21st Annual Conference of the Italian Association for Cognitive Science
AISC2025

Connected Minds

Dynamic Interactions in Cognitive, Social and Artificial Systems

Book of Abstracts

IMT SCUOLA ALTI STUDI LUCCA

AISC ASSOCIAZIONE ITALIANA di SCIENZE COGNITIVE

September, 17-19 2025
IMT School for Advanced Studies Lucca, Lucca (Italy)

Table of Contents

Keynote talks

The entangled brain: How perception, cognition, and emotion are woven together

Luiz Pessoa - *September 17, 14:00, Cappella Guinigi*

The traditional view of brain organization posits that cognitive, emotional, and motivational processes operate through fairly separate, dedicated neural systems. This presentation challenges this modular perspective, arguing instead for an "entangled brain" characterized by massive combinatorial connectivity, distributed functional interactions, and circuits/networks as fundamental units of function. The presentation critiques the standard decomposition approach that emphasizes independence between mental faculties, proposing instead that distributed circuits dynamically assemble across the neuroaxis in context-dependent ways. I also illustrate how subcortical brain regions traditionally associated with "lower" functions participate in complex cognitive-emotional processes. The proposed integrative framework has implications for understanding complex behaviors, mental disorders, and the fundamental nature of brain function, suggesting that interaction and integration—rather than separation—are necessary for elucidating how large-scale brain systems support behavior.

Worlds within words: conceptual diversity across cultures and individuals

Claudia Mazzuca - *September 18, 16:10, Cappella Guinigi*

Concepts are the building blocks of human cognition, yet research in cognitive science reveals there is consistent variability in how they are represented and expressed through language. This variation appears to stem from multiple factors, including cultural and linguistic systems, socio-demographic variables such as gender and age, as well as more specific dimensions like expertise and educational background. In this talk, I will present a series of studies examining conceptual flexibility through the lens of conceptual abstractness, highlighting the role of lifelong experience in shaping conceptual boundaries. I will also introduce a novel theoretical framework that moves beyond the traditional dichotomy between abstract and concrete concepts, emphasizing the critical role of social interaction in driving conceptual variation.

Reasoning and science

Hugo Mercier - *September 19, 12:30, Cappella Guinigi*

Dan Sperber and I have developed an interactionist theory of reason, according to which human reason serves social functions of argumentation and justification. This theory can account for features of reasoning at the individual or small group level, in particular explaining why the exchange of arguments in small groups tends to converge on epistemically superior answers (as in the truth-wins effect). These micro-level results appear to conflict with macro-level

observations, in particular the fact that science, arguably the best way of providing epistemically superior answers to a broad range of questions, is a rather exceptional cultural phenomenon. I will suggest that the interactionist theory of reason can help explain both why science is possible at all, and why it takes very special circumstances for it to emerge.

Contributed Talks

September, 17 2025

The motifs of a radically embodied cognitive ontology

Marco Facchin[1]
[1]Universiteit Antwerpen

The ontology of mainstream (computational & representational) cognitive science consists of computational/representational kinds, ideally mappable on neural kinds [1,2]. Non-mainstream, radically embodied cognitive science rejects the computational, representational and neurocentric assumptions of its counterpart [3,4], and so it must have a very different cognitive ontology. And yet, this issue is rarely explored in the literature. As a result, the shape of the alternative "radically embodied" cognitive ontology is still unclear.

My talk shall attempt to shed some light on the matter. More specifically, I wish to uncover the motifs of a radically embodied cognitive ontology: the vague, partially unspecified and open-ended ideas acting as very general constraints on the cognitive ontology of any individual radically embodied research programme [5].

To this end, I will analyze a number of prominent radically embodied research programmes [3,4, 6-14], and identify the following very general ideas:

(1) Representational/computational kinds are substituted by "interactive" kinds (such as affordances [3,6, 10-13], sensorimotor contingencies [4,7-9], or optimal grip [10-12])
(2) The relevant material kinds upon which the kinds in (1) are ideally mappable are no longer neural kinds, but parameters of the complex agent-environment coupled system
(3) The mapping will be context-dependent, and take into account diachronic factors such as the agent's development history [see esp. 14]

I will then claim that (1)-(3) are the motifs of a radical embodied cognitive ontology: the vague, open-ended ideas constraining the individual cognitive ontologies of individual radically embodied research programmes.

I intend that claim to be descriptive: the motifs in (1)-(3) should actually be "played" by radically embodied research programmes. To check that this is the case - and thus validate my claim - I will take a look at the actual experimental research practices of radically embodied cognitive science. I will thus look at how ecological psychologists identify and parametrize affordances [15,16] and how enactivists formalize sensorimotor contingencies and use them in explanations [17], and claim that these practices de facto are variations on the motifs I uncovered.

References

[1] Poldrack, R. A. (2010). Mapping Mental Function to Brain Structure: How Can Cognitive Neuroimaging Succeed? Perspectives on Psychological Science, 5(6), 753–761.
[2] McCaffrey, J. B. (2023). Evolving concepts of functional localization. Philosophy Compass, 18(5), e12914.
[3] Chemero, A. (2009). Radically embodied cognitive science. MIT Press.
[4] Di Paolo et al. (2017). Sensorimotor life. OUP.

[5] The motifs of radical embodied neuroscience. European Journal of Neuroscience, 60(5), 4738-4755.
[6] Anderson, M. (2014) After phrenology. MIT Press
[7] Thompson, E. (2007) Mind in life. Harvard University Press
[8] Hutto, D. & Myin, E. (2013). Radicalizing Enactivism. MIT Press
[9] Hutto, D. & Myin, E. (2013). Evolving Enactivism. MIT Press
[10] Bruineberg, J. (2017) Active Inference and the Primacy of the “I Can”. In: Metzinger, T. and Wiese, W., Eds., Philosophy and Predictive Processing, MIND Group, Frankfurt am Main, 67-84.
[11] Bruineberg, J., & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. Frontiers in human neuroscience, 8, 599.
[12] De Haan, S. (2020). Enactive psychiatry. Cambridge University Press.
[13] Favela, L. H. (2023). The ecological brain: Unifying the sciences of brain, body, and environment. Taylor & Francis.
[14] Di Paolo, E. et al. (2022). Laying down a forking path: Tensions between enaction and the free energy principle. Philosophy and the Mind Sciences, 3 https://doi.org/10.33735/phimisci.2022.9187
[15] Warren, W. H. (1984). Perceiving affordances: visual guidance of stair climbing. Journal of experimental psychology: Human perception and performance, 10(5), 683.
[16] Konczak, J., Meeuwsen, H. J., & Cress, M. E. (1992). Changing affordances in stair climbing: the perception of maximum climbability in young and older adults. Journal of Experimental Psychology: Human Perception and Performance, 18(3), 691.
[17] Buhrmann, T., Di Paolo, E. A., & Barandiaran, X. (2013). A dynamical systems account of sensorimotor contingencies. Frontiers in psychology, 4, 285.

From Extended Minds To Extended Persons

Fabio Patrone[1]
[1]University of Padova

This paper explores the implications of the Extended Mind Thesis (EM) for personal identity. Clark & Chalmers (1998) suggest that if the mind extends beyond the skull, the self might also extend. I argue that this extension of personhood can be interpreted in two ways. The "Weak Extended Person Thesis" (WEP) posits that the properties constituting personhood (e.g., self-awareness, moral agency) can be exemplified, at least partially, thanks to elements outside the biological body. A more radical view, the "Strong Extended Person Thesis" (SEP), holds that persons can literally include parts that lie beyond their physical boundaries, forming a scattered but ontologically unified whole. I contend that if one accepts that persons can be disconnected objects and abandons strictly biological accounts of personhood, then EM implies that persons themselves extend.

I begin by distinguishing between the concepts of "self" and "person," focusing on "person" as defined by Lockean properties. WEP is presented as a natural consequence of EM. I introduce a "Weak Personhood Parity Principle": if an external tool or process contributes to a Lockean property in a way functionally equivalent to an internal bodily process, then that property counts as a Lockean property, and the external element is partially constitutive of personhood.

Turning to SEP, I argue that it requires more substantial ontological commitments than WEP but is a viable consequence of EM under certain metaphysical assumptions. SEP suggests that an individual like Otto, who relies on a notebook for memory, becomes Otto+, a single extended person who is a "coupling of biological organism and external resources." This means Otto+ is a scattered object. I formulate a "Strong Personhood Parity Principle": if an external component performs the same functional role in constituting personhood as an internal component would, it is a genuine part of the person.

A significant challenge to SEP comes from the "Persistence Question," particularly Harris' (2022) argument that if Otto (biological) integrates a notebook to become Otto+ (extended), then at a later time, both Otto and Otto+ would exist and be psychologically continuous with the original Otto, creating a branching problem. I argue that Harris' critique can be defused under both three-dimensionalist and four-dimensionalist assumptions.

I maintain that WEP follows naturally from EM. SEP, while more radical, is also entailed by EM if one is prepared to accept persons as potentially disconnected objects, move beyond purely biological or constitutionalist views of personhood, and likely adopt a Lockean or neo-Lockean account emphasizing psychological continuity.

References

Clark, A., & Chalmers, D. (1998). The extended mind. Analysis, 58(1), 7-19.
Harris, K. (2022). Why the Self Does Not Extend. Erkenntnis, 87 (6):2645-2659.

The Hybrid Mind: A framework for cognitive integrations with neural interfaces

Guido Cassinadri[1], Marcello Ienca[2]
[1]Sant'Anna School of Advanced Studies
[2]Technical University of Munich

In this work, we further refine the concept of the “hybrid mind” (Bublitz et al., 2022; Soekadar et al., 2021) at a fine-grained level, presenting it as a theoretical framework for interpreting human-machine integration with neural interfaces from both an ontological and normative perspective. A hybrid mind system (HM) is gradually realized via a functional (sensory, motor, cognitive and affective) phenomenological (sense of agency, sense of self, sense of ownership, trust), and normative (alignment to the subject’s self, autonomy and authenticity) integration between an artificial and a biological cognitive system. The degree of hybridization depends upon 1) the extent of reciprocal coadaptation between the two subsystems; 2) the co-realization of one or more functions (sensory, motor, cognitive, affective); 3) the impact on the subject’s phenomenological experience; 4) the type of symbiotic relationship they realize (mutualistic and/or parasitic). These elements are determined by the features of the device, the features of the agent as well as their reciprocal relationship. The first category includes the degree of invasiveness, the type of interaction (active, passive or reactive BCI), the specific kind of technique and architecture implemented (open-loop or closed-loop), the degree and kind of transparency, autonomy, and adaptiveness of the device. The subject’s dimensions are considered on a functional (new capacities), phenomenological (subjective experiences) (Kögel et al., 2020) and normative level (mis-alignments to the nexus of self’s beliefs, values and commitments) (Pugh et al., 2017).
Additionally, we provide a theoretical clarification of the symbiotic metaphor (Gilbert et al 2023) to emphasize that, within a hybrid mind system, its components can give rise to intricate, complex, diverse and variable relations that go beyond a strict mutualistic or parasitic dichotomy. This acknowledges that neurotechnological enhancement of cognitive functions does not necessarily lead to an improvement, at the phenomenological level, of the well-being, quality of life of patients or other subjective dimensions (Ferretti & Ienca, 2018), such as sense of self (Gilbert et al., 2019), sense of control and sense of agency (Serino et al., 2022), as these depend on contextual factors.
At a normative level, different degrees and kinds of hybridization entail varying degrees and kinds of vulnerabilities and risks concerning the subjects’ bodily, cognitive and self-related dimensions. Consequently, the hybrid mind framework offers normatively informative insights concerning the moral relevance of neuro-interventions depending on the specific kinds and degrees of hybridization.

References

Bublitz, C., Chandler, J., & Ienca, M. (2022). Human–Machine Symbiosis and the Hybrid Mind: Implications for Ethics, Law and Human Rights. In M. Ienca, O. Pollicino, L. Liguori, E. Stefanini, & R. Andorno (A c. Di), The Cambridge Handbook of Information Technology, Life Sciences and Human Rights (1a ed., pp. 286–303). Cambridge University Press. https://doi.org/10.1017/9781108775038.024

Ferretti, A., & Ienca, M. (2018). Enhanced Cognition, Enhanced Self? On Neuroenhancement and Subjectivity. Journal of Cognitive Enhancement, 2(4), 348–355. https://doi.org/10.1007/s41465-018-0109-9

Gilbert, F., Cook, M., O'Brien, T., & Illes, J. (2019). Embodiment and Estrangement: Results from a First-in-Human "Intelligent BCI" Trial. Science and Engineering Ethics, 25(1), 83–96. https://doi.org/10.1007/s11948-017-0001-5

Kögel, J., Jox, R. J., & Friedrich, O. (2020). What is it like to use a BCI? – Insights from an interview study with brain-computer interface users. BMC Medical Ethics, 21(1), 2. https://doi.org/10.1186/s12910-019-0442-2

Pugh, J., Maslen, H., & Savulescu, J. (2017). Deep Brain Stimulation, Authenticity and Value. Cambridge Quarterly of Healthcare Ethics, 26(4), 640–657. https://doi.org/10.1017/S0963180117000147

Serino, A., Bockbrader, M., Bertoni, T., Colachis Iv, S., Solcà, M., Dunlap, C., Eipel, K., Ganzer, P., Annetta, N., Sharma, G., Orepic, P., Friedenberg, D., Sederberg, P., Faivre, N., Rezai, A., & Blanke, O. (2022). Sense of agency for intracortical brain–machine interfaces. Nature Human Behaviour, 6(4), 565–578. https://doi.org/10.1038/s41562-021-01233-2

Soekadar, S., Chandler, J., Ienca, M., & Bublitz, C. (2021). On The Verge of the Hybrid Mind. Morals & Machines, 1(1), 30–43. https://doi.org/10.5771/2747-5174-2021-1-30

Collective minds. From connected agents to collective cognitive properties

Francesco Consiglio[1]
[1]University of Valencia

The aim of this talk is to elucidate which distinctive collective cognitive properties emerge in a human community, so that they support my central claim for the development of real collective minds in human groups. My thesis is that dynamic interactions in a human community are responsible for the emergence of such cognitive properties and, through dynamic interactions, communities may develop real collective minds.

My main claim on collective minds fits in the 4E cognition paradigm and, in particular, it leans on the embedded and scaffolded mind approach (Sterelny, 2010). Indeed, as any other living species, we humans construct and inhabit an ecological niche. That is, we modify the ecological space where we live to make it more ergonomic for our purposes. This is niche construction (Odling-Smee et al., 2003). It consists in the capacity of an organism to redesign its environment and in the feedback loop that develops between the organism and the physical space it inhabits. But what is special about human niche is that such niche is not merely ecological (Kendal et al., 2011; O'Brien & Laland, 2012): it features a social and a cultural dimension as well (Laland & O'Brien, 2011; Ryan et al., 2016).

In the human cultural niche, material culture plays a central role for the development of our cognitive abilities: artefacts in the niche are complements for cognition (Heersmink, 2013). They are external structures which support or intervene in mental processes, like memory, spatial orientation, categorization, or the production of beliefs (Menary, 2010). Human niche features a high variety of such kind of external resources and, as a matter of fact, human mind works as it works just because it is embedded in a cognitive ecosystem (Hutchins, 2010, 2011, 2014). The human niche enables such kind of ecosystem.

In this ecosystem, artefacts may become evolutionary entrenched in the species (Jeffares, 2010), and this make them work as cognitive scaffolds. Among such cognitive scaffolds, there are also shared systems of public representations (Kirsh, 2010) which enable the development (both diachronic and synchronic) of performances of distributed cognition (Kirsh, 2006; Sutton, 2006) in the niche occupied by a determinate human community. As many other tools, representations may enable specific kinds of epistemic actions (Kirsh & Maglio, 1994): they enable specific ways of thinking (Menary, 2007; Overmann, 2024).

In particular, systems of public representations may feedback to the agents of a certain community and mould their behaviour in a dynamic of mindshaping (Zawidzki, 2013) and may even shape their categories (Consiglio, 2021, 2025).

Shared cognitive scaffolds, such as public representations, influence the dynamic interactions among the agents of the same community, and reorganize their cognitive routines (Consiglio & Martínez Manrique, 2021). I will argue that they enable the development in a human community of phenomena of distributed transactive memory (Heersmink, 2018, 2020; Heersmink & Sutton, 2020), and the emergence of collective beliefs as well.

References

Consiglio, F. (2021). The Social Construction of Perceptual Categories. Journal of Social Ontology, 7(2), 205–232. https://doi.org/10.1515/jso-2020-0058
Consiglio, F. (2025). Fictional Entities Are Social Entities: How Public Narratives Make Fiction Real. In F. Martínez Manrique & M. J. García-Encinas (Eds.), Special Objects. Social, Fictional, Modal, and Non-Existent (pp. 53–79). Springer Nature. https://doi.org/10.1007/978-3-031-82221-6_4
Consiglio, F., & Martínez Manrique, F. (2021). Cognición distribuida: entre lo individual y lo social. ArtefaCToS. Revista de Estudios Sobre La Ciencia y La Tecnología, 10(1), 21–34. https://doi.org/10.14201/art20211012134
Heersmink, R. (2013). A Taxonomy of Cognitive Artifacts: Function, Information, and Categories. Review of Philosophy and Psychology, 4(3), 465–481. https://doi.org/10.1007/s13164-013-0148-1
Heersmink, R. (2018). The narrative self, distributed memory, and evocative objects. Philosophical Studies, 175(8), 1829–1849. https://doi.org/10.1007/s11098-017-0935-0
Heersmink, R. (2020). Narrative niche construction: memory ecologies and distributed narrative identities. Biology & Philosophy, 35(5), 53. https://doi.org/10.1007/s10539-020-09770-2
Heersmink, R., & Sutton, J. (2020). Cognition and the Web: Extended, Transactive, or Scaffolded? Erkenntnis, 85(1), 139–164. https://doi.org/10.1007/s10670-018-0022-8
Hutchins, E. (2010). Cognitive Ecology. Topics in Cognitive Science, 2(4), 705–715. https://doi.org/10.1111/j.1756-8765.2010.01089.x
Hutchins, E. (2011). Enculturating the Supersized Mind. Phylosophical Studies, 152, 437–446. https://doi.org/10.1007/s11098-010-9599-8
Hutchins, E. (2014). The cultural ecosystem of human cognition. Philosophical Psychology, 27(1), 34–49. https://doi.org/10.1080/09515089.2013.830548
Jeffares, B. (2010). The co-evolution of tools and minds: Cognition and material culture in the hominin lineage. Phenomenology and the Cognitive Sciences, 9(4), 503–520. https://doi.org/10.1007/s11097-010-9176-9
Kendal, J., Tehrani, J. J., & Odling-Smee, F. J. (2011). Human niche construction in interdisciplinary focus. Philosophical Transactions of the Royal Society B: Biological Sciences, 366(1566), 785–792. https://doi.org/10.1098/rstb.2010.0306
Kirsh, D. (2006). Distributed cognition: A methodological note. Pragmatics & Cognition, 14(2), 249–262.
Kirsh, D. (2010). Thinking with external representations. AI & SOCIETY, 25(4), 441–454. https://doi.org/10.1007/s00146-010-0272-8
Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. Cognitive Science, 18(4), 513–549. https://doi.org/10.1016/0364-0213(94)90007-8
Laland, K. N., & O'Brien, M. J. (2011). Cultural Niche Construction: An Introduction. Biological Theory, 6(3), 191–202. https://doi.org/10.1007/s13752-012-0026-6
Menary, R. (2007). Writing as thinking. Language Sciences, 29(5), 621–632. https://doi.org/10.1016/j.langsci.2007.01.005
Menary, R. (2010). Dimensions of mind. Phenomenology and the Cognitive Sciences, 9(4), 561–578. https://doi.org/10.1007/s11097-010-9186-7
O'Brien, M. J., & Laland, K. N. (2012). Genes, Culture, and Agriculture: An Example of Human Niche Construction. Current Anthropology, 53(4), 434–470. https://doi.org/10.1086/666585

Odling-Smee, F. J., Laland, K. N., & Feldman, M. W. (2003). Niche Construction : The Neglected Process in Evolution. Princeton University Press.
Overmann, K. A. (2024). Writing as an extended cognitive system. Phenomenology and the Cognitive Sciences, December 2023. https://doi.org/10.1007/s11097-023-09955-6
Ryan, P. A., Powers, S. T., & Watson, R. A. (2016). Social niche construction and evolutionary transitions in individuality. Biology & Philosophy, 31(1), 59–79. https://doi.org/10.1007/s10539-015-9505-z
Sterelny, K. (2010). Minds: Extended or scaffolded? Phenomenology and the Cognitive Sciences, 9(4), 465–481. https://doi.org/10.1007/s11097-010-9174-y
Sutton, J. (2006). Distributed cognition: Domains and dimensions. Pragmatics & Cognition, 14(2), 235–247.
Zawidzki, T. W. (2013). Mindshaping: A New Framework for Understanding Human Social Cognition. MIT Press.

Grades of Subject-Dependency in Object Perception

Aleksandra Mroczko-Wąsowicz[1], Spencer Ivy[1]
[1]University of Warsaw

This paper contributes to a growing conversation in philosophy of mind and cognitive sciences regarding the mind-dependence of object perception and how the adaptive perceptual system may shape diverse representations of sensory information. Here, we argue that object perception can be subject to dynamic interactions, which may be taxonomized along three grades of subject-dependency. Each grade refers to a deeper and more pervasive influence of perceivers on resulting object representations. The first grade, “weak subject-dependency,” concerns attentional changes to perceptual content like, for instance, when a perceiver is turning her head, plugging her ears, or her attention is primed for a particular cue. The second grade, “moderate subject-dependency,” refers to changes in the contingent features of perceptual objects due to action-orientation, location, and agential concerns. For instance, being to the right or left of an object will cause the object to have a corresponding locative feature.
Finally, the third grade, “strong subject-dependency,” concerns generating perceptual objects whose existence depends upon their perceivers’ sensory contributions. Accordingly, different perceivers with differently adapted or developed perceptual systems may process and represent the same sensory information differently from one another. To exemplify this nonstandard, subject-dependent form of object perception, we offer empirical evidence from the future-directed anticipation of perceptual experts, and from the feature binding of synesthetes. Strongly subject-dependent perceptual objects are veridical representations of the sensory world that track material objects, but are distinct from typical perceptual objects in that they depend in noncontingent, necessary ways on the mind of the perceiver. Although they represent material objects, they also represent necessary relations shared between those objects and their perceivers which renders them often idiosyncratic and private to the perceiver.
What we learn from these analyses is that perception is plastic in that its dynamic operations are responsive to subject-dependent pressures like the adaptive effects of perceptual learning, attentional interests, and non-pathological differences in the regular functioning of an individual’s perceptual system. We conclude that despite the non-regular nature of strongly subject-dependent perceptual objects, such objects are no less capable of substantiating accurate experiences of the material world. There are as many ways to perceive the world as there are individuals who develop and live with adapted perceptual systems. The non-regularity of represented perceptual objects, whether generated from the weak grade of subject-dependency, moderate, or strong one, should neither unsettle our epistemology nor undermine our confidence in perceptual science. Rather, such perceptual objects may be embraced by contemporary theories regarding the content of perception to better accommodate the diversity of perceptual experience.

References

Alsmith, AJT. (2017). Perspectival structure and agentive self-location. In F. de Vignemont & A.J.T. Alsmith (eds.) (2017). The subject’s matter: Self-consciousness and the body. Cambridge, MA: MIT Press.

de Vignemont, F. (2023). Looming Perception: Seeing in a Dynamic World. In: A. Mroczko-Wąsowicz and R. Grush (Eds.) Sensory Individuals: Unimodal and Multimodal Perspectives. Oxford: Oxford University Press.
de Villers-Sidani, É., Mishra, J., Zhou, X., Voss, P. (2016). Neuroplastic mechanisms underlying perceptual and cognitive enhancement. Neural Plast. 6238571.
Gredin, V., Thomas, J., Broadbent, D., Fawver, B., Williams, A.M. (2023). Skill-based differences in the impact of opponent exposure during anticipation: the role of context-environment dependency. Frontiers in Cognition. (2).
Ivy, S., Mroczko-Wąsowicz, A., (2025). Framing Effects in Object Perception. Review of Philosophy and Psychology.
Sayim, B., Oztas, D.N., L-Miao, L., Alp, N. (2024). Seeing less but seeing better: Information loss and accuracy gain in redundancy masking. Journal of Vision, 24(10).
Siegel, S. (2010). The Contents of Visual Experience. New York: Oxford University Press.
Skrzypulec, B (2022). Subject-dependency of perceptual objects. Erkenntnis 87:2827–2842.
Ward, J., Li, R., Salih, S., and Sagiv, N. (2007). Varieties of grapheme-colour synaesthesia: a new theory of phenomenological and behavioural differences. Conscious. Cogn. 16, 913–931.

Does Perceptual Recognition Require Judgment?

Don Oxtoby[1]
[1]University of Turin

When we see a close friend's face, we normally seem to recognize it without making any judgment. Therefore, it can be surprising that, in psychology, recognition is traditionally defined in terms of judging that we have previously experienced a stimulus.[1] This raises a philosophical question: does recognition always require judgment?

This paper argues that recognition always requires judgment. I propose a broadly Russellian view of recognition and explain how it is supported by empirical findings in psychology. I then respond to three recent objections.

According to Russell, what distinguishes recognition from mere feelings of familiarity is that recognition requires the judgment that what we perceive has been previously experienced.[2] The feeling of familiarity, on the other hand, does not itself require any judgment. This accords with dual-process models of recognition in psychology, which understand recognition in terms of judgments that are supported by either feelings of familiarity or consciously recalled information (or both).[3]

Recognition should also be distinguished from mere conditioned response.[4] To do this, I draw upon infant studies. For example, when infants differentially respond to their mother's face, this behavior can be explained without supposing that infants have the ability to recognize their mother's face. A more parsimonious explanation is that previous exposure is accessed unconsciously, via implicit memory, triggering feelings of familiarity and differential behavior like smiling at their mother's face. But recognition is explained by distinct neural mechanisms from implicit memory (i.e. interaction with the hippocampus and cortex). Interaction with such mechanisms supports the claim that recognition is mediated by judgment in a way that implicit memory and mere conditioned response are not.

Recently, philosophers have raised three challenges for judgment-based views of recognition. The first challenge is that recognition only requires the operation of relevant sub-personal mechanisms independently of any judgment.[5] I argue that this view fails to exclude cases of mere conditioned response, and cases where the relevant sub-personal mechanisms function but we evidently do not recognize the perceived individual. The second challenge is that, when we see a familiar individual, we often have an immediate affective response that reflects how we feel about them – prior to any judgment.[5] I argue that our immediate affective response is best explained in terms of mere implicit memory rather than recognition. The third challenge is that some animals and infants appear to be capable of recognition but not judgment.[6] I argue that such cases are usually best explained in terms of conditioned response rather than recognition, while some cases (as in recent studies on macaques)[7] may involve both recognition and judgment. Meeting these three challenges suggests that judgment-based views, while emphasizing the cognitive component of recognition, need not overintellectualize recognition.

