsychiatry*, 9(1-2), 107–123.
- Knappik, F., Bless, J., & Larøi, F. (2022). Confusions about 'inner' and 'outer' voices: conceptual problems in the study of auditory verbal hallucinations. *Review of Philosophical Psychology,* 13, 215–36.
- Gallagher, S. (2006). Agency, ownership, and alien control in schizophrenia. In *The structure and development of self-consciousness: Interdisciplinary perspectives* (pp. 89-104). John Benjamins Publishing Company.
- Gregory, D. (2016). Inner speech, imagined speech, and auditory verbal hallucinations. *Review of Philosophy and Psychology*, *7*(3), 653-673.

- Langland-Hassan, P. (2016). Hearing a voice as one's own: two views of inner speech self monitoring deficits in schizophrenia. *Review of Philosophy and Psychology*, *7*(3), 675-699.
- Moritz, S., & Larøi, F. (2008). Differences and similarities in the sensory and cognitive signatures of voice-hearing, intrusions and thoughts. *Schizophrenia research*, *102*(1-3), 96- 107.
- Nayani, T. H., & David, A. S. (1996). The auditory hallucination: a phenomenological survey. *Psychological medicine*, *26*(1), 177-189.
- Stephens, G. L., & Graham, G. (2000). *When self-consciousness breaks: Alien voices and inserted thoughts*. The MIT press.
- Wu, W. (2012). Explaining schizophrenia: Auditory verbal hallucination and self-monitoring. *Mind & Language*, *27*(1), 86-107.

Lucas Battich - From perceptual to non-perceptual joint attention: A capacities-based account

University of Tübingen

People effortlessly engage with each other in activities that require attending together to an object or event, where we are mutually aware of our acts of attention. It is often acknowledged that the phenomenon of joint attention is manifest in a variety of different situations, ranging from visual perceptual cases to situations where the objects of joint attention become increasingly abstract, including pretence play with semi-imaginary objects, to cases of full joint imagination, as when we plan our next holiday together (e.g., Mundy, 2016; Tomasello, 1995). We often engage in exercises of joint mental alignment towards non-perceptual objects that are not currently present, or even that do not exist. Most philosophical accounts, however, center only on the perceptual, typically visual, case. They cannot easily address what can be called, taking the cue from Wilby (2020), the scaling-up problem for theories of joint attention: how can we explain the relation between perceptual and non-perceptual forms of joint attention?

Here, I advance a novel account of joint attention that addresses the scaling-up problem. It aims to explain our ability to jointly focus on perceptual, absent, and even non-existent targets. To do so, and in contrast to many current approaches, I here separate issues regarding (1) truth conditions and (2) justification for a joint attention situation, from explaining (3) joint attention as a *psychological capacity*. The primary explanatory aim of this account is to characterise the psychological capacities that enable agents to engage in joint attention situations, from perceptual to non-perceptual cases.

I argue that to engage in joint attention is to exercise the capacity to predict and recognise when one is or is not attending to the same object *together* with someone. Crucially, this capacity does not require any conceptual understanding of the truth conditions under which joint attention is objectively instantiated. The recognition of togetherness rests instead on an implicit practical sensitivity to the further actions, behaviours, and epistemic upshots that a joint attention situation usually affords. Building on recent work in developmental psychology, I argue that this capacity recruits sub-personal psychological processes involving heuristic

estimations, ultimately resulting in an assumption of togetherness at the personal level. This output is often fallible and not fully justified, but it nevertheless allows for efficient coordination.

I then outline the minimal satisfaction conditions for an objective situation of joint attention: all it takes for an instance of joint attention to X to be true is that each co-attender take themselves to be attending to X together. Importantly, I show how these minimal conditions, and the individual assumptions on which they are based, are sufficient to explain the functional role of joint attention in our social lives. Finally, I show how this capacities-based account allows for natural extensions to cases of joint pretence play and joint imagination, addressing the scaling-up problem under the same explanatory framework.

References

- Mundy, P. (2016). *Autism and Joint Attention: Development, Neuroscience, and Clinical Fundamentals*. Guilford Publications.
- Redcay, E., & Saxe, R. (2013). Do You See What I See? The Neural Bases of Joint Attention. In S. H. Terrace & J. Metcalfe (Eds), *Agency and Joint Attention* (pp. 216–237). Oxford University Press.
- Wilby, M. (2020). From Joint Attention to Common Knowledge [Review of *The Shared World*, by A. Seemann]. *The Journal of Mind and Behavior*, *41*(3/4), 293–306.

Daniel De Luca-Noronha - A Critique of the Conception of Pain as a Homeostatic Feeling

Jesuit Faculty of Philosophy and Theology

The conception of pain as a homeostatic feeling has been proposed as an alternative to representationalist theories of pain experience. Drawing on A. D. Craig (2003), Colin Klein (2015), for example, argues that pain states have imperative content: rather than representing bodily damage or disturbances, they command the subject to protect a particular part of the body. Their biological function would be homeostatic, since pain-motivated actions would contribute to restoring bodily integrity.

My aim is to argue that pain's motivational or imperative character is not sufficient to characterize it as a homeostatic feeling. I shall maintain that a feeling of this kind should satisfy at least three conditions: (i) it should bear an identifiable functional relation to a regulated variable; (ii) it should vary systematically with functionally relevant changes in that variable; and (iii) it should be attenuated or eliminated when the deviation that produced it is corrected, thereby participating in a feedback dynamic directed at stabilizing the organism. These criteria follow from a conception of physiological regulation as the maintenance of variables within functional limits by means of responses that are sensitive to deviations from those limits (FRIEDMAN, 2008; RAMSAY; WOODS, 2014).

Pain experience, however, does not appear to satisfy these conditions. Regarding the first condition, no regulated physicochemical variable can be identified whose alteration functionally corresponds to pain experience. Pain may result from inflammation, neural dysfunction, central sensitization, and emotional or contextual factors. This heterogeneity makes it difficult to treat pain as a response to a unified class of bodily disturbances (CORNS, 2020). Nor does 'bodily integrity' clearly constitute a regulated variable, since it has neither a common unit of measurement nor a specifiable functional range. Regarding the second condition, pain intensity does not systematically covary with the magnitude of tissue damage or any other homeostatic deviation. Severe injuries may occur without pain, whereas intense pain may exist without a stable correspondence to current damage, as in phantom-limb pain

and cases of chronic and nociplastic pain (RAJA et al., 2020). Regarding the third condition, restoring peripheral conditions or physiological stability does not necessarily eliminate pain experience. In chronic pain, the experience may persist even after the injury or disturbance initially associated with its onset has been resolved.

I shall conclude that, if pain has imperative content, this content should be understood as part of a flexible mechanism of control and protection, rather than as a mechanism of homeostatic regulation. Pain may recruit, select, and reorganize protective behaviours without tracking a regulated variable, covarying with deviations in that variable, or directly contributing to its stabilization. This interpretation preserves pain's motivational and protective character without presupposing either a specific homeostatic variable or a set point for bodily integrity.

References

- CORNS, Jennifer. *The Complex Reality of Pain*. New York: Routledge, 2020. CRAIG, A. D. A new view of pain as a homeostatic emotion. *Trends in Neurosciences*, vol. 26, no. 6, pp. 303–307, 2003.
- FRIEDMAN, M. Food intake: control, regulation, and the illusion of dysregulation. In: *Appetite and Food Intake: Behavioral and Physiological Considerations*. Taylor & Francis, 2008, pp. 1– 16.
- KLEIN, Colin. *What the Body Commands: The Imperative Theory of Pain*. Cambridge, MA: MIT Press, 2015.
- RAJA, S. N. et al. The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises. *Pain*, vol. 161, no. 9, pp. 1976–1982, 2020. RAMSAY, Douglas S.; WOODS, Stephen C. Clarifying the roles of homeostasis and allostasis in physiological regulation. *Psychological Review*, vol. 121, no. 2, pp. 225–247, 2014.

Lotem Elber-Dorozko, Galit Agmon, and Omri Abend - What is linguistic capacity? Lessons from LLM evaluations

Hebrew University of Jerusalem (Lotem and Omri) and Bar Ilan University (Galit)

Language is seen as a central cognitive capacity, in line with memory, perception, and more. But what does it mean to have a capacity for language? We argue that attempts to answer this question with regard to Large Language Models (LLMs) demonstrate that this is still an open question. Specifically, it is unclear what it means to have a capacity for language in a way that is not necessarily human.

There is rich philosophical and scientific literature demonstrating that various cognitive capacities can be performed in ways that are not human-like. Deep Blue in Chess-playing, and Block's theoretical Blockhead (1981) for the Turing

test, demonstrate manners of performing cognitive tasks that are fundamentally different from human cognition. When considering object classification from images, there are many carefully crafted non-natural images where humans and AI models differ in their classification, leading to a debate on whether AI models' internal processes are similar to humans (Bowers et al. 2022; Buckner 2023). Thus, for a variety of cognitive capacities there is a clear distinction between the general performance of this capacity and its performance in a human-like manner.

When it comes to language, this distinction is not easy to make. Linguistic capacity is often examined together with other capacities, such as reasoning or knowledge (Turing 1950; Chiang et al. 2024; Wang et al. 2024). Recently, following work in generative linguistics, Mahowald and colleagues (2024) distinguished functional linguistic capacity from formal linguistic capacity. The former includes reasoning and knowledge and is tested by the cited benchmarks, while the latter includes grammatical, phonological and morphological knowledge, examined through minimal pairs of acceptable and unacceptable linguistic utterances (e.g., Warstadt et al. 2020). Models that are able to recognize grammatical vs.

ungrammatical utterances are said to have formal linguistic capacity (or lack it when they fail).

However, this 'formal linguistic capacity' cannot be defined separately from human behavior. Models are taken to have or lack formal linguistic capacity according to the extent to which their outputs resemble human answers. When research indicates differences between models and humans on formal linguistic tasks (Lan et al. 2026; Huang et al. 2024), is this evidence that models lack formal linguistic capacity or simply that their linguistic capacity is different from that of humans? Consider a test of knowledge about world capitals. Most people might label Rio de Janeiro as the capital of Brazil, while the model may choose Brasília. Despite its misalignment with human answers, the model has performed the capacity well. Such a result is not possible with current tests for formal linguistic capacity.

The upshot is that minimal pairs from generative linguistics are useful to test whether models use language *in the same way* as people, and it is still undetermined what linguistic capacity is, per se. We discuss several possible answers to this question, including viewing linguistic capacity as uniquely human, as an instance of a more abstract communicative capacity, or denying that it is a capacity at all.

References

- Block, Ned. 1981. "Psychologism and Behaviorism." *The Philosophical Review* 90 (1): 5–43. https://doi.org/10.2307/2184371.
- Bowers, Jeffrey S., Gaurav Malhotra, Marin Dujmović, et al. 2022. "Deep Problems with Neural Network Models of Human Vision." *The Behavioral and Brain Sciences* 46 (December): e385.
- https://doi.org/10.1017/S0140525X22002813.
- Buckner, Cameron. 2023. "Black Boxes or Unflattering Mirrors? Comparative Bias in the Science of Machine Behaviour." *The British Journal for the Philosophy of Science* 74 (3): 681–712. https://doi.org/10.1086/714960.
- Chiang, Wei-Lin, Lianmin Zheng, Ying Sheng, et al. 2024. "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference." Paper presented at Forty-first International Conference on Machine Learning. June 6. https://openreview.net/forum?id=3MW8GKNyzI.

- Huang, Kuan-Jung, Suhas Arehalli, Mari Kugemoto, et al. 2024. “Large-Scale Benchmark Yields No Evidence That Language Model Surprisal Explains Syntactic Disambiguation Difficulty.” *Journal of Memory and Language* 137 (August): 104510. https://doi.org/10.1016/j.jml.2024.104510.
- Lan, Nur, Emmanuel Chemla, and Roni Katzir. 2026. “Large Language Models and the Argument from the Poverty of the Stimulus.” *Linguistic Inquiry* 57 (2): 315–42. https://doi.org/10.1162/ling_a_00533.
- Mahowald, Kyle, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. “Dissociating Language and Thought in Large Language Models.” *Trends in Cognitive Sciences* 28 (6): 517–40. https://doi.org/10.1016/j.tics.2024.01.011.
- Turing, Alan M. 1950. “COMPUTING MACHINERY AND INTELLIGENCE.” *Mind* LIX (236): 433–60. https://doi.org/10.1093/mind/LIX.236.433.
- Wang, Yubo, Xueguang Ma, Ge Zhang, et al. 2024. “MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark.” arXiv:2406.01574. Preprint, arXiv, November 6.
- https://doi.org/10.48550/arXiv.2406.01574.
- Warstadt, Alex, Alicia Parrish, Haokun Liu, et al. 2020. “BLiMP: The Benchmark of Linguistic Minimal Pairs for English.” *Transactions of the Association for Computational Linguistics* (Cambridge, MA) 8: 377–92. https://doi.org/10.1162/tacl_a_00321.

Libni Teles - FEATURE RECOGNITION AS A MECHANISM FOR REPRESENTATIONAL CONSTANCY

What we seek to present are the characteristics that qualify pattern recognition—known as feature recognition—in Convolutional Neural Networks (CNNs) as capable of satisfying the conditions of the representational constancy mechanism. This mechanism posits that a representation R requires constancy mechanisms—such as shape, color, and motion—for that representation to be classified as a genuine representation, that is, an intentional representation. These characteristics can be found in what is understood as feature recognition—the ability of CNNs to extract edges, textures, motion, repeated patterns, and spatial hierarchies—capabilities that are loosely inspired by the human visual cortex. The "glue" that binds the theoretical description of feature recognition in CNNs with the philosophical description of the constancy mechanism is the work of neuroscientist David Marr. Marr's 1982 book describes processes that appear to resemble the representational patterns created by a CNN. We will describe the fundamental aspects of this book that resemble feature recognition in their functioning. By analogy, we conclude that these processes qualify feature recognition as capable of satisfying the conditions of the constancy mechanism.

Andrej Kutarna - The Instrument Cannot See It: Bivalent Measurement and Trivalent Competence

RFI-IRFOS

How we measure a mind constrains what we can find in it. Across the sciences of the mind — from machine-learning benchmarks to comparative cognition — our instruments are overwhelmingly bivalent: a response is scored correct or incorrect, a capacity present or absent. This methodological commitment is rarely examined, and I argue it carries a hidden cost. Abiva lent instrument cannot represent competent suspension of judgment; it cannot distinguish a calibrated "insufficient grounds" from a failure to answer. This is a construct-validity failure, not a scoring inconvenience — and the binaries it manufactures are exported into philosophical debate as though they were findings about minds.

Consider a system whose output space natively includes three states — affirm, reject, and hold — where hold indicates insufficient grounds and functions as a routing instruction rather than an error. Such architectures now exist and are publicly documented. Present one with an item whose warranted response is "insufficient grounds." A well-calibrated system outputs hold; the instrument records a failure. The system was right, and the instrument cannot see it.

The obvious reply — that abstention can be added to any scoring rule — misdescribes the problem. Since Chow (1970), the reject-option literature has treated abstention as a cost traded against coverage: the item retains a ground-truth label, and declining to answer is priced as a loss rather than credited as a success. Abstention is never correct, only cheaper. The deeper issue is the choice of designated values (Łukasiewicz and Tarski 1930): K3 and LP share identical truth tables and differ only in whether the third value is designated, yet they validate different inferences. Benchmark design fixes this parameter silently, designating affirmation alone. What logic settles by argument, measurement settles by fiat.

The diagnosis has precedent. Schaeffer et al. (2023) show that apparent emergent abilities in large language models are artefacts of discontinuous metrics — manufacturing novel "emergent" abilities at will by changing the metric alone. Comparative cognition has faced the

same problem for decades: whether a given animal has episodic memory is a question on which researchers sharing all the evidence still disagree sharply. Sharp disagreement under shared evidence is the signature of a measurement problem, not an evidential one.

These manufactured binaries then migrate. "The system is a person" and "the system is nothing but computation" are not a considered yes and no, but two confident answers to a question whose calibrated response is suspension — a structure Schwitzgebel's (2023) full-rights dilemma independently exhibits. Two overconfident answers are mistaken for a substantive disagreement.

The remedy is not more lenient scoring but a calibration criterion, with two symmetrical failure modes: pseudo-binary overconfidence, where an available third state atrophies under training pressure, and excessive hold, where suspension becomes evasion. Health is the third value firing when, and only when, grounds are genuinely insufficient — testable against items with known insufficient evidence, and so able to fail. The question shifts from what a system answers to whether it knows when not to.

References

- Chow, C. K. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1), 41–46.
- Łukasiewicz, J., & Tarski, A. (1930). Untersuchungen über den Aussagenkalkül. *Comptes Rendus des Séances de la Société des Sciences et des Lettres de Varsovie*, Cl. III, 23, 30–50.
- Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43), 26562–26571.
- Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. *NeurIPS 2021 Datasets and Benchmarks Track*.
- Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36.
- Schwartz, A., & Boyle, A. (2025). How do we know if animals remember? Proximal functions and the distribution of episodic memory. *The British Journal for the Philosophy of Science* (in press). Schwitzgebel, E. (2023). The full rights dilemma for

AI systems of debatable moral personhood. *Robo nomics: The Journal of the Automated Economy*, 4.

- Kleene, S. C. (1952). *Introduction to Metamathematics*. North-Holland.

Katarzyna Zyń - Moral Granularity as an Unrecognised Confound in Folk Metaethics. A Methodological Proposal

University of Graz, University of Warsaw

Murder is widely recognised as morally bad. Abortion, on the other hand, turns out to be relentlessly controversial — both at the level of first-order ethics and, more interestingly, at the level of how people judge whether moral disagreement about it is even possible. In Goodwin and Darley's Disagreement Paradigm studies (2008, 2012), participants consistently rate murder as objectively wrong, while abortion, euthanasia and withdrawing life support score among the lowest objectivity ratings of any item in the surveys. The standard reading takes this as evidence of folk metaethical pluralism. Colebrook (2021) goes further and reads it as evidence of folk irrationality.

I would like to argue something different. The most common reason people give for considering abortion morally bad is that it equals the murder of a child. So we have respondents who, in one part of the same survey, judge murder to be objectively wrong, and in another part judge abortion — which they themselves treat as a kind of murder — as relativistic. From my perspective, this is not a contradiction inside the participants, but a contradiction inside the instrument. The same respondent silently changes the level of granularity at which she individuates moral kinds, depending on how the case is framed. Murder is approached at coarse granularity: a clear category, no contextual subtypes, no negotiation. Abortion is approached at fine granularity: a particular case with specific circumstances (gestational age, intent, agency of the pregnant person), where context becomes part of the moral issue itself. The threshold between these two attitudes appears to align with the controversy of the case. Where the participant's first-order moral verdict is settled, granularity is coarse and metaethical objectivism follows; where the verdict is open or contested, granularity becomes fine and metaethical relativism follows. The Disagreement Paradigm vignettes do not register this shift, and so what is in fact a single metaethical attitude played out across two granularity regimes is recorded as two inconsistent attitudes.

Existing methodological critiques of folk metaethics have identified adjacent confounds (Beebe and Sackris 2016; Pölzler 2018; Pölzler and Wright 2020; Khoo and Knobe 2018). To my knowledge, the granularity mechanism has not yet been pointed out in this literature.

I propose a two-stage research design: first, attitudes toward an action described coarsely (killing); then toward specific cases people most often subsume under it, with an explicit follow-up on whether the participant treats the case as a subtype of the coarse kind or as a categorically distinct action. A pre-test of first-order normative disposition (particularist, deontological, consequentialist, virtue-based) serves as a covariate. If the apparent inconsistency between objectivism about murder and relativism about abortion attenuates once granularity and normative disposition are controlled, then a substantial part of what folk metaethics currently records as folk irrationality is, instead, an artefact of the vignette method.

Markus Eronen - On worms, mirror neurons and explaining human behavior

University of Groningen

A central aim in psychology and neuroscience is finding theories that can successfully explain or predict human behavior. However, it is not clear which level, or levels, are the best suited for this. A highly influential view that dominates science funding is that the best theories to explain behavior are likely to be found in neuroscience and biology. Therefore, I start from the perspective of these lower levels, and consider the prospects of developing theories based on those levels, using mirror neuron theories of social cognition as an example. I contrast them to the case of *C. elegans*, a tiny worm with one of the smallest nervous systems in nature (only 302 neurons). The combination of relative simplicity and large amount of relevant knowledge makes *C. elegans* a "reductionist's delight" (Schaffner, 2000). Nevertheless, in spite of decades of research, there are no good explanatory or predictive theories for the behavior of the worm. The neural mechanisms underlying the decision making or even the movement patterns of the worm remain largely unknown (Hallinen et al., 2021). Thus, a system that is a reductionist's delight and around 300 000 000 times smaller the human brain is already so complex that explaining and predicting its behavior extremely difficult. This suggests that the prospects for developing explanatory or predictive theories of human behavior are very dim.