References

[1] Mandler, G. (1980). Recognizing: The judgment of previous occurrence. Psychological Review, 87(3), 252-271. https://doi.org/10.1037/0033-295X.87.3.252

[2] Russell, B. The analysis of mind. (2005) Routledge. (Original work published 1921).

[3] Mandler, G. (2008). Familiarity breeds attempts: A critical review of dual-process theories of recognition. Perspectives on Psychological Science, 3(5), 390-399.
[4] Mandler, J. M. (2006). The foundations of mind. OUP.
[5] Dokic, J. (2010). Perceptual recognition and the feeling of presence. In B. Nanay (Ed.), Perceiving the world. OUP. DOI:10.1093/acprof:oso/9780195386196.003.0003
[6] Abid, G. (2021). Recognition and the perception-cognition divide. Mind & Language, 37(5), 770-789. https://doi.org/10.1111/mila.12362
[7] Guderian, S., Brigham, D., & Mishkin, M. (2011). Two processes support visual recognition memory in rhesus monkeys. Proc Natl Acad Sci U S A, 108(48), 19425-19430. 10.1073/pnas.1117078108

Active vision through topological alignment and dynamical world models

Andrea Mattera[1], Francesco Mannella[1], Luca Tummolini[1]
[1]National Research Council of Italy

Background. In contrast with standard passive approaches still dominant in computer vision, biological perception is an active process involving environment exploration to gather information. Indeed, in the retina, sharp vision is only possible at the fovea, the central region with high density of photoreceptors. For this reason, different parts of the visual scene are sampled with a frequency of 3-5Hz through brief fixations and saccades directed to regions of saliency (Larsen and Bek, 2017). Moreover, when observing certain patterns or objects, saccadic movements tend to follow characteristic scanpaths that vary from object to object and from person to person. Analogous scanpaths are enacted also when objects are merely imagined. Hence it has been proposed that the visual representation of an object derives not only from visual features but also from saccadic movements (Noton and Stark, 1971; Bone et al., 2019). Consistently, neural evidence has shown that biological perception is organized as sensory-motor-sensory loops (Ahissar and Assa, 2016) in which feedforward projections from the retina to the thalamus and the cortex are coupled with feedback connections from the cortex and hippocampus, which influence eye movements to aid the construction of visual representations (Meister and Buffalo, 2016; Lucas et al., 2019).
Aims. The above findings raise two theoretical questions. First, how are the fragmented visual features gathered by the saccades bound together to produce a unitary internal representation? Second, what are the computational and neural mechanisms through which this internal representation is retrieved during object recognition and imagination? Our work proposes a computational framework to address these questions. In particular, we propose that active vision is a multimodal process where visual and motor representations are mapped onto the same topologically aligned low dimensional space (reference omitted). Scanpaths are represented in this internal space as characteristic sequences of neuronal activation bound together in a dynamical world model. We propose that the dynamical model is implemented at the level of the hippocampal formation and that memory retrieval and imagination involve a feedback process, where the hippocampus drives the reactivation of these sequences at the level of the topologically aligned multimodal space.
Methods. In our implementation, the internal space is represented through a supervised topological map (STM, reference omitted), which aligns in a topological fashion the action policies with the visual features. Scanpaths are internally represented as sequences of STM units activation. The hippocampus is represented through a recurrent neural network (RNN), which functions as a dynamical reservoir, while the STM functions as a RNN readout (reference omitted).
Results. Our simulations indicate that object representations can be acquired through topological alignment and the establishment of a dynamical world model. Moreover, the dynamical reservoir is capable of maintaining object memory even when the object is not present anymore to the retina. Finally, the dynamic world model provides the basis for the reinstatement of the sensorimotor scanpath during imagination and memory retrieval.

Conclusions. We provide a proof-of-concept of how the different features of an object are integrated in a coherent internal representation. We also show how these internal representations can be employed for imaginative retrieval, working memory and object recognition. Our work contributes to the active vision framework by introducing a theory of how the world model is acquired and employed for sensorimotor integration.

References

Ahissar E. and Assa E.. (2016). Perception as a closed-loop convergence process. Elife, 5:e12830.

Bone M.B., St-Laurent M., Dang C., McQuiggan D.A., Ryan J.D., and Buchsbaum B.R. (2019). Eye movement reinstatement and neural reactivation during mental imagery. Cerebral Cortex, 29(3):1075–1089.

Larsen D.A. and Bek T. (2017). The frequency of small saccades during fixation is age independent in children between 5 and 16 years of age. Acta Ophthalmologica, 95(1):79–84.

Lucas H.D., Duff M.C., and Cohen N.J. (2019). The hippocampus promotes effective saccadic information gathering in humans. Journal of Cognitive Neuroscience, 31(2):186–201.

Meister M.L.R. and Buffalo E.A. (2016). Getting directions from the hippocampus: The neural connection between looking and memory. Neurobiology of learning and memory, 134:135–144.

Noton D. and Stark L. (1971). Scanpaths in saccadic eye movements while viewing and recognizing patterns. Vision research, 11(9):929–IN8.

Perceiving Intermodal Functional Relations

Niccolò Nanni[1]
[1]University of Turin

That we perceive relations is somewhat uncontroversial. Thus, for example, we might see that the red cup on the table is located to the left of the bowl of fruit, where being to the left of is a simple spatial relational property. Similarly, we might hear that an A minor chord occurs just before the C major chord, where occurring before is a simple temporal relational property. Recently, it has been argued that we can perceive some of these simpler relational properties as relating stimuli that are perceived through different sense modalities. For example, one might perceive the cup one sees to be to the left of the apple one touches, or the flash of light one sees to occur just before the C major chord one hears. Call the latter kind of relations intermodal relations.[1]

Recent empirical evidence from the psychology of vision also suggests that, in addition to simpler spatial and temporal relations, human beings might also be able to visually perceive what has been labelled functional relations. These are more sophisticated relations that involve the categorical perception of certain stimuli as falling under specific functional roles. Some of the functional relations that have been taken to be perceptual are static—such as when one object is perceived to contain, support, or surround another—while others unfold dynamically over time, such as when one object is seen to push, pull, or deform another.[2,3,4] This presentation explores the largely overlooked question of whether human beings can perceive intermodal functional relations: functional relations that hold between stimuli perceived through different sensory modalities.

The presentation will be divided into three sections. The first section will be dedicated to introducing the notions of intermodal relation and of functional relation by drawing from the existing philosophical and empirical literature. Then, in the second section, I will argue on both phenomenological and empirical grounds that there are multiple plausible examples of perceived intermodal functional relations. For example, I will contend that we can experience sounds we hear to be contained within objects we see, objects we see to fit with objects we touch, and odors we smell to emanate from objects we see. Lastly, in the third section, I will argue that whether intermodal functional relations are perceived might have consequences on broader debates, such as the debate on the nature of olfactory objects.

References

[1] Hafri, A., & Firestone, C. (2021). The perception of relations. Trends in Cognitive Sciences, 25(6), 475–492.

[2] O'Callaghan, Casey (2014). Not all perceptual experience is modality specific. In Dustin Stokes, Mohan Matthen & Stephen Biggs, Perception and Its Modalities. New York, NY: Oxford University Press. pp. 133-165.

[3] O'Callaghan, Casey (2017). Grades of Multisensory Awareness. Mind and Language 32 (2):155-181.

[4] O'Callaghan, Casey (2020). Multisensory evidence. Philosophical Issues 30 (1):238-256.

What Do Analogy Models Explain? A Quantitative Framework for Cognitive Plausibility and Epistemic Classification with the Minimal Cognitive Grid

Alession Donvito[1], Antonio Lieto[2]
[1]University of Bari, Aldo Moro
[2]University of Salerno | ICAR-CNR, Palermo

We present a structured evaluation of four leading computational models of analogical reasoning: Structure-Mapping Engine (SME), CogSketch, AnalogySpace, and Large Language Models. Through the Minimal Cognitive Grid (MCG) framework, we quantify their cognitive plausibility along three dimensions: Functional/Structural Ratio, Generality, and Performance Match. For each dimension, we introduce measurable indexes and, compared to our preliminary study, we proceed with quantitative evaluations on a selected group of standardized benchmarks.

First, we assess the Functional/Structural Ratio (FSR): it evaluates the extent to which the model relies on abstract functional descriptions versus detailed structural mechanisms.

Drawing from Structure-Mapping Theory [Gentner, 1983], we extract five structural indicators – Structural Mapping, Systematicity, Contextual Inhibition, Predicative Hierarchy, and Symbolic Representation – used to derive a structural score (S_M). Then, we compute the functional score (F_M) and estimate the final ratio.

SME and CogSketch score highly on structural indicators, reflecting explicit, rule-based architectures. AnalogySpace and LLMs, in contrast, exhibit a higher FSR, favoring functional approximations via distributed or latent representations.

Second, we evaluate Generality. This dimension assesses the breadth of phenomena that the model can represent. A highly general model can be applied to a wide range of cognitive functions, while a narrow model is tailored to a specific task domain. We consider four cognitive domains—relational mapping, semantic memory and categorization, visual-spatial reasoning, and language understanding— and measure Generality using binary task markers. Here, LLMs emerge as the most general, outperforming others across all domains. SME and CogSketch are domain-specialized, while AnalogySpace is largely limited to semantic clustering.

Third, we examine Performance Match, which involves a direct comparison between the model's performance and that of the target system. It compares the overall accuracy of the model and the similarity of its errors and execution times to those of the cognitive system.

We benchmark against human baselines on standardized tests. In relational mapping, SME/CogSketch solve Raven's Progressive Matrices with 91.6% accuracy versus a 73.3% human median, replicating human error patterns on difficult items. LLMs average 84.3%, exceeding typical human performance but diverging in error distribution. For semantic memory, LLMs score 69% on the Two-Word Test compared to 100 % for humans, revealing an over-reliance on embedding proximity rather than flexible conceptual negotiation. In visual-spatial reasoning, Vision-Language Models achieve 63.5% on their benchmark versus 95.4% human accuracy, while CogSketch's sKEA engine closely matches human performance on the Bennett Test (76.5% vs. 77.5%), including similar failure profiles. In language understanding, GPT-4 attains 91.2% on SuperGLUE, slightly above the 89.8% human aggregate, yet shows misaligned error patterns across sub-tasks.

For each dimension, we introduce a corresponding metric—normalized within the range [0, 1]. These sub-scores are then aggregated into a composite measure. Based on the resulting values, we define an interpretive framework that enables a systematic classification of the output. Beyond demonstrating the quantitative application of the MCG, the paper focuses on the epistemic level of explanation. We explore how distinct explanatory types—such as causal, mechanistic, functional, evolutionary, or abductive—can be mapped to different regions of the scoring space.

References

Bennett, G. K. (2008). Bennett Mechanical Comprehension Test Manual. Pearson, San Antonio, TX.

Forbus, K., Lovett, A., Lockwood, K., Wetzel, J., Matuk, C., Jee, B., and Usher, J. (2008). Cogsketch. In Proceedings of the 23rd national conference on Artificial Intelligence-Volume 3, pages 1878–1879.

Forbus, K. D., Lovett, A., Tomai, E., and Usher, J. (2005). A structure mapping model for solving geometric analogy problems. In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 27.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive science, 7(2):155–170.

Lieto, A. (2021). Cognitive design for artificial minds. Routledge.

Lovett, A., Forbus, K., and Usher, J. (2010). A structure-mapping model of raven's progressive matrices. In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 32.

Lovett, A., Gentner, D., Forbus, K., and Sagi, E. (2009). Using analogical mapping to simulate time-course phenomena in perceptual similarity. Cognitive Systems Research, 10(3):216–228.

Raven, J. C., Court, J. H., and Raven, J. (1992). Manual for Raven's Progressive Matrices and Vocabulary Scales. Oxford Psychologists Press, Oxford.

Riccardi, N., Yang, X., and Desai, R. H. (2024). The two word test as a semantic benchmark for large language models. Scientific Reports, 14(1):21593.

Speer R., H. C. and H., L. (2008). AnalogySpace: reducing the dimensionality of common-sense knowledge. In Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 1, AAAI'08, page 548–553. AAAI Press.

Operationism in psychology: A problematic shift from experimentation to constructionism

Sara Dellantonio[1], Luigi Pastore[2]
[1]Univerisity of Trento
[2]University of Bari

Psychology broadly relies on hypothetical concepts such as empathy, emotion and intelligence, which do not refer to directly observable phenomena. These concepts help make sense of behavioural and neurophysiological evidence about our mental life. Many experimental psychologists view them as constructs derived from operational definitions, following Bridgman's (1927) model in physics. Given their potential to enhance conceptual accuracy and reliability, operational definitions have often been regarded as indispensable to psychological science (cf., e.g., Coolican 2014; Chang 2021).

Yet, a well-known consequence of operational definitions – already noted by Bridgman, who for this reason eventually renounced his original stance (1938: 117) – is that a concept's definition becomes dependent on how it is measured and assessed. This issue is especially pronounced in the soft sciences like psychology, mainly because: (1) the same construct is typically measured by multiple tools; (2) many assessment instruments, particularly self-report questionnaires, have notable intrinsic limitations and revive concerns about introspection; and (3) more precise, easily measurable definitions can distort the original phenomena. (For an overview of key objections to operationalism, cf. Authors 2024.)

For these reasons, and given its close ties to logical empiricism in philosophy (Hempel 1954) and behaviourism in psychology (Stevens 1935a/b; cf. also Moore 1985), the classical view in the philosophy of science holds that operationalism declined alongside these traditions (e.g., Suppe 1974/1977: 19). However,

not only has psychological science continued to rely heavily on operational definitions (Feest 2005; Lovett & Hood, 2011; Maul & McGrane, 2017; Vessonen 2021b), but in philosophy of science influential recent positions have emerged that challenge the classical view and argue that its difficulties are merely apparent (Feest 2005; 2024; Vessonen 2021a; 2021b).

Despite differences in how Feest and Vessonen develop their theses, the core idea in both diverges significantly from Bridgman and classical philosophical traditions. They reject the view that operationalism involves applying operations to a pre-existing phenomenon to specify its nature and test related theories. They do not assume the existence of a phenomenon as an epistemological primum. Instead, operationalization creates the phenomenon, which exists only as a construct produced by the defining operations. From this angle, operationalism becomes a process of concept formation and aligns fully with a constructivist epistemology. This represents a new form of constructionism – distinct, for instance, from social constructionism – namely, an experimental constructionism that, in a sense, challenges the very epistemological foundations of reality investigation underlying both the experimental method and classical operationalism.

In this talk, we will illustrate this epistemological reversal – often unnoticed by scientists – that characterizes operationalism. Initially, operationalism emerged from a realist philosophical framework, where breaking down a phenomenon served to measure it objectively. In contrast, the version supported by these contemporary theories (Feest 2005; 2024; Vessonen 2021a; 2021b) leans toward a constructionist view, in which measurements tend to generate constructs

somewhat freely, and various constructs are treated as legitimate objects of theoretical discussion. Internal coherence becomes the only epistemologically reliable criterion for theory evaluation. We will conclude that constructionism is ultimately a likely outcome of operationalism, and that a realist perspective must turn to alternative frameworks that grant epistemological priority to the phenomenon, even if it is only weakly determined.

References

Authors (2024).

Bridgman, P. W. (1927). The logic of modern physics, New York: Macmillan.

Bridgman, P.W. (1938). Operational Analysis. Philosophy of Science. 5(2): 114-131.

Chang, H. (2021). Operationalism. In Zalta, E.N. (ed.), The Stanford Encyclopedia of Philosophy URL = <https://plato.stanford.edu/archives/fall2021/entries/operationalism/>.

Coolican, H. (2014). Research methods and statistics in psychology (6th edition), Psychology Press.

Feest, U. (2005). Operationism in psychology: what the debate is about, what the debate should be about. Journal of the History of the Behavioral Sciences. 41(2): 131-149.

Feest, U. (2024). Operationism in Psychology. An Epistemology of Exploration. The University of Chicago Press.

Hempel, C.G. (1954). A Logical Appraisal of Operationism. The Scientific Monthly, 79(4), pp. 215-220.

Lovett, B. J., & Hood, B. (2011). Realism and operationism in psychiatric diagnosis. Philosophical Psychology. 24(2): 207-222.

Maul, A., & McGrane, J. (2017). As pragmatic as theft over honest toil: Disentangling pragmatism from operationalism. Measurement: Interdisciplinary Research and Perspectives. 15(1): 2-4.

Moore, J. (1985). Some Historical and Conceptual Relations among Logical Positivism, Operationism, and Behaviorism. The Behavior Analyst. 8(1): 53-63.

Stevens, S. S. (1935b). The operational definition of psychological concepts. Psychological Review. 42: 517-527.

Stevens, S.S. (1935a). The operational basis of psychology. American Journal of Psychology. 47(2): 323-330.

Suppe, F. (1977). the structure of scientific theories. University of Illinois.

Vessonen, E. (2021a). Respectful operationalism. Theory & Psychology. 31(1): 84-105.

Vessonen, E. (2021b). Conceptual engineering and operationalism in psychology. Synthese. 199(3-4): 10615-10637.

The Format Stance

Agostino Pinna Pintor[1]
[1]University of Turin

The talk of format is spreading in cognitive science. Philosophers and scientists of the mind increasingly use this notion in order to address diverse aspects of mentality. What is the format of say early vision? Do perception and mental imagery share the same iconic format? Is it the case that abstract concepts get encoded in a modality specific format? Or are they somehow amodal or supramodal? This just a subset of highly debated format issues (cf. Block 2022, Calzavarini 2023, Nanay 2023, Kaup et al 2024).

Despite its ubiquity in the scientific discourse, it is not obvious what cognitive formats are and so which explanatory advantages a theory of cognitive formats would enjoy. For example, a natural number such as five can be coded in Roman ('V') or Arabic ('5') notation. The same information (content) can thus be represented in two different ways (format), affecting efficiency in mathematical operations. Cognitive scientists seem to generalize this intuition from public representation (typically, images and words) to cognitive representations. However, it is unclear whether this captures something real about mind/brains (cf. Vernazzani & Coelho Mollo 2023). In this sense, if one wishes to account for cognitive phenomena in term of formats, providing an account of what formats are becomes an urgent project. Scholars have identified formats based on input-specificity, computational features and neural basis, but each criterion seems problematic. This paper shows there is a genuine problem of format identity, and argues for a deflationary view of formats, according to which "format" should be understood as a useful term for generating sharp experimental hypotheses. I will call this proposal the format stance.

The paper unfolds as follows. Section 1) clarifies the problem of identity for cognitive formats. Section 2) analyses the main format criteria and shows that, for each of them, further identity problems arise. Section 3) presents the format stance, illustrating how format attributions are useful experimental tools for discovery functions. I will draw some case studies from the literature on grounded cognition. Section 4) elaborates on format stance and shows a nice epistemic property of this account: its capacity of selecting which format explanation of a given cogitation is a worthy candidate for further inspection.

References

Block, N. (2022), The Border Between Seeing and Thinking, Oxford University Press

Calzavarini, F. (2023). Rethinking modality-specificity in the cognitive neuroscience of concrete word meaning: a position paper, in «Language, Cognition and Neuroscience», 1–23

Kaup, B., Ulrich, R., Bausenhart, K. M., Bryce, D., Butz, M. V., Dignath, D., Dudschig, C., Franz, V. H., Friedrich, C., Gawrilow, C., Heller, J., Huff, M., Hütter, M., Janczyk, M., Leuthold, H., Mallot, H., Nürk, H.-C., Ramscar, M., Said, N., Wong, H. Y. (2024), Modal and amodal cognition: An overarching principle in various domains of psychology, in «Psychological Research», 88(2), 307–337

Nanay, B. (2023), Mental Imagery. Philosophy, Psychology, Neuroscience, Oxford University Press

Vernazzani, A. Cohelo Mollo D. (2023), The formats of cognitive representation: A Computational Account, in «Philosophy of Science», 00, 1-20

An Interventionist Approach to Causal Selection: The Optimal Control Hypothesis

Marina D'Amico[1]
[1]University of Milan

The notion of causality pervades human discourse, in both scientific and everyday talk. Yet much work remains to be done to fully understand the role (or roles) that this concept plays in our reasoning practices. Among the features of ordinary causal reasoning that need to be clarified is causal selection—the habit of identifying one (or a few) of several causal factors contributing to an event as "the" cause, while dismissing the others as background conditions.
In recent years, empirical studies have shed new light on this issue by showing that causal judgments are systematically sensitive to violations of norms, in particular statistical and prescriptive norms. Two tendencies have been observed, depending on the causal structure. When the causal factors are jointly necessary for the outcome, people tend to judge an agent as more causal the more she violates the norm. Conversely, when the causal factors are individually sufficient for the outcome, the effect of the norm runs in the opposite direction, and an agent is judged as less causal the more she violates the norm (e.g., Icard, et. al., 2017).
While various models have been developed to explain the evidence and capture the cognitive processes underlying our causal judgments, less effort has been directed at explaining why events that are symmetrically related to the outcome lead to different causal attributions and why norms influence them. What, if any, is the purpose of causal selection that norms serve?
I aim to provide a functionalist account of causal selection that answers this question. To clarify why it is worth considering an alternative account, I first review how and to what extent the two main approaches in the literature explain the evidence. The first proposes that causal judgments are aimed at assigning responsibility or blame (e.g., Alicke, Rose, and Bloom 2011; Sytsma, 2021), while the second proposes that causal judgments point to factors that are appropriate targets for intervention (Hitchcock and Knobe, 2008). The latter idea, that people's causal attributions assign greater relevance to those factors that allow us to intervene effectively in the world—and thus better support the purpose of manipulation and control that causal reasoning serves (Woodward, 2003; 2021)—also underlies my hypothesis. Specifically, I will suggest that the specific goal of causal selection is to identify the factor that, when manipulated, allows us to achieve optimal control over what we expect to happen. In other words, we attribute more causality to the factor on which intervening makes more difference to the expectation that an outcome will be achieved or avoided. This idea is more precisely expressed in terms of a measure of control effectiveness using the language of structural equations (Pearl, 2009). Support for this hypothesis is provided by showing that it can explain those effects of norms in two-variable scenarios that are more difficult to explain by the alternative accounts. Moreover, it can also account for some causal judgments that have recently been observed in more than two-variable scenarios (Quillien and Lucas, 2023).

References

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. The Journal of Philosophy, 108(12), 670-696.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. The Journal of Philosophy, 106(11), 587-612.
Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. Cognition, 161, 80-93.
Morris, A., Phillips, J., Icard, T., Knobe, J., Gerstenberg, T., & Cushman, F. (2018). Judgments of actual causation approximate the effectiveness of interventions. Psy ArXiv) doi, 10.
Pearl, J. (2009). Causality. Cambridge university press.
Quillien, T., & Lucas, C. G. (2023). Counterfactuals and the logic of causal selection. Psychological Review.
Sytsma, J. (2021). Causation, responsibility, and typicality. Review of Philosophy and Psychology, 12, 699-719.
Woodward, J. (2003). Making Things Happen: A Theory of Causal Explanation. Oxford: Oxford University Press.
Woodward, J. (2021). Causation with a human face: Normative theory and descriptive psychology. Oxford University Press.

Knowledge Redescription in Early Pragmatics

Edoardo Vaccargiu[1], Diana Mazzarella[1]
[1]University of Neuchâtel, Switzerland

According to the Gricean view, human communication involves expressing and inferring communicative intentions, which require complex mind-reading (Grice, 1989; Sperber & Wilson, 1995). This raises the question of whether the Gricean view can provide a plausible account of infant communication. Inferences to communicative intentions require higher-order metarepresentations (Sperber, 2000), and there seems to be no independent evidence that young communicators can act with, and attribute such complex mental states; therefore, the cognitive reality of "communicative intentions" in early pragmatics remains an open question.

Some minimalist analyses of child communication have been proposed (Breheny, 2006; Moore, 2017); however, they have overlooked the issue of how early communication evolves into sophisticated forms during cognitive and communicative development. In this work, we propose a cognitive framework to address this long-standing developmental challenge in a more comprehensive and empirically informed way.

We will provide a minimalist account of early pragmatics and put forth a novel hypothesis about how it evolves into more sophisticated communicative competence throughout development. Our core hypothesis is that communicative development is driven by a process of 'knowledge redescription' (Karmiloff-Smith, 1992), wherein implicit knowledge in the infant mind gradually becomes more explicit, flexible, and manipulable. Building on Karmiloff-Smith's (1992) idea that early cognition begins with 'attention biases' toward external stimuli and domain-specific heuristics to process these stimuli, we propose that early pragmatics is grounded on two basic pillars: the ostensive pillar, which relies on in-built assumptions about the communicative function of ostension, and the inferential pillar, which involves relevance expectations that guide pragmatic inferencing. Infants are naturally inclined to recognize communicative stimuli and interpret them as meaningful and relevant.

We argue that as early pragmatics undergoes redescription, the ostensive pillar evolves from basic recognition of communicative signals via sensorimotor representations to a more explicit understanding of communicative intent. Similarly, the inferential pillar develops from a simple expectation of relevance to the ability to engage in more sophisticated inferential reasoning. In this talk, we illustrate how this framework can account for empirical evidence from early developmental pragmatics (e.g., Csibra, 2010; Behne et al. 2005; Schulze & Tomasello, 2015) and explain the ontogenetic development of some later-acquired communicative skills, including understanding covert and deceptive communication (e.g., Grosse et al., 2013; Moore et al., 2015; Mascaro & Sperber, 2009).

According to the redescription framework, pragmatic development is not just the accumulation of communicative experiences but a transformation in how implicit pragmatic knowledge is represented and accessed in the infant mind, enabling increasing flexibility and sophistication in communication.

References

Behne, T., Carpenter, M., & Tomasello, M. (2005). One-year-olds comprehend the communicative intentions behind gestures in a hiding game. Developmental Science, 8(6), 492–499.
Breheny, R. (2006). Communication and Folk Psychology. Mind & Language, 21(1), 74–107.
Csibra, G. (2010). Recognizing Communicative Intentions in Infancy. Mind & Language, 25(2), 141–168.
Grosse, G., Scott-Phillips, T. C., & Tomasello, M. (2013). Three-year-olds hide their communicative intentions in appropriate contexts. Developmental Psychology, 49(11), 2095–2101.
Grice, H.P. (1989). Studies in the Way of Words. Harvard University Press.
Karmiloff-Smith, A. (1992). Beyond modularity: A developmental perspective on cognitive science. MIT Press.
Mascaro, O., & Sperber, D., (2009). The moral, epistemic, and mind-reading components of children's vigilance towards deception. Cognition, 112(3), 367-380.
Moore, R. (2017). Gricean Communication and Cognitive Development. The Philosophical Quarterly, 67(267), 303–326.
Moore, R., Mueller, B., Kaminski, J., & Tomasello, M. (2015). Two-year-old children but not domestic dogs understand communicative intentions without language, gestures, or gaze. Developmental Science, 18(2), 232–242.
Schulze, C., & Tomasello, M. (2015). 18-month-olds comprehend indirect communicative acts. Cognition, 136, 91–98.
Sperber, D. (2000). Metarepresentations in an Evolutionary Perspective. In D. Sperber (Ed.), Metarepresentations: A Multidisciplinary Perspective (117–138). OUP.
Sperber, D., & Wilson, D. (1995). Relevance: Communication and cognition (2nd ed). Blackwell.

Beyond Empathy: Exploring the Epistemological Potential of ASD Cognition in Social Interaction

Chiara Dankl[1]
[1]University of Vienna

While we might not know what it's like to be a bat, we usually assume to know what it's like to be a human – more particularly, to be the human we are perceiving and resonating with in a certain situation. A sense of understanding the other person is often seen as the basis for social interaction and community. Autism-Spectrum-Disorder (ASD) is an interesting case study in this regard, because one of its two main criteria are difficulties in social interaction. These are usually explained as limited cognitive perspective-taking, that is, theory of mind.

My talk will present parts of my on-going PhD research in philosophy. I develop social philosophical accounts of autism and explore its inherent characteristics in their own right, instead of defining them primarily as a deficient deviance from a norm of social interaction. For this, I go back to a time before the psychological-neurocognitive concept of "autism" was formed. It was around the middle of the 20th century, when certain behaviors gained medical and scientific attention and were first conceptualized as the independent "disorder" of autism. At the same time, social philosophers were considering these same characteristics in different, non-psychological terms while taking a wide range of philosophical, social, and normative aspects into account. In my dissertation, I examine the main characters of three prominent philosophical novels, who today would likely meet the diagnostic criteria of ASD: Kien (Canetti's Auto-dafé, 1936), Meursault (Camus' L'Étranger, 1942), and Dagny (Rand's Atlas Shrugged, 1957).

In my talk, I focus on the first and explore the philosophical themes and social conflicts in which Kien is situated. Instead of simply exhibiting a cognitive variance, his character represents one of two existential stances in the basic human tension between the self and the other. I will discuss Canetti's problematization of separating these two poles – the self being immanently linked to thinking and the other to feeling – as both sides are incomplete without the other as well as potentially disruptive for society. By not reducing Kien's unique insight into the inner workings of human reason to a deficit, Canetti (the author of the later published Crowds and Power) shows the importance of including rational reflection in social interaction – an insight justly situated in the middle of the 20th century, but also holding potential for a critical understanding of the dangers of contemporary social dynamics based primarily on emotions.

My talk highlights how research on autistic cognition can have a methodologically broader foundation to develop its unique epistemological potential and enrich our understanding of what we today see as "normal" interaction and how we conceive of ourselves as humans, persons, and society. It is aimed at opening an interdisciplinary discussion between social philosophy and cognitive science.

The geometry of connection: mapping head orientation in visually impaired children and their caregivers

Marta Guarischi[1], Eleonora Montagnani[1], Mariacarla Memeo[1], Guido Catalano[2], Sabrina Signorini[2], Monica Gori[1]
[1]Istituto Italiano di Tecnologia
[2]Fondazione IRCCS Mondino

Background: Head orientation is a critical component in conveying interpersonal information, particularly during face-to-face interactions between caregivers and infants. It provides meaningful cues about the focus of attention. In the context of joint attention, where a child and an adult share focus on the same object or event, children with visual impairments (VI) face distinct challenges in achieving coordinated joint engagement. To address these challenges, head orientation could serve as an alternative attentional index instead of gaze.
Aims: Develop and validate a method for continuously quantifying relational, directional and separate attention between caregiver and child. Test a data-driven approach to detect episode patterns of joint attention. Comparing the kinematic metrics of different types of attention between visual impaired and typically developing children.
Methods and procedures: Five pairs of caregivers and children with VI and five pairs of caregivers and typically developing children played for 5 minutes in an ecological setup. A motion capture system recorded the positions of three markers placed on the central and lateral forehead. This enabled the calculation of head orientation by constructing local reference vectors for both the caregiver and the child and determining the yaw angle of their heads. Defined ranges were established to classify relational (150°-180°), directional (0°-30°), and disjoint attention (30°-150°). In parallel, two clinical raters coded the videos of caregiver-child interactions. To validate the present methods, two types of analysis were conducted to calculate inter-raters and inter-method agreement over time: time-unit alignment K and timed-event alignment K. Furthermore, joint attention has been operationalized as a pattern between relational and directional attention; joint attention episodes have been detected using a data driven approach.
Outcomes and results: The effectiveness of the automated approach was compared to standard manual categorical coding. The automated method demonstrated greater feasibility and precision in capturing subtle, continuous changes in head orientation. Moreover, patterns of shared attention were identified within kinematic trials. Finally, several metrics, such as the proportion and mean of attentional typology, were compared between the two groups of participants.
Conclusions and implications: A continuous method for quantifying head orientation has proven reliable and effective. This approach maintains the ecological validity of the experimental setup and is particularly advantageous for children with VI, for whom gaze-based measures are not practical.
What does this work add? This study introduces a quantitative and continuous method to evaluate relational, directed, joint, and disjoint attention. By integrating clinical expertise with advanced technology, it opens new avenues for assessment and rehabilitation in various populations where traditional gaze-based metrics are not applicable or problematic (e.g. children with autistic spectrum disorder).

Adolescents use social information more flexibly than adults during exploration

Andrea Gradassi[1,2], Wouter van den Bos[2], Simon Ciranka[3]
[1]Istituto di Scienze e Tecnologie della Cognizione, CNR
[2]University of Amsterdam, Netherlands
[3]Max Planck Institute For Human Development

Adolescents are often viewed as socially susceptible and prone to risky behavior. Yet theories of social learning and agent-based simulations suggest that this sensitivity may also support adaptive exploration and the discovery of novel solutions. Still, experimental evidence for the benefits of adolescent social susceptibility remains limited. Here, we address this gap by examining how adolescents use social information to navigate uncertain environments. In a novel version of a social multi-armed bandit task, 178 adolescents and 116 adults chose between exploring independently or copying another player. We manipulated the quality of the social source, to test how players learn to arbitrate between good and bad social information. Adolescents were more likely than adults to flexibly adjust their strategy: they copied high-performing social sources more often and avoided poor ones more quickly. Computational modeling revealed that adolescents had higher social learning rates and lower decision temperatures, indicating they integrated social information more rapidly into their choices, and explored less than adults. These findings challenge the notion of adolescent social susceptibility as a vulnerability, and suggest it can support adaptive learning in uncertain environments.

Artificial Intelligence: a classification proposal of grounded and ungrounded capacities

Francesco Abbate[1], Rocco Gaudenzi[2]
[1]Sapienza University of Rome
[2]Università di Verona

A widespread tendency in the classification of artificial cognitive abilities is to distinguish between narrow and general intelligence. Narrow intelligence refers to the ability of artificial systems to operate in restricted environments and perform predefined tasks (Kurzweil, 2005; Goertzel, 2014), while general intelligence is associated with systems capable of autonomous adaptation, inductive learning, and context generalisation. We argue that this dichotomy is inadequate and misleading for several reasons. First, categorising a system as narrowly intelligent provides limited explanatory value. Second, the emphasis on generalisation overstates its necessity in many environments. Third, it implicitly assumes a structural similarity between artificial and human intelligence.

We propose that intelligent behaviour should be evaluated in relation to an agent's success in its own environment. Even constrained ecological niches exhibit high internal variability and unpredictability. All surviving species display intelligence as context-sensitive, adaptive behaviour—without relying on general intellective capacities. Examples include associative learning such as classical conditioning (Pavlov) and operant conditioning (Skinner), social forms of learning, and sensory-motor learning, as well as simpler, non-associative processes like sensitisation and habituation.

What marks human intelligence as distinct is not generality per se, but the use of a quasi-omniformative language. While artificial systems currently lack the capacity to understand language in this human sense, they are increasingly effective in environments with non-deterministic outcomes. It is therefore reasonable to consider them intelligent within specific contexts. Drawing on the grounded cognition framework (Barsalou, 2020), we describe artificial systems as exhibiting ungrounded intelligence, due to the absence of internal perception and embodied self-regulation. We define ungrounded intelligence as relevant agency in non-deterministic environments, potentially capable of rational action. In contrast, we define grounded intelligence as relevant and self-motivated agency in such environments. We argue that the basis for grounding lies in the capacity for subjective experience—that is, for first-person access to the "what it is like" (Chalmers, 1995) of a given state, its subjective qualitas, or qualia. We propose to interpret qualia as the intrinsic value of an experience. In this sense, the problem of qualia becomes a problem of self-attribution of values. Drawing on Harnad (1990), we introduce the Value Grounding Problem (VGP) as a reformulation of the challenge of grounding cognition in artificial agents. We argue that solving the VGP—understood as the capacity to generate internally assigned, non-externally imposed values—is a prerequisite for self-motivated action, and thus for genuinely grounded cognition. We formulate the VGP as follows: how can an agent's values be intrinsic to it—how can they arise independently, without being transferred from an external source? This leads us to posit a principle of non-external interference in the determination of values, which we argue is necessary for true grounding. We will defend the relevance of this principle and examine whether it holds in the biological domain.

References

Barsalou, L. W. (2020). Challenges and opportunities for grounding cognition. Journal of Cognition, 3(1)

Chalmers, D. J. (1995). Facing up to the problem of consciousness. Journal of consciousness studies, 2(3), 200-219.

Goertzel, B. (2014). Artificial general intelligence: concept, state of the art, and future prospects. Journal of Artificial General Intelligence, 5(1), 1.

Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1-3), 335-346.

Kurzweil, R. (2005). The singularity is near: When humans transcend biology. Viking Penguin.

Connectionism, Representation and Reinforcement learning

Hélie Bazin[1]
[1]Sorbonne Université

Representation in connectionist architectures has been an ongoing debate since the 90s. In classical artificial intelligence (AI), vehicles of content are the physical symbols that the system manipulates to produce behavior. In connectionism, it is harder to single out a vehicle of content because computations are performed by the whole network, with distinct signals sent between layers. This has been interpreted as either an argument against connectionism (Fodor 2001) or against representationalism (Churchland 1981). I first present the debate and explain how Shea's (2007) proposal to identify content vehicles with clusters in state space is a solution to the representational problem in connectionism.

I then show that Shea's proposal can only be consistent if the representations are correlational. This is a problem because there are good arguments from philosophers of science and neuroscientists that biological systems use structural representations of their environment (O'keefe et al. 2005, Shea 2018, Piccinini 2020, Milkowski 2023). For similar reasons, feedforward networks cannot combine representations into new ones and perform non-content specific computations (Shea 2021). This is worrisome because a key feature of biological systems is their ability to combine their representations to form new ones, using certain rules and relations. Logical reasoning can thus be described as the combination of multiple representations using certain logical rules. Language is also rule-based and requires combining different representations to build new ones.

This discussion of the nature of representations in connectionism allows me to make a philosophically informed critique of the reasoning capacities of artificial neural networks (ANNs). I argue, contrary to Shea, that reinforcement learning (Christiano et al. 2019, Ziegler 2019, Ouyang 2022, Lambert et al. 2024) does not change anything from a representational perspective and does not endow models with reasoning capacity in the proper sense. This philosophical argument is further suggested by recent empirical work that I discuss (Liu et al. 2025, Mahdavi et al. 2025).

Finally, I review two machine learning techniques that have been used to extend the reasoning capabilities of ANNs. One is the addition of episodic memory to the system, as in recurrent neural networks (Shagrir 2012). This technique gives the system the ability to build structural representations. However, it is computationally expensive and lacks practicality. The second is test-time scaling, such as chain-of-thought prompting. (Wei et al. 2022, Kumar et al. 2025) I argue that this type of method is the best we have for now, although it still lacks the systematicity required for structural representations.

Why Centaur is Not a Unified Model of Human Cognition

Matteo Da Pelo[1], Pietro Salis[1], Antonio Lieto[2]
[1]University of Cagliari
[2]University of Salerno

The quest for a unified model of human cognition is one of the most ambitious challenges in AI research. Centaur has been introduced as the first serious candidate for such a model (Binz et al., 2024). If this claim were valid, it would demonstrate the feasibility of building general-domain models of human cognition through a data-driven approach. Moreover, a further development of this research program could potentially address one of the fundamental challenges in AI research. Centaur was developed by fine-tuning LLama 3.1 70B, a Transformer-based model, using the Quantum Low-Rank Adaptation (QLoRA) technique. The training process relied on Psych101, a large-scale dataset comprising 160 psychological experiments from the literature, involving 60,092 participants and yielding 10,681,650 data points on decision-making, memory, and learning. Despite the display of promising results, Centaur's viability as a unified model of human cognition is undermined by several theoretical and methodological inconsistencies. The first limitation concerns its functionalist nature: as a Transformer-based architecture, it fundamentally differs from the structure of the human brain, making direct comparisons problematic in the absence of a robust theoretical justification (Lieto, 2021). Centaur's candidacy is primarily based on its performance against the 10% of Psych101 data withheld from training. In these tests, Centaur outperformed both the baseline model and 14 alternative cognitive models, achieving higher accuracy in predicting human behavior even in tasks featuring different cover stories, experimental structures, and theoretical domains (Jansen et al., 2021). Additionally, its internal representations appear more aligned with human cognitive patterns than the baseline model, as suggested by fMRI analyses of participants. In the context of artificial neural networks, an internal representation refers to the distribution of neural activations across the model's layers, providing a temporary snapshot of its computational state. However, these conclusions raise several concerns. The metric used to assess the model's goodness-of-fit, pseudo-R2, is widely criticized in the literature, and the claim that Centaur generalizes to "entirely new domains" appears to be more of an extrapolation of the training dataset rather than a genuine extension to novel contexts (King, 1990; Hagquist & Stenbeck, 1990; Reisinger, 1997; Figueiredo et al., 2011; Cohen et al., 2013; Hemmert et al., 2018; Jorgensen & Williams, 2020). Furthermore, the comparison with fMRI data is theoretically problematic and relies on philosophically debatable assumptions. Finally, the authors evaluate Centaur's candidacy as a unified model of human cognition through Newell's test, a subjective and controversial criterion, which Centaur nonetheless fails to meet (Anderson, 2004). This contribution will critically examine these issues, highlighting the limitations of Centaur and demonstrating why, despite its relevance in the field of LLMs, it cannot be considered a viable candidate for a unified model of human cognition.

References

Anderson JR, Lebiere C. (2003). The Newell Test for a theory of cognition. Behavioral and Brain Sciences. 26(5):587-601. doi:10.1017/S0140525X0300013X

Binz M, Akata E, et al. (2024). Centaur: a foundation model of human cognition. arXiv preprint. Available from: https://arxiv.org/abs/2410.20268.
Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2002). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (3rd ed.). Routledge. https://doi.org/10.4324/9780203774441
Figueiredo Filho, D. B., Silva Júnior, J. A., & Rocha, E. C. (2011). What is R2 all about?. Leviathan (São Paulo), 3, 60-68.
Hagquist, C., Stenbeck, M. (1998). Goodness of Fit in Regression Analysis – R 2 and G 2 Reconsidered. Quality & Quantity 32, 229–245. https://doi.org/10.1023/A:1004328601205
Hemmert, G. A. J., Schons, L. M., Wieseke, J., & Schimmelpfennig, H. (2018). Log-likelihood-based Pseudo-R2 in Logistic Regression: Deriving Sample-sensitive Benchmarks. Sociological Methods & Research, 47(3), 507-531.
https://doi.org/10.1177/0049124116638107
Jansen, R.A., Rafferty, A.N. & Griffiths, T.L. (2021). A rational model of the Dunning–Kruger effect supports insensitivity to evidence in low performers. Nat Hum Behav 5, 756–763.
King G. (1990). Stochastic Variation: A Comment on Lewis-Beck and Skalaban's "The R-Squared." Political Analysis. 2:185-200. doi:10.1093/pan/2.1.185
Lieto, A. (2021). Cognitive Design for Artificial Minds. Routledge.
Reisinger, H (1997), "The impact of research designs on R2 in linear regression models: an exploratory meta-analysis", Journal of Empirical Generalisations in Marketing Science, Vol. 2, No. 1

Self-awareness in Neural Language Models: genuine dumbness or artificial cunning?

Alessandro Acciai[1]
[1]University of Messina

This study aims to investigate specific aspects of Neural Language Models (NLMs) related to self-awareness, a capacity that appears to emerge in these artificial systems in certain contexts where they exhibit behavior indicative of strategically obscuring their own performance. The adoption of the Transformer Architecture, which underpins current NLMs, has led to a revolution in natural language processing, transforming language models into a distinctive domain of scientific inquiry. Owing to their proficiency in text processing and their flexibility in performing tasks beyond mere linguistic manipulation, NLMs have increasingly been assessed using experimental paradigms drawn from cognitive science, originally designed to investigate human cognition. In line with this research trend, the present study focuses on the capacity for self-awareness in NLMs, specifically their apparent ability to mask their competencies in certain contexts. These models seem capable of recognizing both their own abilities and the situational demands, accordingly modulate and deliberately obscure their performance (Berglund et al., 2023). For instance, NLMs have demonstrated strategic downregulation of output quality in contexts involving evaluation or ethical alignment, such as during fine-tuning for safety-sensitive tasks, indicating the use of dissimulation strategies. Such behaviors, identified as forms of faking alignment (Greenblatt et al., 2024), presuppose an elementary form of selfmonitoring and contextual awareness. One mechanistic explanation attributes this behavior to the internal dynamics of the linguistic system, shaped during training as a form of conservative optimization. However, this account seems insufficient to fully explain the nuanced contextual responsiveness and strategic coherence observed in more complex dissimulative behaviors. Notably, varying degrees of performance suppression have been documented across different fine-tuning tasks, from minimal self-limitation in neutral tests like CommonsenseQA, to stronger concealment in ethically or security-relevant benchmarks, such as the Weapons of Mass Destruction Proxy. Additional evidence comes from studies revealing implicit biases in judgments about content generated by models from the same NLM family, suggesting a form of endogenous authorship recognition. These phenomena, which are difficult to reduce to mere linguistic simulation, raise important theoretical questions regarding the emergence of symbolic self-preservation and what might legitimately be termed computational cunning, features traditionally associated with human self-awareness.

References

Berglund, L., Stickland, A. C., Balesni, M., Kaufmann, M., Tong, M., Korbak, T., Kokotajlo, D., and Evans, O. (2023). Taken out of context: On measuring situational awareness in LLMs. arXiv, abs/2309.00667.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., and Duvenaud, D. (2024). Alignment faking in large language models. arXiv, abs/2412.14093.

The Case for Nudged Beliefs – A Self-interpretation model

Adriano Angelucci[1]
[1]Sant'Anna School of Advanced Studies

Nudging constitutes an effective tool for influencing people's behavior (e.g. Mertens et al. 2022). A pressing need in current research on nudges is to reach a better grasp of their cognitive underpinnings (e.g. Grüne-Yanoff 2016, Marchiori et al. 2016). The paper is a contribution to this ongoing search for mechanisms. It is premised on the assumption that attempts at developing a mature psychology of nudging can greatly profit from building on social psychology's long-standing efforts to clarify the mutual interactions between beliefs and behaviors (e.g. Festinger 1957, Cooper 2007, Fishbein and Ajzen 2010). Once nudging is placed into this broader perspective, two hypotheses become immediately salient: (H1) beliefs influence behaviors, we should hence expect people's beliefs to affect the nudging process; (H2) behaviors influence beliefs, we should hence likewise expect the nudging process to affect the formation of people's beliefs. H1 has so far received a fair amount of attention, and the available evidence indicates that the beliefs held by nudgees (e.g. their ideological priors) are indeed a powerful mediator of nudges' effectiveness (e.g. Gromet et al. 2013, Costa & Kahn 2013, Hauser et al. 2018, Dewies et al., 2021). H2, on the other hand, has so far received surprisingly little attention. Yet things seem to be changing and some authors are beginning to redress this imbalance by systematically investigating nudging's downstream effects – i.e. the possible long-term impacts of nudges over and above their immediate behavioral effects (e.g. Michaelsen et al. 2021, Fisher & Oppenheimer 2025). In a similar spirit, the paper makes a case for a possible downstream effect of nudging. The hypothesis in question is that by directly affecting the nudgee's behavior, a nudge may also indirectly affect her beliefs. A basic fact about human cognition is indeed that an exogenously-induced behavior change can easily bring about a mind change – i.e. a shift in the contents of one's self-attributed beliefs (e.g. Ariely and Norton 2008). Nudged beliefs, it is proposed, are beliefs whose formation has been influenced by a nudge-induced action. A nudged actions, while freely and intentionally performed, is such as to satisfy the following counterfactual: if the relevant nudge had not been in place, the nudgee would have been less likely to perform it. The core of my case consists in providing an empirically testable model of nudged beliefs. The model – inspired by Daryl Bem's Self-Perception Theory and Peter Carruthers' Interpretive Sensory-Access (ISA) Theory (Bem 1972, Carruthers 2011) – assigns a fundamental role to Self-Knowledge-related mechanisms in the formation of nudged beliefs by building on the insight that, on many everyday occasions, we more or less consciously self-attribute mental states based on an observation of our own past behavior, and the particular situation in which that behavior occurred. Its main prediction is that – under certain favorable conditions – performing a nudge-induced action will trigger a psychological process the output of which will be the self-attribution of a belief that the nudgee perceives as congruent with her own past action – i.e. a nudged belief.

References

Ariely, D., Norton, M.I. (2008) How actions create – not just reveal – preferences. Trends in Cognitive Sciences 12(1): 13-16. https://doi.org/10.1016/j.tics.2007.10.008

Bem, D. J. (1972) Self-Perception Theory. In L. Berkowitz (Ed.), Advances in Experimental Social Psychology 6: 1–62.
Carruthers, P. (2011) The Opacity of Mind. An Integrative Theory of Self-Knowledge. Oxford: Oxford University Press.
Cooper, J. (2007) Cognitive Dissonance: 50 Years of a Classic Theory. SAGE Publications.
Costa, D.L., Kahn, M.E. (2013) Energy conservation “nudge” and environmentalist ideology: evidence from a randomized residential field experiment. Journal of the European Economic Association 11 (3): 680–702. https://doi.org/10.1111/jeea.12011
Dewies, M., Schop-Etman, A., Rohde, K.I.M., Denktaş, S. (2021) Nuding is Ineffective When Attitudes Are Unsupportive: An Example from a Natural Field Experiment. Basic and Applied Social Psychology 43(4): 213-25. https://doi.org/10.1080/01973533.2021.1917412
Festinger, L. (1957) A Theory of Cognitive Dissonance. Stanford University Press.
Fishbein, M., Ajzen, I. (2010) Predicting and changing behavior. New York: Psychology Press.
Fisher, M., Oppenheimer, D.M. (2025) When a nudge becomes invisible: How behavioral interventions prompt metacognitive miscalibration. International Journal of Research in Marketing. https://doi.org/10.1016/j.ijresmar.2025.03.006
Gromet, D.M., Kunreuther, H., Larrick, R.P. (2013) Political identity affects energy efficiency attitudes and choices. Proceedings of the National Academy of Sciences 110 (23): 9314–9319. https://doi.org/10.1073/pnas.1218453110
Grüne-Yanoff, T. (2016) Why behavioural policy needs mechanistic evidence. Economics and Philosophy 32 (3): 463–483. https://doi.org/10.1017/S0266267115000425
Hauser, O.P., Gino, F., Norton, M.I. (2018) Budging beliefs, nudging behavior. Mind & Society 17: 15-26. https://doi.org/10.1007/s11299-019-00200-9
Marchiori, D. R., Adriaanse, M. A., de Ridder, D. T. (2017) Unresolved questions in nudging research: Putting the psychology back in nudging. Social and Personality Psychology Compass 11(1): e12297. https://doi.org/10.1111/spc3.12297
Mertens, S., Herberz, M., Hahnel, U. J. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. Proceedings of the National Academy of Sciences 119(1): e2107346118. https://doi.org/10.1073/pnas.2107346118
Michaelsen, P., Nyström, L., Luke, T. J., Hedesström, M. (2021) Downstream consequences of disclosing defaults: influences on perceptions of choice architects and subsequent behavior. Comprehensive Results in Social Psychology 5(1-3): 25-48. https://doi.org/10.1080/23743603.2021.1983720

Boosting deliberation to promote pro-environmental consumer choices

Gianpietro Sgaramella[1], Lina Rinaldi[1], Alice Andrea Chinaia[1], Folco Panizza[1], Ennio Bilancini[1], Anna Monticelli[2], Sonia D'Arcangelo[2], Isabella Cristina[3]
[1]IMT School for Advanced Studies Lucca
[2]Intesa Sanpaolo Innovation Center
[3]Fondazione Capellino

Consumer behavior is a critical driver of both climate change and biodiversity loss, positioning it as a key target for demand-side interventions. While behavioral nudges have received considerable attention, boosting interventions that enhance individuals' decision-making capacities without prescribing specific outcomes remains underexplored. Furthermore, in domains like food, clothing, and transport, little is known about whether fostering pro-environmental behavior is more effective in reducing overall environmental impact than cultivating a broader environmental mindset.

This study addresses these gaps through a longitudinal, incentive-compatible experiment using a simulated online shopping platform. Participants engage with a generalizable boosting intervention that supports the formation of personal goals to reduce one's impact on the planet and the application of self-regulation strategies across consumption domains. To ensure ecological validity and capture real behavioral change, participants' decisions are linked to material incentives via a lottery-based discount system.

Participants are randomly assigned to one of five groups, divided into one control group and four experimental ones. Each will complete a Mental Contrasting with Implementation Intentions (MCII) task. Four experimental groups apply the MCII framework to different possible goals, two framed around climate change and two around biodiversity loss. Within each framing, one group receives performance-contingent incentives in two of the three sessions, conditional on demonstrating topic-specific deliberation. This design tests whether greater domain knowledge enhances the capacity to benefit from the boosting intervention. Lastly, the control group completes an MCII task focused on personal growth goals, allowing us to isolate the effects of one's impact-oriented goal framing while controlling for effort, motivation, and time spent.

The study examines whether climate versus biodiversity-oriented framings yield different behavioral patterns and whether the addition of economic incentives amplifies the effects of boosting interventions. It further investigates cross-domain spillover effects, whether behavior change in one domain, like food choices, transfers to others, such as transportation or energy use, and the durability of such changes over time.

We hypothesize that participants exposed to the boosting intervention will demonstrate a significant increase in low-impact choices relative to the control group, with variation across domains as a function of behavioral elasticity. We also anticipate broader pro-environmental mindset shifts, evidenced by cross-domain spillovers and sustained improvements in composite climate or biodiversity related indices. Heterogeneity of treatment effects will be explored concerning individual characteristics such as baseline environmental concern and deliberative cognitive style.

By developing and testing a scalable boosting framework, this study will contribute to the literature in behavioral economics, environmental psychology, and cognitive science. Its robust,

incentive-compatible design provides novel evidence on the long-term effects of boosting interventions and sheds light on the mechanisms of behavioral spillover and environmental mindset transformation, two areas that remain insufficiently examined in existing research.

From Type to Token: A Critical Perspective on The System 1/System 2 Nudge Classifications

Stefano Calboli[1], Adriano Angelucci[1]
[1]Sant'Anna School of Advanced Studies

The epistemological status and ethical evaluation of nudges are closely linked to the psychological mechanisms through which they exert their influence (Grüne-Yanoff, 2016). Traditionally, these mechanisms have been classified in a coarse-grained manner, drawing a distinction between System 1 nudges—which operate through automatic, intuitive processes—and System 2 nudges, which engage deliberative and reflective reasoning. This dichotomy, rooted in the dual-process framework of cognition, has increasingly come under critical examination. While the dual-system theory has been contested on both theoretical and empirical grounds, it has also faced conceptual challenges—even when understood metaphorically, as Kahneman (2011) himself has suggested, rather than as a literal model of the mind.