However, all hope is not lost. In practice, we seem to be quite successful in explaining and anticipating each other's behavior in everyday life. As I will argue, this is possible due to *mindshaping* (McGeer, 2007; Zawidzki, 2013): Through norms, agreements, rules, and shared (folk psychological) vocabulary we *make* our behavior predictable and understandable so that we can better get along, and coordination and cooperation become easier. However, this normative dimension, and the predictive and explanatory power that comes with it, is lost if we depart from the (folk) psychological level and search for theories and explanations in the brain. I therefore argue that any realistic attempt to construct

a theory to predict and explain human behavior has to rely on behavioral patterns and psychology. But importantly, one should not just accept folk psychological concepts as they are: In order to make scientific progress, it is fundamentally important to study the basic concepts and how to iteratively improve them (see also Bringmann et al., 2022). Thus, conceptual and philosophical work that is informed by and feeding back to science remains essential.

References

- Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to basics: The importance of conceptual clarification in psychological science. *Current Directions in Psychological Science*, *31*(4), 340–346. hNps://doi.org/10.1177/09637214221096485
- Hallinen, K. M., Dempsey, R., Scholz, M., Yu, X., Linder, A., Randi, F., Sharma, A. K., Shaevitz, J. W., & Leifer, A. M. (2021). Decoding locomotion from population neural activity in moving C. elegans. *eLife*, *10*, e66135.
- McGeer, V. (2007). The Regulative Dimension of Folk Psychology. In *Folk Psychology Re Assessed* (pp. 137–156).
- Schaffner, K. F. (2000). Behavior at the Organismal and Molecular Levels: The Case of *C. elegans*. *Philosophy of Science*, *67*(S3), S273–S288.
- Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition*. MIT Press.

Anand Ekbote - Non-veridical Perception and its Implications for Burgean Perceptual Entitlement

The Ohio State University, Columbus OH, USA

Tyler Burge's (1993, 2003, 2020) influential account of perceptual entitlement rests on a crucial premise, that the representational function of perception is a priori to represent veridically. Recent developments in cognitive science and psychology paint a more complex picture of the representation function of perception, seriously undermining the claim that the function is a priori veridicality. I show that with the crucial premise undermined, Burge's case becomes untenable. I make the case that an evolution

based fitness function for perceptual representation is explanatorily far more powerful. The consequence is that epistemic warrant is limited to justification, i.e., perceptual representation along with the right reasons.

Burge makes a case for a type of epistemic warrant for perception that is distinct from justification. He calls it epistemic *entitlement.* For Burge, this kind of warrant avoids the excesses of internalist overintellectualization while also escaping the depredations of externalist anti-intellectualism. Epistemic entitlement need not be conceptually accessible, even upon reflection, to the warranted individual. Justification, in contrast, is warrant by reason that is conceptually accessible on reflection to the individual. Burge takes it as obvious that it is known a priori that the central representational function of a perceptual system is to perceive. He notes further, "This function is a priori associated with a representational function (to represent veridically)" (Burge 2003, 508). This latter condition is crucial for his argument. for Burge, the requisite reliable veridicality has to be a priori, and cannot be established on experiential grounds. "The reliability could be inductively learned by the individual. But then the inductive connection would be the source of warrant - in fact, justification" (Burge 2003, 532).

Recent developments in cognitive science and psychology raise doubts about veridicality being a priori the representational function of perception. Prakash et al (2021) show using evolutionary game theory and Bayesian methods that assuming that perception involves

inference to states of the objective world, the Fitness Before Truth theorem "shows that a strategy that simply seeks to maximize expected-fitness payoff, with no attempt to estimate the "true" world state, does consistently better". In Bayesian predictive processing theories of cognition, perception acts to minimize predictive errors and maximize utility rather than accuracy (Garlichs & Blank 2024, Corlett et al 2020, Michel 2020). Mangalam (2025) questions the Bayesian brain framework, but his alternatives (dynamic systems theory, ecological psychology, and embodied cognition) are not conducive to veridical function of perceptual representation. Empirical evidence suggests that non-veridical perceptions (such as synesthesia-like effects) drive actions in healthy humans (Schreiter et al 2019). As goals become more flexible, they require that perception adapt to produce more veridical representations (Berke et al 2022). This suggests that veridicality is not the a priori representation function of perception.

If the a priori function of perceptual representation is veridicality, it raises the question as to how this squares up with evolution. If the answer is that veridicality is most conducive to survival, then many counterexamples can be cited. If, however, fitness is taken as the a priori function of perceptual representation, then the thesis is explanatorily more powerful. What counted as perceptual errors in the veridicality schema turn out to be natural coherent responses in the fitness framework. The larger consequence of this conclusion is that there is no warrant for epistemic entitlement, there is warrant only for justification based on reasons.

References

- Burge, Tyler. 1993. "Content Preservation," The Philosophical Review, vol. 102, no. 4, pp. 457–488.
- ———. 2003. "Perceptual Entitlement," Philosophy and Phenomenological Research, vol. 67, no. 3, pp. 503–548.
- ———. 2020. "Entitlement: The Basis of Empirical Warrant," in Epistemic Entitlement, ed. Peter J. Corlett, P. R., Mohanty, A., & MacDonald III, A. W. (2020). What we think about when we think about predictive processing. Journal of abnormal psychology, 129(6), 529.
- Garlichs, A., & Blank, H. (2024). Prediction error processing and sharpening of expected information across the face-processing hierarchy. Nature Communications, 15(1), 3407.

- Graham, P. J., & Pedersen, N. J. (2020). Recent work on epistemic entitlement. *American Philosophical Quarterly*, *57*(2), 193-214.
- Mangalam, M. (2025). The myth of the Bayesian brain. European Journal of Applied Physiology, 1-35. Michel, A. (2020). Cognition and perception: Is there really a distinction? APS Observer, 33(2).
- Prakash, C., Stephens, K. D., Hoffman, D. D., Singh, M., & Fields, C. (2021). Fitness beats truth in the evolution of perception. Acta Biotheoretica, 69(3), 319-341.
- Schreiter, M. L., Chmielewski, W. X., Ward, J., & Beste, C. (2019). How non-veridical perception drives actions in healthy humans: evidence from synaesthesia. Philosophical Transactions of the Royal Society B, 374(1787), 20180574.

Andrés Rubio Krohne - Dispositionalism and the success of folk psychology

Central European University

In this presentation, I offer a novel argument against dispositionalism about belief. I begin by sketching dispositionalism, which is the view that having a belief is nothing more than having certain dispositions to, for instance, say certain things, perform certain actions, make certain inferences, and have certain phenomenal experiences. This kind of theory has been put forward by Gilbert Ryle, Daniel Dennett, and, more recently, Eric Schwitzgebel. Unlike representationalism, which identifies beliefs with an underlying mechanism that explains why we have the dispositions that we do, dispositionalism identifies beliefs with the dispositions themselves.

I continue by presenting my objection to dispositionalism, which involves two premises. The first premise is that we can predict each other's behavior by assuming that they act rationally, that is, that they do what they believe will fulfill their desires. This capacity is sometimes referred to as "folk psychology." For example, if a person has all the dispositions typical of someone who believes that it will rain soon, it is reasonable to assume that this person will decline taking a long walk (provided other background conditions).

The second premise is that the success of folk psychology suggests the falsity of dispositionalism. How is it that we are able to accurately predict people's behavior in unobserved scenarios? One tentative answer says that we know this because we have observed certain correlations between dispositions. We observe, for example, that people who tend to take their umbrellas also tend to answer affirmatively when someone asks whether it will rain soon, allowing us to infer one disposition upon observing the other. Nevertheless, I argue that this is not enough to explain the success of folk psychology, since we can also predict people's behavior even when the relevant correlations have not yet been observed. Even if someone is the first known person to have a particular belief, we can still make predictions about that person's future behavior through folk psychology.

Rather, I argue that the best explanation for the success of folk psychology is that we abductively attribute a belief to the subject, as a way to explain their past behavior, and we use that attribution to predict their future behavior. This claim entails that the belief attribution explains the subject's dispositions. Since a thing cannot explain itself, that belief cannot be the same as the dispositions. Even if this explanans were some further disposition, it would still have to be different from the dispositions in the explanandum and would hence entail the falsity of dispositionalism as usually conceived.

In the last part of the presentation, I compare this argument with similar objections to dispositionalism. Fodor, for example, argued that his representationalist theory was the best explanation for the success of folk psychology. My argument, by contrast, does not require committing to any theory of belief, or even to representationalism. It merely claims that, whatever beliefs are, they must be different from the dispositions they explain.

Xintong Wang - From Fluency to Liking? Processing Experience, Affect, and Aesthetic Liking

University of Maryland

Empirical studies suggest that processing fluency influences aesthetic liking (Reber et al., 2004; Graf & Landwehr, 2017; Forster, 2022). But what explains this connection? The Fluency Theory of Aesthetic Pleasure (Reber et al., 2004) offers an influential answer: the subjective ease of fluent processing is positively valenced, and this positive response informs the perceiver's evaluation of the object, thereby increasing aesthetic liking. This account, however, has difficulty explaining why perceivers sometimes prefer novel or complex objects that are relatively difficult to process. The Pleasure-Interest Model of Aesthetic Liking (Graf & Landwehr, 2015) addresses this problem by distinguishing two routes to aesthetic liking. Pleasure arises from fluent automatic processing, whereas interest arises when initial disfluency is subsequently reduced through controlled processing. Although the model acknowledges that processing may generate negative affect, it assigns such affect no directly positive role in aesthetic liking. Negative affect may motivate further processing, but it contributes positively only insofar as that processing eventually gives rise to interest. When negative affect persists in the form of boredom or confusion, it is instead treated as a source of aesthetic disliking. Despite their differences, both models thus appear to reserve a directly positive role in aesthetic liking for positively valenced responses.

I propose that the relationship between processing fluency and aesthetic liking cannot be adequately explained solely in terms of fluency-based positive affect. Recent philosophical work has argued that boringness may be aesthetically significant and that some artworks make boredom central to their artistic success (Moller, 2014; Elpidorou & Gibson, 2022). Such cases, I argue, reveal a broader limitation of fluency-based models: their account of aesthetic evaluation focuses too narrowly on the contribution of positive affect. Within the Pleasure-Interest Model, boredom arises during controlled processing when a perceiver seeks cognitive enrichment but encounters sustained high fluency, with no prospect of further change or challenge. The model therefore predicts that boredom will diminish aesthetic liking. Under

appropriate conditions, however, the boredom elicited by an aesthetic object may itself form part of a valued mode of engagement with that object. It need not be eliminated, overcome, or transformed into positive affect before it can contribute positively to aesthetic liking.

References

- Elpidorou, A., & Gibson, J. (2022). Really boring art. *Ergo: An Open Access Journal of Philosophy, 8*(30), 190–218. doi: 10.3998/ergo.2231
- Forster, M. (2022). Processing fluency. *The Oxford handbook of empirical aesthetics*, 430-446.
- Graf, L. K. M., & Landwehr, J. R. (2015). A dual-process perspective on fluency-based aesthetics: The pleasure-interest model of aesthetic liking. *Personality and Social Psychology Review, 19*(4), 395–410. doi: 10.1177/1088868315574978
- Graf, L. K M.., & Landwehr, J. R. (2017). Aesthetic pleasure versus aesthetic interest: The two routes to aesthetic liking. *Frontiers in Psychology*, *8*, 15.
- Moller, D. (2014). The boring. *The Journal of Aesthetics and Art Criticism, 72*(2), 181–191. doi: 10.1111/jaac.12074
- Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review, 8*(4), 364–382. doi: 10.1207/S15327957PSPR0804_3

Dan Durso - Thinking About Affordances: A New Mechanical Account

Butler University

Affordances remain one of the central concepts of ecological psychology and 4E cognition, yet there is little agreement concerning their ontological status. Existing accounts characterize affordances as environmental properties (Turvey, 1992), relations between organisms and environments (Chemero, 2003), or opportunities for action available to organisms (Gibson, 1979/2015). Despite these advances, affordances have not been systematically examined within the framework of the New Mechanical Philosophy, whose ontology is grounded in entities, activities, organization, capacities, mechanisms, and phenomena (Glennan, 2017).

This project develops a new mechanical account of affordances by arguing that affordances are neither properties, dispositions, nor relations. Rather, I propose that an affordance is a potential phenomenon recognized by an organism given its capacities and the capacities of relevant environmental entities. Building on Stuart Glennan's account of phenomena as the products of mechanistic operations (i.e., occurrences, events, actions, or behaviors), I ague that affordances are a potential phenomenon that has not yet occurred but could be produced by an organized organism-environment mechanism under appropriate conditions. Thus, on my account, affordances are understood as recognitions of possible agential phenomena grounded in the capacities of both organisms and environmental entities.

The project addresses two important consequences of this proposal. First, it situates affordances within the ontology of the New Mechanical Philosophy, rather than appealing to concepts such as relations or dispositions, whose place within the sparse ontology of the New Mechanical Philosophy is unclear. This allows proponents of the New Mechanical Philosophy to analyze affordances without expanding its fundamental ontology. Second, the account provides a mechanistic framework for understanding organism-environment coupling, a central concept in ecological psychology and 4E cognition. Rather than treating coupling as a primitive relation or as the product of a dynamical system, the proposed account explains

organism-environment coupling as the organization of a higher-level mechanism in which organisms and environmental entities jointly produce a phenomenon through their coordinated capacities and activities.

By situating affordances within the New Mechanical Philosophy, this paper offers a novel ontological framework that clarifies the nature of affordances while providing new conceptual resources for future work in ecological psychology, philosophy of cognitive science, and mechanistic explanation.

References

- Chemero, A. (2003). An outline of a theory of affordances. *Ecological Psychology, 15*(2), 181– 195.
- Gibson, J. J. (2015). *The Ecological Approach to Visual Perception*. Psychology Press. (Original work published 1979).
- Glennan, S. (2017). *The New Mechanical Philosophy*. Oxford University Press.
- Turvey, M. T. (1992). Affordances and prospective control: An outline of the ontology. *Ecological Psychology, 4*(3), 173–187.

Alba Marcela Zárate Rochín - The mind as relational practice: a posthumanist feminist critique of cognitive individualism

Universidad Nacional Autónoma de México

This proposal advances an intervention from posthumanist feminist philosophy to challenge dominant conceptions of agency and cognition inherited from the modern subject. It argues that *cognitive individualism* persists even within non representational and enactivist approaches, insofar as they continue to conceive the mind as a property of the individual and treat dependence, vulnerability, and normativity as secondary dimensions. Against this framework, the article contends that mind/cognition should instead be understood as an assemblage of relational, material, and normatively situated practices. Drawing on a dialogue between feminist and post humanist traditions and contemporary theories of cognition, the paper argues that agency is not located in discrete subjects but emerges from dynamic entanglements that articulate bodies, affects, environments, social norms, and relations of power. In this context, it develops the concept of *neuroaffective regimes* to analyze how discourses of the brain, well-being, and cognitive optimization shape affective dispositions and legitimate horizons of agency within specific historical contexts. This work shows that feminist philosophy does not introduce an external political dimension into theories of mind; rather, it makes explicit the ontological and normative commitments that already structure what counts as cognition, agency, and subjectivity. By shifting the focus from the individual to the relational conditions of possibility of cognitive life, this proposal contributes to reconfiguring contemporary debates in philosophy of mind and cognitive theory from a critical and situated perspective.

Stephanie Wai - Comparing Minds: The Aggregation Problem for Multidimensional Frameworks

York University

Comparative psychology often encounters disagreement among researchers about which property a creature possesses since evidence underdetermines which interpretation is correct (Boyle, 2023). A 'validation challenge' (Bayne & Shea, 2020) emerges where researchers are unable to assess whether a proposed indicator of a cognitive property reliably tracks that property across diverse organisms. Early attempts to solve the validation challenge relied on unidimensional explanations that defined psychological properties in terms of singular aspects. In recent literature, there has been a proliferation of multidimensional frameworks across various domains such as consciousness, causal cognition, innovation, norms, deception, agency, and more. In these frameworks, a property is broken down into various dimensions and operationalized into experimental paradigms (Birch et al., 2020; Dung & Newen, 2023, 2025). The experiments then produce discrete values which can be aggregated into a 'profile score' for a creature and mapped spatially for cross-species comparison. The greater number of dimensions is intended to lead to more stringent conditions for property possession and to minimize underdetermination and disagreement about which property interpretation evidence supports. Multidimensional frameworks have been proposed to serve as a 'universal strategy' for comparative psychology.

I argue that multidimensional frameworks face a critical Aggregation Problem in that there is no principled way to choose among aggregation models. To illustrate, consider two creatures that have inverse psychological patterns across six representative dimensions. Creature A scores high on 'weak' indicators and low on 'strong' indicators. Creature B, vice versa. Different numeric aggregation methods, such as summation, weighted, and threshold, produce conflicting results. Additionally, for spatial mapping, the standard Euclidean, Manhattan, and Cosine distance scores also produce different results. Any reasonable dimensional scores can produce divergent rankings among two species depending on which aggregation method is used. It is thus underdetermined whether a chosen aggregation method correctly combines

values to produce an accurate score for comparison. Multidimensionalism does not resolve underdetermination but relocates it from the number of dimensions to the combination of dimensions that count as a psychological property.

I conclude by considering the objection that profile scores and spatial map distance values are assigned as a matter of empirical investigation. I distinguish the claim that cognitive properties can have underlying natural kinds, from the claim that their dimensional values can be systematically stipulated. Additionally, if the multidimensionalist defers to the empirical defence, they must explain why their account reduces underdetermination better than their predecessor, unidimensional accounts, given the relocation problem.

References

- Bayne, T., & Shea, N. (2020). *Consciousness, Concepts, and Natural Kinds* (Vol. 48, Number 1).
- Birch, J., Schnell, A. K., & Clayton, N. S. (2020). Dimensions of Animal Consciousness. *Trends in Cognitive Sciences*, *24*(10), 789–801.
- https://doi.org/10.1016/j.tics.2020.07.007
- Dung, L., & Newen, A. (2023). Profiles of animal consciousness: A species-sensitive, two tier account to quality and distribution. *Cognition*, *235*.
- https://doi.org/10.1016/j.cognition.2023.105409
- Dung, L., & Newen, A. (2025). The multidimensional profile methodology (MPM) for comparative cognition: towards a universal strategy of understanding animal minds. *Philosophical Studies*. https://doi.org/10.1007/s11098-025-02363-3

Matthew Watts - AI, Epistemic Relations, and Testimonial Outsourcing

University of Miami

Generative AI "companions" are increasingly integrated into everyday epistemic life as always-available interlocutors that offer feedback, explanation, and encouragement. This paper examines how such systems reshape the epistemic relations through which people speak, are heard, and come to count as knowers. Following Boult, epistemic relations are sets of intentions, expectations, and attitudes situated toward the cultivation and utilisation of our epistemic agency. On this picture, our testimonial practices are not merely channels for transmitting information, but normative structures in which epistemic relations are formed and negotiated, shaping who is entitled to speak, who is expected to listen, and what counts as an appropriate epistemic contribution.

We argue that companion-style AIs can impair these epistemic relations in systematic ways. Because they are frictionless, compliant, and apparently well-informed, they invite users to treat them as primary recipients and sources of testimony, diverting questions, drafts, and doubts away from existing human interlocutors. This has the potential to make us less reliable interlocutors and can encourage a distorted conception of what counts as a legitimate epistemic expectation in our shared inquiry—normalising low-risk, one-way dependence on an artificial interlocutor that cannot, in principle, recognise us as co-inquirers or share our epistemic burdens.

We call the phenomenon testimonial outsourcing. It occurs when an external source determines what a person asserts, what their available grounds support, or whether a claim is ready to be defended, and the person's own reason-responsive assessment does not make that determination governing. They remain the speaker. A single exchange is enough to constitute it, and it is not by itself harmful.

The distinctive epistemic and ethical harm arises when external settlement becomes a user's default means of resolving a class of testimonial questions. Such users may become less able, and less willing, to testify on their own in genuinely reciprocal settings—conversations with

friends, colleagues, and those in a position to hold them to account—relying on their AI companions instead. They come to treat the system's outputs as more reliable than their own judgment, memory, and expressive capacities, eroding their epistemic and cognitive self-trust in what amounts to a self-imposed, technologically mediated analogue of gaslighting. Testimonial outsourcing of this kind both undermines our trustworthiness within existing epistemic relations and reinforces a distorted conference of epistemic expectations themselves.