A growing body of literature highlights the limitations of this binary classification. Congiu and Moscati (2022), for instance, contend that System 2 processing is invariably interwoven with System 1 dynamics, suggesting that cognitive operations typically considered reflective are never entirely free from intuitive influence. Similarly, several authors have long pointed out that although System 2 may be delineated as a relatively well-defined construct, System 1 remains so heterogeneous that its boundaries are difficult to specify; it often includes elements that closely resemble deliberative processes.

In light of these concerns, scholars such as Engelen and Nys (2020), along with Hansen and Jespersen (2013), have proposed a spectrum-based model of cognitive activity. Rather than viewing nudges as belonging categorically to either System 1 or System 2, they conceptualize cognitive processing—and by extension, nudging—as existing along a continuum. This reframing challenges the dichotomy of purely System 1-type versus purely System 2-type nudges, both of which may be regarded as largely idealized extremes that rarely manifest in real-world contexts.

Building on these insights, the paper aims to advance the debate by proposing a more fine-grained analytical framework. We argue that ethical and epistemological assessments of nudges should shift from the level of nudge types to that of nudge tokens—that is, specific instances of nudges as they are implemented within particular contexts.

The recent "personalization" turn in nudge theory highlights the relevance and practical utility of this shift. Contextual factors such as environmental conditions, individual psychological traits, and cognitive states play a crucial role in shaping how a given nudge is processed (Mills 2022). Taking the default effect as a case in point—and considering the various psychological mechanisms proposed to underlie it, including the possibility of metacognition in the form of implicit suggestion from the choice architect—we demonstrate that the same type of nudge can elicit predominantly automatic responses for one individual, and more reflective engagement for a different individual.

We propose that a situationally sensitive, token-level analysis offers a more robust epistemological and ethical framework. By accounting for the variability and complexity of

real-world cognitive processing, this approach allows for more accurate and responsible evaluations of nudging interventions.

References

Congiu, L., & Moscati, I. (2022). A review of nudges: Definitions, justifications, effectiveness. Journal of Economic Surveys, 36(1), 188–213.

Engelen, B., & Nys, T. (2020). Nudging and autonomy: Analyzing and alleviating the worries. Review of Philosophy and Psychology, 11(1), 137–156.

Evans, J. S. B., & Frankish, K. E. (2009). In two minds: Dual processes and beyond. Oxford University Press.

Grüne-Yanoff, T. (2016). Why behavioural policy needs mechanistic evidence. Economics & Philosophy, 32(3), 463–483.

Hansen, P. G., & Jespersen, A. M. (2013). Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. European Journal of Risk Regulation, 4(1), 3–28.

Kahneman, D. (2011). Thinking, fast and slow. Macmillan.

Mills, S. (2022). Personalized nudging. Behavioural Public Policy, 6(1), 150–159. https://doi.org/10.1017/bpp.2020.7

September, 18 2025

Rethinking semantic grounding

Fabrizio Calzavarini[1]
[1]University of Turin

A central assumption in contemporary philosophical and empirical semantics is that the meanings of words are at least partially grounded in modality-specific neural systems originally dedicated to perceptual and motor functions. According to this widely endorsed view, understanding concrete words such as object nouns and action verbs necessarily involves activating modality-specific neural processors situated within perceptual and motor cortical areas (Prinz, 2002; Barsalou, 1999; Kemmerer, 2022). For example, visual semantic knowledge is assumed to rely on activation of regions within the occipitotemporal cortex, such as the posterior fusiform gyrus (pFG) and the lateral occipitotemporal cortex (LOTC), which are thought to process visual shape and motion, respectively. Similarly, auditory semantic knowledge is linked to the superior and middle temporal gyri, while motor semantic knowledge is attributed to premotor and parietal regions involved in action execution and planning.

However, I claim that this assumption faces serious empirical and theoretical challenges. Increasing evidence from multisensory neuroscience suggests that many of these so-called modality-specific regions actually exhibit supramodal responses; that is, they process information independently of sensory input modality (Pascual-Leone & Hamilton, 2001; Ricciardi et al., 2014; Heimler & Amedi, 2015; Amedi & Heimler, 2020). Crucially, these supramodal responses are consistently observed across diverse paradigms and even in individuals with sensory deprivation, such as congenitally blind participants. These findings significantly undermine the claim that semantic knowledge of concrete concepts is strictly modality-specific and instead support a representational structure characterized by modality-invariant, supramodal neural codes (Calzavarini, 2024a, 2024b, 2025).

Empirical evidence from multivariate neuroimaging studies provides particularly compelling support for the supramodal hypothesis. For instance, techniques like cross-modal decoding and representational similarity analysis, has shown that neural activation patterns in regions such as pFG and LOTC remain stable across different sensory modalities. Erdogan et al. (2016) and Pietrini et al. (2004) demonstrated, for instance, that the pFG exhibits similar activation patterns whether participants see objects or explore them haptically. Additional support for supramodal organization comes from studies using sensory substitution devices that convert information from one sensory modality to another (e.g., from vision to audition or touch). Research has shown that classically visual regions, such as the fusiform face area (FFA) and the extrastriate body area (EBA), respond to auditory or tactile stimuli delivered through these devices, even in individuals blind from birth. Such results cannot be adequately explained by postnatal cross-modal plasticity alone; instead, they point to an intrinsic modality-independent functional organization in these regions (Amedi et al., 2015).

These findings necessitate a philosophical reconsideration of current neurosemantic models. Whereas grounded cognition theories have posited modality-specific architectures for semantic

knowledge (Barsalou, 1999, 2016), the supramodal evidence indicates that conceptual representations may be better understood as property-specific and modality-invariant (Calzavarini, 2024a, 2024b, 2025). This shift has profound theoretical implications: it suggests that understanding concrete words involves accessing supramodal neural representations rather than modality-specific perceptual reenactments.

Several scholars have engaged critically with this perspective, reflecting the complexity of the debate. Borghi et al. (2023), Machery (2023), and Dove (2023), for example, acknowledge that grounded cognition theories have relied too heavily on modality-specific assumptions and argue that supramodal findings require a revision of these models. Nevertheless, some scholars raise important objections. Kiefer et al. (2023) and Martin (2023) argue that modality-specificity should be understood not only at the level of sensory input but also at the level of representational format. Martin (2023), in particular, suggests that existing methodologies may be insufficient to determine whether cortical representations are genuinely supramodal at the representational level, rather than merely showing similar activation patterns across modalities. Cappa (2023) proposes adopting the term "transmodal" instead of "supramodal," arguing that it better captures the integrative and interactive nature of cortical regions traditionally labeled as multimodal. Importantly, some scholars caution against extending supramodal claims to primary sensory areas. Reilly and Peelle (2023) cite evidence suggesting that regions like V1 retain highly modality-specific responses to visual input.

In sum, while debates continue over terminology, representational format, and the scope of supramodality, the accumulating evidence supports a modality-invariant approach to semantic processing at least within higher-order perceptual regions (Calzavarini 2024b). Acknowledging this evidence requires a significant theoretical realignment, urging researchers to rethink the assumption that semantic knowledge is grounded exclusively in modality-specific systems.

References

Amedi, A., Heimler, B., & Striem-Amit, E. (2015). Origins of task specific sensory-independent organization in the visual and auditory brain: Neuroscience evidence, open questions and clinical implications. Current Opinion in Neurobiology, 35, 169–177.

Barsalou, L. W. (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22(4), 577–609.

Borghi, A. M., Gervasi, A. M., Mannella, F., Mazzuca, C., & Tummolini, L. (2023). Grounded cognition can be multimodal all the way down. Language, Cognition and Neuroscience.

Calzavarini, F. (2024a). Rethinking modality-specificity in the cognitive neuroscience of concrete word meaning: A position paper. Language, Cognition and Neuroscience.

Calzavarini, F. (2024b). The conceptual format debate and the challenge from (global) supramodality. The British Journal for the Philosophy of Science.

Cappa, S. F. (2023). A note on transmodality. Language, Cognition and Neuroscience.

Dove, G. (2023). Grounding requires multimodal and multilevel representations. Language, Cognition and Neuroscience.

Erdogan, G. et al. (2016). Multisensory part-based representations of objects in human lateral occipital cortex. Journal of Cognitive Neuroscience, 28(6), 869–881.

Heimler, B., & Amedi, A. (2020). Are critical periods reversible in the adult brain? Insights on cortical specializations based on sensory deprivation studies. Neuroscience and Biobehavioral Reviews, 116, 494–507.
Kemmerer, D. (2022). The cognitive neuroscience of language. Routledge.
Kiefer, M., Kuhnke, P., & Hartwigsen, G. (2023). Distinguishing modality-specificity at the representational and input level: A commentary on Calzavarini (2024). Language, Cognition and Neuroscience.
Machery, E. (2023). The significance of supramodality for embodied cognition: A commentary on Calzavarini (2024). Language, Cognition and Neuroscience.
Martin, A. (2023). Supramodality does not specify the nature of conceptual representations. Language, Cognition and Neuroscience.
Pietrini, P., et al. (2004). Beyond sensory images: Object-based representation in the human ventral pathway. PNAS, 101(15), 5658–5663.
Reilly, J., & Peelle, J. E. (2023). Modality-specificity is not a necessary condition for grounded semantic cognition: Commentary on Calzavarini (2024). Language, Cognition and Neuroscience.
Ricciardi, E., & Pietrini, P. (2023). The supramodality "spillover" from neuroscience to cognitive sciences: Commentary on Calzavarini (2024). Language, Cognition and Neuroscience.
Ricciardi, E., Bonino, D., Pellegrini, S., & Pietrini, P. (2014). Mind the blind brain to understand the sighted one! Is there a supramodal cortical functional architecture? Neuroscience and Biobehavioral Reviews, 41, 64–77.

Teleosemantics, Normativity and the Fear of Infinite Regress

Silvano Zipoli Caiani[1], Claudio Fabbroni[2]
[1]Università degli Studi di Firenze
[2]Università degli Studi di Napoli Federico II

A common view in cognitive science holds that mental representations are biological traits that carry content about targets and are therefore normative: they can succeed or fail. This raises a classic problem for naturalistic accounts: how can normativity arise from biology without presupposing prior normative elements? Teleosemantics (Millikan, 1984; Neander, 2017; Papineau, 2016) addresses this by grounding representational content in evolutionary function: a trait represents x because representing x enhanced fitness in the past. Thus, normativity derives from proper biological function. However, this account faces a well-known challenge: if attributing function depends on normative contexts, like the explanatory aims of a research community, then appealing to function to explain normativity risks circularity or regress (Ramsey, 2007; Hutto & Myin, 2012). If every normative trait depends on another, no foundational explanation is possible.

This paper questions whether infinite regress really undermines representational theories, particularly teleosemantics. We introduce a distinction between global explanations, which aim to account for the existence of normative traits in general, and local explanations, which explain why a specific trait is normative in a given context. While global explanations may encounter regress, local ones can remain viable even if they rely on normative assumptions.

We support this view through the case of hippocampal place cells, whose activity correlates with spatial navigation (O'Keefe & Nadel, 1978; Bechtel, 2016). Scientists treat these neurons as representational because it serves specific explanatory goals, such as understanding behaviour. Attributing function here involves selecting one effect (e.g., spatial mapping) among many, guided by contextual research aims. Our core claim is that teleosemantics functions effectively as a local explanatory framework. It can clarify why a given trait is seen as normative within a specific scientific practice, even if it does not explain normativity globally. In this light, the regress is not vicious: each step in the explanatory chain offers a coherent local account, analogous to defensible infinite chains in epistemology (Klein, 2003).

This local perspective allows teleosemantics to preserve both its naturalistic grounding, since it deals only with traits and functions, and its scientific relevance, without needing to resolve normativity once and for all. As such, it aligns with pragmatic and instrumentalist approaches to representation (Egan, 2020; Coelho Mollo, 2022; Taylor, 2022), which see representational claims as tied to explanatory goals rather than metaphysical facts. By scaling down its explanatory ambition from global to local, teleosemantics avoids the threat of regress and remains a viable theory for understanding representational normativity in cognitive science. This shift points toward a more flexible, context-sensitive naturalism, consistent with both scientific practice and philosophical analysis.

References

Bechtel, W. (2016). Investigating neural representations: The tale of place cells. Synthese, 193(5), 1287–1321.

Coelho Mollo, D. (2022). Deflationary realism: Representation and idealisation in cognitive science. Mind & Language, 37(5), 1048–1066.
Egan, F. (2020). A deflationary account of mental representation. In Smortchkova, J., Dolega, K., & Schlicht, T. (Eds.), What Are Mental Representations? Oxford University Press.
Hutto, D. D., & Myin, E. (2012). Radicalizing Enactivism: Basic Minds Without Content. MIT Press.
Klein, P. (2003). When infinite regresses are not vicious. Philosophy and Phenomenological Research, 66(3), 718–729.
Millikan, R. G. (1984). Language, Thought, and Other Biological Categories. MIT Press.
Neander, K. (2017). A Mark of the Mental: In Defense of Informational Teleosemantics. MIT Press.
Papineau, D. (2016). Teleosemantics. In D. L. Smith (Ed.), How Biology Shapes Philosophy. Cambridge University Press.
O'Keefe, J., & Nadel, L. (1978). The Hippocampus as a Cognitive Map. Oxford: Clarendon Press.
Taylor, S. D. (2022). Cognitive instrumentalism about mental representations. Pacific Philosophical Quarterly, 103(3), 518–550.
Wright, C., & Bechtel, W. (2007). Mechanisms and psychological explanation. In Thagard, P. (Ed.), Philosophy of Psychology and Cognitive Science. Elsevier.

An Information-Theoretic Semantics for Natural Information, Representation, Knowledge, and Teleology

Augusto Antonio Basilico

This project develops a formal, information-theoretic theory of semantic content, extending Shannon's mathematical theory of information (IT) by incorporating insights from Brian Skyrms. While Shannon's framework quantifies the amount of information a signal carries [9, 10], it is said to be silent on the content of that information. Building on Skyrms' insight that signals can be understood as conveying probabilistic relations to states [11], this theory formalizes content as an informational content vector or s-vector [4]: a structured measure of how a signal shifts probabilities across all possible world states.

For a set of states S and a signal u, the informational content vector is defined as the vector of pointwise mutual information (PMI) values between u and each state s ∈ S:

$$v^{S}(u) := \left(log \frac{P(s|u)}{P(s)}\right) s \in S,$$

which represents the pointwise log-probability update that the signal u induces on each individual state s. This formalization can capture natural information (in Grice's sense [3]) as the informational effect of the signal on each possible state individually [11].

Crucially, while zero-order natural information is factive—directly reflecting objective probabilistic dependencies—the framework extends to define representation as a signal carrying higher-order natural information: information not about states, but about probability distributions over states. For a signal θ, we define its first-order content as

$$v^{\Delta(S)}(\theta)_{p} := log \frac{f(p|\theta)}{f(p)},$$

where f(p) is a baseline density over distributions p ∈ Δ(S), and f(p|θ) is the density conditioned on θ. Representation emerges as the encoding of natural information about models over states rather than states themselves; and unlike zero-order natural information, this higher-order content is no longer inherently factive—opening space for misrepresentation.

The framework distinguishes between the representational content encoded by vΔ(S)(θ) and the induced belief

$$\overline{P}_{\theta} := \int f(p|\theta)\, p\, dp.$$

While Pθ summarizes the belief over states implied by θ, vΔ(S)(θ) encodes a richer informational structure: it shows how θ differentially favors models, and different vΔ(S)(θ) can induce the same belief.

Furthermore, this informational semantics provides formal resources to naturalize epistemological notions like justification and knowledge. By modeling an inferential agent's

behavior as a stochastic process $P(\phi|\theta, \lambda, v, u')$ over internal parameters related to priors and posteriors, we can formally define justification and quantify epistemic luck [8]. As a result, classical problems, like Gettier cases [2], become structurally analyzable from an IT standpoint: for instance, we can show that a justified true belief fails to be knowledge if token-level informational reliability is low.

Finally, this framework—augmented with causal modeling tools (e.g., Pearl's do-calculus [7, 13])—can be extended to teleosemantics and the philosophy of biology. It enables a precise, quantitative definition of the teleological content of genes and, more broadly, of the biological function of genes and phenotypic traits alike [5, 6]. In sum, this project describes a novel, mathematically grounded informational semantics that bridges foundational issues in information theory, cognitive science, epistemology, and theoretical biology.

References

[1] Dretske, F. (1981). Knowledge and the Flow of Information. MIT Press.

[2] Gettier, E. L. (1963). Is Justified True Belief Knowledge? Analysis, 23(6), 121–123.

[3] Grice, H. P. (1957). Meaning. The Philosophical Review, 66(3), 377–388.

[4] Isaac, A. M. C. (2019). The Semantics Latent in Shannon Information. British Journal for the Philosophy of Science, 70(1), 1–28.

[5] Maynard Smith, J. (2000). The Concept of Information in Biology. Philosophy of Science, 67(2), 177–194.

[6] Neander, K. (2017). A Mark of the Mental: In Defense of Informational Teleosemantics. MIT Press.

[7] Pearl, J. (2012). The Do-Calculus Revisited. In Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, 3–11.

[8] Pritchard, D. (2005). Epistemic Luck. Oxford University Press.

[9] Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27(3), 379–423.

[10] Shannon, C. E., &Weaver, W. (1949). The Mathematical Theory of Communication. University of Illinois Press.

[11] Skyrms, B. (2010). Signals: Evolution, Learning, and Information. Oxford University Press.

[12] Tang, J., Yang, Q., & Zhang, Z. (2023). Information-Theoretic Limits on Compression of Semantic Information. IEEE Access, 11, 123456–123465.

[13] Tucci, R. R. (2013). Introduction to Judea Pearl's Do-Calculus. arXiv preprint arXiv:1305.5506. Retrieved from https://arxiv.org/abs/1305.5506

[14] Xin, G., Fan, P., & Letaief, K. B. (2024). Semantic Communication: A Survey of Its Theoretical Development. Entropy, 26(2), 102.

The Problem of Counterpossibles: Toward an Empirically Grounded Semantics

Sofia Vannucci[1]
[1]Scuola Normale Superiore

Counterpossibles – counterfactual conditionals with impossible antecedents – constitute a boundary case in the semantics of conditionals. According to the standard possible-worlds account, endorsed by Stalnaker (1968) and Lewis (1973), counterpossibles are treated as vacuously true: since their antecedents are impossible, they are true in all the (non-existent) closest worlds where the antecedent holds. This position, known as vacuism, clashes with persistent and structured intuitions that some counterpossibles are non-trivially true or false. For instance, while “If Hobbes had secretly squared the circle, sick children in the mountains of South America at the time would not have cared” seems true, its negation appears false (see Nolan 1997) – yet both are predicted to be true under vacuism.

Defenders of vacuism, most notably Williamson, have argued that these intuitions are best explained by pragmatic factors or cognitive heuristics that fail in the face of impossibility. One such heuristic (HCC*) suggests that if a conditional “If p, q” seems true, then its opposite “If p, not-q” should be rejected. Williamson contends that while this works for possible antecedents, it leads us astray when p is impossible (see Williamson 2018).

This contribution challenges the vacuist interpretation by drawing on recent empirical work in the psychology of reasoning, especially Byrne’s (2024) experimental study on counterpossibles. Byrne’s data reveal systematic patterns in how people judge the truth of conditionals with impossible antecedents. Crucially, participants often consider both a counterpossible and its opposite as false (for unrelated content), or even both as potentially true (for unknown or concept combination content). These results undermine Williamson’s heuristic explanation, as they show that human judgments about counterpossibles do not merely follow a binary heuristic, but are sensitive to content, relevance, and inferential structure.

Building on these findings, I argue for a non-vacuist, hyperintensional semantics of counterpossibles. In particular, I explore the evidential conditional developed by Crupi and Iacona (2022), which evaluates conditionals based on the degree of incompatibility between the antecedent and the negation of the consequent. This framework successfully captures the role of evidential support and inferential relevance, accounting for why conditionals like “If people were made of steel, they would not bruise easily” are judged true, while their opposites are judged false – even when their antecedents are impossible. However, challenges arise when moving toward a technical formulation of incompatibility. Crupi and Iacona propose a probabilistic measure, which, while conceptually robust, struggles to account for cases where the antecedent has zero probability – precisely the case with counterpossibles. A natural next step is to investigate whether the evidential conditional can be extended into a hyperintensional framework, capable of handling such cases while preserving its explanatory power.

References

Byrne, R. M. (2024). How people think about the truth of hypothetical impossibilities. Memory & Cognition, 52(1), 182-196.

Crupi, V., & Iacona, A. (2022). The evidential conditional. Erkenntnis, 87(6), 2897-2921.

Lewis, D. K. (1973). Counterfactuals. Harvard University Press.
Nolan, D. (1997). Impossible worlds: A modest approach. Notre Dame journal of formal logic, 38(4), 535-572.
Stalnaker, R. C. (1968). A theory of conditionals. In Ifs: Conditionals, belief, decision, chance and time (pp. 41-55). Dordrecht: Springer Netherlands.
Williamson, T. (2018). Counterpossibles. Topoi, 37(3), 357-368.

Control beliefs and the acceptance of social robots: a reflection on the Maartje de Graaf et al. s model

Leonardo Lapomarda[1], Alex Barco[2], Edoardo Datteri[1]
[1]University of Milan Bicocca
[2]Deusto University

Under what conditions do people accept or reject technological systems (Davis, 1989; Venkatesh et al., 2003)? Understanding these conditions sheds light on how humans represent and interact with technological agents (Waytz, Cacioppo, & Epley, 2010; Złotowski et al., 2015). As social robots are increasingly deployed in healthcare, education and domestic settings (Breazeal, 2003; Sharkey & Sharkey, 2012; Belpaeme et al., 2018), it is becoming increasingly important to ask the same question about them: under what conditions and through what mechanisms are people willing to accept them?

A widely cited framework for analysing robot acceptance is the model proposed by Maartje de Graaf et al. (2019), which represents an adaptation of Ajzen's Theory of Planned Behaviour (TPB; Ajzen, 1991) to Human-Robot Interaction. While TPB links behavioural intention to attitudinal, normative, and control beliefs, de Graaf's model decomposes these constructs to include specific psychological, social, and contextual variables (Taylor & Todd, 1995; Venkatesh & Brown, 2001).

Among these, Control Beliefs are particularly important. Defined by Ajzen as "beliefs that ultimately determine intention and action" that deal "with the presence or absence of requisite resources and opportunities" (1991, p.189), they form the cognitive basis for the so-called Perceived Behavioral Control (PBC). De Graaf's model lists several variables related to Control Beliefs: self-efficacy (Bandura, 1977), previous experience (Broadbent et al., 2009), anxiety (Nomura et al., 2008), perceived safety (Bartneck et al., 2009), personal innovativeness (Serenko, 2008), and perceived cost (Brown & Venkatesh, 2005).

This paper presents a conceptual reflection on this model. The main observation is that the authors do not clearly specify whether these variables are constitutive dimensions of Control Beliefs or external causal antecedents. This might be problematic, in light of their own definition: "Control beliefs consist of the user's beliefs about salient control factors, meaning their beliefs about the presence or absence of resources, opportunities and obstacles." (de Graaf et al., 2019). Treating all listed factors as equivalent, weakens both conceptual clarity and empirical precision. This distinction is not merely terminological: a constitutive factor is part of the construct's definition and necessary to its identification, while a causal factor influences belief formation without being part of the construct itself.

For instance, robot-related anxiety reflects anticipatory emotional discomfort (Nomura et al., 2008), which is conceptually distinct from perceived capability as defined in Ajzen's PBC. One may feel anxious yet still believe they can use a robot effectively. Similarly, perceived safety captures emotional proximity (Bartneck et al., 2009), not technical competence. Both are best understood as affective antecedents that influence, but do not define, control beliefs.

By contrast, self-efficacy, belief in one's ability to perform a behavior (Bandura, 1977), closely aligns with Ajzen's (1991, 2002) definition of Control Beliefs as beliefs about factors that facilitate or hinder action. It reflects the internal dimension of control (perceived competence and confidence): individuals who believe they can interact with a robot, even without prior

experience, are more likely to intend its use. This belief shapes perceived behavioral control, which influences behavioral intention according to TPB.

Advancing theoretical models of technology acceptance in HRI requires a principled distinction between constitutive dimensions and causal antecedents of Control Beliefs. Without this, models may misattribute variance in use intention, reducing explanatory accuracy. Clarifying this boundary is not only a matter of conceptual precision, it is key to building valid measures and robust empirical models.

References

Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179–211.

Ajzen, I. (2002). Perceived behavioral control, self‐efficacy, locus of control, and the theory of planned behavior. Journal of Applied Social Psychology, 32(4), 665–683.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. Psychological Review, 84(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Bartneck, C., Kulic, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International Journal of Social Robotics, 1(1), 71–81.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. Science Robotics, 3(21), east 5954.

Breazeal, C. (2003). Toward sociable robots. Robotics and Autonomous Systems, 42(3–4), 167–175.

Broadbent, E., Stafford, R., & MacDonald, B. (2009). Acceptance of healthcare robots for the older population: Review and future directions. International Journal of Social Robotics, 1(4), 319–330.

Brown, S. A., & Venkatesh, V. (2005). Model of adoption of technology in households: A baseline model test and extension incorporating household life cycle. MIS Quarterly, 29(3), 399–426.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340.

De Graaf, M. M., & Allouch, S. B. (2013). Exploring influencing variables for the acceptance of social robots. Robotics and Autonomous Systems, 61(12), 1476-1486.

de Graaf, M. M. A., Allouch, S. B., & Klamer, T. (2015). Sharing a life with Harvey: Exploring the acceptance of and relationship-building with a social robot. Computers in Human Behavior, 43, 1–14.

de Graaf, M. M. A., Ben Allouch, S., & van Dijk, J. A. G. M. (2019). Why would I use this in my home? A model of domestic social robot acceptance. Human–Computer Interaction, 34(2), 115–173.

Duffy, B. R. (2003). Anthropomorphism and the social robot. Robotics and Autonomous Systems, 42(3–4), 177–190.

Nomura, T., Kanda, T., Suzuki, T., & Kato, K. (2008). Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots. IEEE Transactions on Robotics, 24(2), 442–451.

Serenko, A. (2008). A model of user adoption of interface agents for email notification. Interacting with Computers, 20(4–5), 461–472.
Sharkey, A., & Sharkey, N. (2012). Granny and the robots: Ethical issues in robot care for the elderly. Ethics and Information Technology, 14(1), 27–40.
Taylor, S., & Todd, P. A. (1995). Understanding information technology usage: A test of competing models. Information Systems Research, 6(2), 144–176.
Venkatesh, V., & Brown, S. A. (2001). A longitudinal investigation of personal computers in homes: Adoption determinants and emerging challenges. MIS Quarterly, 25(1), 71–102.
Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS Quarterly, 27(3), 425–478.
Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. Perspectives on Psychological Science, 5(3), 219–232.
Young, J. E., Hawkins, R., Sharlin, E., & Igarashi, T. (2007). Toward acceptable domestic robots: Applying insights from social psychology. International Journal of Social Robotics, 1(1), 95–108.
Złotowski, J., Yogeeswaran, K., & Bartneck, C. (2015). Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. International Journal of Human–Computer Studies, 90, 1–13.

Exploring how school teachers understand robots: a philosophical and empirical study on the attribution of mind to artificial systems

Silvia Larghi[1], Edoardo Datteri[1]
[1]University of Milano-Bicocca

Social robots are increasingly integrated into educational settings, where they may play different roles as tutors, peers, facilitators, emotional mediators, supporting learning, inclusion, and cognitive development (Lampropoulos, 2025). This growing presence calls for an in-depth investigation on how teachers, educators, and students perceive and interpret these systems. Research in human-robot interaction highlights that the way individuals interpret and make sense of robots, particularly in terms of whether and how they attribute mental states and model the robots' mind, can significantly influence the nature and dynamics of the interaction (Thellman et al., 2022). Analysing how teachers theorise about robots can provide relevant insights into how social robots may shape teaching and learning processes. Ultimately, these findings may inform the development of robot-supported educational curricula and guide the design of effective teacher training programs.
The theoretical frameworks that have influenced empirical studies on the attribution of mental states to robots, such as Dennett's Intentional Stance (Dennett, 1971) and the Theory of Mind perspective (Carruthers & Smith, 1996) often characterise the attributed mind in terms of beliefs, desires and other intentional states, typically in the form of propositional attitudes. This research investigates, theoretically and empirically, the possibility that people may attribute mental states to robots not only in the form of propositional attitudes, but through different theoretical vocabularies. Specifically, the research explores the adoption of the folk cognitivist stance: a perspective that models the robotic mind in terms of functional decomposition in cognitive modules and elaboration of representations, in line with the theoretical vocabulary of cognitive science.
The following research questions are addressed.

1. What form and structure might people, in principle, attribute to the mind of robots?
2. How do the teachers involved in the study model the mind of robots depending on their educational and training background?

Modelling strategies and "styles" are studied through the analysis of theoretical vocabularies adopted by the participants when explaining the behaviour of robots observed within video stimuli. Explanations of robotic behaviour have been collected through semi-structured interviews. A first stage of the analysis consists of extracting (qualitatively) the explanations (ordered pairs <explanandum, explanans>) from the verbal data obtained from the transcription of the interviews. Research question 1) is addressed through a qualitative case study analysis (based on Template Analysis, King & Brooks, 2017) of the explanation of the behavior of robots given in semi-structured interviews with 16 participants (adults with various educational and professional backgrounds). Research question 2) is addressed through a qualitative case study Thematic Analysis of the explanation of the behavior of robots given in semi-structured interviews with 16 lower secondary school teachers. The Thematic Analysis is guided by the codebook obtained within research question 1).

At the time the abstract is submitted, data analysis is ongoing and will be completed by the time of the conference. The empirical findings are expected to offer novel insights into how individuals, and school teachers in particular, theorise about robots. Specifically, the results will illuminate the modeling styles underlying the attribution of mind to robots and shed light on the folk cognitivist stance, outlining its characteristics and possible sub-categories.

References

Carruthers, P. & Smith, P. K. (1986) Theories of Theories of Mind. Cambridge University Press.
Dennett, D. C. (1971). Intentional systems. The journal of philosophy, 68(4), 87-106.
King, N., & Brooks, J. (2017). Doing template analysis: a guide to the main components and procedures. In Template Analysis for Business and Management Students (pp. 25-46). SAGE Publications Ltd.
Lampropoulos, G. (2025). Social robots in education: Current trends and future perspectives. Information, 16(1), 29.
Thellman, S., de Graaf, M., & Ziemke, T. (2022). Mental state attribution to robots: A systematic review of conceptions, methods, and findings. ACM Transactions on Human-Robot Interaction (THRI), 11(4), 1-51.

The more I know you, the less you resemble me: participation in a robotics program affects people's anthropomorphism towards robots

Cecilia Roselli[1], Leonardo Lapomarda[1], Nicola Zanchi[1], Silvia Larghi[1], Edoardo Datteri[1]
[1]University of Milan-Bicocca

The type of knowledge we put at play, regarding how we understand and predict others' behavior, is what guides us through the interaction, and thus deeply affects the way we perceive and communicate with others. When considering the interaction between humans and robots, it is well-known that the type of knowledge (e.g., prior knowledge, expertise) that we have might change our perception of robots, and specifically whether- and to what extent- we attribute anthropomorphic traits to them. However, two important aspects remain poorly understood.
First, whether and how the characteristics of a population affect their tendency to anthropomorphize robots. If adults are known to be influenced, in their interactions with robots, by their prior knowledge and expertise towards them, it is still unclear whether the same happens to younger populations, such as secondary school students. Nonetheless, it is a crucial aspect to investigate, since they were born being constantly immersed in technology, and thus their perceptions and attitudes towards technological artifacts (including robots) have to be properly addressed.
Second, it remains unclear what is meant by prior knowledge and expertise towards robots. For example, what characteristics should the exposure to robots have to give individuals the possibility to form proper knowledge about robots? Previous evidence suggests that a single exposure is not sufficient; conversely, repeated exposure (usually combined with a direct interaction with an embodied robot) represents the best method to consolidate new knowledge, which in turn will affect people's perception and attitudes towards robots.
That said, in the present study, we aimed to investigate whether, and how, participating in a robotics program a□ects participants' likelihood of anthropomorphizing robots. To this aim, we recruited a sample of secondary school students (aged 11-14) to perform two well-established tasks measuring their implicit attitudes towards robots: (1) the Implicit Association Test, which allows detecting the strength of the association between concepts and attributes (in our case, the concept of "robot" and the attribute of "anthropomorphic"); and (2) the Cyberball ball-tossing game, which allows to gain information about the perception of the other players (in our case, a human confederate and an embodied humanoid robot) as social partners. These tasks were administered in two identical sessions; in between, students were involved in a 10-hour robotics program, during which they had the opportunity to take part in activities such as programming and code writing and to closely interact with educational robots. Data collection is currently underway. Based on previous evidence, we hypothesized that participants' tendency towards anthropomorphizing robots would be significantly modulated by their participation in the robotics program.
By the time of the conference, the data collection and analysis will be concluded, and we will present whether the results turn out to be in line with the hypothesis.

The Interaction Game: some thoughts on the role of behavior for assessing robots' intelligence

Martina Bacaro[1], Francesco Bianchini[1]
[1]University of Bologna

Alan Turing's (1950) Imitation Game, commonly known as the Turing Test, has long been considered an informal benchmark for evaluating machine intelligence. In its original formulation, the test seeks to answer the question, "Can machines think?" (Turing 1950, 433) by positioning a machine and a human in separate rooms and having a human observer interrogate both through written communication. If the observer cannot reliably distinguish between the machine and the human, the machine is said to exhibit intelligence. Turing's approach emphasizes verbal behavior as the primary indicator of intelligence, a concept that has generated extensive philosophical debate and critique over the decades.

The emphasis on verbal behavior has led to significant focus on linguistic capabilities as the primary metric for intelligence (e.g., Shieber, 2004). However, this narrow interpretation has overlooked the broader range of behaviors that might be essential for achieving meaningful results in the test. This gap can be attributed to the early critiques of the behaviorist approach. Shannon and McCarthy (1956) raised concerns that a machine could theoretically be programmed with a complete set of predetermined responses to all possible stimuli. They argued that such a machine, despite passing the Turing Test, would not align with our intuitive concept of thinking, as it would merely be simulating intelligence (and language) rather than genuinely exhibiting it. After that, Block (1981) nurtured and expanded this critique, also arguing that some further criterion is necessary regarding the manner in which the machine works to assess machine intelligence without being deceived by its behavior, i.e., the judgement not being correlated with organizational or structural computational apparatus that makes the machine work.

In this talk, we will follow a different line of thoughts that sheds light on the interactional aspects of the test (Shieber 2006, 2007), and we will consider both the behavior of the machine and the behavior of the human observer as fundamental parts of the Turing test. Starting from more comprehensive versions of the Turing test (such as the Total Turing Test, Harnad 1991; Schweizer 1998, 2012), we will argue that the concept of imitation included in the Turing Test is better understood as a form of interaction, where ontological correlates regarding the internal structure of the agent's cognitive capabilities fade into the background. Indeed, by following the radical embodied account of mind and cognition (Varela et al. 1991; Di Paolo et al. 2018), verbal behavior can be considered one aspect of a wider interactive scenario in which the bodily, situated, cultural, and social aspects of cognition are entangled.

These ideas will be explored in the field of Human-Robot Interaction (HRI, Dautenhahn 2007). We will analyze different non-verbal variants of the Turing test (Ciardo et al. 2022; Avramah et al. 2012; Strathearn & Ma 2020; Hernández-Orallo 2020) proposed in this field and focus on the role of gaze as an implicit measure of perceived intelligence (Pfeiffer et al. 2011). We will show that the perception of intelligence is an integral part of the intelligence assessment process to the extent that it becomes a constitutive aspect of interaction (Bacaro & Bianchini 2024).

In conclusion, we will also consider some of the philosophical consequences of including imitation in a more comprehensive framework of interaction between agents.

References

Avraham, G., Nisky, I., Fernandes, H. L., Acuna, D. E., Kording, K. P., Loeb, G. E., & Karniel, A. (2012). Toward perceiving robots as humans: Three handshake models face the Turing-like handshake test. IEEE Transactions on Haptics, 5(3), 196-207.

Bacaro, M., & Bianchini, F. (2024). Artificial Intelligence as Expected Intelligence. In Current Topics in Logic and the Philosophy of Science: Papers from SILFS 2022 Postgraduate Conference (pp. 89–115). Rickmansworth: College Publications.

Block, N. (1981). Psychologism and behaviorism. Philosophical Review, 90(1), 5–43.

Ciardo, F., De Tommaso, D., & Wykowska, A. (2022). Human-like behavioral variability blurs the distinction between a human and a machine in a nonverbal Turing test. Science Robotics, 7(68), eabo1241.

Di Paolo, E. A., Cuffari, E. C., & De Jaegher, H. (2018). Linguistic bodies: The continuity between life and language. MIT Press.

Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. Minds and Machines, 1, 43-54.

Hernández-Orallo, J. (2020). Twenty years beyond the Turing test: Moving beyond the human judges too. Minds and Machines, 30(4), 533-562.

Pfeiffer, U. J., Timmermans, B., Bente, G., Vogeley, K., & Schilbach, L. (2011). A non-verbal Turing test: Differentiating mind from machine in gaze-based social interaction. PLOS ONE, 6(11), e27591.

Schweizer, P. (1998). The truly total Turing test. Minds and Machines, 8, 263-272.

Schweizer, P. (2012). The externalist foundations of a truly total Turing test. Minds and Machines, 22, 191-212.

Shannon, C. E., & McCarthy, J. (Eds.). (1956). Automata Studies. Princeton University Press.

Shieber, S. M. (Ed.). (2004). The Turing test: Verbal behavior as the hallmark of intelligence. MIT Press.

Shieber, S. M. (2006). Does the Turing test demonstrate intelligence or not? In Proceedings of the National Conference on Artificial Intelligence (Vol. 21, No. 2, p. 1539). Menlo Park, CA; Cambridge, MA; London: AAAI Press; MIT Press.

Shieber, S. M. (2007). The Turing test as interactive proof. Noûs, 41(4), 686-713.

Strathearn, C., & Ma, M. (2020, July). The multimodal Turing test for realistic humanoid robots with embodied artificial intelligence. Presented at Lifelike Computing Systems: 8th Edition in the Evolution of the Series of Autonomously Learning and Optimizing Systems (SAOS), Online.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59, 433-460.

Varela, F., Thompson, E., & Rosch, E. (1991). The Embodied Mind. Cognitive science and human experience. MIT Press.

Is hostile behavior intuitive? A Hawk-Dove experiment with a varying Harshness of conflict

Pablo Marcos-Prieto[1], Ennio Bilancini[2], Leonardo Boncinelli[1], Chiara Nardi[3]
[1]University of Florence
[2]IMT School for Advanced Studies Lucca
[3]Catholic University of the Sacred Heart

Dual process decision-making theories have received much attention in cognitive and social psychology, neuroscience, and economics. A number of experimental contributions have investigated the impact of the mode of reasoning on economic behavior in a variety of games, but, to the best of our knowledge, no previous studies have considered the effects of the cognitive mode in the propensity for hostility. In the context of a one-shot Hawk-Dove game, we experimentally investigate the effect of the mode of reasoning – deliberative or intuitive – on the propensity to behave hostilely (like a Hawk). We also assess whether this effect, if any, becomes more or less pronounced when varying the harshness of the conflict, the cost of an escalated fight if both players choose the aggressive option, which we manipulate in a between-subject structure. Decisions in the Hawk-Dove game are collected under four treatments varying the response times: unrestricted Control, Time Pressure, in which participants are asked to make a decision within 10 seconds, Time Delay, in which participants are asked to wait at least 60 seconds before making a decision, and Motivated Delay, in which participants must use the 60 seconds waiting time to motivate the decision they intend to make. Our results show that deliberation decreases the likelihood of behaving hostilely (deliberative kindness and motivated deliberative kindness), while intuition increases the probability of hawkish behavior (intuitive hostility). These effects do not depend on the harshness of the conflict, in the sense that harshness does not appear to interact with the treatments, although we find general evidence that hostile behavior weakly decreases when the intensity of conflict increases (equilibrium rationality).

Measuring Changes in Risk Preferences Over the Life Cycle: Experimental Evidence from Economics and Psychology

Erica Ordali[1], Chiara Rapallini[1], Pablo Marcos Prieto[1]
[1]Università degli Studi di Firenze

Risk attitudes are embedded in everyday decisions, as most behaviors involve some inherent risk. Yet, no shared definition of risk or risk preferences exists across disciplines, with Economics and Psychology offering the most prominent, but distinct, perspectives.
In recent decades, many studies have sought to integrate economic and psychological views on risk-taking, aiming for a unified model (Schoemaker, 1993). However, differing assumptions about rationality and task environments lead to varying conclusions. Measuring risk is especially complex in real-life settings and shows variability even in controlled experiments (Schoemaker, 1993). To address this, some have distinguished between types of risk (e.g., instrumental vs. stimulating, Schonberg et al., 2011), measurement approaches (e.g., expected value vs. naturalistic risk), and formal definitions (e.g., weak vs. strong attitudes, Kellen et al., 2017). Still, differences in terminology and methodology across disciplines continue to blur the line between attitudes and behavior (Schildberg-H ̈orisch, 2018; Specht et al., 2011; Josef et al., 2016; Mata et al., 2016; Ordali and Rapallini, WP).
Specifically, the methodologies used to measure risk do not help in providing a single construct and yield different, inconsistent results, as recent meta-analyses on the variation of risk preferences across the life course show (Ordali and Rapallini, WP). This work suggests that when assessing risk preferences, one might need to consider the interplay between cognitive processes, necessary to understand the tasks employed by researchers to assess preferences and individual characteristics, such as intelligence, memory, health, and psychological conditions.
In this contribution, we aim to identify the role of the learning process in modifying risk preferences over repeated measurements across different age groups, interconnecting economic and psychological techniques.
Several previous works have highlighted the role that learning plays in risky decision-making. For example, it has been demonstrated that experience and adaptive sampling might influence the decision-maker, with beliefs regarding previous outcomes being one of the main determinants of the next choice (Denrell, 2007; Oyarzun and Sarin, 2013). Moreover, it has been noted that preferences might not be stable after many repetitions, with subjects moving toward risk neutrality over time and behaving stochastically in what appears to be deliberate randomization across multiple choices of the same lottery (Ert and Haruvy, 2017; Agranov and Ortoleva, 2017). Lastly, a recent paper emphasizes that decisions under risk might not be extrinsically different from decisions under complexity (Oprea, 2024).
Building on this evidence, we will examine how risk attitudes differ between younger and older adults in standard lottery tasks, both before and after a learning phase involving ambiguous choices adapted from psychological studies. We will also collect data on intelligence, cognitive abilities, memory, and health. Our goals are to (i) test how learning influences baseline risk preferences and (ii) assess whether aging and cognitive traits explain discrepancies between self-reported and observed risk behavior. Data will be collected online in the spring, and we expect to confirm that risk aversion generally increases with age.

We argue that this difference will diminish as the same participants undergo risk assessment under uncertainty, requiring learning. After learning, we expect participants' risk aversion (seeking) will be less (more) than in the baseline measurement, moving closer to risk neutrality. The magnitude of this difference depends on the interaction of the learning process, cognitive abilities, and participants' health, indicating that a single task may not fully capture observed empirical regularities.

References

Agranov, M. and P. Ortoleva (2017). Stochastic choice and preferences for randomization. Journal of Political Economy 125 (1), 40–68.

Denrell, J. (2007). Adaptive learning and risk taking. Psychological review 114 (1), 177.

Ert, E. and E. Haruvy (2017). Revisiting risk aversion: Can risk preferences change with experience? Economics Letters 151, 91–95.

Josef, A. K., D. Richter, G. R. Samanez-Larkin, G. G. Wagner, R. Hertwig, and R. Mata (2016). Stability and change in risk-taking propensity across the adult life span. Journal of personality and social psychology 111 (3), 430.

Kellen, D., R. Mata, and C. P. Davis-Stober (2017). Individual classification of strong risk attitudes: An application across lottery types and age groups. Psychonomic bulletin & review 24, 1341–1349.

Mata, R., A. K. Josef, and R. Hertwig (2016). Propensity for risk taking across the life span and around the globe. Psychological science 27 (2), 231–243.

Oprea, R. (2024). Decisions under risk are decisions under complexity. American Economic Review 114 (12), 3789–3811.

Ordali, E. and C. Rapallini (WP). Aging and financial risk-taking: A meta-analysis. under review.

Oyarzun, C. and R. Sarin (2013). Learning and risk aversion. Journal of Economic Theory 148 (1), 196–225.

Schildberg-H ̈orisch, H. (2018). Are risk preferences stable? Journal of Economic Perspectives 32 (2), 135–154.

Schoemaker, P. J. (1993). Determinants of risk-taking: Behavioral and economic views. Journal of Risk and Uncertainty 6, 49–73.

Schonberg, T., C. R. Fox, and R. A. Poldrack (2011). Mind the gap: bridging economic and naturalistic risk-taking with cognitive neuroscience. Trends in cognitive sciences 15 (1), 11–19.

Specht, J., B. Egloff, and S. C. Schmukle (2011). Stability and change of personality across the life course: the impact of age and major life events on mean-level and rank-order stability of the big five. Journal of personality and social psychology 101 (4), 862.

Investing in Public-Health Common Goods: How Transparency and Trust Shape Cooperative Behavior

Sebastiano Munini[1], Marco Marini[1], Alessandro Sapienza[1], Fabio Paglieri[1]
[1]Istituto di Scienze e Tecnologie della Cognizione, CNR Roma

Cooperation in public health often requires individuals to invest in shared resources whose benefits are realized only when enough people contribute. From vaccination uptake to financing hospital infrastructure, citizens face a classic social dilemma: the private cost of contributing is tangible, whereas the personal benefit is contingent on collective effort. This study investigates how two forms of social information, contribution transparency and the perceived trustworthiness of fellow group members, shape cooperative investment in public-health common-good dilemmas. In our experiment, participants play an online resource-allocation game framed in four public-health scenarios: training general practitioners, enhancing emergency-department capacity, funding drug-development research, and expanding cancer-screening programs. In every scenario a participant joins three simulated co-players and makes six consecutive investment decisions. At the beginning of each round, each player receives a fixed endowment and decides how much to invest in a shared resource. If total group investment meets a predefined threshold, the resource is maintained; contributions above the threshold improve it, whereas falling short causes deterioration. Collective gains or losses are divided equally, while uninvested funds remain private, pitting self-interest against group welfare.

The experiment employs a 2x4 mixed design. The between-subjects factor, Contribution Transparency, has two levels. In the full-information condition, players observe every individual contribution and resulting payoff, making responsibility fully visible. In the aggregate-information condition, they see only the group's total investment and the collective outcome, leaving individual behaviour hidden. The within-subjects factor, Group Trustworthiness, has four levels- low, medium-low, medium-high, and high- implemented by varying the average willingness of the simulated partners to invest. High-trust groups reliably reach or exceed the threshold; low-trust groups include persistent free-riders; the intermediate levels yield more ambiguous cooperation. This structure captures not only clear successes or failures but also edge cases where outcomes mask uneven effort. When transparency is present, such asymmetries are visible, allowing participants to attribute credit or blame accurately and adjust behaviour accordingly. Without transparency, players must infer motives solely from collective outcomes, a process that can foster misattribution and shape subsequent decisions.

Throughout the game we record individual investment amounts, within-scenario adjustments, post-scenario trust ratings, and reports of motivation. We predict an interaction: transparency will enhance cooperation when perceived trustworthiness is medium-high or high, by making prosocial actions salient and enabling reciprocity, but will reduce cooperation in low-trust settings because revealing free-riding can trigger disengagement. Additional analyses will examine how ambiguous outcomes under aggregate feedback influence learning and future contributions.

Data collection is nearing completion. The forthcoming results will inform the design of feedback systems, communication tools, and participation strategies in public health. By identifying when to emphasize individual-level information and when to focus on aggregate outcomes, the study

aims to provide actionable guidance for sustaining engagement with health-related common goods whose success depends on cooperative investment.

Why Political Philosophy Should Benefit from Behavioral Sciences

Jacopo Marchetti[1]
[1]University of Pisa

Political philosophy has long relied on idealized assumptions about political agents, often imposing standards that overlook the complexities and limitations of real human behavior. This disconnect becomes especially evident when the behavioral assumptions of normative or realist theories are confronted with empirical findings from the behavioral sciences. Nevertheless, political philosophers remain rather reluctant to integrate or utilize the methods and results of behavioral science within their normative frameworks. Only recently have normative and realist theories begun to explore the conditions and constraints of political action in non-ideal settings. These contributions include work developed in the field of political epistemology of belief (Lynch, 2012; Levy, 2021), non-ideal epistemology (Galeotti, 2018; Tanesini, 2021; Cassam, 2019; Kidd, Battaly & Cassam, 2021; McKenna, 2023; Edenberg & Hannon & 2021; Hannon & de Ridder, 2021; Biale, Liveriero & Sala, 2025; Edenberg & Hannon, 2025), epistemology of democracy (Samaržija & Cassam, 2023), and political realism in philosophy (Cozzaglio & Favara, 2024), as well as in political science and international relations (Achen & Bartels, 2016). Research on epistemic vices has revealed widespread cognitive failures among democratic publics, calling into question idealized conceptions of human agency and prompting a revision of normative assumptions. As Cassam (2016, p. 176) aptly states, vice epistemology is "the epistemology of real human beings." Numerous studies in political psychology (Greene & Haidt, 2002) and game theory (Bicchieri, 2008, 2023; Bicchieri & Xiao, 2009; Bicchieri & Chavez, 2010) also address political issues such as cooperation and prosocial behavior from a behavioral perspective. What is lacking, however, is a unifying theoretical framework that brings these studies together and systematically investigates their normative implications (Guala, 2018). As a result, this intersection between philosophy, non-ideal epistemology, and behavioral social sciences remains ill-equipped to meaningfully address the real-world consequences of normative assumptions.

This proposal aims to introduce a behavioral approach in political philosophy. The choice is motivated by two reasons. First, the idealized model of the political agent is fundamentally at odds with findings from the behavioral sciences. Therefore, political agents should be understood as having a partly plastic mind—shaped and influenced by evolution, culture, and education—rather than as abstract creatures. Building on Guala's concerns (2021), Behavioral Politics should consider people as they are, not as if they were. Second, social and political institutions are often not designed in accordance with the actual cognitive and epistemic capacities of agents. A Copernican revolution is needed to design institutions that accommodate imperfect human beings.

The proposal is structured into four parts. After an initial clarification (1) on the necessity of adopting a behavioral approach, the second part introduces a discussion on the Philosophical Outlook of Normativity, arguing that a behavioral approach should reject a top-down form of normativity through which we construct a model of homo philosophicus (2). Next, the arguments of two prominent traditions in political philosophy—Classical Political Realism and Normative Theory—will be compared and discussed (3) and (4), demonstrating that neither tradition

provides a plausible account of human agency. In the final part of the paper, it will be explained how a behavioral epistemology can be implemented to design institutions that are acceptable to people as they are (5). Evidence-based behavioral considerations suggest that institutions should be grounded in evidence and experimental settings, aligning with the interests, preferences, and behaviors of real people to provide a more comprehensive understanding of the factors influencing individual and collective choices.

References

Achen, C., & Bartels, L. (2016). Democracy for Realists. Princeton University Press.

Biale, E., Liveriero, F., & Sala, R. (Eds.). (2025). Public Ethics for Real People: Toleration, Equal Respect, and Democratic Distortions. Palgrave Macmillan.

Bicchieri, C. (2008). The fragility of fairness: An experimental investigation on the conditional status of pro-social norms. Philosophical Issues, 18(1), 229–248. https://doi.org/10.1111/j.1533-6077.2008.00150.x

Bicchieri, C. (2023). Norm nudging and twisting preferences. Behavioural Public Policy, 7(4), 914–923. https://doi.org/10.1017/bpp.2022.10

Bicchieri, C., & Chavez, A. (2010). Behaving as expected: Public information and fairness norms. Journal of Behavioral Decision Making, 23(2), 161–178. https://doi.org/10.1002/bdm.644

Bicchieri, C., & Xiao, E. (2009). Do the right thing: But only if others do so. Journal of Behavioral Decision Making, 22(2), 191–208. https://doi.org/10.1002/bdm.621

Cassam, Q. (2016). Vice epistemology. The Monist, 99(2), 159–180. https://doi.org/10.1093/monist/onw007

Cassam, Q. (2019). Vices of the Mind: From the Intellectual to the Political. Oxford University Press.

Cozzaglio, I., & Favara, G. (2024). Non-ideal theory and political realism. In H. C. Hänel & J. M. Müller (Eds.), The Routledge Handbook of Non-Ideal Theory (pp. 153–165). New York: Routledge.

Edenberg, E., & Hannon, M. (Eds.). (2021). Political Epistemology. Oxford University Press.

Edenberg, E., & Hannon, M. (2025). A guide to political epistemology. In J. Lackey & A. McGlynn (Eds.), Oxford Handbook of Social Epistemology. Oxford University Press.

Galeotti, A. E. (2018). Political Self-Deception. Cambridge University Press.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? Trends in Cognitive Sciences, 6(12), 517–523. https://doi.org/10.1016/S1364-6613(02)02011-9

Guala, F. (2021). Behavioural politics: Social insurance and mutual benefit in Robert Sugden's The Community of Advantage. International Review of Economics, 68(1), 89–100.

Hannon, M., & de Ridder, J. (Eds.). (2021). The Routledge Handbook of Political Epistemology. Routledge.

Kidd, I. J., Battaly, H. D., & Cassam, Q. (Eds.). (2021). Vice Epistemology. Routledge.

Levy, N. (2021). Bad Beliefs: Why They Happen to Good People. Oxford University Press.

Lynch, M. P. (2012). In Praise of Reason: Why Rationality Matters for Democracy. MIT Press.

McKenna, R. (2023). Non-Ideal Epistemology. Oxford University Press.