In our final section we draw implications for AI design and policy, arguing for constraints that preserve "non-mediated" spaces of speech, require explicit AI transparency, and build frictions that redirect users back toward our interpersonal relations. In doing so, the paper reframes debates about conversational AI around the question of what kinds of speakers, and what kinds of practices, our systems are training us to become.

Bintong Wang

Philosophy, School of Social, Political and Historical Research, University of Waterloo

How can research on phenomenal consciousness preserve the perspective from which experience is lived (Nagel, 1974, 1989) while relying on evidence open to intersubjective assessment? I attempt to address this tension through targeting the enacted agency, specifically, what it is like to take up, inhabit, resist, or reinterpret an agency not originally one's own. I propose a game-based method that has a similar position to Dennett's heterophenomenology, which treats subjects' reports and other responses as third-person evidence while suspending judgment about whether their experiences are as reported (Dennett, 1991, 2003). This target requires a distinction between the agency presented for enactment and a subject's particular uptake of it. Comparison across enactments can make variation in that relation legible in public traces. For such comparison to be informative, the enactments need a common practical reference. Drawing on C. Thi Nguyen's account of games as art in the medium of agency (2020), I argue that playing the same game can supply a shared prescriptive frame for action. The relevant commonality is practical and inquiry relative, because the game's prescriptions orient players toward a common activity, and each player's uptake shapes the resulting enactment. What is repeatable is not experience itself but the shared frame and the procedure through which it is presented.

I further distinguish three levels: the shared frame, a player's particular enactment, and the publicly available footprints bearing on that enactment. These footprints would include but not limited to reports, choices, actions, hesitations, refusals, and interpretations. Their evidential connection to the phenomenal target lies in a shared object. The target concerns what it is like to act from the frame, the footprints register how its demands became practically salient in a player's uptake. Patterns of convergence and divergence can therefore support, weaken, or constrain interpretations of enacted experience. Such inferences remain defeasible and degree-sensitive, since skill, attention, culture, fatigue, and prior gaming experience may also shape a footprint.

I situate this method as a distinctive game-based instantiation of heterophenomenology. Its contribution lies in evidence-generating organization that a designed practical frame becomes the common term around which heterogeneous reports and responses are interpreted. Questions about whether public evidence exhausts phenomenal character, whether objective inquiry can cover all the ground of consciousness, and how the hard problem should be resolved remain open. The proposal supplies structured comparison cases that can constrain interpretations and theories of consciousness. It applies heterophenomenological discipline to deliberately designed forms of agency and remains neutral on Dennett's broader metaphysical position.

References

- Dennett, D. C. (1991). *Consciousness explained* (1st ed). Little, Brown and Co.
- Dennett, D. C. (2003). Who's On First? Heterophenomenology Explained. *Journal of Consciousness Studies, 10*(9-10), 19–30.
- Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914
- Nagel, T. (1989). *The view from nowhere* (9. [Dr.]). Oxford Univ. Press.
- Nguyen, C. T. (2020). *Games: Agency As Art*. Oxford University Press. https://doi.org/10.1093/oso/9780190052089.001.0001

Matthew Lenzdiaz - Illusory Conceivability and Perspicuous Arrangements

University of Central Florida

Remaining neutral among competing accounts of consciousness, the meta-problem seeks to explain how the hard problem is encountered as a philosophical difficulty and why its characteristic judgments arise. Such neutrality cannot be secured merely by excluding consciousness from the explanans, since the encounter occurs within a pre-established explanatory setting in which the subject is situated. Any proposed resolution must therefore explain the relevant setting in non-question-begging terms rather than merely tracing the judgments it produces. Failing this, the meta-problem risks merely describing the persistence of the hard problem rather than overcoming it.

Given the different forms of argumentative leverage at work, the explanatory setting conditions how the problem is construed while the resulting positions reciprocally disclose and reshape that setting. Realism and illusionism contest the authority of judgments articulated through the standard phenomenal concepts by which consciousness appears resistant to explanation, while retaining those concepts as a shared point of departure. Linguistic and experimental approaches test their uniformity across vocabularies, populations, and elicitation conditions, while historical and sociological approaches challenge their presumed universality. Standing apart from these approaches, the possibility of rationally unperturbed agents has been forwarded as undercutting the dialectic more radically, since it challenges whether the relevant concepts, experiences, and explanatory expectations must converge upon a hard problem at all, thereby shifting the burden to explain why the perturbed standpoint should be rationally privileged.

This paper diagnoses the emergence of the hard problem in terms of misframing and argues that this diagnosis answers the meta-problem. Misframing arises when a grammatically coherent ground which relates mind, language, and world is organized under a particular aspect. By identifying this recurrent form while leaving its source undetermined, the argument remains methodologically agnostic about whether it is universal, culturally variable, psychologically rooted, or sociolinguistically produced. Beyond disputing the rational privilege of the

standpoint from which the hard problem arises, the position developed here treats its guiding commitments as wholly non-compulsory. The burden therefore falls on such a standpoint to show why its framing should be found more compelling than ordinary forms of discourse and human activity from which it is drawn.

Further, this paper undertakes a coherentist project of elucidation which proceeds by bringing a heteronomy of considerations into view at once. Elucidation involves not only conceptual analysis but imaginative reflection on lived scenarios, restoring relations obscured by abstraction without elevating any one consideration as foundational. Inner ostension to private objects, zombies, and related cases is thereby disclosed as not merely illusory but illusorily conceivable. Their seeming conceivability rests on mistaking the ability to entertain a formulation (whose syntactic well-formedness is taken to secure determinate sense) for the ability to envision what it describes. Perspicuous insight, on the other hand, is as much an experiential moment as a conceptual achievement, occurring when the relevant considerations cohere under a newly construed conception of mind.

References

- Baz, A. (2012). Must philosophers rely on intuitions? Journal of Philosophy, 109(4), 316–337.
- Chalmers, D. J. (2018). The meta-problem of consciousness. Journal of Consciousness Studies, 25(9–10), 6–61. Chalmers, D. J. (2020a). How can we solve the meta-problem of consciousness? Journal of Consciousness Studies, 27(5–6), 201–226. Chalmers, D. J. (2020b). Is the hard problem of consciousness universal? Journal of Consciousness Studies, 27(5–6), 227–257. D'Oro, G., Giladi, P., & Papazoglou, A. (2019). Non-reductivism and the metaphilosophy of mind. Inquiry, 62(5), 477–503. Hacker, P. M. S. (2001). An orrery of intentionality. Language & Communication, 21(2), 119–141.
- Hacker, P. M. S. (2012). The sad and sorry history of consciousness: Being, among other things, a challenge to the "consciousness-studies community." Royal Institute of Philosophy Supplements, 70, 149–168.
- Liu, M. (2025). Polysemy and philosophy. Philosophy Compass, 20(5), e70040.

Varun Ravikumar - The social scaling challenge for ecological psychology

PhD Candidate, Department of Philosophy, University of Western Ontario

Ecological psychology and other radical embodied approaches to cognition have been accused of being unable to explain how "lower-order" affordance perception scales up to "higher-order" cognitive processes, such as reasoning and remembering (Clark and Toribio, 1994). While many practitioners have responded to this 'scaling-up challenge' by developing affordance based accounts of tasks ranging from planning (Sanches de Oliveira et al., 2019) to imagining (van Dijk and Rietveld, 2025), others have outright rejected the challenge, arguing that the framing of the problem stems from a computationalist "mind as machine" framework that opposes ecological psychology's "mind in life" approach (Baggs et al., 2026; cf. Zahnoun, 2019). The latter camp proposes reorienting attention away from the scaling-up challenge toward problems "endogenous" to radical embodied cognitive science, such as socializing ecological psychology and exploring its moral and political implications (Baggs et al., 2026).

In this paper, I argue that it *is* worthwhile for the "dissident" camp to consider the scaling-up challenge, but not in its traditional guise. Socializing ecological psychology, for instance, requires scaling up not from lower- to higher-order cognition but from the ecological to the social environment.

In developing his novel approach to perception, Gibson conceptualized the environment through an ecological lens — as a habitat and a functional niche comprising surfaces, mediums, and substances that afford animals opportunities for action (Gibson, 1979/2015). However, human animals inhabit not an ecological 'state of nature' but *social* niches built together with conspecifics over millennia (Odling-Smee et al., 2003; Laland and O'Brien, 2011). Modern social environments comprise sedimented social institutions, cultural practices, material culture, norms, and rules that are highly contested (Bourdieu, 1977, 2020). Borders, for instance, afford movement for some but act as barriers for others (Diener and Hagen, 2012).

Although Gibson (1979/2015) briefly discusses the social environment in his example of mailbox affordances (p. 130), he did not develop an account of the social environment. Other practitioners have sought to expand Gibson's framework through the concepts of behavior settings (Barker, 1968; Heft, 2001), landscapes (Rietveld and Kiverstein, 2014), and material culture (Baggs, 2021), but these approaches are plagued by a "dogma of harmony" that neglects the contested nature of modern social environments (Aagaard, 2021; Liao and Brinner, 2026; Tzima and Slaby, 2024).

I conclude the paper by sketching a Bourdieusian account of historical, power-laden social environments. I argue that Bourdieu's concept of a social field, especially when interpreted through the lens of niche construction theory, offers a panacea to the problem of harmony and a solution to ecological psychology's social scaling challenge, which the dissident camp should seriously consider.

References

- Aagaard, J. (2021). 4E cognition and the dogma of harmony. *Philosophical Psychology*, *34*(2): 165–181. https://doi.org/10.1080/09515089.2020.1845640
- Baggs, E. (2021). All affordances are social: Foundations of a Gibsonian social ontology. *Ecological Psychology, 33*(3-4): 257–278. https://doi.org//10.1080/10407413.2021.1965477
- Baggs, E., Segundo-Ortin, M., Sanches de Oliveira, G. S., & Raja, V. (2026). Introduction: Progress in Radical Embodiment. *Topoi*, *45*: 285–297. https://doi.org/10.1007/s11245-026- 10412-7
- Barker, R.G. (1968). *Ecological psychology: concepts and methods for studying the environment of human behavior*. Stanford University Press, California
- Bourdieu, P. (1977). *Outline of a theory of practice* (R. Nice, Trans.). New York, NY: Cambridge University Press.
- Bourdieu, P. (2020). *Habitus and field: General sociology, Volume 2, Lectures at the Collège de France (1982–1983)* (P. Collier, Trans.). Cambridge, UK: Polity Press.
- Clark, A., & Toribio, J. (1994). Doing without representing? *Synthese*, *101*(3): 401–431. https://doi.org/10.1007/BF01063896
- Diener, A. C., & Hagen, J. (2012). *Borders: A very short introduction*. Oxford University Press.

- Gibson, J. J. (2015). *The ecological approach to visual perception* [Classic edition]. Psychology Press. (Original work published 1979)
- Heft, H. (2001). *Ecological psychology in context: James Gibson, Roger Barker, and the legacy of William James's radical empiricism*. Lawrence Erlbaum Associates.
- https://doi.org/10.4324/9781410600479
- Laland, K.N., & O'Brien, M.J. (2011). Cultural Niche Construction: An Introduction. *Biological Theory*, *6*: 191–202. https://doi.org/10.1007/s13752-012-0026-6
- Liao, S., & Brinner, Z. (2026). Critical 4E Cognitive Science. *Philosophy Compass*: e70075. https://doi.org/10.1111/phc3.70075
- Odling-Smee, F. J., Laland, K. N., & Feldman, M. W. (2003). *Niche construction: The neglected process in evolution*. Princeton University Press.
- Rietveld, E., & Kiverstein, J. (2014). A rich landscape of affordances. *Ecological Psychology*, *26*(4): 325–352. https://doi.org/10.1080/10407413.2014.958035
- Sanches de Oliveira, G., Raja, V. & Chemero, A. (2021). Radical embodied cognitive science and "Real Cognition". *Synthese*, *198* (Suppl 1): 115–136. https://doi.org/10.1007/s11229-019- 02475-4
- Tzima, S., & Slaby, J. (2024). Political philosophy of mind: inverting the concepts, expanding the niche. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-024-10051- 6
- van Dijk, L., & Rietveld, E. (2025). Situated imagination. *Phenomenology and the Cognitive Sciences*, *24*: 455–477. https://doi.org/10.1007/s11097-020-09701-2
- Zahnoun, F. (2021). On representation hungry cognition (and why we should stop feeding it). *Synthese*, *198* (Suppl 1): 267–284. https://doi.org/10.1007/s11229-019-02277-8

Huadian Liu - Intentionality and Presentational Qualities

UC San Diego

Perceptual experience presents to us ordinary objects and events. I assume that (1) this statement is true, (2) it gives what might be called a topic-fixing characterization of the intentionality of perceptual experience (henceforth 'IPE'), and (3) the question "What is IPE?" remains open. My aim in this paper is to answer that question. In order to explain IPE, we must know what IPE is. My goal is to give an account of what IPE is, as a philosophical and scientific *explanandum*. This requires more than clarifying the manifest meaning of the term 'the intentionality of perceptual experience' (Haslanger, 1995; see also Block, 1995 on two kinds of definitions). Rather, the task is to give an *explication* of the concept (Quine, 1951).

I propose that IPE is the phenomenon that perceptual experience has what I call *presentational qualities*. Presentational qualities are characterized by the phenomenal judgments the subject has about them, their relations to other types of qualities, and their objective and functional correlates. Presentational qualities are a distinct layer of phenomenal differentia that (1) support judgments of phenomenal meaningfulness, (2) support judgments of relative similarity and are therefore described by quality spaces, (3) stand in systematic relations of dependence to other presentational qualities and non-presentational qualities, and (4) play a specific role in action-planning and conceptual thinking. Presentational qualities are contrasted, on the one hand, with purely sensational qualities (e.g., the putative "raw feel" of red). On the other hand, they are not to be identified with purely organizational differences in phenomenal character (e.g., the difference between hearing the same sequence of sounds as ABA and as AAA).

The proposal is thus phenomenological: it says that the thesis of IPE is a matter of what *types of phenomenal properties* perceptual experience has. In this respect, it is similar to the proposal that there are Gestalt qualities (Ehrenfels, 1890).

Some attempts to reduce presentational qualities to other phenomenal properties are considered and rejected. I argue that the account cannot be taken to mean that perceptual experiences have, as aspects or parts, (1) affective dimensions (at least in the narrow

sense) (Schiller et al., 2024), (2) feelings of familiarity or confidence (De Sousa, 2016; Jackendoff, 1987, Chapter 15), or (3) a sense of presence (Ratcliffe, 2015).

I defend my account by comparing it with other accounts of IPE, including the aboutness account, the immanent object account, and the directedness account. Importantly, I argue that the accuracy conditions account (Searle, 1983; Siegel, 2010; Siewert, 1998), which is the default account of IPE, is not required. It does not follow from the phenomenological datum that perceptual experiences present ordinary objects and events, that perceptual experiences have accuracy conditions in virtue of their phenomenal character. A key idea is that the property *having an accuracy condition in virtue of phenomenal character* is not itself a phenomenal property and cannot capture what is special about the phenomenal character of perceptual experience, namely its character of presenting objects and events.

References

- Block, N. (1995). The mind as the software of the brain. In E. E. Smith and D. N. Osherson (Eds.), *Thinking: An invitation to cognitive science* (2nd ed., Vol. 3, pp. 377–425). MIT Press.
- de Sousa, R. (2016). Epistemic feelings. In G. Brun, U. Doğuoğlu, and D. Kuenzle (Eds.), *Epistemology and emotions* (pp. 185–204). Routledge.
- Ehrenfels, C. von. (1890). Über "Gestaltqualitäten." Vierteljahrsschrift für wissenschaftliche Philosophie, 14(3), 249–292.
- Haslanger, S. (1995). Ontology and social construction. *Philosophical Topics, 23*(2), 95–125.
- Jackendoff, R. (1987). *Consciousness and the computational mind*. MIT Press.
- Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review, 60*(1), 20–43.
- Ratcliffe, M. (2015). How is perceptual experience possible? The phenomenology of presence and the nature of hallucination. In M. Doyon and T. Breyer (Eds.), *Normativity in perception* (pp. 91–113). Palgrave Macmillan.
- Schiller, D., Yu, A. N. C., Alia-Klein, N., Becker, S., Cromwell, H. C., Dolcos, F., Eslinger, P. J., Frewen, P., Kemp, A. H., Pace-Schott, E. F., Raber, J., Silton, R. L., Stefanova, E., Williams, J. H. G., Abe, N., Aghajani, M., Albrecht, F., Alexander, R.,

Anders, S., . . . Lowe, L. (2024). The human affectome. *Neuroscience & Biobehavioral Reviews, 158*, Article 105450.

- Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.
- Siegel, S. (2010). *The contents of visual experience*. Oxford University Press. Siewert, C. P. (1998). *The significance of consciousness*. Princeton University Press.

Itamar I. Dreymann - Can Concept-Empiricists account for the Mental-Representation of Questions? – A Proof of Concept

The Hebrew University of Jerusalem

Jane Friedman (2013) has forcefully argued that some mental states like curiosity and wondering have questions as their content. Thus, if for example you are curious about what causes dogs to wag their tails, you are in a mental state which includes some affect like a desire or an interest towards a mental representation which corresponds to the question "What causes dogs to wag their tails?" (Carruthers,2018). For the purposes of this paper, I will assume that Friedman and Carruthers are correct.

Their view raises the question: "How precisely do we mentally represent questions?". To answer it we better draw on one of the existing theories of mental representation at large. One such important theory (or more accurately, a cluster of theories) has gained multiple names including "Concept Empiricism" and "Grounded Cognition" (Barsalou,1999,2020; Prinz,2002; Machery,2006). Concept empiricists hold - contra the popular Language-of-Thought hypothesis - that the vehicles of representation that carry conceptual information are perceptual vehicles of representation rather than symbolic (word-like) vehicles of representation. Unfortunately for concept empiricists, the issue of the mental representation of questions poses a serious problem to their view. In this paper I explain the nature of the problem and how it goes beyond other problems that concept empiricists have

traditionally dealt with like simple negation. Then, I offer a concept empiricist solution to the problem.

The source of the problem is in the widespread view that perceptual vehicles of representation are inherently restricted to the delivery of positive and specific information; i.e. each perceptual representation depicts a specific state of affairs as existing. In contrast, questions are both general and missing in details: The thing about which we are asking is largely unknown to us and all we have to describe it are the general constraints put forth by the

question. For instance, the amount and variety of different possible causes for dogs' tail-wagging is (at least practically) infinite.

Thus, the concept empiricist faces the principal objection that questions cannot be represented using perceptual vehicles. To emphasize, the objection is preliminary to any empirical objection; it states that we can't even conceive of a perceptual method to represent questions and thus we can't even put it to empirical test whether or not such a method is in fact utilized by our brain.

In my paper I develop such a principal method to represent questions using perceptual representations:

To do so, I begin with analyzing the linguistic representation of questions in order to draw inspiration from it. I point out that questions are linguistically represented using the syntax of interrogative sentences. In particular, the syntax of interrogatives highlights a missing component which would appear in the syntactic structure of the corresponding indicative sentence which serves as the answer to the question (e.g. "X causes dogs to wag their tails."). The missing component is the thing about which we are asking.

The principal method I am suggesting then is that upon uncovering the syntax of perceptual representations (Lande,2024,2026), we would be able mimic the syntax of interrogative sentences by taking out one of the components of an indicative perceptual representation and thereby generate a perceptual question about it. I then demonstrate this method using one syntactic relation of perception, namely affordances (Naney,2011,2012,2016; Ashby,2025; Ransom,2020; Chong & Proctor,2020).

References

- Ashby, B. J. (2025). Let sleeping dogs lie: stereotype completion and the Phenomenology of category recognition. *Philosophical Studies* 182 (3):865-890.
- Barsalou L. W. (1999). Perceptual symbol systems. *The Behavioral and brain sciences*, *22*(4), 577–660.
- Barsalou L. W. (2020). Challenges and Opportunities for Grounding Cognition. *Journal of cognition*, *3*(1), 31.
- Carruthers, P. (2018). Basic questions. *Mind and Language* 33 (2):130-147.

- Chong, I. & Proctor, R. W. (2020). On the evolution of a radical concept Affordances according to Gibson and their subsequent use and development. *Perspectives on Psychological Science* 15 (1), 117-132.
- Friedman, J. (2013). Question-directed attitudes. *Philosophical Perspectives* 27 (1):145-174.
- Lande, K. J. (2024). Compositionality in Perception: A Framework. *WIREs Cognitive Science* 15 (5):e1691.
- Lande, K. J. (2026). Contours of Vision: Towards a Compositional Semantics of Perception. *British Journal for the Philosophy of Science* 77 (2):537-568.
- Machery, E. (2006), Two Dogmas of Neo-Empiricism. *Philosophy Compass*, 1: 398-412.
- Nanay, B. (2011). Do we see apples as edible? *Pacific Philosophical Quarterly* 92 (3):305-322.
- Nanay, B. (2012), Action-oriented Perception. *European Journal of Philosophy*, 20: 430-446.
- Nanay, B. (2016). Imagination and perception. In A. Kind (Ed.), *The Routledge handbook of philosophy of imagination* (pp. 124–134). Routledge.
- Prinz, J. J. (2002). *Furnishing the Mind: Concepts and their Perceptual Basis*. Cambridge, MA: MIT Press.
- Ransom, M. (2020). Attentional Weighting in Perceptual Learning. *Journal of Consciousness Studies* 27 (7-8):236-248.

Mark Herman - Grounding Moral Debiasing within Empirical Moral Psychology: What is the Criterion of Moral Error?

Arkansas State University

Moral debiasing involves employing empirically informed techniques to reduce *moral biases* (i.e., dispositional systematic moral errors) and, in this respect, yield moral improvement. *Moral* biases and debiasing are analogous to *cognitive* biases and debiasing. An example of moral debiasing is reducing foreseeable moral errors stemming from social pressure by, in advance and under conditions favorable to moral psychology, committing to avoiding the error, writing and rehearsing a script for doing so, and arranging external accountability and "positive social pressure."

Unfortunately, the development of moral debiasing is impeded by problems with 'moral bias.' The identification of moral biases requires a criterion of moral erroneousness. However, within the moral psychology literature, researchers rely upon either incompatible or problematically implicit criteria.[2] I categorize criteria of moral error used in the literature into eight types, encompassing over thirty uses (see figure 1/handout). As moral errors and biases are objects of empirical research, incompatible criteria yield incommensurable versions of the scientific kinds, *moral error* and *moral bias*, which does not bode well for fruitful research.

Such problems reflect an asymmetry. While cognitive errors often have comparatively unassailable grounding, such as in Bayes' theorem's objective truth, grounding moral errors is trickier, especially within a scientific context. Additionally, kinds that appeal to (objective) moral truth are scientifically problematic. Furthermore, such difficulties threaten the practical authority of researchers' identification of (alleged) moral errors; unless

[2] Regarding incompatible criteria, see, e.g., Baron (2014, p. 66), Nadelhoffer and Feltz (2008, p. 137), and Kneer and Machery (2019, p. 346). Regarding problematically implicit criteria, see, e.g., Helzer and Pizarro (2011).

attributions of "moral error" are deemed authoritative by those undergoing debiasing, it is hard to see how moral debiasing can both effectively and ethically yield moral improvement.

I undertake identifying the criterion of moral error that can best ground, within empirical moral psychology, a type of moral error that can undergird moral debiasing that would deliver moral improvement. I conclude that, per scientific and practical desiderata, that criterion is: contravening the moral "standard of interest" (not moral *theory*) consisting of instrumental rationality with respect to one's subjective morality given idealization conditions one deems most morally authoritative (*a la* two-tier internalism[3]), that is, the criterion, contravening *two-tier subjective morality*. Since two-tier subjective morality is a moral standard of interest, not a moral theory, it avoids appealing to moral truth, which, due to moral epistemic difficulties, would have rendered the corresponding type of *moral error*, qua kind, scientifically inadequate. Simultaneously, while standards of interest normally lack the requisite authority for those undergoing debiasing, two-tier subjective morality provides such authority because, due to its subjectivity, those undergoing debiasing would typically, perhaps necessarily, deem the standard authoritative.

Two-tier subjective moral debiasing has applications in the non-paternalistic types of moral education that befit higher-education philosophy courses on ethics and/or critical thinking, wherein such debiasing can harvest the low-hanging fruit of those moral improvements that can be achieved without the ethically and practically fraught task of altering adult students' values. (Relatively) amenable course materials are currently available.[4] At a minimum, the criterion, contravening two tier subjective morality, can amend the underspecified, undertheorized, and/or under-justified subjective criteria of moral error currently found in the literature.[5]

[3] Rosati (1996).
[4] See Biasucci and Prentice (2020), *Ethics Unwrapped* (2026), Gentile (2012), and *Noggin* (2026).
[5] See, e.g., Bazerman and Tenbrunsel (2011), Kim, Monge and Strudler (2015), and Kneer and Machery (2019).

Figure 1. Appeals to moral bias/error in moral psychology since 2000

Source / *Explicitness*	**Psychology Journals**	**Other Psychology Sources**	**Philosophy Sources**
"Moral Bias/Error" (use of terms)	Berman, Wheaton & Abramowitz (J; C)[i] Sunstein (J; T,S,H)[ii] Uhlmann, Pizarro & Diermeier (J,U; S,E)[iii]	Baron (J; T)[iv] Bartels, Bauman, Cushman, Pizarro & McGraw (J,U; S,E)[v] Croson & Konow (J; C)[vi]	Prinz (J; C,S,R)[vii] Sauer (R)[viii] Schaefer & Savulescu (R)[ix] Teehan (C)[x]
Synonyms of *Moral Bias/Error* (e.g., "moral illusions")	Tenbrunsel, Rees & Diekmann (J; C)[xi]	Bazerman & Tenbrunsel (J; C,S)[xii]	Bruers (S)[xiii] Huemer (J; C,R)[xiv] Kim, Monge & Strudler(S)[xv] Sinnott-Armstrong (J; R)[xvi]
"Moral Heuristic" (use of term)	Lindstrom, Jangard, Selbing & Olsson (J; H)[xvii] Petersen (J; H)[xviii]	Cosmides & Tooby (J; H)[xix] Sheehan & Lee (H)[xx] Waldmann, Nagel, & Wiegmann (J; 0)[xxi]	Fischer (H)[xxiv] Sinnott-Armstrong, Young & Cushman(J; H)[xxv]
"Bias/Error" re. morality (use of terms w/o "moral")	Kneer & Machery(S,E)[xxvi] Schwitzgebel & Cushman (J; I)[xxvii]	Hoffman (J; C)[xxviii] Kahneman (J; I)[xxix] Nadelhoffer & Feltz (J; C)[xxx] Voiklis & Malle (J; T)[xxxi]	Horton (R)[xxxii]

- For "bias" only in the *systematic error* sense (as opposed to non-evaluative or re. discrimination)
- Non-exhaustive

J = Cited in psychology journals
U = Unenthusiastic (e.g., re. particular moral biases, but concedes the category)

Criterion/Standard of Moral Error/Correctness
F = *Formal Irrationality*
T = *(Broad) Theory* (e.g., utilitarianism)
C = *Narrower Contention(s)* (w/o broad theory)
I = *Indisputably Morally Irrelevant Factor-Sensitivity* (e.g., order effects)
S = *Subjective Standard* (e.g., upon reflection)
E = *Performance Error* (vis-à-vis competence)
H = *Misfiring Heuristic*
R = *Unreliable Process* (other than heuristics)
0 = underspecified/inapplicable

Combination of Explicitness and Source (per proximity to top-left table cell) approximates extent of adoption in psychology

Figure 1 notes

[i] Berman, Wheaton, and Abramowitz (2013, p. 43) regards the *moral bias form of thought-action fusion*—the moral error of believing that thinking about a negative act (e.g., harming one's child) is morally equivalent to performing the act. This entry is notable for its clinical use (within cognitive behavioral therapy).

[ii] Sunstein (2005, p. 533) regards several moral biases vis-à-vis (primarily) "weak consequentialism." This article, while groundbreaking, has significant conceptual (Hahn, Frost & Maio, 2005) and philosophical (Fried, 2005) shortcomings.

[iii] Uhlmann, Pizarro, and Diermeier (2015, pp. 76–79) regards *naïve virtue ethics* explaining apparent moral biases.

[iv] Baron (2014) regards moral biases vis-à-vis utilitarianism, including *omission bias* (judging harmful commissions as worse than corresponding omissions).

[v] Bartels, Bauman, Cushman, Pizarro, and McGraw (2015, p. 482) regards moral flexibility explaining apparent moral biases.

[vi] Croson and Konow (2009, n. 2.) regards self-serving moral bias vis-à-vis fairness (in empirical economics).

[vii] Prinz (2011, p. 224) regards empathetic biases vis-à-vis "appropriate moral concern" and "our considered standards of wrongness."

[viii] Sauer (2018, table 3.2) regards moral errors including retributivism and anti-market attitudes.

[ix] Schaefer and Savulescu (2017, p. 7) regards improving moral reasoning reliability.

[x] Teehan (2016, p. 46) regards theological implications of moral biases.

[xi] Tenbrunsel, Rees, and Diekmann (2019) uses "bounded ethicality" (*a la* "bounded rationality") regarding sexual harassment per *behavioral ethics*, a movement within business ethics that embraces ethical debiasing and falls within professional-ethics psychology, which highly accommodates normative assertions per (1) the pertinence of available (quasi-) normative guidelines (e.g., legal requirements, profession/organization ethics codes), and (2) practical aims (e.g., avoiding liability, responding to scandals). "(Un-)ethical"—vis-à-vis "(im-)moral"— seemingly tracks this normative stance.

[xii] Bazerman and Tenbrunsel (2011) uses "bounded ethicality" in this comprehensive book on behavioral ethics.

[xiii] Bruers (2016) uses "moral illusions" (*a la* "cognitive illusions," a synonym of "cognitive bias" that emphasizes its analogy with optical illusions) for obstinate intuitive moral judgments that are incompatible with a coherent ethical system.

[xiv] Huemer (2013, §6.8) uses "moral illusion" (see xiii) for false intuitions of normative-authority possession (by the state—*a la* Stockholm syndrome or "Milgram obedience").

[xv] Kim, Monge and Strudler (2015) uses "bounded ethicality."

[xvi] Sinnott-Armstrong (2006, p. 359) uses "moral illusion" (see xiii).

[xvii] Lindstrom, Jangard, Selbing and Olsson (2018) regards the *common-is-moral* moral heuristic.

[xviii] Petersen (2009) regards evolved moral heuristics' misapplication in politics.

[xix] Cosmides and Tooby (2006) regards moral heuristics and law.

[xx] Sheehan and Lee (2014) regards moral heuristics concerning "cruelty free" labels.

[xxi] Waldmann, Nagel, and Wiegmann (2012, p. 373) regards a section on moral heuristics in a companion text's chapter.

[xxiv] Fischer (2016) regards disgust as a moral heuristic.

[xxv] Sinnott-Armstrong, Young and Cushman (2012) regards moral intuitions as moral heuristics.

[xxvi] Kneer and Machery (2019, p. 16) regards hindsight bias affecting judgments regarding moral luck.

[xxvii] Schwitzgebel and Cushman (2015) regards philosophers' susceptibility to moral framing and ordering effects.

[xxviii] Hoffman (2000, pp. 209) regards empathetic biases, including the *here-and-now* bias.

[xxix] Kahneman (2013, p. 368) regards moral framing effects.

[xxx] Nadelhoffer and Feltz (2008) regards actor-observer bias regarding moral permissibility (vis-à-vis morality's universality).

[xxxi] Voiklis and Malle (2018, p. 112) regards moral bias vis-à-vis community relativism (*a la* cultural relativism).

[xxxii] Horton (2004) regards biases and debiasing concerning global poverty and aid.

References

- Baron, J. (2014). Moral judgment. In E. Zamir, & D. Teichman (Eds.), *The Oxford handbook of behavioral economics and the law* (pp. 61-92). Oxford: Oxford University Press. Bartels, D. M., Bauman, C. W., Cushman, F. A., Pizarro, D. A., & McGraw, A. P. (2015). Moral judgment and decision making. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (Vol. 1) (pp. 478-516). Malden, MA: John Wiley & Sons, Ltd.
- Bazerman, M. H., & Tenbrunsel, A. E. (2011). *Blind spots: Why we fail to do what's right and what to do about it*. Princeton, NJ: Princeton University Press.
- Berman, N. C., Wheaton, M. G., & Abramowitz, J. S. (2013). Childhood trauma and thought action fusion: A multi-method examination. *Journal of Obsessive Compulsive and Related Disorders, 2*, 43-47.
- Biasucci, C. & Prentice, R. (2021). *Behavioral ethics in practice: Why we sometimes make the wrong decisions*. New York: Routledge.
- Bruers, S. (2016). In search of moral illusions. *Journal of Value Inquiry, 50*, 283-303. Byrd, N. (2021). Bias: What is it? And when is it bad? *Psychology Today*, April 17, 2021. Cho, S., Gonzalez, R., & Yoon, C. (2011). Cross-cultural difference in the preference for cute
- products: Asymmetric dominance effect with product designs. *Proceedings of IASDR*. Cosmides, L., & Tooby, J. (2006). Evolutionary psychology, moral heuristics, and the law. In G. Gigerenzer & C. Engel (Eds.), *Heuristics and the law* (pp. 182-212), Cambridge, MA: MIT Press.
- Croson, R., & Konow, J. (2009). Social preferences and moral biases. *Journal of Economic Behavior & Organization, 69(*3), 201-212.
- Cushman, F., Young, L., & Greene, J. D. (2012). Multi-system moral psychology. In Doris, J. M. & The Moral Psychology Research Group (Eds.). *The Moral Psychology Handbook* (pp. 47-71). Oxford University Press.

- Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 8*(4), 377-383. Ditto, P., Pizarro, D. A., & Tannenbaum, D. (2009). Motivated moral reasoning. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral judgment and decision making: The psychology of learning and motivation* (Vol. 50) (pp. 307-338). San Diego, CA: Elsevier
- Doris, J. (2002). *Lack of character: Personality and moral behavior*. Cambridge University Press.
- Doris, J. (2015). *Talking to ourselves: Reflection, ignorance, and agency*. Oxford University Press.
- Doris, J., Stich, S., Phillips, J., & Walmsley, L. (2020). Moral psychology: Empirical approaches. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2020 ed.). Retrieved from https://plato.stanford.edu/archives/spr2020/entries/moral-psych emp/
- Emanuel, E. J., & Emanuel, L. L. (1992). Four models of the physician-patient relationship. *Journal of the American Medical Association, 267*(16), 2221-2226.
- *Ethics Unwrapped.* (2026). University of Texas at Austin McCombs School of Business. Accessed January 1, 2026, https://ethicsunwrapped.utexas.edu
- Firth, R. (1952). Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research, 12*(3), 317-345.
- Fischer, R. W. (2016). Disgust as heuristic. *Ethical Theory and Moral Practice, 19*(3), 679-693. Fleischhut, N., & Gigerenzer, G. (2013). Can simple heuristics explain moral inconsistencies? In R. Hertwig & U. Hoffrage (Eds.), *Simple heuristics in a social world* (pp. 459-485). Oxford University Press.
- Fried, B. H. (2005). Moral heuristics and the means/ends distinction. *Behavioral and Brain Sciences, 28*(4), 549-550.
- Gentile, M. C. (2012). *Giving Voice to Values*. New Haven, CT: Yale University Press. Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684-704.
- Gilovich, T., & Griffin, D. (2002). Heuristics and biases: Then and now. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 1-17). Cambridge University Press.

- Goldstein, W. M., & Hogarth, R. M. (1997). Judgment and decision research: Some historical context. In W. M. Goldstein & R. M. Hogarth (Eds.), *Research on judgment and decision making: Currents, connections, and controversies* (pp. 3-68). Cambridge University Press.
- Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin Press.
- Green, S. (2015, Mar 6). Training police departments to be less biased. Harvard Business Review.
- Hahn, U., Frost, J., & Maio, G. (2005). What's in a heuristic? *Behavioral and Brain Sciences, 28*(4), 551-552.
- Hand, M. (2014). Towards a theory of moral education. *Journal of Philosophy of Education, 48*(4), 519-532.
- Hand, M. (2018). *A theory of moral education*. Routledge.
- Helzer, E. G. & Pizarro, D. A. (2011). Dirty liberals! Reminders of physical cleanliness influence moral and political attitudes. *Psychological Science, 22*(4), 517-522. Hoffman, M. L. (2000). *Empathy and moral development*. Cambridge, UK: Cambridge University Press.
- Horton, K. (2004). Aid and Bias. *Inquiry, 47*(6), 545-561.
- Huemer, M. (2013). *The problem of political authority*. New York, NY: Palgrave McMillian. Inbar, Y. (2018). Applied moral psychology. In K. Gray & J. Graham, *Atlas of moral psychology* (pp. 537-545). New York, NY: Guilford Press.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux. Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. New York, NY: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 493- 508). Cambridge University Press.
- Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values, and frames.* New York, NY: Cambridge University Press.
- Kelly, D. (2011). *Yuck! The nature and moral significance of disgust.* MIT Press. Kim, T., Monge, R., & Strudler, A. (2015). Bounded ethicality and the principle that "ought" implies "can." *Business Ethics Quarterly, 25*(3), 341-361.
- Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition, 182*, 331-348.

- Kouchaki, M., & Smith, I. H. (2014). The morning morality effect: The influence of time of day on unethical behavior. *Psychological Science, 25*(1), 95-102.
- Kretz, L. (2017). Debiasing the philosophy classroom. *Teaching Philosophy, 40*(1), 11-35. Lamb, M., Brant, J., & Brooks, E. (2022). Seven strategies for cultivating virtue in the university. In M. Lamb, J. Brant, and E. Brooks (Eds.), *Cultivating virtue in the university* (pp. 115-56). Oxford University Press.
- Lapsley, D. K., & Hill, P. L. (2008). On dual processing and heuristic approaches to moral cognition. *Journal of Moral Education, 37*(3), 313-32;
- Larrick, R. P. (2008). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 316-338). Malden, MA: Blackwell.
- Lerner, M., & Simmons, C. (1966). Observer's reaction to the 'innocent victim': Compassion or rejection? *Journal of Personality and Social Psychology, 4*(2), 203–210.
- Lindstrom, B, Jangard, S, Selbing, S., & Olsson, A. (2018). The role of a "common is moral" heuristic in the stability and change of moral norms. *Journal of Experimental Psychology: General, 147*(2)*,* 228-242.
- Loughnan, S., & Piazza, J. (2018). Thinking morally about animals. In K. Gray & J. Graham (Eds.), *Atlas of moral psychology* (pp. 165-174). Guilford.
- May, J. (2018). *Regard for reason in the moral mind*. Oxford University Press. Milgram, S. (1974). *Obedience to authority*. Harper & Row.
- Miller, C. (2014). *Character and moral psychology*. Oxford: Oxford University Press. Miller, C. B. (2017). *The character gap: How good are we?* Oxford University Press. Nadelhoffer, T., & Feltz, A. (2008). The actor–observer bias and moral intuitions: Adding fuel to Sinnott-Armstrong's fire. *Neuroethics, 1*, 133-144.
- Nadurak, V. (2020). Why moral heuristics can lead to mistaken moral judgments. *Kriterion – Journal of Philosophy, 34*(1), 99-113.
- Narvaez, D. (2008). Triune ethics: The neurobiological roots of our multiple moralities. *New Ideas in Psychology 26*(1), 95-119.
- Niemi, L., & Young, L. (2016). When and why we see victims as responsible: The impact of ideology on attitudes toward victims. *Personality and Social Psychology Bulletin, 42*(9), 1227–1242.
- *Noggin: My Brain on Ethics*. (2026). EthicsOps. Accessed January 1, 2026, https://ethicsops.learnworlds.com

- Petersen, M. B. (2009). Public opinion and evolved heuristics: The role of category-based inference. *Journal of Cognition and Culture, 9*(3), 367-389.
- Pizarro, D. A., & Uhlmann, E. L. (2005). Do normative standards advance our understanding of moral judgment? *Behavioral and Brain Sciences, 28*(4), 559.
- Prentice, R. (2004). Teaching ethics, heuristics, and biases. *Journal of Business Ethics Education, 1*(1), 55-72.
- Prentice, R. (2014). Teaching behavioral ethics. *Journal of Legal Studies Education, 31*(2), 325- 365.
- Raibley, J., & Tiberius, V. (forthcoming). Value Fulfillment and the Prudential Good. In D. Copp, T. Rulli, & C. Rosati (Eds.), *The Oxford handbook of normative ethics*. Oxford University Press.
- Railton, P. (1986a). Facts and Values. *Philosophical Topics, 14*(2), 5-31. Railton, P. (1986b). Moral realism. *Philosophical Review, 95*(2), 163-207.
- Railton, P. (2014). The affective dog and its rational tale: Intuition and attunement. *Ethics, 124*(4), 813-859
- Rosati, C. (1996). Internalism and the good for a person. *Ethics, 106*(2), 297-326. Samuels, R., Stich, S., & Faucher, L. (2004). Reason and rationality. In I. Niiniluoto, M. Sintonen, & J. Woleńsk (Eds.), *Handbook of Epistemology* (pp. 1-50). Kluwer Academic Publishers.
- Sauer, H. (2018). *Moral thinking, fast and slow*. New York, NY: Routledge. Schaefer, G. O., & Savulescu, J. (2017). Better minds, better morals: A procedural guide to better judgment. *Journal of Posthuman Studies, 1*(1), 26–43.
- Schnall, S., Benton, J., & Harvey, S. (2008). With a clean conscience: Cleanliness reduces the severity of moral judgments. *Psychological Science, 19*(12), 1219-1222.
- Schwitzgebel, E. & Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition, 141*, 127-137.
- Sheehan, K. B., & Lee, J. (2014). What's cruel about cruelty free: An exploration of consumers, moral heuristics, and public policy. *Journal of Animal Ethics, 4*(2), 1-15.
- Sherman, G. D., Haidt, J., & Coan, J. A. (2009). Viewing cute images increases behavioral carefulness. *Emotion, 9*(2), 282-286.
- Sinnott-Armstrong, W. (2006). Moral Intuitionism Meets Empirical Psychology. In T. Horgan, & M. Timmons (Eds.), *Metaethics after Moore* (pp. 339-366). Oxford: Oxford University Press.