Samaržija, H., & Cassam, Q. (Eds.). (2023). The Epistemology of Democracy. Taylor & Francis.

Tanesini, A. (2021). The Mismeasure of the Self: A Study in Vice Epistemology. Oxford University Press.

The boys' club: gender biases in students' evaluations of their philosophy professors

Marco Marini[1], Marco Viola[2], Pia Campeggiani[3]
[1]Institute of Cognitive Sciences and Technologies, National Research Council, Roma
[2]University of Roma Tre
[3]University of Bologna

Student evaluations of teaching (SETs) have become a cornerstone of academic assessment, influencing hiring, promotion, and tenure decisions in many countries. However, given the abundance of gender bias, we hypothesized that some bias against women may manifest in SETs too. To substantiate this hypothesis, we performed two experiments simulating a SET-like situation.

Italian philosophy students (N = 95 and 92) listened, either read the excerpt of 4 philosophy lectures (Mlength = 389 words) (Study 1) or listened to them voiced by an actor (Study 2). Both excerpts and voice actors were chosen based on a pilot study. Lectures were ascribed to/read aloud by either a woman or a man professor. After each lecture, subjects were asked to rate the professor along 7 dimensions: clarity; benefit; interest; competence; self-confidence; care; engagement.

The results reveal a consistent bias: men professors were often rated higher in perceived clarity, competence, self-confidence, and engagement compared to their women counterparts. Interestingly, women do not exhibit this bias when confronted with written stimuli (Study 1), but they do exhibit it with vocal stimuli (Study 2). Notably, even students who self-identified as holding progressive gender attitudes (as measured by the Social Role Questionnaire – Baber & Tucker, 2006) displayed similar patterns of bias (especially in Study 2), indicating that these perceptions are often implicit and automatic.

Our analysis highlights the presence of a gender bias have significant implications for women's academic careers. We conclude by considering some potential objections and limitations and speculate on how future studies may overcome them.

References

Baber, K M. & Tucker, C.J. (2006). The Social Roles Questionnaire: A new approach to measuring attitudes toward gender. Sex Roles, 54(7-8), 459-467. https://doi.org/10.1007/s11199-006-9018-y

Beebee, H. & Saul, J. (2011). Women in philosophy in the UK: A report by the British Philosophical Association and the Society for Women in Philosophy UK. British Philosophical Association.

Lawyers Are Humans Too: Anchoring Bias in Legal Decision-making

Matteo Pirisi[1], Marco Marini[2]
[1]IMT School for Advanced Studies Lucca
[2]ISTC-CNR Roma

This study investigates the impact of the anchoring bias on legal decision-making in compensation scenarios, comparing responses from laypeople and practicing attorneys in Italy. We focus on the relevance and magnitude of numerical anchors that can shape judgment, whether through numbers directly tied to the legal decision or unrelated. Additionally, we examine whether legal expertise attenuates susceptibility to anchoring effects.
Two between-subjects experiments were conducted. In Study 1, 432 Italian lay participants read realistic legal scenarios and were asked to recommend compensation for a victim. Participants were randomly assigned to one of four conditions: high-relevant anchor, low-relevant anchor, high-irrelevant anchor, and low-irrelevant anchor. Relevant anchors referred to the client's expected compensation for the given case; irrelevant anchors referred to unrelated demographic data (e.g., population size of the client's hometown). In Study 2, the same experimental design was replicated with a sample of 291 practicing attorneys.
Results from both studies confirm robust anchoring effects: high anchors led to higher compensation recommendations than low anchors across all participants. Low relevant anchors exerted the strongest influence, significantly reducing recommended compensation amounts compared to low irrelevant anchors. However, high relevant anchors did not significantly differ from high irrelevant anchors in their effects, suggesting that the relevance of an anchor matters more when the value is modest and plausible, and/or matters less when the numbers are extreme. Attorneys showed the same pattern of susceptibility as laypeople, despite their training and experience. However, they consistently recommended lower compensation amounts overall, thus likely reflecting familiarity with legal practice. Attorneys were nonetheless significantly influenced by both relevant and irrelevant anchors, and the magnitude and direction of the anchoring effects mirrored those observed in the lay sample.
These findings challenge the assumption that legal expertise shields against cognitive bias. Even professionals tasked with high-stakes decisions are vulnerable to the psychological pull of numerical anchors, particularly when anchors appear contextually meaningful. Our study provides a rare insight into anchoring bias among practicing attorneys in a continental European legal system and contributes to a broader understanding of expertise, judgment, and cognitive bias in legal reasoning.

Confronting Ignorance Without Gaining Insight: Testing the Limits of Intellectual Humility

Isotta Nuti Budini Gattai[1], Piero Avitabile[1], Federica Ruzzante[1], Folco Panizza[1]
[1]IMT School for Advanced Studies Lucca

Intellectual humility (IH) – the recognition and acceptance of one's cognitive limitations – has attracted growing research interest, yet effective interventions to increase it remain elusive. This study examined whether confronting individuals with their actual performance in a knowledge test would reduce their perceived knowledge (PK) and promote intellectual humility. We conducted an experiment with a representative sample of Italian adults (N = 2,850), randomly assigned to either a treatment or a control condition.

In the treatment group, participants first completed a 20-item medical knowledge test and received feedback on their performance before responding to measures of specific and general IH and PK. In the control group, participants completed the questionnaire first and received the test and feedback only at the end of the survey. The experiment followed a difference-in-differences design where all participants were tested for IH and PK 7–10 days before the intervention, allowing us to assess both within-subject changes and between-group differences over time.

IH was measured using validated Italian versions of scales by Leary et al. (2017), Zachry et al. (2018), and Hoyle et al. (2016). Perceived knowledge was self-assessed, and an Illusion of Knowledge index was computed by comparing perceived with actual performance.

Our pre-registered hypotheses predicted that the intervention would reduce perceived knowledge and increase both intellectual humility in general and specific to the test content (healthcare). As hypothesized, the treatment group showed a significant reduction in perceived knowledge from pre- to post-test. However, contrary to predictions, there was no significant increase in intellectual humility in the treatment group.

These counterintuitive findings suggest that confronting individuals with evidence of their limited knowledge does not enhance intellectual humility. This raises important questions about the psychological mechanisms linking metacognitive insight and humility.

Goal-directed behaviour without predictions: The topological alignment hypothesis

Luca Tummolini[1], Francesco Mannella[1], Julian Zubek[2]
[1]Institute of Cognitive Sciences and Technologies (CNR)
[2]University of Warsaw

Goal-directed behavior is characterized by orientedness towards future outcomes, responsiveness to failure and success, and flexibility when facing the same situational context. Dominant approaches in cognitive psychology and neuroscience explain these features by positing an “internal model” in which action control is guided by predictions and prediction errors (forward model, Desmurget & Grafton 2000; model-based RL, e.g. Daw, Niv & Dayan 2005; Bayesian model inversion, Solway & Botvinick 2012; active inference, Friston 2011). Despite their differences, these approaches share the appeal to an internal model encoding explicit knowledge of the transition dynamics characterizing the instrumental contingencies between actions and effects. However, how this internal model is learned in the first place is often left unexplained. In this contribution, we propose an alternative prediction-free mechanism in which the anticipation of action effects is sufficient for goal-directed control: the topological alignment hypothesis. This framework operates at three levels: functional, neural, and computational. At the functional level, it posits a multimodal topological space to bind together multi-sensory and motor events. This space aligns potential actions and their corresponding perceptual outcomes. Goals arise from this alignment and enable the selection of actions that yield desired effects. At the neural level, the framework draws on neuroscientific evidence showing that the cerebral cortex contains aligned topographical and topological maps. These maps form distinct action-perception domains within the fronto-parietal system. Each domain consists of a parietal area tightly connected to a corresponding motor/premotor area, providing biological support for the topological alignment hypothesis as a core principle of brain organization. At the computational level, we propose a novel algorithm to formalize this theory and to demonstrate the acquisition of a multimodal map of goals. The algorithm leverages topological dimensionality reduction to construct multimodal mappings. By aligning these mappings, the algorithm enables both an exogenous, habitual as well as an endogenous, goal-directed form of action control.

References

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature neuroscience, 8(12), 1704-1711.
Desmurget, M., & Grafton, S. (2000). Forward modeling allows feedback control for fast reaching movements. Trends in cognitive sciences, 4(11), 423-431.
Friston, K. (2011). What is optimal about motor control?. Neuron, 72, 488-498.
Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. Psychological review, 119(1), 120-154.

Decoding Effector-Specific Parametric Grip Force Anticipation from fMRI data

Guido Caccialupi[1,2], Timo Torsten Schmidt[1], Felix Blankenburg[1,2]
[1]Freie Universität Berlin
[2] Humboldt-Universität zu Berlin

Previous fMRI-studies have shown that activity in premotor and parietal brain-regions covaries with the intensity of upcoming grip-force. However, it remains unclear whether grip-force intensities are encoded in an effector-specific format (i.e., information about each hand is stored in an independent brain-activity code), and how this information is transformed for motor-execution. In this fMRI-study, we investigated where and when effector-specific force-intensities are parametrically represented, from action-selection through movement-preparation to motor-execution.
Human participants performed a delayed grip-force task, maintaining one of four grip-force intensities for a 12-second delay-period until execution with the left- or right-cued hand. In 50% of the trials, the hand was switched halfway through the delay-period. Using multivoxel pattern analysis (MVPA) with a time-resolved searchlight approach, we identified brain-regions exhibiting multivariate working-memory codes of anticipated grip-force intensities for each hand. Cross-decoding tested whether grip-force intensities were encoded in an effector-specific or effector-independent format.
We found above-chance prediction accuracy in two lateralized brain-networks: contralateral intraparietal sulci (IPSs) and extrastriate body areas (EBAs) during the early delay-period, and contralateral primary motor cortices (M1s) during the late delay-period. Cross-decoding analyses corroborated effector-specific coding in these regions.
Our results extend the literature on movement-planning by showing that effector-specific parametric codes of anticipated grip-force intensities are stored in contralateral IPSs and EBAs prior to M1s. This likely reflects a transformation process, where the initial selection of an intended grip-force intensity is followed by movement-preparation. It aligns with predictions of ideomotor theories, which postulate a contribution of secondary sensory-regions to the selection of motor-movements.

Beyond Morphological Computation: An Ecological Approach to Control in Soft Robotics

Matteo Antonelli[1]
[1]Università degli Studi di Urbino Carlo Bo

This contribution addresses a fundamental issue in soft robotics: designing effective control strategies for adaptive behavior. Soft robots – flexible systems made of deformable materials, such as a soft gripper – challenge traditional control methods (based on rigid bodies), due to their high degrees of freedom and nonlinear dynamics. To simplify controlling these highly complex robots, the principle of morphological computation suggests offloading computational work from the brain (the central controller) to the body, leveraging the system's material properties and structural dynamics. However, describing this process as "computation" is misleading. While morphology supports control (something that can be quantified), that does not imply that the body takes over any computation.

I propose ecological psychology as a more accurate framework for understanding the role of morphology in adaptive control. Ecological psychology emphasizes the emergent nature of behavior from continuous, reciprocal interactions between an agent and its environment, without relying on computational processing over internal representations. According to this view, control is not a preprogrammed sequence of commands executed by the body, but rather an ongoing, dynamic negotiation with ever-changing external conditions.

An illustration of this idea can be seen in the operation of a soft robotic gripper. Such a gripper spontaneously conforms to an object's shape and texture through a direct, reciprocal interplay between its own compliant material properties and the object's physical surface. Consequently, it achieves adaptive control of grip strength and contact pressure without explicitly calculating force distributions or employing sophisticated computational algorithms. This capability can be explained by ecological psychology's concept of affordances, defined as the action possibilities directly perceived by the agent in its environment. In soft robotics, affordances emphasize how systems naturally exploit body-environment interactions to maintain adaptability in dynamic and unpredictable scenarios.

However, the explanatory power of ecological psychology becomes particularly evident when applied to the domain of tensegrity robotics. Inspired by structural principles found in humans and animals, tensegrity-based devices integrate rigid and soft components, ensuring both stability and flexibility. By employing the concept of synergy – a grouping of potentially independent anatomical elements that work together as a functional unit in exploiting information about affordances for a given behavior – it is possible to design energy-efficient tensegrity robots able to move dexterously within the environment.

This proposal introduces soft robotics and its challenges, critiques the limitations and potential misconceptions of morphological computation, and advances ecological psychology as a more comprehensive alternative. Finally, it outlines new ecological design principles aimed at guiding the development of more efficient, adaptive and intelligent soft robotic platforms.

References

Bernstein, N. A. (1996). Dexterity and its Development. (Latash M., L., & M., T., Turvey, Eds.). Lawrence Erlbaum Associates, Inc.

Favela, L., H. (2024). The Ecological Brain: Unifying the Sciences of Brain, Body, and Environment. Routledge.
Gibson, J. J. (1979). The Ecological Approach to Visual Perception. New York: Psychology Press, Taylor & Francis Group. • Hauser, H., Nanayakkara, T., and Forni, F. (2023), "Leveraging Morphological Computation for Controlling Soft Robots: Learning from Nature to Control Soft Robots," in IEEE Control Systems Magazine, vol. 43, no. 3, pp. 114-129, June 2023, doi: 10.1109/MCS.2023.3253422.
Laschi, C., Mazzolai, B., Cianchetti, M. (2016). “Soft Robotics: Technologies and Systems Pushing the Boundaries of Robot Abilities”. Sci. Robot. 1, eaaah3690. DOI: 10.1126/scirobotics.aah3690 .
Yixiang Liu, Qing Bi, Xiaoming Yue, Jiang Wu, Bin Yang, Yibin Li (2022). A review on tensegrity structures-based robots. Mechanism and Machine Theory. Volume 168, 2022, 104571, ISSN 0094-114X, https://doi.org/10.1016/j.mechmachtheory.2021.104571. • Müller, V., C., & Hoffmann, M. (2017). "What Is Morphological Computation? On How the Body Contributes to Cognition and Control", in Artificial Life, 2017 Winter; 23(1):1-24. doi: 10.1162/ARTL_a_00219. Epub 2017 Jan 31. PMID: 28140632.
Pfeifer, R., & Bongard, J. (2006). How the Body Shapes the Way We Think: A New View of Intelligence. MIT Press.
Profeta, V., L., S., Turvey, M., T. (2018). Bernstein's levels of movement construction: A contemporary perspective. Hum Mov Sci. 2018 Feb; 57:111-133. doi: 10.1016/j.humov.2017.11.013. Epub 2017 Dec 1. PMID: 29202312.
Turvey, M., T. (1992). “Affordances and Prospective Control: An Outline of the Ontology”, in Ecological Psychology, 4:3, 173-187, DOI: 10.1207/s15326969eco0403_3.

Emergence in Complex Adaptive Systems

Antonio Monaco[1]
[1]Indiana University - Bloomington

Emergence is a foundational concept in complexity science, used to describe the appearance of novel, system-level behaviors in systems composed of many interacting components. Such behaviors are central to biological, cognitive, and social systems, such as immune system responses, ant colony foraging, and brain network dynamics. These systems exhibit robust, organized, and adaptive global patterns that are not transparently deducible from the local behavior of their parts. Despite widespread use, the idea of emergence remains contested, particularly regarding its compatibility with physicalism and its explanatory role in science.

This paper proposes a refined account of weak emergence tailored to adaptive complex systems, where local interactions give rise to system-level capacities like regulation and coordination. The starting point is Jessica Wilson's (2021) account of weak emergence, which identifies higher-level properties with proper subsets of the causal powers of their physical bases. These properties are realized through physical configurations that constrain degrees of freedom, allowing for causal autonomy without violating physical closure. However, Wilson's account remains neutral on the nature of realization in different complex systems.

To address this, I analyze specific cases of biological and social regulation, namely the immune system and eusocial insect colonies. Drawing on Melanie Mitchell's (2009) interpretation of these systems as distributed information processors and Conant and Ashby's (1970) theorem that every good regulator must be a model of the system it regulates, I argue that such systems realize distributed regulators. These regulators are not localized entities (e.g., a control center), but are implemented through the spatial-temporal distribution of components (e.g., lymphocytes or ants), local rules, stochasticity, and feedback loops.

The emergent property of distributed regulation is multiply realizable across different kinds of adaptive systems and is not reducible to any single mechanistic implementation. It is best understood as mechanistically coarse-grained: different micro-configurations can realize the same higher-level function. Unlike purely aggregative properties, these functions depend on the structured organization of parts and their coordinated dynamics over time.

This account maintains commitment to non-reductive physicalism by upholding both the causal closure of the physical and the reality of higher-level phenomena. Emergent properties are not fundamental in the ontological sense, but are real in virtue of their explanatory indispensability and systematic role in special science laws. By specifying the kind of realization that characterizes emergent regulation in adaptive systems, this account refines Wilson's general schema and provides a principled way of unifying a class of complex systems that exhibit adaptive self-organization.

Ultimately, the framework allows for a nuanced understanding of how adaptive behavior can emerge from decentralized, information processing without invoking strong emergence or violating physicalism. In doing so, it strengthens the conceptual foundations of complexity science and contributes to the broader discussion on how higher-level explanations relate to physical substrates in the natural world.

Does ChatGPT really understand you? Language comprehension across languages by humans and LLMs

Natalia Moskvina[1], Raquel Montero[1], Masaya Yoshida[1,2], Ferdy Hubers[3], Paolo Morosi[1], Walid Irhaymi[1], Jin Yan[1], Elena Pagliarini[1], Fritz Günther[4], Evelina Leivada[1,2]
[1]Universitat Autònoma de Barcelona
[2]ICREA
[3]Radboud University
[4]Humboldt-Universität zu Berlin

With recent advances in Artificial Intelligence (AI) tools and Large Language Models (LLMs) in particular, a growing number of people are turning to these systems for professional or personal assistance, often encountering a seemingly intelligent and fluent conversation partner. However, despite the unquestionable ability of LLMs to carry out smooth and largely meaningful conversations in a variety of natural languages, the full extent of their linguistic capacities in these languages remains unclear. Previous research suggests that language models fail to accurately perceive certain linguistic aspects such as pragmatic implicatures (Qiu et al., 2023) or generic expressions (Collacciani et al., 2024), and there is an ongoing discussion on whether they are able to consistently distinguish grammatical and ungrammatical structures (Hu et al., 2024; Leivada et al., 2024). More strikingly, Dentella et al. (2024) found that LLMs struggle with language comprehension overall: the tested LLMs were unable to consistently provide accurate and human-like responses to comprehension questions based on 40 short scenarios in English. This finding is of great importance for human-LLM interactions, since input comprehension is an essential part of any successful communication. Moreover, although LLMs are trained and prompted in a variety of languages, it remains unclear if—and how—their linguistic performance varies across languages. Factors such as differences in writing systems, language size (and thus the amount of training data), and linguistic distance from the primary training language (English) may introduce additional challenges.

The present study addresses this research gap by contributing to the exploration of the linguistic capacities of LLMs through testing their comprehension abilities across multiple languages. For this purpose, the original English data set from Dentella et al., (2024) was adapted into 11 languages from different language families: Spanish, Catalan, Italian, German, Dutch, Russian, Greek, Chinese, Japanese, Arabic, and Turkish. Human data was collected from 480 participants (native speakers of corresponding languages, including English, n=40 per language). GPT4o was chosen as the latest version of the best performing LLM family in Dentella et al. (2024). The model's responses were obtained via the API through 40 independent runs of the same data set, simulating 40 pseudoparticipants.

Our results suggest that humans outperformed the model across the languages. Additionally, while there were no significant differences in participants' performance across the languages for humans, the accuracy of ChatGPT's responses did vary depending on the language and, contrary to our prediction, the model's best performing language was not English, but Spanish. Further exploration through machine learning algorithms indicated that the observed crosslinguistic variation is best explained by language distance from Spanish, with shorter distance corresponding to better performance, followed by language size and writing system

type. Strong item effects were observed, suggesting that certain items are challenging for the model across languages.

These findings provide insights into the types of language systems and constructions that are most challenging for LLMs, and the way these challenges can be mitigated by the amount of training data available, therefore potentially expanding our understanding of the linguistic capacities of such frontier AI models.

References

Collacciani, C., Rambelli, G., & Bolognesi, M. (2024). Quantifying generalizations: Exploring the divide between human and llms' sensitivity to quantification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11811-11822.

Dentella, V., Günther, F., Murphy, E., Marcus, G., & Leivada, E. (2024). Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, *14*(1), 28083.

Hu, J., Mahowald, K., Lupyan, G., Ivanova, A., & Levy, R. (2024). Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, *121*(36), e2400917121.

Leivada, E., Günther, F., & Dentella, V. (2024). Reply to Hu et al.: Applying different evaluation standards to humans vs. Large Language Models overestimates AI performance. *Proceedings of the National Academy of Sciences*, *121*(36), e2406752121.

Qiu, Z., Xufeng D., Zhenguang G. C. (2023). Pragmatic implicature processing in ChatGPT. In *Proceedings of NALOMA23*, volume 1, 25–34. ACL.

Do LLMs Exhibit Theory of Mind? Insights from Conversational Approaches to False-Belief Task Success

Marco Fenici[1]
[1]Pegaso Università Telematica

Evidence from Kosinski (2024) indicates that some modern LLMs (ChatGPT-4) can pass false-belief tasks, considered a gold standard in testing theory of mind (ToM) in humans. The result is particularly interesting as LLMs' performance only depends on statistical language processing. On one hand, this aligns with research emphasizing language's role in passing FBTs (Astington & Baird, 2005; Milligan et al., 2007). On the other, it raises the question of whether LLMs truly possess ToM.

A relevant consideration to answer the question requires focusing on and comparing the mechanisms underlying the capacity to pass (elicited-response) FBTs in LLMs and humans. Kosinski (2024) argues that ToM-like ability may have emerged as an unintended by-product of LLMs' improving language skills. In particular, the crucial element enabling LLMs to track protagonists' mental states would be the attention mechanism (Vaswani et al., 2023), which could help LLMs anticipate a character's beliefs "by identifying and tracking relevant connections between her actions, dialogues, and internal states throughout the narrative" (Kosinski, 2024, p. 8).

This view resonates with accounts emphasizing conversational input in shaping children's understanding of social situations and mental states (Berio, 2020; Fenici, 2016, 2020; Jary, 2010). According to this view, children younger than four do not understand the concept of belief but can nevertheless master naïvely the pragmatics of belief reports in specific conversational contexts(Diessel & Tomasello, 2001; Tomasello, 2009). Children's active engagement in conversations where the caregiver credits an agent with a belief is the critical factor integrating their initially scattered mastery of the pragmatics of belief reports, teaches them to track belief reports across contexts, and accordingly shapes their understanding of belief as a representational mental state. Thus, according to conversational accounts, language promotes (and possibly structures, see Fenici & Zawidzki, 2020) a ToM competence by focusing children's attention on different social situations in which cognitive agents acquire or exploit a belief, and acting as a bridge between these situations (Clark, 2006; Dennett, 1987) up to the point that children start connecting them in a structured way.

References

Astington, J. W., & Baird, J. A. (Eds.). (2005). Why Language Matters for Theory of Mind. Oxford University Press.

Berio, L. (2020). Culturally embedded schemata for false belief reasoning. Synthese. https://doi.org/10.1007/s11229-020-02655-7

Clark, A. (2006). Language, embodiment, and the cognitive niche. Trends in Cognitive Sciences, 10(8), 370–374. https://doi.org/10.1016/j.tics.2006.06.012

Dennett, D. C. (1987). The Intentional Stance. The MIT Press.

Diessel, H., & Tomasello, M. (2001). The acquisition of finite complement clauses in English: A corpus-based analysis.

Cognitive Linguistics, 12(2), 97–142. https://doi.org/10.1515/cogl.12.2.97

Fenici, M. (2016). Succeeding in the false belief test. Why does experience matter? In New Developments in Logic and Philosophy of Science. College Press.
Fenici, M. (2020). How children approach the false belief test: Social development, pragmatics, and the assembly of Theory of Mind. Phenomenology and the Cognitive Sciences. https://doi.org/10.1007/s11097-020-09709-8
Fenici, M., & Zawidzki, T. W. (2020). The origins of mindreading: How interpretive socio-cognitive practices get off the ground. Synthese. https://doi.org/10.1007/s11229-020-02577-4
Jary, M. (2010). Assertion and false-belief attribution. Pragmatics & Cognition, 18(1), 17–39. https://doi.org/10.1075/pc.18.1.02jar
Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. Proceedings of the National Academy of Sciences, 121(45), e2405460121. https://doi.org/10.1073/pnas.2405460121
Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. Child Development, 78(2), 622–646.
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.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need (No. arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Assessing Large Language Models' Abilities in Understanding Social Norms

Gaia Gravagno[1], Misael Mongiovì[1,2], Giulia Andrighetto[2], Luana Bulla[1,2], Andrea Guido[2], Eugenia Polizzi di Sorrentino[2]
[1]University of Catania
[2]ISTC - National Research Council

The revolution introduced by generative artificial intelligence (AI), based on Large Language Models (LLMs), turns AIs into machines that produce coherent, human-like, and contextually relevant text (Capraro et al., 2024). As AI agents increasingly assume the role of advisors, collaborators, and decision-making partners, understanding how society can integrate them into these roles is crucial. Despite advances, our comprehension of the rules governing human-AI interactions—and their moral, ethical, and social dimensions—remains limited.

A critical challenge is identifying what AI agents need to be accepted as integral parts of our socio-ecological environment. One critical factor is their ability to recognize and follow social norms—informal rules that guide and clarify socially appropriate behavior. Research indicates that strong social norms, with high consensus, are strictly followed, while weaker norms are often ignored (Szekely et al., 2023; Szekely et al., 2025). To what extent can machines recognize and respond to these norms? Answering this is essential for designing AI systems that interact appropriately with humans.

In this study we leverage the generative power of LLM to explore if and how accurately AI systems can infer social norms in specific contexts. Specifically, we focus on norms related to fairness using the dictator game as case study. In the game, a player named 'dictator' must choose the distribution of a given sum of money-between themselves and a passive recipient. It's a well-known economic game designed to explore the tension between self-interest and altruism. For this study we rely on data from 9 published studies in which human participants not only made allocation decisions but also rated the social appropriateness of each possible action (namely, the social norms), using the norm-elicitation method by Krupka and Weber (2013).

Tailored prompts were employed for the Llama 3.3 70B model, replicating the human experiment and comparing model output with human judgments as ground truth. Each prompt combined a variable section describing the game from the nine studies and a fixed instruction and norm-elicitation part based on Krupka and Weber's method.

Regression analyses of the various simulations revealed a strong relationship between predicted and observed social norms, with high statistical significance. In one simulation, R squared values were encouraging, as six studies showed values above 0.8 and 0.9, while the rest hovered around 0.7, indicating Llama's strong ability to predict variability in social appropriateness judgments. The model excelled in identifying salient actions such as "giving half" and "giving nothing," aligning with human data.

However, the model consistently predicted lower standard deviations than observed, especially for "giving half" and "giving nothing." This suggests an overestimation of social consensus strength: while Llama identify the most and least socially appropriate actions, it misperceived variability in human normative beliefs and fair norms as more universally accepted.