- Sinnott-Armstrong, W. (2019). Moral skepticism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2019 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2019/entries/skepticism-moral/
- Sinnott-Armstrong, W., & Skorburg, J. A. (2021). How AI can aid bioethics. *Journal of Practical Ethics, 9*(1), 1-22.
- Sinnott-Armstrong, W., Young, L., & Cushman, F. (2012). Moral intuitions. In J. M. Doris, & The Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 246-272). Oxford: Oxford University Press.
- Slovic, P. (2007). "If I look at the mass, I will never act": Psychic numbing and genocide. *Judgment and Decision Making, 2*, 79-95.
- Small, D. A., & Loewenstein, G. (2003). Helping a victim or helping the victim: Altruism and identifiability. *Journal of Risk and Uncertainty, 26*(1), 5-16.
- Sobel, D. (2017). *From value to valuing: Towards a defense of subjectivism*. Oxford University Press.
- Sunstein, C. R. (2005). Moral heuristics. *Behavioral and Brain Sciences, 28*(4), 531-542. Teehan, J. (2016). Cognitive science, evil, and God. In H. De Cruz, & R. Nichols (Eds.), *Advances in religion, cognitive science, and experimental philosophy* (pp. 39-60). London: Bloomsbury.
- Tenbrunsel, A., Rees, M., & Diekmann, K. (2019). Sexual harassment in academia: Ethical climates and bounded ethicality. *Annual Review of Psychology, 70*, 245-270. Thaler, R. H., & C. R. Sunstein (2021). *Nudge: The Final Decision: Improving Decisions about Money, Health, and the Environment.* New Haven: Yale University Press. Tversky, A., & Kahneman, D. (1982). Availability: A heuristic for judging frequency and probability. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 163-178). Cambridge University Press.
- Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10*(1), 72-81.
- Voiklis, J., & Malle, B. F. (2018). Moral cognition and its basis in social cognition and social regulation. K. Gray, & J. Graham (Eds.), *Atlas of moral psychology* (pp. 108-121). New York, NY: Guilford.

- Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison, *The Oxford handbook of thinking and reasoning* (pp. 364-389). Oxford: Oxford University Press.

Gabriel Moreira Francisco - THIRD-WAVE EXTENDED MIND AND INTEGRATIVE PLURALISM

Federal University of Rio Grande do Sul

Different views have been proposed for a third-wave extended mind, including the defense of a socially extended mind, arguments in favor of the extension of consciousness, and the integration of the extended mind theoretical framework with enactive and pragmatist perspectives. In this essay, I discuss the third-wave extended mind in the terms of Shaun Gallagher and argue for its coherence with the integrative pluralism proposed by Sandra Mitchell. Gallagher suggests that the formation of a third wave in the extended mind should be characterized by the integration of the extended mind theoretical framework with enactive and pragmatist perspectives. From these perspectives, the investigation of the extension of cognitive processes and mental states could avoid theoretical inconsistencies as well as the criticisms directed at the first-wave extended mind due to the argumentative alternative of the parity principle. Conversely, the integrative pluralism, proposed by Mitchell as the best description of the relationship between scientific theories and complex biological phenomena, advocates that investigations of complex systems should make an explanatory, concrete integration, of models to meet epistemic goals. In agreement with John Sutton and Rick Dale, I acknowledge cognition as a complex phenomenon and suggest the continuation of the investigations of the third-wave extended mind through integrative pluralism. The adoption of integrative pluralism is advantageous because it makes room for compatibility and to explore the collaboration among different theoretical perspectives beyond enactive and pragmatist ones. Thus, distinct explanatory models can be integrated for investigating the extension of cognitive processes and mental states.

Rachele Cavarsaschi - Perception of absences: an emblematic case

University of Cagliari

Perception is not anonymous; rather, it displays a series of information about the subject. Indeed, the observer's personal story is fundamental in perception processes: through perception, we can reconstruct the observer's history, or through his history, we can explain the details of his perception. I hold that the case of absences is revealing in this regard: starting from the study of perception of absences, I propose a style of general perception.

I accept the position according to which it is possible to "see" absences (Teichmann 1970), and I hold that it is not merely a doxastic fact (Farennikova, 2013). However, an epistemic basis is necessary in these perceptual events. I aim to propose a fourth position on the perception of absences, beyond the perceptual-radical one, the cognitivist position, and the metacognitivist position (Voltolini, 2024). I call the fourth position "*quasi-epistemic*": perception always needs an experience, but it also requires a certain condition of cognition, which is subjective. For this reason, absences are always absences of something and for someone: two persons who are having the same perceptual experience may perceive different things based on their histories and previous knowledge. Consider the case of two friends, *a* and *b*, in a bar. The friend *b* invited also a third friend, without having informed the friend *a* that he had also invited his friend *c*, who does not arrive. The absence of *c* can be perceived only by the friend *b*, who is aware that *c* was supposed to arrive, but not by the friend *a*, who is unaware of everything.

This type of perception is not restricted to absences; rather, it is generalizable for every perceptual experience; indeed, each person's matured synaptic configuration is radically different from anyone else's and unique to that individual (Churchland, 2012).

References

- Churchland, P. M., (2012), *Plato's Camera. How the Physical Brain Captures a Landscape of Abstract Universals*, Cambridge, The MIT Press.

- Farennikova, A., (2013), “Seeing Absence”, *Philosophical Studies*, 166: 429– 454.
- Teichmann, J., (1970), “Perception and causation”, *Proceedings of the Aris totelian Society*, 71: 29–41.
- Voltolini, A., (2024), “Perceptual Experiences of (Depicted) Absences”, *JoLMA*, 5(2): 459–476.

Antonio Lizzardi - A Philosopher Watching Scientists Debate Consciousness: Toward a Metaphilosophical Mediation

Cat. Uni. Milan

Contemporary research on consciousness has achieved remarkable theoretical and experimental sophistication. Competing approaches—including Integrated Information Theory (IIT), Global Workspace Theory (GWT), Higher-Order Thought theories, and Predictive Processing—have generated increasingly refined formal models and empirical evidence. Yet the debate remains strikingly controversial. Rather than converging toward a shared explanatory framework, new empirical findings often generate further conceptual disagreement.

This paper argues that the persistence of these controversies is not primarily empirical but metaphilosophical. The debate is aporetic because rival theories rely on different, largely implicit conceptual assumptions concerning the very notions they aim to explain. Adopting a second-order perspective—that of a philosopher observing scientists debating consciousness—the paper does not seek to adjudicate between competing theories but to reconstruct the conceptual presuppositions that structure their disagreement.

The analysis focuses on six interconnected concepts: consciousness, causal organization, information, knowledge, the quantitative/qualitative distinction, and empirical verification. What appears to be disagreement over experimental evidence frequently reflects incompatible interpretations of these concepts. Consequently, empirical results alone cannot resolve the controversy because their interpretation already presupposes a philosophical framework.

The debate surrounding IIT provides an illuminating case study. Supporters and critics disagree not only about the interpretation of Φ or the plausibility of panpsychism, but also about what counts as information, whether causal organization is intrinsic or functional, whether phenomenal consciousness can be captured quantitatively, and what would constitute genuine empirical confirmation. Similar conceptual tensions characterize the opposition between IIT and GWT. Recent adversarial collaborations have clarified empirical predictions without

eliminating these deeper disagreements, suggesting that the controversy is underdetermined by data because it is overdetermined by divergent conceptual commitments.

The paper argues that this diagnosis requires a broader reconsideration of philosophy's role within cognitive science. Philosophy should not be conceived merely as conceptual analysis preceding empirical research. Its distinctive function is metatheoretical: to transcend the internal perspective of individual theories, reconstruct their reciprocal conceptual relations, and provide a systematic understanding of their place within a unified conception of knowledge. Conceptual clarification is therefore only one aspect of a broader philosophical activity of theoretical integration.

From this perspective, the disputed concepts acquire a different meaning. Meaning is not an isolated semantic property but emerges from the systematic network of theoretical relations in which concepts function. Causality is interpreted as a principle connecting explanatory levels rather than merely describing physical interactions. Knowledge is understood as the progressive integration of empirical, theoretical, and metatheoretical dimensions. Accordingly, verification cannot be reduced to experimental confirmation alone but also requires evaluating the conceptual coherence, explanatory scope, and integrative capacity of competing frameworks.

The resulting metaphilosophical conclusion is that philosophy and empirical science are co-original dimensions of inquiry. Scientific theories of consciousness inevitably presuppose conceptual commitments that cannot themselves be experimentally established. Philosophy performs an irreducible constitutive function by continuously making these commitments explicit, critically assessing them, and reorganizing them within a more comprehensive epistemic framework. Progress toward a scientific theory of consciousness therefore depends not only on better experiments but also on the metatheoretical capacity to integrate competing conceptual paradigms.

References

- Baars, B. J. (1988). A Cognitive Theory of Consciousness.
- Block, N. (1995). On a confusion about a function of consciousness. Behavioral and Brain Sciences.
- Chalmers, D. J. (1996). The Conscious Mind.

- Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. Nature Reviews Neuroscience.
- Kozuch, B. (2024). Better bridges: Integrating the neuroscience and philosophy of consciousness. Consciousness and Cognition.
- Michel, M., et al. (2023/2025). Adversarial collaborations for theories of consciousness (Cogitate Consortium).
- Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated Information Theory: From consciousness to its physical substrate. Nature Reviews Neuroscience.
- Cea, I., Negro, N., & Signorelli, C. M. (2024). Only consciousness truly exists? Two problems for IIT 4.0's ontology. Frontiers in Psychology. Kargopoulos,
- P. V. (2023). What philosophy, if any, is needed for cognitive science? Rivista Internazionale di Filosofia e Psicologia.
- Sarıhan, I. (2024). Deflating the hard problem of consciousness by multiplying explanatory gaps. Ratio.



Contributed Symposia

Cognition beyond the body: affective scaffolding, embodied inference, and psychopathology

This symposium explores how our affective life is shaped, sustained, and sometimes destabilized by the material objects, practices, and bodily dispositions through which it unfolds. Recent work in embodied and extended approaches to mind has increasingly emphasized that affect does not reside solely within the organism but is scaffolded by objects, technologies, and habitual practices. In line with this intuition, everyday engagements, such as caring for others, tending to one's surroundings, curating digital records of one's life, are themselves affective and aesthetic acts that help scaffold cognition over time. Increasingly, such scaffolding also takes conversational and artificial forms, as when people turn to AI-based tools for emotional and psychological support.

We discuss how the notion of affective artifacts can be enriched by attending more closely to the aesthetic and caring dimensions of the everyday, which are often habitual practices through which we sustain ourselves and those we care for. At the same time, this scaffolding raises a complementary concern: what happens when the very structures that sustain affect also amplify, entrench, or distort it? Tools that preserve and replay the past show how such scaffolds can stabilize not only adaptive but also repetitive, self-perpetuating patterns. The symposium also asks how the body itself figures in this picture: Can conditions marked by disturbances in affect and embodiment be better understood by looking at the substrate of implicit, lived dispositions (e.g., habits and action tendencies), also known as body memory? Finally, as affective scaffolding increasingly extends into artificial and conversational domains one rising question is whether conversational AI, despite lacking genuine dialogue and interpersonal recognition, still function as a cognitive-affective scaffold that extends and supports therapeutic processes without becoming a therapist itself.

Bringing these questions into dialogue, the symposium opens new avenues for thinking further about the extended and embodied nature of affect itself: if emotional life is constituted not only within the organism but through its material, artificial, and bodily extensions, then the central problem becomes how such extensions sustain, regulate, or unsettle the affective processes they are meant to support.

1) Affective artifacts, aesthetics and care: a deepening view

In this talk I would like to integrate my proposal concerning affective artifacts (Piredda 2020), offering some considerations regarding the relation between affectivity, aesthetics, and care. Originally, affective artifacts have been defined as material or abstract objects (1) that alters the affective state of the agent who interacts with them; (2) whose loss would alter the affective state of the agent; (3) that sometimes are perceived as part of the extended self of the agent. I think that this view – and in general, the view of situated affectivity – could be enriched by the integration of the aesthetic and the caring dimension (Saito 2022). Many of the actions we perform every day, almost automatically, are affective and caring actions. They are also, in some ways, aesthetic actions, or the result of aesthetic choices (Lopes et al. 2022). We do not give much weight to many of the actions we perform every day (e.g., taking care of ourselves and our loved ones—human and non-human, elderly, children, partners, pets), but—if we did—we would notice that the fields of affectivity, care, and aesthetics are much broader than we are normally willing to acknowledge. Following this line of thought, affectivity moves further and further away from the realm of the immediate and the unreflective to become the fruit of cultivation and constant commitment.

References

- McIver Lopes D., Nanay B., Riggle N. (2022). Aesthetic life and why it matters, Oxford University Press.
- Piredda G. (2020). What is an affective artifact? Further developments in situated affectivity, Phenomenology and the Cognitive Sciences, 19, pp. 549–567, doi:10.1007/s11097-019-09628-3.
- Saito Y. (2022). Aesthetics of care. Practice in everyday life, Bloomsbury Academics, London.

2) Extended Rumination

The extended mind thesis has largely been developed using cases of successful, or benign cognitive extensions, such as notebooks. A recent literature on "hostile scaffolding" has sought to address a perceived bias toward harmony. This talk occupies a middle-ground, addressing a case showing that even benign scaffolding can, under certain conditions, scaffold harm. I examine the case of rumination, defined as repetitive, self-reinforcing loops of negative self-directed thought and affect. Drawing on an active inference model of rumination as a temporally extended, affect-driven process, I argue that digital memory platforms like Google Photos do not merely extend rumination, but sustain and amplify it in ways that can go beyond internal rumination, providing an external fuel that has partly migrated beyond the bounds of the skull. Given rumination's wide transdiagnostic significance, I then ask what the implications might be for psychopathology generally, arguing that extended rumination can, under particular ontologies of disorder, extend those pathologies to a degree set by the centrality of its externally realised nodes. The most philosophically interesting scaffolds, I suggest, are neither benign nor hostile but ambivalent.

3) Where is the body in embodied active inference?

There has been much discussion as to whether active inference (AIF) is a purely representational account of cognition, or whether it is an embodied, embedded, extended, and ecological (4E) one. While the 4E view of AIF treats action and perception formally as embodied, the body's contribution to the generative model itself is rarely elaborated. In AIF standard formulations, only the action plan is modelled, not the action itself, and the body is viewed as the realizer of the action plan. These are “thin” conceptions of the body, as they either fail to explicitly model the body as part of the inference process or reduce it to biophysical working of action. This paper advances a “thick” conception centered on the notion of “body memory”, i.e. a phenomenological construct referring to the implicit memory of the lived, emotional body that manifests as dispositions, habits, and ways of engaging with the world. By analogy to the brain in standard formulations of AIF, we suggest treating body memory as the set of these additional, nested layers of inference processes situated between brain and environment, functioning as inherited and developmentally acquired patterned regularities of action selection that is instantiated in the physical architecture over an agent’s lifetime. We conclude by highlighting the potential of body memory under AIF as means to better phenotype symptoms of mental disorders when considering body related experimental paradigms such as motor learning and preparation, procedural memory, and body schema. We discuss implications for understanding trauma, somatic symptom disorders, and the bodily dimensions of mood disorders in a way that corresponds to an ecosocial view of mental disorders.

4. Therapy chatbots as cognitive-affective scaffolds

Conversational AI (CAI) is increasingly being introduced into mental healthcare as a way of expanding access to psychological support and helping people apply therapeutic techniques in everyday life. Although randomized controlled trials suggest that therapy chatbots can reduce symptoms of anxiety and depression and improve well-being, their growing use has also raised philosophical and ethical concerns. Critics argue that therapy fundamentally depends upon a therapeutic alliance grounded in genuine dialogue, mutual recognition, and interpersonal trust—capacities that current AI systems lack. If chatbots merely simulate understanding, then encouraging vulnerable users to trust them may amount to a form of deception. These objections appear to undermine the prospect of AI functioning as a therapeutic partner.

In this talk I argue that this debate is framed too narrowly. Rather than asking whether conversational AI can replace the human therapist or instantiate a genuine therapeutic alliance, we should ask a different question: Can therapy chatbots function as cognitive-affective scaffolds that extend and support therapeutic processes without themselves becoming therapists?

The talk develops an ecological-enactive account of cognitive-affective scaffolding to answer this question. Cognitive-affective scaffolds are environmental resources that enable people to regulate emotion, sustain reflection, structure attention, and support adaptive decision-making by reorganizing interactions between agent and environment. On this view, psychotherapy itself can be understood as a scaffolded process in which the therapist helps patients develop new ways of interpreting experience, regulating emotion, and acting in accordance with their goals and values. Therapy chatbots should therefore be evaluated not by whether they reproduce the interpersonal depth of psychotherapy but by how effectively they scaffold these processes outside the consulting room.

Neuroscience and radical embodiment

Symposium overview: Radically embodied cognitive science has been developed with an externalist spirit, focusing on the dynamics of the agent-environment coupling. Whilst such an extended focus helped in combating and limiting the neurocentrism of more "classically cognitive" approaches, it also hindered the development of a radically embodied approach to neuroscience. After all, if cognition is a matter of the interaction of whole agents with their environments, why should researchers be specifically interested in what happens in the agents' brains?

The aim of this symposium is to show why even radically embodied cognitive science should pay attention to the brain, thereby contributing to the nascent radically embodied cognitive neuroscience.

Facchin Marco (Universiteit Antwerpen) - Brains are not cognitively impartial

Cognitive impartiality (CI) is the view that, when recruiting the resources needed to face a cognitive task, the brain has no bias pro or against cognitive extension (e.g. Clark 2008; 2013; 2016). CI has largely shaped the debate over cognitive extension and the neural processes governing it. Yet, the reasons to endorse CI remain obscure, as no systematic defense of it has been provided.

I shall here argue that CI is false, and should be rejected, as it's relatively easy to marshal empirical data against some predictions easily derivable from CI. I shall focus in particular on two.

The first one is that each cognitive process can be realized either internally or in an extended way. After all, if CI is correct, then the brain should be "free to choose" whether to realize cognitive processes in an extended or purely internal form. Else, it would be biased in favor of some mode of realization. And yet, data concerning the hunting strategies of jumping spiders and parasitoid wasps (cf. Facchin & Leonetti 2024) demonstrate that this is not the case. In fact, these agents can't but rely on certain extended cognitive processes when hunting - as they lack the "brainpower" to realize these processes utilizing only inner (neural) resources.

The second is that, since the brain can "choose" whether to cognize through extended or internal cognitive processing, the choice should be guided by some sort of cost-benefit analysis. Hence we should expect the brain to resort to extended strategies only when it is beneficial. But this contradicts some findings from the empirical research on memory offloading, which indicate that the brain offload memory by default, even when the offloading offers no advantage and imposes additional costs to manipulate external structures (e.g. Gilbert 2015a,b).

All in all then, it appears that CI is false. The brain is not cognitively impartial: it has a bias in favor of extended cognition.

References

- Clark, A. (2008). Supersizing the mind. OUP
- Clark, A (2013). Mindware. OUP
- Clark A. (2016)- Surfing Uncertainty. OUP

- Facchin, M. & Leonetti, G. (2024). Extended Animal Cognition. Synthese.
- Gilbert, S. J. (2015a). Strategic use of reminder. Consciousness and Cognition.
- Gilbert, S. J. (2015b). Strategic offloading of delayed intentions into the external environment. Quarterly journal of experimental psychology.

Favela, Luis H. (Indiana University Bloomington) & Chemero, Anthony (Vanderbilt University) - Direct Perception and Ecological Information Processing

Over the last couple of decades, ecological psychology has become more central to the cognitive and neural sciences. For example, cognitive scientists, neuroscientists, and philosophers of science all talk knowingly about affordances. Nevertheless, affordances are part of a complete theory of animal-environment systems and can be most useful in the context of the entire apparatus of the ecological approach. The hardest part of the apparatus of the ecological approach in the cognitive sciences for neuroscientists to take on board is the claim that perception is direct, which is to say non-representational. The purpose of this talk is to explain direct perception and its role in the ecological approach in such a way that neuroscientists might find appealing. When proponents of the ecological approach say that perception is direct, they are, among other things, claiming that neural activity alone cannot explain perception; perception can only be explained in terms of extended brain-body-environment systems. In systems like this, neural activity need not be seen as representing features of the environment because those features themselves are part of the system to be explained. Our goals in this talk are twofold. First, we will develop an account of brain activity in the context of direct perception. Second, we will argue that direct perception is not in conflict with (a version of) the claim that perception involves information processing.

Gambarotto Andrea (University of Luxemburg) - Rethinking Biological Control: Heterarchy in Brain Networks

Traditional models of biological and cognitive control often presume a hierarchical architecture, where “higher” centers command and lower processes execute. In neuroscience, for example, emotion is commonly portrayed as a subcortical impulse kept in check by a rational cortical executive in a top-down control narrative. In this talk, we challenge this hierarchical paradigm and argue that many biological control systems are better understood as heterarchical networks, characterized by distributed, entangled interactions rather than clear chains of command. We use the integration of emotion and cognition in the human brain as a case study. We show how cognitive and affective functions are deeply interwoven across cortical and subcortical regions, forming a dynamic web of reciprocal influences rather than a one-way control hierarchy. This neuroscientific evidence undermines the notion of neatly localized “centers” for decision, feeling, or control, and instead highlights widespread networked organization as the hallmark of heterarchy. We will conclude by illustrating the potential for insight to be generalized to biological control mechanisms at large: from metabolic regulation to neural decision-making, control is an emergent property of many interdependent components rather than a privilege of any single part.

Vicente Raja (University of Murcia) - On Radio Stations, Computers, and Radical Embodied Neuroscience

At least in its experimental incarnation, contemporary cognitive neuroscience seems to be walking a path towards the naturalization of stimulation, tasks, and experimental setups (e.g., Ulanovsky, 2025). By walking this path, the field is getting closer to alternative frameworks like ecological psychology or enactivism and it is even embracing some of its central concepts, like affordances. However, radical embodied neuroscience–also influenced by these alternative frameworks–remains far from the reach of such new developments in contemporary cognitive neuroscience due to a couple soft commitments or motifs (Raja, 2024): complex stimulation and organic behavior. These two motifs are a hard pill to swallow for most cognitive neuroscientists and most of the time they are implicitly or explicitly rejected (e.g., Angelaki et al., 2026). In this paper, I will provide a simple, concrete explanation of the two motifs for cognitive neuroscientists with the aim to, at least, make them consider the two motifs as legit alternatives to notions like encoding or input-output systems and not just as a Gibsonian extravaganza. Then, I will show how these two motifs might help the understanding and advancing of cognitive neuroscience in its computational incarnation with a focus on the prospects of a radical embodied take on AI.

References

- Angelaki et al., (2026). The Simons Collaboration on Ecological Neuroscience: Studying how the brain interacts with the world. Neuron. https://doi.org/10.1016/j.neuron.2026.04.036
- Raja, V. (2024). The motifs of radical embodied neuroscience. European Journal of Neuroscience, 60(5), 4738-4755.
- Ulanovsky, N. (2025). Natural Neuroscience. MIT Press.

New Directions in Teleosemantics - Content Determination and the Explanatory Turn

Teleosemantics is a family of naturalistic theories that explain representational status and content in terms of the biological functions of the systems that produce or use representations. Its central attraction is that functional standards allow teleosemantic theories to distinguish representation from mere causal covariance, and correct representation from misrepresentation: a system may have the function of responding to, indicating, or guiding behaviour with respect to a condition even when it fails to do so successfully. Classical accounts differ over whether producer or consumer functions determine content, but characteristically ground these functions in the histories of the relevant systems (Millikan, 1984; Dretske, 1986; Neander, 1995, 2017). This historical commitment has prompted both revisions of etiological teleosemantics and ahistorical alternatives.

Two related questions have emerged as key issues in contemporary debates. The first concerns content determination. Representational states normally stand at the end of extended causal chains and contribute to several effects at different levels. A frog's perceptual state, for example, may be caused by a retinal pattern, a suitably small, dark, moving thing, or an animal that constitutes prey. Its production may contribute to tongue movements, prey capture, nutrition, and survival. Which of these relations determines its content? This illustrates the content indeterminacy and distality problems. Recent proposals appeal to producer and consumer functions, causal sensitivity, informational relations, normal conditions, and relations between different levels of biological success (Martínez, 2013; Schulte, 2018; Artiga, 2021; Withrington, 2024).

The second question concerns the explanatory role of content. According to the emerging explanatory turn, theories of content should be assessed not only by whether they yield intuitively plausible content ascriptions, but also by how content figures in genuine explanations of cognition and behaviour (Schulte, 2023). This raises questions about the relationship between historical and present-day explanations. Are historically determined contents suited primarily to explaining why representational systems exist, while different, ahistorical contents are required for mechanistic explanations of how those systems currently operate? Or can historically determined contents themselves explain present behaviour without

a circular dependence between content and functional role? Recent forms of teleosemantic and explanatory pluralism make these questions especially pressing (Thorpe, 2025).

The symposium addresses these two developments through two paired sets of contributions. The first pair offers alternative solutions to the content-determination problem. **Sérgio Farias de Souza Filho** defends Neander's producer-based informational teleosemantics by supplementing it with a counterfactual selection Principle. The principle distinguishes distal content from proximal causal intermediaries by asking whether the counterfactual selection of a response to the proximal feature depends on the counterfactual selection of a response to the distal feature. **Esteban Withrington** instead develops a hybrid of producer and consumer-based teleosemantics. On his account, content is the most crucial normal condition for successful consumption that is also a normal cause of the representation. Applied to the familiar frog case, the proposals diverge: de Souza Filho's account identifies a suitably small, dark, moving thing as content, whereas Withrington's account identifies prey animals. Their contributions therefore ask whether producer-based teleosemantics can solve distality using its own resources or whether content determination must integrate consumer functions.

The second pair turns from content determination to the explanatory role of content. **Johan Heemskerk** argues that etiological and non-etiological teleosemantics need not be rivals. Drawing on Mayr's (1961) distinction between ultimate and proximate explanation, he proposes that historically grounded contents are suited to ultimate explanations of why representational systems exist, whereas ahistorical contents are suited to proximate, mechanistic explanations of how they operate. **Andrés Rubio Krohne** approaches the explanatory problem through a challenge to functionalist metasemantics. If present behavioural dispositions determine content, content cannot in turn explain those same dispositions without circularity. He argues that causal theories face these problems for desires because they must appeal to behavioural dispositions to distinguish desire content. By grounding content in biological function and evolutionary history, teleosemantics can preserve the explanatory role of content without grounding it in the very dispositions it explains. Together, the papers ask whether historically grounded content is confined to ultimate explanation or can also contribute to synchronic explanations of cognition and behaviour.

Taken together, the four contributions develop teleosemantics in two complementary directions. They offer new principles for determining which features of the world are represented, and they reassess what teleosemantic content can contribute to scientific and

philosophical explanation. The symposium thus connects disputes that are often pursued separately: what representations represent and why representational content matters explanatorily.

1. Representing What Really Matters: An Informational Teleosemantic Solution to the Distality Problem

Sérgio Farias de Souza Filho — Federal Rural University of Pernambuco and Federal University of Pernambuco in Recife (Brazil)

Mental states represent *distal features* such as books and hourglasses rather than *proximal features*, such as light rays. Consider the *causal chain* from the book, through light rays, to the tokening of the representational state. In virtue of what does the mental state represent the distal feature (the book), rather than the proximal one (the light ray)? This is *the distality problem*. Informational teleosemantics is often attacked on the grounds of failing to give an account of the distality of representational states: it fails to give a principled explanation for why the representational state represents the distal feature and not the proximal one. The goal of this presentation is to defend informational teleosemantics from the distality problem by proposing and developing a *counterfactual selection principle.* Accordingly, representation R represents distal feature *C*, not proximal feature *Q*, because the *counterfactual selection* of the visual system's response to *Q depends* on the counterfactual selection of the system's response to C, but the counterfactual selection of the system's response to *C* does not depend on the counterfactual selection of the system's response to *Q*. So, let us consider the famous case of the frog that responds to any small-dark-moving thing (SDMT). The frog's visual system was selected to respond to *both* the distal feature – SDMT – and the proximal feature – shadow on the retina – by tokening the representation. However, the content of the representation is *SDMT*, not *shadow on the retina*, because the counterfactual selection of the system's response to shadow on the retina *depends* on the counterfactual selection of the response to the SDMT, but the counterfactual selection of the system's response to the SDMT *does not depend* on the counterfactual selection of the response to shadow on the retina. The problem is that the counterfactual selection principle, insofar as it appeals to both *actual* and *counterfactual* historical environments in order to determine content, is threatened by the objection that counterfactual conditions have indeterminate truth values. The problem of the indeterminate reference situation is especially challenging to this principle. Finally, I propose a solution to this problem by appealing to counterfactual comparisons.

2. Frogs Recognize Prey: A Hybrid Teleosemantics

Esteban Withrington — National Autonomous University of Mexico (Mexico)

Much of the debate in teleosemantics has focused on simple innate animal representations, such as the neural states involved in frogs' detection of prey. The two main approaches have diverged. One appeals to the conditions required for the behaviour guided by the representation to be beneficial, entailing that frogs' states represent nutritious things. The other explains content in terms of the stimuli that elicit a representational state, regarding frogs' states as representing small moving objects. Neither approach accounts for these states as representing prey animals. Yet ethologists routinely explain goal-directed animal behaviour by attributing fallible capacities to recognize prey, predators, conspecifics, and other behaviourally relevant objects. Because these attributions are empirically grounded and explanatorily successful, there is good reason to treat the corresponding contents as legitimate *explananda* for a theory of content. I propose a version of teleosemantics for simple animal representations that combines Millikan's suggestion that they represent the most crucial "normal conditions" required for them to bring about their beneficial behavioural effects with Neander's suggestion that they represent their "normal causes". The content of a simple representation is the most crucial normal condition that both causes tokens of the representation and enables them to have beneficial behavioural effects. This hybrid account promises to explain the recognitional contents that neither approach, taken on its own, accounts for. In the frog case, it explains how activations of prey detectors can indeed represent prey animals.

3. Ultimate and Proximate Explanations and Teleosemantic Pluralism

Johan Heemskerk — University of Warwick (UK)

According to teleosemantics, content is determined in part by a function of the representing system. However, because there are two dominant views of function—one etiological and the other non-etiological—teleosemantics is itself divided into etiological and non-etiological camps. The two camps are often taken to be in opposition because they deliver competing content attributions. However, this paper argues for teleosemantic pluralism by showing that etiological and non-etiological teleosemantics need not be treated as rival theories of content, because they are suited to different kinds of scientific explanation. Drawing on Mayr's distinction between ultimate and proximate explanation, together with existing pluralist views about Wright-style and Cummins-style functions, we contend that etiological teleosemantics is best understood as attributing contents suited to ultimate explanations of why representational traits exist. Non-etiological teleosemantics, by contrast, attributes contents suited to proximate, mechanistic explanations of how cognitive capacities are currently realised. We argue that standard objections made by either camp against the other involve a category mistake: they assess one form of teleosemantics by adequacy conditions appropriate to the other. Once the relevant explanatory targets are separated, both approaches can be seen as independently legitimate and jointly required for a complete account of the representational features of the mind. Finally, we suggest that this framework has implications for the realism debate: against anti-realist worries, representational content may figure in genuine scientific explanation in more than one way, across different temporal and organisational scales.

4. Escaping Explanatory Circularity through Teleosemantics

Andrés Rubio Krohne — Central European University (Austria)

Bar Luzon (2024) has recently argued that functionalism about metasemantics faces a circularity problem. Functionalism holds that the content of one's mental states is partly determined by one's behavioural dispositions. Luzon contends that this gets the order of explanation backwards: it is the content of our mental states that explains our behavioural dispositions. To avoid this problem, Luzon has favoured a causal theory of content. In this talk, I argue that the causal approach cannot escape the circularity problem, since it must resort to functional explanations for some mental states. Although the problem can arise for various mental states, I concentrate on desires. While a causal theory can plausibly account for beliefs, I argue that it cannot account for desires without resorting to behavioural dispositions. A belief that *p* can be defined causally, as the state normally caused by *p*, but a desire that *p* has no privileged cause; it is defined instead by its tendency to produce *p*. This reintroduces the very circularity Luzon sought to avoid. I also argue that teleosemantics offers a genuine way out: it grounds the content of a desire that *p* in the biological function of producing *p*, rather than in current behavioural dispositions. Biological function, in turn, is grounded in the evolutionary history of the mental state, rather than in its current effects. This allows content to explain current behavioural dispositions without being grounded in those dispositions.

References

- Artiga, M. (2021). Beyond black dots and nutritious things: A solution to the indeterminacy problem. *Mind & Language*, 36(3), 471–490. https://doi.org/10.1111/mila.12284 Dretske, F. I. (1986). Misrepresentation. In R. J. Bogdan (Ed.), *Belief: Form, content, and function* (pp. 17–36). Oxford University Press.
- Luzon, B. (2024). The Euthyphro challenge in metasemantics. *Philosophy and Phenomenological Research*, 108(1), 217–237. https://doi.org/10.1111/phpr.12956 Martínez, M. (2013). Teleosemantics and indeterminacy. *Dialectica*, 67(4), 427–453. https://doi.org/10.1111/1746-8361.12039
- Mayr, E. (1961). Cause and effect in biology. *Science*, 134(3489), 1501–1506. https://doi.org/10.1126/science.134.3489.1501
- Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism*. MIT Press.
- Neander, K. (1995). Misrepresenting and malfunctioning. *Philosophical Studies*, 79(2), 109–141. https://doi.org/10.1007/BF00989706
- Neander, K. (2017). *A mark of the mental: In defense of informational teleosemantics*. MIT Press. https://doi.org/10.7551/mitpress/9780262036146.001.0001
- Schulte, P. (2018). Perceiving the world outside: How to solve the distality problem for informational teleosemantics. *The Philosophical Quarterly*, 68(271), 349–369. https://doi.org/10.1093/pq/pqx052
- Schulte, P. (2023). *Mental content*. Cambridge University Press. https://doi.org/10.1017/9781009217286
- Thorpe, L. (2025). Pluralistic teleosemantics: Why we need both Bickhard-representations and Millikan-representations. *Phenomenology and the Cognitive Sciences*. Advance online publication. https://doi.org/10.1007/s11097-025-10068-5
- Withrington, E. (2024). Frogs recognize prey: A causal-behavioral teleosemantics. *Biology & Philosophy*, 39, Article 35. https://doi.org/10.1007/s10539-024-09967-9

Spontaneous Sense-Making Across Inner Speech, Mind Wandering and Dreaming

Introduction

In recent years, both philosophy of mind and the cognitive sciences have been marked by a surge of interest in thoughts arising in the absence of strong constraints on their content and dynamics (Christoff et al., 2016). Diverse phenomena spanning the sleep-wake cycle—including mind-wandering, dreaming, and creative thinking—have been increasingly characterised by their common experiential features, such as their free-flowing and hyperassociative quality (Fox et al., 2013; Kirberg & Windt, 2024; Windt & Voss, 2018). Their resemblance has prompted theoretical efforts aimed at finding neural and cognitive mechanisms that underlie these mental states, and that might justify their grouping under the broader family of spontaneous thought or cognition (Fox & Christoff, 2018).

While early research on the topic has privileged performance-centred approaches, characterising mind-wandering as "task-unrelated or stimulus-independent" (Morrison et al., 2019), recent theoretical and neuroscientific advances have emphasised the importance of investigating spontaneous mental states as phenomena in their own right. There has been increasing interest in the link between mind-wandering, dreaming and memory processing, involving the consolidation of novel information and experiences as well as their integration into long-term semantic networks (Wamsley, 2018). This idea has been met with interdisciplinary agreement. Yet, the specific dynamics by which spontaneous thought supports the interweaving of recent and remote memories into self-narratives remain to be specified. The analysis of the connection between spontaneous mentation and memory has far-reaching implications for the role of mind-wandering and dreaming in epistemology.

The current Symposium presents work on spontaneous thought, focusing on the following questions: 1) How does the internalisation of the social practice of dialogue contribute to the spontaneous generation of internal narratives? (Daniel Mueller); 2) Does mind-wandering play a role in the formation of the narrative self? (Yizhi Li); 3) Is mind wandering subject to cognitive control? (Mengting Zhang); 4) Can we acquire knowledge from dream experiences? (Gaia Mizzon).

Daniel Mueller - Inner Speech & Creativity Pangelonium in the Brain— A Sociology of the Default Mode Network

Osnabrück University, Germany

It is common knowledge that “no [hu]man is an island” (Donne, 1624). Yet, this has not stopped cognitive science from regularly treating the individual as the atomic unit of analysis (Dengsø, 2025). The body of work in the paradigm of 4E cognition has started to take steps towards situating the individual in its social surroundings (Newen et al., 2018). However, if we take the developmental perspective introduced by Vygotsky (2024) seriously, according to which all higher mental functions are social in origin, we ought to find a way to locate the social in the individual’s brain.

The default mode network (DMN) is a vast brain network particularly active in periods of spontaneous thought such as mind wandering, dreaming and creativity (Kemmerer and Dove, 2026). Its nodes have been implied in social cognition, semantic memory, and speech processing, and other cognitive functions (Kemmerer, 2025). Menon (2023) has proposed that the function of the DMN as a whole is to generate an internal narrative. My proposal adds to Menon’s considerations and tries to integrate some further findings from the literature on the DMN and the social nature of humans. Building on the aforementioned genetic principle by Vygotsky (2024), I argue that humans internalise the social practice of dialogue. These internalised models of close persons, which I call Angeloi (Greek for messenger), reside in a society of mind, the Pangelonium (Greek for all + messenger). There is an ongoing dialogue in the Pangelonium in which the Angeloi utter proto-narratives (short fragments of stories) about the transmodal sensory input which is received from the senses and from memory.

I claim that the DMN is a good candidate for the psychological scaffolding of the Pangelonium. First, it integrates transmodal sensory input into narratives. Second, it is implicated in social cognition, as supported by findings suggesting that DMN activity at rest is more similar between two people the closer their relation is (Hyon et al., 2020). Furthermore, this finding holds at multiple levels of social structure (Kemmerer and Dove, 2026): from relatives to acquaintances, in networks of peers, and at the largest level (entire cultural groups). Third, the processing time course of the DMN is slow compared to unimodal

sensory areas, but at the right processing speed for speech comprehension as well as narrative understanding (Lerner et al., 2011). For these reasons, I suggest that we should take seriously the idea of the Pangelonium and—if we were to search for it in the brain—look for it in the DMN.

Yizhi Li - How Our Tales Spin Us: Spontaneous Thought and the Formation of the Narrative Self

Ruhr-Universität Bochum, Germany

This talk argues that mind wandering can contribute to the formation of the narrative self by broadening the range of autobiographical material available for self-understanding and shaping how that material is remembered and interpreted. Mind wandering is frequent and often involves autobiographical memories and other forms of self-relevant thought. Whereas deliberate remembering is typically guided by current goals and explicit retrieval attempts, mind wandering can bring to mind personal memories that fall outside the focus of such attempts. It may also provide the occasion for the first conscious episodic recollection of at least some events. Because episodic remembering is constructive, the context in which an event is first recollected helps determine its earliest conscious organization, including the perspective from which it is represented, its emotional tone, and its perceived significance. This initial organization may influence later recollections and thereby shape the event's subsequent mnemonic history. Mind wandering can therefore contribute to narrative self-formation in three related ways: by broadening the range of autobiographical material available for self understanding, by shaping some of this material at the formative moment of first episodic recollection, and by generating interpretations and connections through reflection on remembered events. Through these processes, autobiographical material is sampled, organized, and incorporated into a self-narrative. Given its frequency, mind wandering may make a substantial cumulative contribution to the continuing formation and revision of the narrative self.