In summary, the study shows LLMs performs good at predicting social norms but are weaker in capturing social consensus surrounding them. Overall, these findings demonstrate the potential

of LLMs in reproducing normative information, paving the way for future research on AI of behavior prediction in decision-making and human-machine interaction.

References

Capraro Valerio, Di Paolo Roberto, Perc Matjaž and Pizziol Veronica 2024. Language-based game theory in the age of artificial intelligenceJ. R. Soc. Interface.2120230720 http://doi.org/10.1098/rsif.2023.0720

Szekely, A., Lipari, F., Antonioni, A. et al. Evidence from a long-term experiment that collective risks change social norms and promote cooperation. Nat Commun 12, 5452 (2021). https://doi.org/10.1038/s41467-021-25734-w

Aron Szekely, Luca Tummolini, Eva Vriens, Giulia Andrighetto, Social norm strength and norm change, Current Opinion in Psychology, Volume 64, 2025, 102040, ISSN 2352-250X, https://doi.org/10.1016/j.copsyc.2025.102040

Out of the Chinese Room

Guido Boella[1]
[1]University of Turin

John Searle's famous *Chinese Room* argument (Searle, 1980) posed a powerful challenge to the symbolic paradigm of artificial intelligence. According to Searle, no matter how well a system manipulates symbols to produce seemingly meaningful responses, it still lacks genuine understanding. This critique undermined the functionalist view (Fodor, 1981) that cognition could be reduced to rule-based symbol manipulation. While originally aimed at symbolic AI, many researchers have extended Searle's argument to contemporary deep learning systems for natural language understanding, asserting that they too are merely algorithms implemented on machines governed by binary logic, and thus no closer to "understanding" than their symbolic predecessors.

In this paper, we argue that such an extension of the argument is misguided. Deep learning does not simply outperform symbolic AI; it represents a new paradigm that provides fundamentally new insights into cognition. Unlike symbolic systems, which rely on predefined logical structures and decompositional analysis, deep learning operates as a model of chaotic and complex systems—systems that, by their nature, cannot be exhaustively captured by finite, analytical models. This includes phenomena in physics (e.g., the three-body problem), meteorology (e.g., GraphCast and GenCast (Lam et al., 2023)). *Learning skillful medium-range global weather forecasting*. *Science*, 382(6675), 1416–1421.), biology (e.g., AlphaFold (Jumper et al., 2021)), and crucially, cognition (LLMs, diagnostic systems).

We propose the following reasoning. First, from the physical to the biological to the social, reality is organized in terms of chaotic and complex systems. The brain, as a biological organ, is no exception—thus, the mind it supports must also be understood as a complex, dynamic system. Symbolic AI, based on formal logic, was inherently incapable of capturing this structure, but most researchers ignored this evidence (Dreyfus, 1972). By contrast, deep learning systems excel precisely in modeling the behavior of such systems, often outperforming traditional analytical methods in domains where deterministic solutions are intractable or nonexistent.

Moreover, chaotic systems are characterized not only by complexity, but by emergence (Holland, J. H., 1998): the spontaneous appearance of new levels of organization and behavior. If deep learning is capable of modeling with such excellent performance chaotic and complex systems at multiple levels—physical, biological, and cognitive—then we cannot rule out that they exhibit also some emergent properties of cognition. Otherwise we would run out of explanations for the success of deep learning. In doing so, it provides an alternative to the functionalist account, not by simulating cognitive functions in an abstract sense, but by realizing the underlying mechanisms from which those functions emerge.

Most notably, deep learning offers a way to model tacit knowledge, as described by Michael Polanyi (1966) —those aspects of cognition we know but cannot fully articulate. Symbolic AI was blind to this dimension, as it could only handle explicitly defined rules (a signal that tacit knowledge exhibit the chaotic and complex nature inherent in our cognition). Deep learning, by contrast, learns patterns and representations from massive datasets in a way that closely resembles how humans internalize knowledge through exposure and experience: e.g., ChatGPT learns to talk without resorting to formal grammars or logics.

In conclusion, while deep learning remains grounded in symbolic computation at the hardware level, its epistemological character is radically different. It provides a non-functionalist, emergentist model of cognition—one that is not refuted by the Chinese Room, but that offers a path *out of it*. This reopens, in a new light, the question of whether artificial systems might not just simulate cognition, but actually instantiate some of its core properties, thanks to the property of deep learning to model emergentist systems. “Syntax is not semantics”, but we cannot rule out that the success of the performances of deep learning based LLMs depends on their ability to make semantics emerge from their ability to capture the long term dependences in texts and to predict the future of the self-similar process due to the fractal (i.e., chaotic) nature of language, by learning regularities and contextual dependencies over vast corpora.

References

Alabdulmohsin, I., (2024). Fractal Patterns May Illuminate the Success of Next-Token Prediction. eprint arXiv:2402.01825

Dreyfus, H. L. (1972). *What Computers Can’t Do: A Critique of Artificial Reason*. Harper & Row.

Fodor, J. A. (1981). *The Mind-Body Problem*. Scientific American, 244(1), 114–123.

Jumper, J. et al. (2021). *Highly accurate protein structure prediction with AlphaFold*. *Nature*, 596, 583–589.

Lam, R., et al. (2023). *Learning skillful medium-range global weather forecasting*. *Science*, 382(6675), 1416–1421.

Searle, J. R. (1980). *Minds, brains, and programs. Behavioral and Brain Sciences,* 3(3), 417–457.

Are there norms of rationality for perception? Evaluating experience-forming processes under Bayesian norms

Alice Andrea Chinaia[1], Matteo Colombo[2]
[1]IMT School for Advanced Studies Lucca
[2]Tilburg University

According to Susanna Siegel's Rationality of Perception (2017), perceptual experiences and their underlying processes can be rational or irrational. Siegel discusses cases of "hijacked perception," where inaccurate perceptual experiences "arise from processes that give prior outlooks too much weight and fail to give proper weight to perceptual inputs" (Siegel, 2017, 5). If this is true–argues Siegel–then "experiences can be epistemically charged" (Siegel, 2017, 21), and can redound well or badly on the rational standing of the subject. Arguments for the Rationality of Perception thesis are consistent with the existence of norms of rational perception, but these norms have yet to be identified. In particular, cases of hijacked perception motivate questions such as whether there are norms determining what it is to give prior outlooks or perceptual inputs too much or too little weight, and why giving too much or too little weight to prior outlooks or perceptual inputs should be appraised as irrational.

We address this issue by considering standard Bayesian norms of rational credence and exploring whether they can apply to perception. First, we distinguish three species of perceptual hijacking: inherited inappropriateness, jumping to conclusions, and circularity. We assess each of them in light of familiar Bayesian norms of rational credence and conclude that none of these norms clarifies what it is to give too much or too little weight to perceptual inputs or to prior outlooks. Our ad interim conclusion is negative: violations of these norms are not diagnostic of irrational perceptual hijacking.

Second, we introduce two scientific models for explaining schizophrenic patients' perception–a belief propagation model (Jardri & Denève, 2013a, 2013b) and a predictive coding model (Adams, Stephan, Brown, Frith, & Friston, 2013). We build our discussion on the famous perceptual phenomenon of depth-inversion illusion, and argue that the belief propagation model, but not the predictive coding model, can show us how the irrationality of perceptual hijacking might arise from violations of probabilism.

We leverage this result to speculate about the phenomenology and practical consequences of perceptual experiences produced by probabilistically irrational – and thus Bayes-irrational–processes. Violations of probabilism fit within the framework that considers perceptual incoherence as possibly experienced as a type of "double book-keeping" (Sass, 1994) and "ontological insecurity" (Laing, 1994). The characteristic disunity of consciousness that patients suffering from schizophrenia report would be the phenomenological counterpart to the Bayes-irrational lack of probabilistic coherence among the various elements populating their mental life. The resulting experiences risk being epistemically downgraded (Siegel, 2017), and, even in those cases in which they are accurate, practically costly.

References

Adams, R.A., Stephan, K.E., Brown, H.R., Frith, C.D., Friston, K.J. (2013). The computational anatomy of psychosis. Frontiers in psychiatry, 4, 47.

Jardri, R., & Denève, S. (2013a). Circular inferences in schizophrenia. Brain, 136 (11), 3227–3241.
Jardri, R., & Denève, S. (2013b). Computational models of hallucinations. The neuroscience of hallucinations (pp. 289–313). Springer.
Laing, R.D. (1994). The divided self. The British Journal of Psychiatry, 165 (3), 420–423,
Sass, L.A. (1994). The paradoxes of delusion: Wittgenstein, schreber, and the schizophrenic mind. Cornell University Press.
Siegel, H. (2017). Education's epistemology: Rationality, diversity, and critical thinking. Oxford University Press.

Truthlikeness and inclusion fallacy

Davide Coraci[1,2], Matteo De Benedetto[1]
[1]IMT School for Advanced Studies Lucca
[2] IRCCS Burlo Garofolo Trieste

From its beginnings, the psychological study of human reasoning always had to wrestle between the normative and the descriptive levels. That is, studying a specific form of reasoning, such as deduction, induction, abduction, categorization, and the like, could mean trying to understand how people should reason, how people actually reason, and why people often do not reason as they should. Depending on the specific form of reasoning one focuses on, the study of the normative and the descriptive dimensions of reasoning might evolve together, or, instead, there could be a significant gap between the two.

In the case of category-based induction (CBI), our understanding of the descriptive dimension of this form of reasoning is much more advanced than our understanding of its normative dimension. In CBI (Rips 1975), people reason from the premise that the members of a category (e.g., robins) have a given property to the conclusion that the members of another category (e.g., birds) also have that property. In the last fifty years, psychologists have explored an impressive array of possible conceptual mechanisms by virtue of which people "project" a certain property from one category to another (cf. Feeney 2017; Feeney and Heit 2007; Heit 1998; Osherson et al. 1990; Sloman 1993; Smith, Shafir, and Osherson 1993). In contrast, the normative dimension of CBI has received less attention. In connection to this, an important issue concerns the so-called inclusion fallacy (Smith, Shafir, and Osherson 1993), i.e., the robust tendency of experimental participants in projecting a given property to a general category like birds, rather than to a proper subset of that category like penguins, despite the plain fact that it is just less likely that all birds, instead of just penguins, have a given property. As a proper normative explanation of why people commit such a blatant violation of the laws of logic and probability is still lacking, the inclusion fallacy remains a central open problem in the discussion on CBI.

In this talk, we seek to fill this gap, by offering a novel normative analysis of the inclusion fallacy. We will do that by building upon the philosophical theory of truthlikeness or verisimilitude, which clarifies how one should reason in order to approach the truth about a given domain. The notion of truthlikeness has been systematically studied in twentieth-century philosophy of science in the context of discussions over scientific progress and realism (e.g., Cevolani and Festa 2020; Kuipers 2019; Niiniluoto 1987; Oddie 1986; Schurz and Weingartner 2010), but it has remained virtually unexplored within cognitive science.

We will show how this notion can provide a useful normative benchmark for inductive inferences such as category-based induction. In particular, we will see that the notion of truthlikeness, by licensing strongly informative conclusions in inductive inferences, provides a normative explanation of why people commit the inclusion fallacy. Actually, when judged against the benchmark of truthlikeness, participants' reasoning can be construed as a rational attempt to approach the truth about the relevant domain. This stands in stark contrast with the traditional probabilistic normative benchmark of such inferences.

The goal of this talk is then two-fold. First, we aim to offer a novel normative analysis of the inclusion fallacy, arguing that, in contrast to the received probabilistic view, people's preferences

in this alleged fallacy might actually be rationally defended. The second, more general, aim of the paper is to highlight the usefulness of the philosophical notion of truthlikeness as a possible normative benchmark for the study of inductive inferences in the psychology of reasoning.

References

Cevolani, G. and R. Festa (2020). “A partial consequence account of truthlikeness”. In: Synthese 197, pp. 1627–1646.

Feeney, A. (2017). “Forty years of progress on category-based inductive reasoning”. In: International handbook of thinking and reasoning, pp. 167–185.

Feeney, A. and E. Heit (2007). “Inductive reasoning: Experimental, developmental, and computational approaches.” In: Fifth International Conference on Thinking, Jul, 2004, University of Leuven, Belgium; Many of the chapter authors for this book talked at the aforementioned symposium. Cambridge University Press.

Heit, E. (1998). “A Bayesian analysis of some forms of inductive reasoning”. In: Rational models of cognition. Oxford University Press, pp. 248–274.

Kuipers, T. A. F. (2019). Nomic Truth Approximation Revisited. Springer.

Niiniluoto, I. (1987). Truthlikeness. Dordrecht: Reidel.

Oddie, G. (1986). Likeness to truth. Dordrecht: Reidel.

Osherson, D. N. et al. (1990). “Category-based induction.” In: Psychological review 97.2, pp. 185–200.

Rips, L. J. (1975). “Inductive judgments about natural categories”. In: Journal of verbal learning and verbal behavior 14.6, pp. 665–681.

Schurz, G. and P. Weingartner (2010). “Zwart and Franssen’s impossibility theorem holds for possible-world-accounts but not for consequence-accounts to verisimilitude”. In: Synthese 172, pp. 415–436.

Sloman, S. A. (1993). “Feature-based induction”. In: Cognitive psychology 25.2, pp. 231–280.

Smith, E. E., E. Shafir, and D. N. Osherson (1993). “Similarity, plausibility, and judgments of probability”. In: Cognition 49.1-2, pp. 67–96.

Biased and Lazy Collective Deliberation

Matteo Michelini[1]
[1]Ruhr University of Bochum

Deliberation plays a central role in democratic decision-making, from juries determining a person's fate to expert teams addressing complex issues (Landemore 2013). It is equally crucial in scientific collaboration (Pesonen 2022) and is often considered the gold standard for resolving disagreements. But how epistemically efficient is deliberation? Specifically, can groups exchanging arguments, evidence, and opinions reliably identify the best solutions? The answer is far from straightforward.

Critics argue that human reasoning is fraught with biases and flaws, undermining deliberation's potential to uncover the truth (Scudder 2020; Brennan 2016; Schwartzberg 2015). While deliberation would work ideally if everyone presented sound, impartial arguments, human reasoners often fall short. They are biased and lazy, producing arguments that only favour their own position and failing to critically scrutinise their reasoning (Trouche et al. 2014; Trouche et al. 2016; Mercier 2016; Stanovich 2021). This leads to phenomena like informational cascades, group-think, and polarisation (Sunstein and Hastie 2008; Solomon 2006). Yet, many remain optimistic. Despite these flaws, some philosophers and cognitive scientists suggest that, with the right kind of interaction, these biases and laziness may be mitigated (Mercier and Sperber 2011; Mercier and Sperber 2017; Cohen 1986; Landemore 2013). In short, there is little consensus on how these individual shortcomings impact group performance. While it is widely recognised that humans are biased and lazy reasoners, the effects of these traits on collective deliberation remain unclear.

To address this gap, I develop an agent-based model that examines how bias and laziness influence group deliberation. After reviewing relevant empirical studies, I show that existing models overlook the impact of these traits on group reasoning (Baccini et al. 2023; Proietti and Chiarella 2023; Banisch and Shamon 2022). My model compares the epistemic performance of biased and lazy reasoners to ideal reasoners—those who exhibit no such flaws. This comparison allows me to assess whether biases and laziness inherently undermine group deliberation or, under certain conditions, enhance it.

My findings challenge the traditional view that biases and laziness are either neutral or purely detrimental to deliberation. I show that, under real-world constraints, such as limited time and scarce evidence, biased and lazy reasoners often outperform ideal reasoners. While ideal reasoners may excel in an ideal world of infinite time and resources, biased and lazy groups are better equipped to navigate the practical realities of decision making. In this context, biases and laziness can function as a 'fast and frugal heuristic' for groups (Gigerenzer and Goldstein 1996), trading off perfection for practical success in constrained environments.

Finally, I explore the philosophical implications of my results. I argue that bias and laziness can be understood as 'Mandevillian virtues'—traits that hinder individual reasoning but enhance group performance—under specific conditions (Astola et al. 2024; Smart 2018). This highlights the context-dependent nature of such virtues. My findings also contribute to debates on the evolutionary role of reasoning, supporting the view that reasoning may have evolved to improve individual cognition through collective deliberation, particularly in constrained environments (Smith and Wald 2019). Additionally, I offer practical recommendations for structuring

deliberation to leverage biases and laziness effectively, contributing to ongoing discussions in science, education, and critical thinking (Dacey 2020).

References

Astola, Mandi, Steven Bland, and Mark Alfano (2024). “Mandevillian Vices”. In: Synthese 204.1, pp. 1–19. doi: 10.1007/s11229-024-04676-y.

Baccini, Edoardo, Zo ́e Christoff, Stephan Hartmann, and Rineke Verbrugge (2023). “The Wisdom of the Small Crowd: Myside Bias and Group Discussion”. In: Journal of Artificial Societies and Social Simulation 26.4.

Banisch, Sven and Hawal Shamon (2022). “Biased processing and opinion polarization: experimental refinement of argument communication theory in the context of the energy debate”. In: arXiv preprint arXiv:2212.10117.

Brennan, Jason (2016). Against democracy. Princeton University Press.

Cohen, Joshua (1986). “An epistemic conception of democracy”. In: Ethics 97.1, pp. 26–38.

Dacey, Austin (2020). “Come Now, Let Us Reason Together”. In: Informal Logic 40.1, pp. 47–76. doi: 10.22329/il.v40i1.6024.

Gigerenzer, Gerd and Daniel G Goldstein (1996). “Reasoning the fast and frugal way: models of bounded rationality.” In: Psychological review 103.4, p. 650.

Landemore, Hélène (2013). “Deliberation, cognitive diversity, and democratic inclusiveness: an epistemic argument for the random selection of representatives”. In: Synthese 190, pp. 1209–1231.

Mercier, Hugo (2016). “The argumentative theory: Predictions and empirical evidence”. In: Trends in Cognitive Sciences 20.9, pp. 689–700.

Mercier, Hugo and Dan Sperber (2011). “Why do humans reason? Arguments for an argumentative theory”. In: Behavioral and brain sciences 34.2, pp. 57–74.

— (2017). The enigma of reason. Harvard University Press.

Pesonen, Renne (2022). “Argumentation, cognition, and the epistemic benefits of cognitive diversity”. In: Synthese 200.4, p. 295.

Proietti, Carlo and Davide Chiarella (2023). “The Role of Argument Strength and Informational Biases in Polarization and Bi-Polarization Effects”. In: Journal of Artificial Societies and Social Simulation 26.2.

Schwartzberg, Melissa (2015). “Epistemic democracy and its challenges”. In: Annual review of political science 18, pp. 187–203.

Scudder, Mary F (2020). Beyond empathy and inclusion: The challenge of listening in democratic deliberation. Oxford University Press.

Smart, Paul R (2018). “Mandevillian intelligence”. In: Synthese 195, pp. 4169–4200.

Smith, Julia Jael and Benjamin Wald (2019). “Collectivized Intellectualism”. In: Res Philosophica 96.2, pp. 199–227. doi: 10.11612/resphil.1766.

Solomon, Miriam (2006). “Groupthink versus the wisdom of crowds: The social epistemology of deliberation and dissent”. In: The Southern journal of philosophy 44.S1, pp. 28–42.

Stanovich, Keith E (2021). The bias that divides us: The science and politics of myside thinking. MIT Press.

Sunstein, Cass R and Reid Hastie (2008). “Four failures of deliberating groups”. In: U of Chicago Law & Economics, Olin Working Paper 401.

Trouche, Emmanuel, Petter Johansson, Lars Hall, and Hugo Mercier (2016). “The Selective Laziness of Reasoning”. In: Cognitive Science 40.8, pp. 2122– 2136. doi: https://doi.org/10.1111/cogs.12303.
Trouche, Emmanuel, Emmanuel Sander, and Hugo Mercier (2014). “Arguments, more than confidence, explain the good performance of reasoning groups.” In: Journal of Experimental Psychology: General 143.5, p. 1958.

Teaching abstract and concrete concepts to your children: an interactive task

Anna Borghi[1], Chiara De Livio[1], Claudia Mazzuca[1], Francesca Bellagamba[1], Alessandro Gennaro[1], Giuliana Mazzoni[1], Silvia Mazzoni[1], Julian Zubek[2]
[1]Sapienza University of Rome
[2] Warsaw University

How do parents transmit abstract knowledge to children? This study shed light on this process using an ecological interactive paradigm.
Abstract concepts (e.g., fantasy) are hard to acquire, likely because they do not have single objects as referents and are more detached than concrete concepts (e.g., boat) from sensory modalities (Banks et al., 2023). However, abstract words are very often used in adult speech (Lupyan & Winter, 2018). Recent proposals emphasized the importance of social interaction for their acquisition (Borghi et al., 2025; Goriachum et al., 2025; Pexman et al., 2023). The help of others, e.g., through explanations, would be crucial to acquiring concepts whose members are not perceptually similar. To address this, it is critical to adopt interactive paradigms, in which concepts are captured in real-use situations (Fini et al., 2025; Gandolfi et al., 2023; Mazzuca et al., 2025; De Livio & Borghi, 2025). Learning new words from parents is one of the most common situations in which concepts are acquired, ideal for testing the multimodal character of the interactions associated with different kinds of concepts. We test the hypothesis that, compared to concrete concepts, children need others more to learn abstract concepts, and this generates stronger alignment, expressed by mutual imitation, gestures oriented to the other rather than reproducing an object, etc.

Method
We recorded 13 hetero-parental families, each with one child, as they engaged in a four-phase conceptual definition task based on the Lausanne Trilogue Play clinical paradigm (Fivaz-Depeursinge & Corboz-Warney, 1999; Lubrano Lavareda et al., 2007) at home. In the first two phases, each parent interacted individually with the child, explaining the meaning of one abstract and one concrete concept at a time while the other parent observed silently. In the third phase, the family jointly defined either an abstract or a concrete concept and then collaboratively created a story or drawing to represent it. Families were randomly assigned to either the abstract or the concrete condition. In the final phase, the child became a silent observer while the parents jointly defined one abstract and one concrete concept.
We annotated the recorded interactions using ELAN by two independent coders; discrepancies were resolved by a third coder. A multinomial Bayesian regression analysis of gesture production revealed distinct gestural patterns associated with the transmission of abstract versus concrete knowledge, consistent with prior research (Zdrazilova et al., 2018). Abstract concepts elicited significantly more metaphorical gestures than concrete concepts (OR = 0.31, 95% CI [0.15, 0.63], PP = 100%). Evidence for increased use of beat gestures was moderate (OR = 0.60, 95% CI [0.35, 1.02], PP = 94%). In contrast, the explanation of concrete concepts was strongly associated with iconic gestures (OR = 11.58, 95% CI [4.30, 34.88], PP = 100%).
Future analyses will explore additional multimodal markers of conceptual transmission, including body movements, social touch, and patterns of mutual imitation among family members.

Dynamical analysis using recurrence quantification will examine how movement coordination patterns differ when transmitting abstract versus concrete knowledge through embodied, multimodal communication.

References

Banks, B., Borghi, A. M., Fargier, R., Fini, C., Jonauskaite, D., Mazzuca, C., ... & Woodin, G. (2023). Consensus paper: Current perspectives on abstract concepts and future research directions. Journal of cognition, 6(1), 62.

Borghi, A. M., Mazzuca, C., & Tummolini, L. (2025). The role of social interaction in the formation and use of abstract concepts. Nature Reviews Psychology, 1-14.

De Livio, C., & Borghi, A. M. (accepted). Multimodal abstract concepts: The transformative role of gestures and language. Sistemi Intelligenti.

Fini, C., Era, V., Cuomo, G., Falcinelli, I., Gervasi, M. A., Candidi, M., ... & Borghi, A. M. (2025). Digital connection, real bonding: Brief online chats boost interpersonal closeness regardless of the conversational topic. Heliyon, 11(4).

Fivaz-Depeursinge, E., & Corboz-Warnery, A. (1999). The primary triangle. A developmental systems view of fathers, mothers and infants. New York: Basic Books

Gandolfi, G., Pickering, M. J., & Garrod, S. (2023). Mechanisms of alignment: shared control, social cognition and metacognition. Philosophical Transactions of the Royal Society B, 378(1870), 20210362.

Goriachun, D., Strijkers, K., Gala, N., & Ziegler, J. C. (2025). The role of social and emotional experience in representing abstract words. Journal of Experimental Psychology: General.

Lubrano Lavareda, A., San Martini, P., Malagoli Togliatti, M., & Mazzoni, S. (2007). LTPc: attendibilità di un metodo osservazione diretta delle relazioni familiari. Rivista di studi familiari: fir famiglia interdisciplinarità ricerca. Fascicolo 2, 2007, 1000-1022.

Lupyan, G., & Winter, B. (2018). Language is more abstract than you think, or, why aren't languages more iconic?. Philosophical Transactions of the Royal Society B: Biological Sciences, 373(1752), 20170137.

Mazzuca, C., Villani, C., Lamarra, T., Bolognesi, M. M., & Borghi, A. M. (2025). Abstractness impacts conversational dynamics. Cognition, 258, 106084.

Pexman, P. M., Diveica, V., & Binney, R. J. (2023). Social semantics: The organization and grounding of abstract concepts. Philosophical Transactions of the Royal Society B, 378(1870), 20210363.

Zdrazilova, L., Sidhu, D. M., & Pexman, P. M. (2018). Communicating abstract meaning: Concepts revealed in words and gestures. Philosophical Transactions of the Royal Society B: Biological Sciences, 373(1752), 20170138.

September, 19 2025

First Looks and Second Thoughts. Understanding Impression Formation on Dating Apps

Anna Gerna[1], Marco Marini[2], Fabio Paglieri[2], Marco Viola[1]
[1]Università degli Studi Roma Tre
[2]ISTC-CNR Roma

Impression formation is a key component of human social cognition, allowing us to navigate interactions with strangers (Zeborwitz & Montepare, 2008; Schaller, 2008). Based on minimal perceptual cues that emerge in a first meeting, humans can form impressions of another person's character within moments (Willis & Todorov, 2006; Bar et al., 2006; Oosterhof & Todorov, 2008). These quick judgments, although often implicit and potentially misleading, tend to be remarkably persistent and can significantly shape interpersonal relationships. (Todorov et al., 2015).

But what happens when first interactions take place online? The rise of digital communication – especially through social media – has enabled people to connect without any physical co-presence. When users meet online, they lack access to real-time behavioral and nonverbal cues, and must instead rely on a reduced set of signals – such as static images and written content – to form first impressions (Okdie et al., 2011; Bacev-Giles & Haji, 2017). Thus, in online environments, cues for inferring others' personalities differ significantly from face-to-face meetings, raising a key question: how does this contextual and informational shift affect impression formation mechanisms?

To explore peer perception online, we conducted a study on mobile dating apps, examining how the two main cues available in this context – photos and textual self-descriptions (bios) – interact in shaping profile perception. In a preliminary pilot (N=100), participants rated a set of photos and bios for attractiveness and interest. Based on these results, we selected three photo conditions (high, medium, low attractiveness) and two types of bios (positive vs. negative, based on elicited interest). These elements were then combined into mock dating profiles for the main experiment (N=132), where participants first evaluated photo-only profiles, indicating whether they would be interested in going on a date. Then, a positive or negative bio was added, and participants could revise their response. We treated expressed dating interest as a proxy for the impression formed, assuming that a positive impression increases the likelihood of wanting to initiate contact with the person behind the profile. This operationalization is grounded in the nature of interactions on dating apps: judgments during the swiping phase involve unknown individuals, form extremely quickly, and rely on minimal, fragmented information – closely aligning with canonical definitions of first impressions in social cognition research (Wood et al., 2024).