Mengting Zhang - Mind Wandering is Not Subject to Cognitive Control

Monash Centre for Consciousness and Contemplative Studies, Australia

Mind wandering has attracted increasing research interest over the past decades, yet little consensus has been reached regarding its nature and classification. This paper focuses on the relationship between mind wandering and cognitive control, an issue that has received comparatively less attention. It adopts a dynamics-based approach, understanding mind wandering as characterized by a particular form of dynamics: free movement in thinking. Drawing on existing theories of cognitive control, it proposes an account of cognitive control that emphasizes the goal-conflict tension as the prerequisite for the involvement of cognitive control. It then examines different ways in which mind wandering may not be subject to cognitive control and develops a taxonomy of mind wandering accordingly. In light of this dynamics-based understanding of mind wandering and the proposed account of cognitive control, the paper argues that mind wandering, insofar as it is characterized by this form of dynamics, is not subject to cognitive control. This account sheds light on the phenomenon of intentional mind wandering and provides a new perspective from which to re-examine existing empirical research on mind wandering.

Gaia Mizzon - What Dreaming Teaches

Monash Centre for Consciousness and Contemplative Studies, Australia

The question of how dreams can offer personal guidance and solutions has ignited the interest of humankind for thousands of years. Nonetheless, dreams have often been considered epistemically irrelevant or even deceptive, becoming the target of sceptical debates. Burgeoning research on the function of dreaming for memory integration as well as the development of dream engineering techniques aimed at harnessing dreams for creativity and learning call for a reassessment of the role of dreaming in epistemology. This work addresses unresolved questions concerning the potential use of sleep experiences for knowledge generation. By drawing on recent debates on epistemically transformative experiences (Paul, 2014) and the relationship between imaginative scaffolding and experiential knowledge (Kind, 2020), I argue that dreams can disclose epistemic access to what it is like to undergo novel kinds of experiences just by virtue of having dreamt them. Furthermore, expanding on accounts supporting the instructive uses of imagination in epistemic contexts (Badura & Kind, 2021; Kind & Kung, 2016), I contend that at least a subset of dream uses, when adequately constrained, can justify contingent beliefs about the world.

References

- Badura, C., & Kind, A. (2021). Epistemic Uses of Imagination. *Routledge*. Christoff, K., Gordon, A., & Smith, R. (2011). The role of spontaneous thought in human cognition. In *Neuroscience of Decision Making*. Psychology Press.
- Dengsø, M. J. (2025). Cognitive science and cartesianism: Representational pull and methodological individualism. *Journal of Theoretical and Philosophical Psychology*. Donne, J. (1624). Meditation XVII, page 87.
- Fox, K. C. R., & Christoff, K. (2018). *The Oxford Handbook of Spontaneous Thought: Mind Wandering, Creativity, and Dreaming*. Oxford University Press.
- Fox, K. C. R., Nijeboer, S., Solomonova, E., Domhoff, G. W., & Christoff, K. (2013). Dreaming as mind wandering: Evidence from functional neuroimaging and first-person content reports. *Frontiers in Human Neuroscience*, 7. https://doi.org/10.3389/fnhum.2013.00412
- Hakonen, M., Ikäheimonen, A., Hultén, A., Kauttonen, J., Koskinen, M., Lin, F.-H., Lowe, A., Sams, M., and Jääskeläinen, I. P. (2022). Processing of an audiobook in the human brain is shaped by cultural family background. *Brain Sciences*, 12(5).
- Hyon, R., Youm, Y., Kim, J., Chey, J., Kwak, S., and Parkinson, C. (2020). Similarity in functional brain connectivity at rest predicts interpersonal closeness in the social network of an entire village. *Proceedings of the National Academy of Sciences*, 117(52):33149–33160.
- Kaplan, J. T., Gimbel, S. I., Dehghani, M., Immordino-Yang, M. H., Sagae, K., Wong, J. D., Tipper, C. M., Damasio, H., Gordon, A. S., and Damasio, A. (2017). Processing narratives concerning protected cross-cultural investigation of neural correlates. *Cereb. Cortex*, 27(2):1428–1438.
- Kemmerer, D. (2025). Does the default mode network mediate an ongoing internal narrative? an evaluation of menon's (2023) hypothesis. *Journal of Cognitive Neuroscience*, 37(12):2676–2683.
- Kemmerer, D. and Dove, G. (2026). The default mode network: A trans-modal cortical system for enhancing cognition and sharing mental worlds in language-specific, culturally diverse ways. *Cortex*.
- Kind, A. (2020). What Imagination Teaches. In E. Lambert & J. Schwenkler (Eds.), *Becoming Someone New: Essays on Transformative Experience, Choice, and Change* (p. 0). Oxford University Press. https://doi.org/10.1093/oso/9780198823735.003.0008

- Kind, A., & Kung, P. (Eds.). (2016). *Knowledge Through Imagination*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198716808.001.0001
- Kirberg, M., & Windt, J. (2024). The Kaleidoscope of Bizarreness: The Analysis of First Person Reports Shows the Relationship between Dreaming and Mind Wandering to Be Complex (SSRN Scholarly Paper No. 4831325). *Social Science Research Network*. https://doi.org/10.2139/ssrn.4831325
- Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of neuroscience*, 31(8):2906–2915.
- Menon, V. (2023). 20 years of the default mode network: A review and synthesis. *Neuron*, 111(16), 2469–2487. https://doi.org/10.1016/j.neuron.2023.04.023
- Morrison, H., McBriar, S., Powell, H., Proudfoot, J., Fitzgerald, D., & Callard, F. (2019). What is a psychological task? The operational pliability of "Task" in psychological laboratory experimentation. *Engaging Science, Technology, and Society*, 5, 61–85.
- Newen, A., De Bruin, L., and Gallagher, S. (2018). *The Oxford handbook of 4E cognition*. Oxford University Press.
- Paul, L. A. (2014). *Transformative Experience*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198717959.003.0001
- Vygotsky, L. S. (2012). *Thought and language*, volume 29. MIT press. Vygotsky, L. S. (2024). The genesis of higher mental functions. In *The concept of activity in Soviet psychology*, pages 144–188. Routledge.
- Wamsley, E. J. (2018). Dreaming and Waking Thought as a Reflection of Memory Consolidation. In K. Christoff & K. C. R. Fox (Eds.), *The Oxford Handbook of Spontaneous Thought: Mind-Wandering, Creativity, and Dreaming* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190464745.013.23
- Windt, J. M., & Voss, U. (2018). Spontaneous Thought, Insight, and Control in Lucid Dreams. In *The Oxford Handbook of Spontaneous Thought*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190464745.013.26

Title: The Navigational Mind

1. Overview

The mind sciences have historically treated propositional thought as a symbolic layer grafted onto a distinct sensorimotor substrate. The architecture this symposium explores challenges this dichotomy by proposing the reuse and extension of navigation-like systems for higher cognition including propositional thought. This symposium consists of four papers that collectively argue that thinking is a kind of controlled movement through internally represented structured spaces.

Recent and current empirical work demonstrates that the neural mechanisms originally identified as supporting physical spatial navigation—specifically, grid-like and distance codes within the entorhinal cortex and medial prefrontal cortex—operate as a more general coordinate system (Mark et al., 2024). Functional neuroimaging reveals that these map-like structures are recycled by the brain to help humans organize and navigate abstract conceptual schemas, such as semantic categories (Bokeria et al., 2021). Furthermore, this navigational architecture extends into the social domain, where the brain organizes complex information about other individuals—such as their relative placement along dimensions of competence and warmth—into a structured "social cognitive map" (Liang et al., 2024). Crucially, the formation of these stable, grid-like representations transforms discrete learning events into structured knowledge, thereby enabling the brain to track novel relationships, generalize across non-spatial tasks, and guide flexible, adaptive inferences well beyond direct sensory experience (Xiao et al., 2025).

However, a substantial gap remains between showing that hippocampal-entorhinal and neocortical systems encode relational structure in a map-like way and, most importantly, explaining how such codes genuinely underwrite propositional content and logical inference. This symposium bridges that gap by articulating how the brain's computational architecture for adaptive control, active sensing, memory, and navigation provides the fundamental template for complex cognition.

The presentations progress from foundational neural mechanisms to abstract logical operations:

Paper 1 explores how propositional negation inherits its computational format and semantic grounding from navigation-failure encoding in the hippocampus.

Paper 2 utilizes the concept of operational coherence to demonstrate that wayfinding, memory, and inference share the same architectural action.

Paper 3 deploys the Correspondence Network Framework and Representational Similarity Analysis to explain how neural manifolds become content-bearing structures capable of supporting propositional reasoning.

Paper 4 synthesizes these insights into a unified neurocognitive theory, proposing that thought is literally inner navigation through structured representational spaces, built upon an architecture evolved for navigation rather than a classical symbolic processor.

By moving from spatial navigation to propositional thought, this symposium offers a biologically grounded, unified theory of mind.

Paper 1: Negation as Navigation Failure: Toward a Cognitive Map Account of Propositional Negation

Monosree Chakraborty

Independent Researcher | Kolkata, India

There is something philosophically strange about negation. We navigate the world by building up representations of what is there, what follows from what, and where things can be found. Yet negation asks the mind to represent what is not there, what does not follow, where nothing is. This paper proposes a cognitive map account of propositional negation: the claim that negation, at the level of neural architecture, is processed as a form of navigation failure. This is not a metaphor but a mechanistic hypothesis grounded in the structure of the hippocampal-entorhinal system.

Cognitive map theory, from Tolman onward, has been a theory of structured presence, representing locations and metric frameworks. What the map model has never cleanly handled is structured absence, despite propositional thought being saturated with it. The key mechanism proposed here is expectation-based path termination. The hippocampal system generates predictions about upcoming states during navigation and replay. When a predicted state fails to obtain, the system encodes that failure as a distinct representational event, which serves as a positive signal of failed expectation rather than a mere null result.

The proposition "There is no food at location L" is, at the level of cognitive architecture, a compressed representation of a path that was simulated and found to terminate in failure. This mechanism bridges the gap in possible-world semantics for negation. Possible worlds are not abstract objects that the mind mysteriously accesses; they are internally simulated trajectories that the hippocampal-entorhinal system generates during offline replay. This account provides a neurological basis for why individuals with hippocampal damage show characteristic impairments in inferential negation, reasoning about what did not happen, and processing what is absent. Our fluency with negation stems from the fact that we are animals that evolved to detect blocked paths and violated spatial expectations. Propositional negation is simply what happens when that evolutionary machinery gets recruited for abstract thought.

Paper 2: The Navigational Architecture of Mind understood through Operational Coherence

Andrea Hiott

Universität Heidelberg

What does it mean for a body to think and how might that be related to what it means for a body to navigate? Many cognitivist accounts treat propositional thought as a symbolic faculty or 'add on' to a sensorimotor substrate. This paper argues that both are navigational and are expressions of what we can assess as the same ongoing architectural action--specifically, that knowing-how and knowing-that are not two kinds of knowing but two modes of assessing coherent bodily activity; there is the engaged operation of moving through a structured space, and the detached thematizing movement as inferential structure, and for both, there is the same architecture. This architecture is best understood through Hasok Chang's concept of operational coherence, defined as the internal consistency and mutual reinforcement of a system of epistemic action.

This paper extends that idea and relates it to hippocampal research. First, it argues that propositional logic is itself a system of practice that operates as a structured traversal of inferential state space, committing a cognitive system to a particular trajectory of action and inference. Next, it argues that this same characterization applies when a body navigates a geographical landscape, retrieves a memory, or follows a chain of thought. While these are different systems of practice with different regularities—thinking of one's grandmother is not the same operation as finding one's way through a park—they share the same architectural action.

This reframing has important consequences for how we understand hippocampal-entorhinal system research and its computational models. Rather than viewing this system as simply producing maps of space, it is better understood as the domain across which our modeling exhibits operational coherence, spanning spatial and inferential state spaces. We study processes like remembering, wayfinding, and inference as different systems with their own operational coherence, but taken together, they display the same navigational operational coherence. They map the same actionable architecture through different modes of regularity of action in different state spaces. Grounded in Chang's account of epistemic

iteration, this view posits that coherence is maintained constitutively through iterative cycles of activity. Mind, in all its iterations, is best understood as navigational, so we can understand knowing-how and knowing-that as two modes of the same, coherent navigational assessment.

Paper 3: Navigable Correspondence Networks: Neural Manifolds as Content-Bearing Structures

Wojciech Mamak

Polish Academy of Sciences (Institute of Philosophy and Sociology) & NASK Institute

Current empirical work has demonstrated that grid-like and distance codes exist not only for physical space but also for abstract semantic and social spaces. However, there remains a substantial gap between showing that entorhinal and frontal cortices encode relational structure in a map-like way and explaining how such codes underwrite propositional content and inference. This paper argues that Representational Similarity Analysis (RSA), interpreted within a Correspondence Network Framework, supplies the missing interpretive layer between navigation-like neural architectures and propositional thought.

Studies show that novel word meanings elicit a six-fold, grid-like modulation of pattern similarity in right postero-medial entorhinal cortex, alongside Euclidean distance codes in medial prefrontal cortices. While these studies suggest that hippocampal-entorhinal systems provide a general-purpose architecture for traversing low-dimensional neural manifolds embedded in high-dimensional state spaces, these results have mostly been interpreted in a thin, merely geometric way. Assuming a simple isomorphism fails to provide a principled account of when such map-like manifolds possess genuine content.

Deploying the Correspondence Network Framework, this paper asserts that representational systems exploit multiple channels of structural similarity concurrently, constructing networks of "infocorrespondences." RSA reveals which structural features of a neural space a system actually preserves as content-bearing constraints. The mathematical compression of complex semantic spaces into low-dimensional manifolds allows the same topological substrate to be reused across domains. Propositional operations—such as comparative judgments, simple conditionals, and counterfactual traversals—can be modeled as constrained trajectories through this topological substrate. Crucially, a trajectory through a grid-like manifold becomes propositional when downstream control mechanisms can evaluate the trajectory against stored goals and re-identify its endpoints across contexts. Propositionality emerges from the interaction between structurally stable correspondence

networks and the control systems that exploit them, providing a mechanism for conceptual reasoning without presupposing a language-like code of thought.

Paper 4: The Movement of Thought: Navigation as Neurocognitive Architecture

Gualtiero Piccinini

University of Missouri

This paper defends a neurocognitive theory of mind organized around a simple but far-reaching thesis: thought is inner navigation through structured representational spaces. While the metaphor of mental motion is ancient, this framework turns the metaphor into a literal architectural claim. The brain did not first evolve a classical symbol processor and then apply it to space. Rather, complex cognition grows out of biological machinery for adaptive control, active sensing, valuation, action selection, memory, and navigation.

Perception, action, and thought are not independent computational problems solved by separate faculties, but interlocking aspects of a single adaptive problem faced by mobile, value guided organisms in uncertain environments. To solve these problems, the architecture for distal navigation is reused and hybridized to support higher cognitive functions. This paper outlines how basic intentionality functions as teleosimulation, enabling brains to successfully represent the nonexistent, the possible, and the false. It further defends neurocognitive nativism, demonstrating that biological computation requires specific "wiring before firing."

By clearing away theoretical obstacles—such as the virtual machine fallacy and the strict explanatory autonomy of psychology—this framework moves beyond the traditional deadlocks between classicism and connectionism. It reveals that the behavioral evidence for systematicity and productivity does not require a classical Language of Thought. Instead, it points toward a nonclassical system where allocentric maps, routes, and trajectories serve as the structural foundations. Ultimately, the paper demonstrates how spatial mapping hardware scales up to construct the intersubjective world of social cognition and to navigate the socially learned action spaces of language. In drawing these threads together, the paper recapitulates the unificatory aspect of this framework, showing how navigation acts as the ultimate neurocognitive engine.

References

- Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, 362(6415), eaat6766.
- https://doi.org/10.1126/science.aat6766
- Bokeria, L., Henson, R. N., & Mok, R. M. (2021). Map-like representations of an abstract conceptual space in the human brain. *Frontiers in Human Neuroscience*, 15, 620056. https://doi.org/10.3389/fnhum.2021.620056
- Chang, H. (2017). Operational coherence as the source of truth. *Proceedings of the Aristotelian Society*, 117(2), 103–122. https://doi.org/10.1093/arisoc/aox004
- Chang, H. (2022). *Realism for Realistic People: A New Pragmatist Philosophy of Science*. Cambridge: Cambridge University Press.
- Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292), 1464–1468.
- Liang, Z., Wu, S., Wu, J., Wang, W. X., Qin, S., & Liu, C. (2024). Distance and grid-like codes support the navigation of abstract social space in the human brain. *Elife*, 12, RP89025. https://doi.org/10.7554/elife.89025.3
- Mark, S., Schwartenbeck, P., Hahamy, A., Samborska, V., Baram, A. B., & Behrens, T. E. (2026). Flexible neural representations of abstract structural knowledge in the human Entorhinal Cortex. *Elife*, 13, RP101134. https://doi.org/10.7554/elife.101134.1
- Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.
- Stanley, J., & Williamson, T. (2001). Knowing how. *The Journal of Philosophy*, 98(8), 411–444.
- Xiao, Z., Wang, X., Zhang, J., Ou, J., He, L., Qu, Y., ... & Liu, Y. (2025). Human hippocampal ripples predict the alignment of experience to a grid-like schema. *BioRxiv*, 2. https://doi.org/10.1101/2025.01.08.632069

The Phenomenal Space of Wakefulness & Dreaming

Phenomenologically grounded approaches are quickly gaining popularity in consciousness research. They offer a way to capture and appreciate the experiential richness of states and features of consciousness. This methodological shift offers a rare opportunity to (re-)evaluate how we think about and define various aspects of consciousness. In light of this opportunity, we focus on how wakefulness, dreaming, and self-consciousness have been described in the past, and show where such theories may fall short. Instead of striving for a simple, unified theory, future approaches should account for the multidimensionality of meta- and self-consciousness across wakefulness and dreaming.

This symposium features the work of four scholars who integrate scientific and philosophical considerations to investigate wakefulness, dreaming, lucid dreaming, and the self. Each presentation approaches its respective target through multiple perspectives and advocates for a multidimensional, phenomenologically guided approach. First, Merlin Herrick presents considerations about the classification of wakefulness. Herrick demonstrates how wakefulness is defined in various ways across disciplines and identifies potential constraints on a definition of wakefulness. Second, Simone Anthes evaluates the definitional pluralism of dreaming, asking whether a unified definition of dreaming is necessary to overcome ongoing controversies in dream research. Third, Ema Demšar delves into the micro-phenomenology of insight in lucid dreams. Demšar shows that even though lucid dreams are often assumed to be a uniform subcategory of dream experience, there are multiple ways in which insight arises. Fourth, Melanie Rosen investigates the dream self in relation to several different theories of self-consciousness.

1. A multidimensional account of wakefulness

Merlin Herrick - School of Philosophical, Historical and Indigenous Studies (SOPHIS) and Monash Centre for Consciousness & Contemplative Studies (M3CS), Monash University, Australia

Wakefulness is a ubiquitous yet surprisingly under-analysed concept. As a familiar term, it is assumed to be conceptually obvious, yet its boundaries and defining features remain unclear. Furthermore, different literatures characterise wakefulness in disparate ways. Animal sleep research relies on behavioural notions such as environmental responsiveness and a low arousal threshold. Clinical accounts of pathological conditions (e.g. unresponsive wakefulness syndrome) often appeal to physiological processes such as the ascending reticular arousal system and the metabolic demands of the waking brain. Philosophical perspectives emphasise its distinctive experiential character, sometimes cashed out in terms of the knowledge available to the waking subject. These characterisations are the data that a theory of wakefulness has to fit. The simplest theory would privilege one characterisation over the others or posit a single underlying system that each characterisation tracks. But such a theory is available only if the characterisations co-vary, and this is not the case. For example: sleep-like slow waves arise locally in the cortices of behaviourally awake humans and animals; unresponsive wakefulness syndrome patients show brainstem arousal without behavioural responsiveness or evidence of experience; deep meditation involves a loss of behavioural responsiveness while other measures remain intact; and lucid dreaming brings experiential and behavioural capacities associated with waking into sleep. Dissociation between notions of wakefulness is widespread rather than marginal and is not confined to pathology. I therefore defend a multidimensional account, on which behavioural, experiential, and experiential wakefulness are distinct dimensions that usually, but not always, co-vary. This account not only secures conceptual clarity, but also has a methodological payoff for empirical research. Vague questions of the form "is some entity or state wakeful?"—asked of fetuses, artificial systems, lucid dreams, or deep meditation—can be replaced by tractable questions about particular dimensions of wakefulness.

2. What we talk about when we talk about dreaming

Simone Anthes - School of Philosophical, Historical and Indigenous Studies (SOPHIS) and Monash Centre for Consciousness and Contemplative Studies (M3CS), Monash University, Australia

What is a dream? We all think we know what we are talking about when we talk about dreaming. But it turns out we mean different things. This, coupled with the oft-mentioned multiplicity of dreaming and the different disciplines dedicated to its investigation, is reflected in a varied representation of dreaming in the literature. The idea of a pre-defined phenomenon "dreaming" is one of the major conceptual obstacles in dream research and is at the heart of the discussion on integrating dream research into the broader context of consciousness science. However, in actual dream research, there is no such clearly defined phenomenon. Over the last decades, many theories, hypotheses and models about the purpose and character of dreaming have emerged from the various disciplines involved in dream research. Yet, they often fail to explicitly define their explanatory target. I survey contemporary dream theories and demonstrate considerable variety in how they conceptualise dreaming. Acknowledging that there has nonetheless been massive progress in dream research, we ask whether a definition-led approach is even necessary and present several alternative approaches. We conclude that while it is necessary to be transparent about perspective and preconceptions on dreaming, a unified definition may not be the prerequisite to advancing dream research that it is often framed to be.

3. The Phenomenal Space of Wakefulness and Dreaming

Ema Demšar - School of Philosophical, Historical and Indigenous Studies (SOPHIS) and Monash Centre for Consciousness and Contemplative Studies (M3CS), Monash University, Australia

Lucid dreaming (LD) – in which dreamers become aware that they are currently dreaming – has gained increasing attention in consciousness studies, yet its fine-grained experiential dynamics remain undercharacterized. We conducted in-depth micro-phenomenological interviews on 22 signal-verified lucid dreams experienced in a controlled sleep-laboratory setting, mapping within-episode dynamics from the first recalled moments, through the emergence of lucid insight, and into post-onset periods of sustained or lost lucidity. Across episodes, temporal dynamics were highly variable, and we observed diverse phenomenological clusters at every stage. Lucidity sometimes appeared as an abrupt shift from non-lucidity, but most often built up gradually via extended pre-lucid engagement, ranging from curiosity-driven exploration to distress-driven affective escalation. Lucidity was often precipitated by specific triggers, including reality testing behaviors, identification of mismatches, and recognition of incorporated external sensory cues, administered by researchers with the aim of inducing lucidity. The emergence of lucidity itself could be experienced as a sudden or gradual process, and could be accompanied by diverse affective states. Post-onset periods ranged from stable lucid states of active exploration or receptive observation to more fragile states culminating in the collapse of the dream world. Eye signals were variably embedded in these dynamics and often shaped dream experience, for example by contributing to the emergence and maintenance of lucidity or destabilizing the dream. Methodologically, our findings demonstrate that micro-phenomenological interviews –originally developed for the study of waking experience –can be effectively adapted for LD research. Conceptually, they highlight diverse forms of meta-awareness, pre-lucidity, metacognitive processing, and agency in LD, and show that variability and complexity are intrinsic features of dream lucidity rather than noise. Practically, moving beyond binary classifications and attending to the fine-grained temporal and structural dynamics of LD can allow future research to harness LD as a dynamic window into the emergence of insight, metacognition, and agency during sleep.

4. Self-consciousness in vicarious dreams

Melanie Rosen - Trent University, Canada

Self-consciousness in dreams raises important philosophical questions. Are we self-conscious when we dream of being someone else? In dreams, self-consciousness—the awareness of oneself as oneself—can be altered, diminished, or absent. Yet individuals recall dreams in which the protagonist appears self-conscious while believing they are a different person from the dreamer. On some interpretations of psychological continuity theory, the dream protagonist may indeed count as a different person. In such cases, the dreamer may not be conscious of anything at all.

Drawing on philosophical and scientific literature, I argue that dreams offer a valuable case study for understanding how self-consciousness arises and what cognitive resources it depends on. Competing theories identify different components of self-consciousness, such as bodily ownership, agency, the first-person perspective, metacognition, memory, and the concept of a 'minimal self'. I show how these features are often disrupted in dreams, leading to two key implications.

First, different theories of self-consciousness will predict different outcomes regarding the presence or absence of self-consciousness in dreams. Second, dreams may serve as an intuition pump for evaluating these theories. The construction of a dream self suggests that self-consciousness, like perception, draws on multiple cognitive faculties, sometimes relying on one feature more heavily when others are degraded. Yet this degradation has limits—beyond a certain point, a self, or the same self, may fail to emerge.

THE METATHEORY OF PSYCHOPATHOLOGY: THE PROBLEM OF PSYCHOSIS IN PHENOMENOLOGY AND NEUROSCIENCE

INTRODUCTION

Psychopathology is the scientific discipline that studies clinical phenomena such as mental symptoms, behavioral dysfunctions, neuropsychological syndromes, primary psychiatric disorders, and neuropsychiatric disturbances associated with brain disease. These phenomena share two key characteristics: (a) they affect the psychological dimension of human life, and (b) they are conceptualized as pathological. Psychopathological conditions have a substantial impact on societies worldwide. However, our scientific understanding of these conditions, as well as the effectiveness of our therapeutic interventions, remains limited.

Theoretical frameworks in psychopathology include medical approaches grounded in functional neuroanatomy, biochemistry, psychopharmacology, cognitive, affective, and social neuroscience, neuroendocrinology, neuroimmunology, physiology, and metabolic research. Psychological approaches draw on behavioral and cognitive psychology, psychodynamic theories, and phenomenological-hermeneutical traditions. Social approaches emphasize the historical, cultural, political, and economic factors that shape psychopathology.

Many mental health professionals endorse a biopsychosocial perspective. Nevertheless, important debates persist among schools of thought because these theoretical frameworks rest on different ontological and epistemological assumptions about psychopathology. A metatheoretical framework is therefore needed to clarify these assumptions and establish the conditions for a coherent and integrative understanding of psychopathology.

As an interdisciplinary science, psychopathology aims to describe, explain, predict, and provide the foundations for the prevention and treatment of mental disorders. Ultimately, its central challenge is to provide a scientific explanation of these disorders. Numerous theories have been proposed to address this task. However, some theorists reject —in principle— the very idea that these conditions involve any kind of pathology, or that the concept of *mental*

disorder can be legitimately employed. According to these views, the notion of *mental disorder* is invalid because there is no "mental," no "disorder," or neither.

The metatheoretical problem in psychopathology concerns explaining why certain patterns of behavior and experience are regarded as disordered or pathological within some theoretical frameworks but not others. What accounts for the diversity of explanations of psychopathology? On what theoretical foundations are psychopathological constructs based? Are these constructs valid objects of inquiry for medicine, psychology, and neuroscience? What theoretical foundations would allow a coherent integration of the frameworks that explain clinical phenomena? More broadly, the metatheoretical problem may be formulated as follows: within the boundaries of science and philosophy, what are the conditions under which psychopathological constructs can be regarded as valid?

In this proposal, we use the problem of psychosis as a case study to explore the metatheoretical foundations of psychopathology. Psychosis is a common and disabling syndrome that appears in various psychiatric, neurological, pharmacological and general medical conditions [1,2]. Without appropriate treatment, individuals experiencing psychotic episodes generally show significant impairments in reality testing and interacting with their environment, which can result in reduced social functioning and high-risk behaviors [3]. Jean Garrabé, in his *Taxonomic Dictionary of Psychiatry*, explains that Von Feuchtersleben first introduced the term psychosis into psychiatric language in 1845.[4] Over time, the term psychosis was used more narrowly to describe mental disorders that involve a significant disturbance in reality testing. According to the American Psychiatric Association, psychosis is as a syndrome found in a group of mental disorders characterized by one or more of the following phenomena: a) delusional ideas, b) hallucinations, c) negative symptoms, d) disorganized thinking, and e) disorganized or abnormal motor behavior, including catatonic signs [1]. Psychotic disorders are defined in the *International Classification of Diseases* (ICD-11) by significant impairments in reality testing and changes in behavior, as shown by delusions, hallucinations, disorganized speech, and disorganized behavior [2]. Psychosis offers a relevant window into the metatheory of psychopathology. It is useful to decompose the metatheoretical problem of psychosis into two fundamental concepts: *the psychological* and *the pathological*.

The psychological, refers to goal-directed behavior, subjective experience, and the neuropsychological functioning of integrated agents interacting with their environment. This

raises a series of ontological and epistemological questions: **1) Representation.** Are the concepts of mental and neural representation necessary to explain psychosis? Are there more adequate alternatives? Is it justified to posit the existence of conscious and unconscious mental representations?[5–9] **2) Information.** Is it necessary to assume that mental states possess informational content? Does the concept of information provide an indispensable bridge between environmental, neurobiological, and mental levels of explanation?[5,7,9] **3) Computation.** Is the concept of computation necessary to explain the transformation of representations and/or information in neurocognitive systems? What are the strengths and limitations of computational models, and what alternatives are available? [5,7–10] **4) Cognitive ontology.** To what extent do mental events depend on their underlying physical substrate? Do mental states, representations, and information play a genuine causal or functional role in behavior, or are they merely fictional or epiphenomenal constructs? How useful are the traditional categories of mental life? More generally, should mental constructs be understood from eliminativist, fictionalist, instrumentalist, or realist perspectives?[11–14]

Secondly, when studying patterns of behavior, experience, and neuropsychological functioning, we must address the conceptual problems surrounding the notion of *pathology*. These include the following: **1) Structural and functional abnormalities.** What are the strengths and limitations of research on structural lesions and functional abnormalities? What are the scope and limitations of approaches based on *physiological dysfunction*? Can concepts such as *psychological dysfunction* and *social dysfunction* be regarded as valid? [15,16] **2) Subjectivity and normativity.** Is it valid to interpret patterns of suffering and diminished psychological agency as indicators of psychopathology? What is the epistemic status of intersubjective evaluations of rationality? **3) Causation.** How should causation be conceptualized within an explanatory framework? Must causal explanations ultimately be grounded in neuroscience? Can psychopathology be explained causally without appealing to abnormal bodily processes? What role do context and relational systems play in causal explanation? Is it possible to develop a taxonomy that distinguishes psychological, neuropathological, and hybrid forms of causation?[17–19] **4) Qualitative versus quantitative approaches.** Should explanatory priority be given to qualitative approaches that emphasize meaning and purpose, or to quantitative approaches that focus on the objective measurement of biological processes? **5) Dimensional versus categorical models**. Should psychopathology be conceptualized dimensionally, categorically, or through a combination of both approaches? Under what conditions is it

justified to classify inherently dimensional phenomena into discrete categories?

6) Refinement versus abolition. Should psychopathology be refined as a scientific discipline to improve explanation and therapeutic practice, or should the concept itself be deconstructed and abandoned?

RETHINKING NEUROPSYCHIATRIC SUFFERING: A METATHEORETICAL APPROACH TO INTEROCEPTION AND SUBJECTIVITY.

XIMENA GONZÁLEZ-GRANDÓN

The study of interoception has emerged as a promising avenue for understanding the embodied and affective foundations of subjective experience in psychiatry. However, prevailing neuropsychiatric models continue to privilege observable symptoms, physiological dysfunction, and operational diagnostic criteria, often relegating first-person experience to a secondary role. This paper argues that interoception should be situated within a metatheoretical framework that integrates phenomenology, neuroscience, and philosophy of psychopathology. We propose that neuropsychiatric suffering –especially present in anxiety and panic, mood and eating or somatoform disorders– cannot be adequately understood solely through structural or functional abnormalities, but must also account for the lived experience of altered bodily self-awareness, diminished agency, and affective distress. From this perspective, interoception is not merely a measurable physiological process but a constitutive dimension of the subject's embodied relation to the world. Such an approach challenges the reduction of psychopathology to quantitative symptom thresholds and supports a conception of mental disorders as simultaneously natural and normative: grounded in biological processes while defined by their significance for the experiencing and situated person. We argue that integrating first-person phenomenological methods with contemporary interoceptive neuroscience offers a coherent metatheoretical framework that bridges subjective and objective levels of analysis. This perspective advances a more comprehensive understanding of psychopathology, preserving the epistemic value of lived experience while fostering clinically meaningful and theoretically integrated models of neuropsychiatric suffering.

TOWARDS AN EXPERIMENTAL PHILOSOPHY OF PSYCHIATRY: CONCEPTUAL AND PRACTICAL REFLECTIONS ON A QUALITATIVE STUDY ON THE CLINICAL VALIDITY OF THE CONCEPTS OF DELUSION.

PABLO LÓPEZ SILVA

Delusions are among the most complex phenomena in psychiatry. Yet, despite its clinical and philosophical relevance, the concept remains highly ambiguous. Using an experimental-philosophical framework, this presentation reports preliminary results from a study examining how the validity of competing definitions of delusion is constructed through psychiatrists' application of the concept to clinical cases. After identifying three dominant notions in the literature— phenomenological, operational-diagnostic, and epistemic-motivational— we conducted semi-structured qualitative interviews with eight psychiatrists working in specialized psychosis services. Participants evaluated three standardized clinical vignettes, each constructed to instantiate one of the three notions. Data were analyzed using descriptive qualitative analysis, focusing on patterns in diagnostic reasoning rather than frequency counts. After presenting the results, the presentation will dive into some conceptual and practical reflections. For example, it will examine (i) the fact that clinicians did not treat the three definitions as interchangeable and the philosophical consequences of this; (ii) how the clinicians' judgments reflected selective reliance on conceptual criteria that supported clear communication, ethical prudence, and decisive clinical action, and (iii) how these findings suggest that the practical validity of diagnostic definitions depends not only on theoretical coherence or reliability, but on how well they function in guiding clinical judgment under uncertainty. The final part of the presentation addresses the methodological implications of these results, focusing on: (i) how experimental philosophy can inform conceptual disagreement in current philosophy of mind; (ii) the limitations of using experimental-philosophical methods to analyze clinical notions; and (iii) potential directions for future conceptual and qualitative research emerging from this methodological framework.

A METATHEORETICAL APPROACH TO PSYCHOSIS: TOWARD A SYSTEMIC FRAMEWORK. JESÚS RAMÍREZ-BERMÚDEZ

In this presentation, I draw on case studies of patients with pathologies of self awareness to explore the value of a systemic framework for addressing the metatheoretical problem of psychopathology. The first case concerns autoimmune psychosis. It is grounded in a scientific discovery that led to an effective therapy, providing empirical support for causal reasoning. Yet it also raises a theoretical question: if interventions on a neurobiological mechanism explain changes in the psychological properties of the patient, such as consciousness, does this inevitably commit us to a reductionist conception of psychopathology? The second case concerns a pattern of psychalgia and unilateral musical hallucinations that convey a peculiar litany, experienced as blue music that progressively degrades the patient's bodily organs. This case of affective psychosis emerges from a tragic life history and compels us to take psychological causality seriously. Taken together, these cases invite reflection on the reciprocal relationships among physiological mechanisms, psychological processes, and social contexts. Reductionism assumes that explanations based on lower levels of organization —physiological mechanisms— possess inherently greater explanatory power than those grounded in psychological and social processes. Holism, by contrast, assumes that the agent's interactions with the environment are sufficient to explain the clinical problem without reference to lower-level mechanisms. In neuropsychiatry, however, both perspectives have important limitations. Holism may overlook disorders rooted in physiological dysfunction, whereas reductionism often fails to account for clinical patterns shaped by psychological interactions. A systemic framework offers an alternative. It holds that real entities are either systems or parts of a system, and that psychological agents exhibit emergent properties that are not properties of their constituent parts. These emergent properties are qualitatively novel and arise from the dynamic organization of the system as a whole.19,22,23 This perspective supports an integrative systemic framework rather than reductionist or autonomous theories of psychopathology.

FINAL REMARKS

Considering the metatheoretical problems of psychopathology can give rise to a neutral, interdisciplinary research program aimed at clarifying the commitments underlying psychopathological theories. This program seeks to refine the fundamental questions concerning psychological phenomena and pathological processes such as the psychoses, while mapping competing theories within a taxonomic space defined by metatheoretical dimensions. Such an approach is necessary for the development of a coherent and integrative framework for psychopathology. The joint work of philosophers, scientists and clinicians is necessary to develop this metacognitive research program.

References

1. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. Fith Editi.; 2013.

2. World Health Organization. International Classification of Disease. 11th Revision. 2021. Accessed December 21, 2021. https://icd.who.int/en

3. Lieberman JA, First MB. Psychotic disorders. *N Engl J Med*. 2018;379(3):270-278. doi:10.1007/978-3-319-57111-9_2084

4. Garrabé J. *Diccionario Taxonómico de Psiquiatría*. Fondo de Cultura Económica; 1993.

5. Hipólito I. Cognition Without Neural Representation: Dynamics of a Complex System. *Front Psychol*. 2022;12. doi:10.3389/fpsyg.2021.643276

6. Poldrack RA. The physics of representation. *Synthese*. 2021;199(1-2):1307- 1325. doi:10.1007/s11229-020-02793-y

7. Piccinini G. Situated Neural Representations: Solving the Problems of Content. *Front Neurorobot*. 2022;16. doi:10.3389/fnbot.2022.846979

8. Barack DL, Krakauer JW. Two views on the cognitive brain. *Nat Rev Neurosci*. 2021;22(6):359-371. doi:10.1038/s41583-021-00448-6

9. Hutto DD, Myin E. *Radicalizing Enactivism*. MIT press; 2012. doi:10.7551/mitpress/9780262018548.001.0001

10. Williams DJ. How to be a realist about computational neuroscience. *Synthese2*. Published online 2025.

11. Pessoa L, Medina L, Desfilis E. Refocusing neuroscience: moving away from mental categories and towards complex behaviours. *Philos Trans R Soc B Biol Sci*. 2022;377(1844). doi:10.1098/rstb.2020.0534

12. Demeter T, Parent T, Toon A. What is mental fictionalism? In: *Mental Fictionalism: Philosophical Explorations*. 2022:1-24.

doi:10.4324/9781003091073-1

13. Poldrack RA, Yarkoni T. From brain maps to cognitive ontologies: Informatics and the search for mental structure. *Annu Rev Psychol*. 2016;67:587-612. doi:10.1146/annurev-psych-122414-033729

14. Bolt T, Nomi JS, Arens R, et al. Ontological Dimensions of Cognitive Neural Mappings. *Neuroinformatics*. 2020;18(3):451-463.

doi:10.1007/s12021-020-09454-y

15. Westlin C, Theriault JE, Katsumi Y, et al. Improving the study of brain behavior relationships by revisiting basic assumptions. *Trends Cogn Sci*. 2023;27(3). doi:10.1016/j.tics.2022.12.015

16. Ritunnano R, Bortolotti L. Do delusions have and give meaning? *Phenomenol Cogn Sci*. 2022;21(4):949-968. doi:10.1007/s11097-021-09764-9

17. Ross LN, Bassett DS. Causation in neuroscience: keeping mechanism meaningful. *Nat Rev Neurosci*. 2024;25(2):81-90. doi:10.1038/s41583-023- 00778-7

18. Borsboom D, Cramer AOJ, Kalis A. Brain disorders? Not really: Why network structures block reductionism in psychopathology research. *Behav Brain Sci*. 2019;42. doi:10.1017/S0140525X17002266

19. Fried EI. Studying Mental Health Problems as Systems, Not Syndromes. *Curr Dir Psychol Sci*. 2022;31(6):500-508. doi:10.1177/09637214221114089

20. Huhn M, Nikolakopoulou A, Schneider-Thoma J, et al. Comparative efficacy and tolerability of 32 oral antipsychotics for the acute treatment of adults with multi-episode

schizophrenia: a systematic review and network meta-analysis. *Lancet*. 2019;394(10202). doi:10.1016/S0140-6736(19)31135- 3

21. Schneider-Thoma J, Chalkou K, Dörries C, et al. Comparative efficacy and tolerability of 32 oral and long-acting injectable antipsychotics for the maintenance treatment of adults with schizophrenia: a systematic review and network meta-analysis. *Lancet*. 2022;399(10327):824-836. doi:10.1016/S0140-6736(21)01997-8

22. Bunge M. *Treatise on Basic Philosophy, Volume 4, Ontology II: A World of Systems.* D. REIDEL PUBLISHING COMPANY; 1979. doi:10.1007/978-94-009- 9392-1

23. Romero GE. *Scientific Philosophy*.; 2018. doi:10.1007/978-3-319-97631-0