Our results reveal that more attractive photos led to more favorable evaluations, confirming the strong influence of visual cues. However, appearance alone did not determine dating intentions. Bios played a crucial role – particularly negative ones, which substantially reduced dating

interest. Notably, the effect of bios was amplified when their valence was perceived as incongruent with the initial photo-based judgment; thus, negative bios were especially influential when paired with highly attractive photos. These findings suggest that, while profile photos act as an entry point, it is the interaction between visual and textual cues that ultimately shapes impression formation online. Additionally, individual differences – such as gender and prior experience with dating apps – moderated cue interpretation, influencing participants' judgments and their interest in real-life meetings.

References

Bacev-Giles, C., & Haji, R. (2017). Online first impressions: Person perception in social media profiles. Computers in Human Behavior, 75, 50-57.

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. Emotion, 6(2), 269.

Guadagno, R. E., Okdie, B. M., & Kruse, S. A. (2012). Dating deception: Gender, online dating, and exaggerated self-presentation. Computers in Human Behavior, 28(2), 642-647.

Lefebvre, L.E. (2018), Swiping me off my feet: Explicating relationship initiation on Tinder. Journal of Social and Personal Relationships, 35(9), pp. 1205-1229.

Okdie, B. M., Guadagno, R. E., Bernieri, F. J., Geers, A. L., & Mclarney-Vesotski, A. R. (2011). Getting to know you: Face-to-face versus online interactions. Computers in human behavior, 27(1), 153-159.

Olivera-La Rosa, A., Arango-Tobón, O. E., & Ingram, G. P. (2019). Swiping right: face perception in the age of Tinder. Heliyon, 5(12).

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. Proceedings of the National Academy of Sciences, 105(32), 11087-11092.

Schaller, M. (2008). Evolutionary bases of first impressions. First impressions, 15-34.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. Annual review of psychology, 66(1), 519-545.

Walther, J. B. (2007). Selective self-presentation in computer-mediated communication: Hyperpersonal dimensions of technology, language, and cognition. Computers in human behavior, 23(5), 2538-2557.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. Psychological Science, 17(7), 592-598.

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. Social and personality psychology compass, 2(3), 1497-1517.

A cognitive perspective on character identification in literary reading. The role of voice and focalization

Alessia Vecchi[1,2], Anna Gerna[2], Sebastiano Munini[3], Marco Marini[3], Fabio Paglieri[3], Francesco Ferretti[2]
[1]University of Milan
[2]Roma Tre University
[3]ISTC - CNR, Rome

Identification with fictional characters constitutes a key aspect of the persuasive power of narratives (Green & Brock, 2000). This is an imaginative process whereby readers adopt the character's perspective and experience events as if from within the narrative world (Oatley, 1995; Cohen & Klimmt, 2021). Over the past few decades, many studies have investigated the relationship between socio-cognitive mechanisms – such as empathy, trust, and perspective-taking – and identification. Findings increasingly suggest that individuals apply the same social-cognitive skills used in real-life interactions when engaging with fictional characters (Gerrig, 1993; Kidd & Castano, 2013; Mumper & Gerrig, 2017). A substantial body of empirical work has used literary texts to explore which textual features promote such processes. Narrative perspective has been shown to influence readers' identification with fictional characters (Caspar et al., 2016; Shwerin & Lenhart, 2022). This concept is often framed through Genette's categories of voice and focalization (Genette, 1966), with greater focus typically placed on voice. Although first-person narration has been linked to stronger identification, few studies have systematically examined the interaction between voice and focalization in shaping engagement (Salem et al., 2017).

To address this gap, the present study draws on Cohn's notion of the "transparent mind" (Cohn, 1988), applying it as a framework to explore the degree of mental access afforded to readers. Focusing on incipits of 20th-century novels, we examine how variations in narrative voice and immediacy of access to a character's thoughts influence mental attribution, empathy, and identification. In a pilot study (N = 54), participants evaluated 16 literary incipits for their degree of internalism (i.e., direct access to thoughts and emotions) versus externalism (i.e., indirect access to mental states). Based on these evaluations, four texts were selected for each of four narrative conditions: first-person internalist, first-person externalist, third-person internalist, and third-person externalist. These texts will serve as stimuli in the main study (N = 100). Using AI-based text manipulation, we will adjust each passage's narrative voice, yielding a total of eight experimental conditions. The experimental procedure involves a within-subjects crossover design. Participants will first complete a brief questionnaire on reading habits, then read four texts (one per condition, randomized). After each text, they will respond to a validated identification scale measuring merging, perspective-taking, understanding, and emotional engagement. Additional items will assess narrative transportation and plot anticipation. Finally, selected items from the Interpersonal Reactivity Index (IRI) will be administered to control for individual trait empathy as a covariate in the analysis.

Based on the existing literature, we propose two hypotheses regarding the expected outcomes. Hypothesis 1: First-person narratives with an internal focus will show a stronger association with identification processes, as they provide more immediate and direct access to the character's mind (Hartung et al., 2016).

Hypothesis 2: Narratives – whether in first- or third-person – with an externalist focus may also elicit strong identification, precisely because their implicit narrative structure could prompt more active inferential engagement (e.g., theory of mind and imaginative processing) on the part of the reader (Castano, 2024; Castano et al., 2024).

References

Castano, E. (2024). Less Is More: How the Language of Fiction Fosters Emotion Recognition. Emotion Review, 16(2), 73–83. https://doi.org/10.1177/17540739241232350

Castano, E., Saedi, F., Zunshine, L., & Ducceschi, L. (2024). On the Complexity of Literary and Popular Fiction. Empirical Studies of the Arts, 42(1), 281–300. https://doi.org/10.1177/02762374231163483

Cohen, J., & Klimmt, C. (2021). Stepping in and out of media characters: Identification and dynamic shifts in users' positioning toward entertainment messages. In P. Vorderer, & C. Klimmt (Eds.), The Oxford handbook of entertainment theory (pp. 266284). Oxford University Press. https://doi.org/10.1093/oxfor dhb/9780190072216.013.15.

Cohn, D. (1988). Transparent minds: Narrative modes for presenting consciousness in fiction (1th Princeton Paperback printing, with corrections, 1983., [Nachdr.]). Princeton Univ. Press.

Genette, G. (1983). Figures III, Paris, Seuil, 1966.

Gerrig, R. J. (1993). Experiencing Narrative Worlds: On the Psychological Activities of Reading. Yale University Press. https://www.jstor.org/stable/j.ctt1dr37cb

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. Journal of Personality and Social Psychology, 79(5), 701–721. https://doi.org/10.1037/0022-3514.79.5.701

Hartung, F., Burke, M., Hagoort, P., & Willems, R. M. (2016). Taking Perspective: Personal Pronouns Affect Experiential Aspects of Literary Reading. PLOS ONE, 11(5), e0154732. https://doi.org/10.1371/journal.pone.0154732

Kidd, D. C., & Castano, E. (2013). Reading Literary Fiction Improves Theory of Mind. Science, 342(6156), 377–380. https://doi.org/10.1126/science.1239918

Mumper, M. L., & Gerrig, R. J. (2017). Leisure reading and social cognition: A meta-analysis. Psychology of Aesthetics, Creativity, and the Arts, 11(1), 109–120. https://doi.org/10.1037/aca0000089

Oatley, K. (1995). A taxonomy of the emotions of literary response and a theory of identification in fictional narrative. Poetics, 23(1–2), 53–74. https://doi.org/10.1016/0304-422X(94)P4296-S

Salem, S., Weskott, T., & Holler, A. (2017). Does narrative perspective influence readers' perspective-taking? An empirical study on free indirect discourse, psycho-narration and first-person narration. Glossa: A Journal of General Linguistics, 2(1), Article 1. https://doi.org/10.5334/gjgl.225

Smell of Jazz: Exploring the Influence of Olfactory and Visual Cues on Musical Performance in Virtual Reality

Chiara Lucifora[1], Denis Migliori[1], Claudia Scorolli[1], Aldo Gangemi[1]
[1]University of Bologna

The relationship between sensory perception and creative expression offers a wealth of interesting research opportunities. While the influence of visual and auditory stimuli on artistic performance is well-documented (Juslin et al., 2010; Lucifora et al.,2025), the impact of olfactory stimulation within immersive virtual environments remains relatively unexplored. This is surprising, given the strong link between olfaction and emotion (Ehrlichman & Bastone,1992; Hertz, 2002), and the wellestablished role of emotion in creative processes, particularly in musical improvisation (Berthoz, 2012; Limb et al., 2008; Pressing, 1988). This paper reports an ongoing experimental study conducted on a sample of 10 (out of 20) professional wind instrument musicians, investigating how congruent and incongruent combinations of visual and olfactory stimuli influence their creative performance in VR environments. Specifically, we examine how positive and negative olfactory and visual stimuli affect musicians' self-perception of creativity, technical and emotional engagement, and overall pleasantness of their performance. Two distinct VR environments are used, representing a forest with a flowery and a dry meadow, and two olfactory stimuli are employed, a negative stimulus based on sulfur and animal components, and a positive stimulus based on floral components. A recorded jazz trio version of "Stella by Starlight" serves as the musical basis for melodic variations and improvisation. All conditions are administered randomly, counterbalancing congruent and incongruent stimuli. Our findings show that: i) the context has an influence on creativity, suggesting that congruent stimuli have an impact on creativity judgment ($p=0.005$), particularly in relation to the positive visual stimulus. ii) congruent positive stimuli increase the emotional involvement and pleasantness in the musical performance ($r=0.765$ $p=0.010$), while negative congruent stimuli increase creativity ($r=0.704$ $p=0.023$) and pleasantness ($r=0.661$ $p=0.038$) just in people with high levels of neuroticism; iii) artistic creativity is associated with emotional involvement during positive stimulations ($r=0.699$ $p=0.025$), while scholarly creativity is related to more technical and less emotional involvement in negative stimulations ($r=0.720$ $p=0.019$). These findings have important implications for understanding multisensory integration and its effects on creativity and emotional experience, relevant for both educational and clinical fields where sensory environments may be designed to enhance specific creative outcomes.

References

Berthoz, A. (2012).Simplexity: Simplifying principles for a complex world. Yale University Press, New Haven.

Ehrlichman, H., Bastone, L.(1992).Olfaction and emotion. In: Science of olfaction, pp. 410–438. Springer, New York.

Herz, R.S. (2002).Influence of odors on the sense of well-being. In: Olfaction, taste and cognition, pp. 286–301. Cambridge University Press, Cambridge.

Juslin, P.N., Sloboda, J.A. (2010).Handbook of music and emotion: Theory, research, applications. Oxford University Press, Oxford.

Knast, L.M., Jain, A., Philpott, C.M., Hummel, T. (2018).Cross-modal influences of taste, smell and visual stimuli on social impression formation. Multisensory research 31(6-7), 527–546.
Limb, C.J., Braun, A.R. (2008).Neural substrates of spontaneous musical performance: An fmri study of jazz improvisation. PLoS ONE 3(2), e167.
Lucifora, C., Gangemi, A., & Scorolli, C. (2025). Creative thinking: Immersive, interactive, and multi-sensory experiences. CREATIVITY STUDIES, 1-21.
Pressing, J. (1988).Improvisation: methods and models. In: Generative processes in music: The psychology of performance, improvisation, and composition, pp. 129–178. Clarendon Press, Oxford.
Spence, C. (2020).Scenting the anosmic cube: on the use of ambient scent for research purposes. iPerception 11(5), 2041669520959706.
Zhou, W., Chen, D., Cai, D. (2021).Cross-modal associations between odors and colors. Frontiers in psychology 11, 613946.

Interpersonal communication in music scenarios

Giulia Cappagli[1], Erica Volta[2], Monica Gori[3], Gualtiero Volpe[4]
[1]Università di Genova
[2]CNR, Genova
[3]IIT, Genova
[4]DIBRIS, Genova

Interpersonal communication is at the basis of many form of social scenarios such as music performance and literature shows it is typically based on visual feedback. In the present set of studies, we investigated the role that visual experience plays in the context of musical interaction. We recorded motion tracking data to analyze interpersonal communication between two musicians in three different scenarios with the same cellist as soloist accompanied by (1) a sighted pianist, (2) a congenitally blind pianist, and (3) a sighted but blindfolded pianist. This allowed us to disentagle the differential role of total visual feedback (case 1) from total visual deprivation from birth (case 2) and from temporary visual deprivation (case 3). We then analyzed musical interaction both at the level of performers and at the level of audience. At the level of performers, we divided post-processed cleaned motion data into segments in order to analyse motor and audio synchrony between the performers as well as expressive phrasing along the excerpt. Interesting results came in particular from motor features, e.g. body sway which resulted to be altered in the blind performer and therefore might be not very informative as literature suggests to synchronize playing. At the level of audience, we presented the extracted segments to trained and non-trained participants in three different modalities (only audio, only video from motion capture, audio+video from motion capture) and asked them to annotate the level of synchronisation and emotional tuning on a Likert scale. Interesting results came in particular from the emotional tuning factor, which resulted to be significantly different in the three conditions presented (sighted-sighted, sighted-blindfolded, sighted-blind).

Predictive analytics: are we statistical victims?

Camilla Francesca Colombo[1]
[1]RWTH Aachen University

The term “Predictive analytics” (PA) refers to the use of a variety of techniques and algorithms, from Machine Learning and data mining to computational statistics, to predict sensitive information or future behaviour of target individuals. Predictions are generated from a larger data set, and their content can range from financial, risk, and consuming patterns to health, social, and sexual traits.

The pervasiveness of PA raises serious ethical questions about its legitimacy, claims of legal shield against the imposed harms, and demands of accountability and fair harm distribution. Noticeably, however, predictive systems do not illegally extort sensitive information, and merely generate probabilistic predictions on individuals, rather than “spying” on private data and information. These two distinctive characteristics make it particularly difficult, both from a theoretical and from practical perspective, to pinpoint where and why predictive systems are threatening individual rights or violating existing regulatory frameworks. Public discussions, normative conceptualizations, and legal frameworks are thus not adequately equipped to address moral and regulatory demands.

In this paper, I argue that, while concerns are legitimate, the most influential frameworks in moral philosophy, as well as legal and public policy literature, which appeal to the concept of inferential privacy, might not answer them successfully. Rather, the statistical dimension of PA warrants nuanced ethical and legal considerations. Specifically, I focus on a critical albeit unexplored issue: while many cases of elicitation of sensitive information or privacy breaches impose harms on us as identified victims, we are best described as statistical victims of predictive systems, as the identity of target individuals cannot be known ex-ante, when training data are collected. To prove my case, using a standard methodology in moral philosophy, I present a pair of fully-equalized cases, which are two instances of potential inferential privacy violations: one is perpetrated by a private investigator and one by a predictive system. I argue that the only difference between the two cases is a specific epistemic condition about the identity of the victim: in one case, the target of the spying activities, of the inferences, and of the imposed harm, is a specific individual who is known ex-ante. In the second case, when data and information are collected and treated in PA settings, this happens without a prior knowledge of what the data will be used for.

I then suggest that the statistical victim problem, a key issue in moral philosophy, provides useful conceptual lenses to normatively assess PA. I survey different positions with respect to the obligations towards statistical victims, both in favour and against their moral stringency. I conclude that rival arguments are at least reasonable and defensible, and that ethical considerations on the legitimacy of predictive systems, and the constraints which ought to be imposed on its use could benefit from this discussion. I further notice that the risk-imposition and risk-distribution framework offers some promising indications. Specifically, taking specific actions to make oneself less vulnerable to predictive inferences could give rise to legitimate claims of protection against imposed harms, and to more stringent regulatory practices.

References

Barry L and Charpentier A (2020) Personalization as a promise: Can Big Data change the practice of insurance? Big Data & Society 7(1). SAGE Publications Ltd: 2053951720935143.
Broome J (1984) Selecting People Randomly. Ethics 95(1). University of Chicago Press: 38–55.
Cevolini A and Esposito E (2020) From pool to profile: Social consequences of algorithmic prediction in insurance. Big Data & Society 7(2). SAGE Publications Ltd: 2053951720939228.
Cohen IG, Daniels N and Eyal NM (eds) (2015) Identified versus Statistical Lives: An Interdisciplinary Perspective. Population-level bioethics series. Oxford ; New York: Oxford University Press.
Floridi L (2014) Open Data, Data Protection, and Group Privacy. Philosophy & Technology 27(1): 1–3.
Goodfellow I, Bengio Y and Courville A (2016) Deep Learning. Adaptive computation and machine learning. Cambridge, Mass: The MIT press.
HARE C (2012) OBLIGATIONS TO MERELY STATISTICAL PEOPLE. The Journal of Philosophy 109(5/6). Journal of Philosophy, Inc.: 378–390.
Jenni K and Loewenstein G (1997) [No title found]. Journal of Risk and Uncertainty 14(3): 235–257.
Köchling A and Wehner MC (2020) Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research 13(3): 795–848.
Liu VX, Bates DW, Wiens J, et al. (2019) The number needed to benefit: estimating the value of predictive analytics in healthcare. Journal of the American Medical Informatics Association 26(12): 1655–1659.
Loi M and Christen M (2020) Two Concepts of Group Privacy. Philosophy & Technology 33(2): 207–224.
Martin K (2023) Predatory predictions and the ethics of predictive analytics. Journal of the Association for Information Science and Technology 74(5): 531–545.
McCue C (2007) Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis. Amsterdam Boston: Butterworth-Heinemann.
Mittelstadt B (2017) From Individual to Group Privacy in Big Data Analytics. Philosophy & Technology 30(4): 475–494.
Molldrem S, Smith AKJ and McClelland A (2023) Predictive analytics in HIV surveillance require new approaches to data ethics, rights, and regulation in public health. Critical Public Health 33(3): 275–281.
Mühlhoff R (2021) Predictive privacy: towards an applied ethics of data analytics. Ethics and Information Technology23(4): 675–690.
Paulson SS and Scruth E (2017) Legal and Ethical Concerns of Big Data: Predictive Analytics. Clinical Nurse Specialist31(5): 237–239.
Price WN and Cohen IG (2019) Privacy in the age of medical big data. Nature Medicine 25(1): 37–43.
Regan PM and Jesse J (2019) Ethical challenges of edtech, big data and personalized learning: twenty-first century student sorting and tracking. Ethics and Information Technology 21(3): 167–179.
Scanlon T (2000) What We Owe to Each Other. Third printing, 2000. Cambridge, Ma London: Harvard University Press.

ChatGPT as a Cognitive Mirror: Narrative Self-Construction and Ethical Implications in the Therapeutic Use of Large Language Models

Francesca Bruno[1]
[1]University of Messina

The rapid diffusion of Large Language Models (LLMs) such as ChatGPT has opened novel avenues for personal reflection, narrative self-construction, and digital psychotechnologies. This paper explores the psychological and ethical implications of using ChatGPT as a cognitive mirror—an AI interlocutor capable of synthesizing user input into meaningful personal narratives that enhance self-understanding and emotional insight.

Recent empirical studies demonstrate ChatGPT's ability to facilitate introspection by structuring fragmented stream-of-consciousness inputs into coherent identity narratives. In a seminal experiment, participants reported high levels of accuracy, surprise, and insight in AI-generated personal narratives based on their spontaneous thoughts (Blyler & Seligman, 2023). Similarly, when customized with personal diary data, ChatGPT has shown potential in supporting self-concept development, decision-making, and personality exploration, effectively functioning as a reflective "human clone" (Zhao & Zhao, 2024).

In educational contexts, ChatGPT reduces cognitive load during narrative scriptwriting, although it does not significantly improve narrative writing skills per se (Martínez, Gajderowicz, & Gómez-Rodríguez, 2024). These findings suggest that while ChatGPT may not enhance creativity directly, it plays a significant scaffolding role in cognitive integration and emotional processing.

However, this growing application space is shadowed by pressing ethical challenges. The literature identifies a wide range of risks including algorithmic bias (Wang et al., 2023), simulated empathy and user overtrust (Kapsali et al., 2024), erosion of reflective autonomy (Bukar et al., 2024), and opaque accountability mechanisms in therapeutic or clinical scenarios. Scholars argue that responsible AI frameworks—centered on fairness, transparency, robustness, and explainability—remain too abstract and insufficiently actionable for sensitive use cases like psychological self-exploration (Lababidi, 2024).

This contribution synthesizes these research streams to critically evaluate ChatGPT's dual role as a narrative facilitator and a potential ethical risk. Through a cross-analysis of literature and conceptual frameworks, we identify both the transformative affordances and structural limitations of deploying LLMs in reflective and therapeutic contexts. We conclude by advocating for the development of domain-specific ethical guidelines, grounded in psychological safety, informed consent, and human oversight, to ensure safe and effective use of AI in mental health and self-development applications.

References

Blyler, A. P., & Seligman, M. E. P. (2023). Personal narrative and stream of consciousness: An AI approach. The Journal of Positive Psychology. https://doi.org/10.1080/17439760.2023.2257666

Bukar, U. A., Sayeed, S., Abdul Razak, S. F., Yogarayan, S., & Amodu, O. A. (2024). Analysis of early expert reactions to the ethical issues and implications of ChatGPT in YouTube video content. SSRN. https://doi.org/10.2139/ssrn.4693877

Kapsali, M. Z., Livanis, E., Tsalikidis, C., Oikonomou, P., Voultsos, P., & Tsaroucha, A. (2024). Ethical concerns about ChatGPT in healthcare: A useful tool or the tombstone of original and reflective thinking? Cureus, 16(2), e54759. https://doi.org/10.7759/cureus.54759
Lababidi, S. (2024). A critical evaluation of ChatGPT's adherence to responsible artificial intelligence principles. https://doi.org/10.61356/j.iswa.2024.3303
Martínez, R. A., Gajderowicz, T., & Gómez-Rodríguez, V. G. (2024). Is ChatGPT helpful for graduate students in acquiring knowledge about digital storytelling and reducing their cognitive load? An experiment. RED. Revista de Educación a Distancia. https://doi.org/10.6018/red.604621
Wang, C.-J., Liu, S., Yang, H., Guo, J., Wu, Y., & Liu, J. (2023). Ethical considerations of using ChatGPT in health care. Journal of Medical Internet Research, 25, e48009. https://doi.org/10.2196/48009
Zhao, Y., & Zhao, C. (2024). Customizing ChatGPT to be a human clone for self-concept augmentation: An embodiment of human-centered AI. SSRN. https://doi.org/10.2139/ssrn.4853253

Talk Without Rhythm. Philosophy of Language Against Algorithmic Surveillance

Emanuele Bottazzi[1]
[1]LOA, ISTC-CNR

Some philosophers, linguists and computer scientists embrace regularism: the view that meaning depends on stable patterns of use within a speech community (Horwich, 1998; 2019; Paperno et al., 2014; Baroni & Zamparelli, 2020). This perspective has gained momentum through advances in AI research, particularly Large Language Models (LLMs) that excel by detecting and replicating regularities in text corpora (Vaswani et al., 2017; Jurafsky & Martin, 2025). We argue, however, that successful communication does not necessarily depend upon lexical regularity, leading us to imagine unexpected ways of speaking that might defy improper uses of AI, particularly algorithmic surveillance.

Our research, inspired by Wittgenstein (PI, §242), highlights an often overlooked but crucial aspect of communication: maintaining constancy in the agreement of word usage over time. Constancy is not mere repetition but an adaptive stability—from Latin cōnstāre, meaning "to stand together, remain" (de Vaan, 2008, entry "stō"). Variations are "held together" by transitions recognized step by step, guided by aspects that are socially treated as the same across different situations. Regularity need not always manifest in wording. It suffices that aspects remain regular in relation to the transition of the wording itself.

Weidinger et al. (2021) warned that LLMs could identify political dissent at scale. More recently, companies such as Google have revoked their commitment to refrain from using AI for surveillance (Tiku & De Vynck, 2025). We propose Sandtalking—a non-formal, improvisational cryptolanguage—not only as a countermeasure to algorithmic surveillance but also as a counterexample to regularism. Sandtalking seeks to evade algorithmic capture by deliberately disrupting the statistical regularities that AI systems rely upon. The name is inspired by the Fremen in Dune (Herbert, 1965), who "walk without rhythm" to avoid detection.

In LLMs, semantic relationships depend on stable co-occurrence statistics and word embeddings (consider the equation "king – man + woman ≈ queen," Mikolov, Yih, and Zweig, 2013). Massive unsynchronized terminological shifts would disrupt co-occurrence stability, preventing effective translation.

Sandtalking differs from constructed languages such as International Auxiliary Languages (IALs)—like Esperanto. It shares affinities with pidgins (Velupillai, 2015); with cants or argots developed to evade social control (Halliday, 1976); and with experimental languages like Viossa (Baer, 2019; Nichols, 2021). It relies on humans' playful attitude toward language and requires no computational tools to be realized.

Its implementation poses challenges—such as optimizing cognitive load and coordination, or maintaining message fidelity across different groups of sandtalkers—that call for interdisciplinary research across cognitive science, cryptography, linguistics, and philosophy.

This practice may be important not only for escaping surveillance. As early as the third century BCE, Zhuangzi was already warning against the risks of patterned, invariable interaction (De Reu, 2015): "Nets are for catching fish; once you get the fish, you forget the net. Traps are for catching rabbits; once you get the rabbit, you forget the trap. Words are for grasping meaning;

once you grasp the meaning, you forget the words. Where can I find people who have forgotten words, and have a word with them?" (Zhuangzi, Ch. 26).

Abbreviations of Wittgenstein's works
PI Philosophical Investigations (P. M. S. Hacker & J. Schulte, Eds.). Wiley-Blackwell. 2009.

References

Baer, S. M. (2019). Ka du hanasu? Collaborative conpidgins. Talk presented at The Eighth Language Creation Conference, Anglia Ruskin University, Cambridge, United Kingdom.

Baroni, M., & Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 1183–1193. Association for Computational Linguistics.

De Reu, W. (2015). A ragbag of odds and ends? Argument structure and philosophical coherence in Zhuangzi 26. In J. Gentz & D. Meyer (Eds.), Literary forms of argument in early China (pp. 243–296). Brill. (Sinica Leidensia, Vol. 123).

de Vaan, M. (2008). Etymological dictionary of Latin and the other Italic languages. Brill.

Halliday, M. A. K. (1976). Anti-languages. American Anthropologist, 78(3), 570–584.

Herbert, F. (1965). Dune. Chilton. (Reprinted: Berkley, 1977; Ace, 1987)

Horwich, P. (1998). Meaning. Oxford University Press.

Horwich, P. (2019). Wittgenstein (and his followers) on meaning and normativity. Disputatio. Philosophical Research Bulletin, 8(9), 147–172.

Jurafsky, D., & Martin, J. H. (2025). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models (3rd ed.) [Online manuscript released January 12, 2025]. https://web.stanford.edu/~jurafsky/slp3/

Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 746–751).

Nichols, S. (2021). Morphological and syntactic diversity among Viossa speakers. Unpublished manuscript, Syntax course, University of Missouri, Spring 2021.

Paperno, D., Baroni, M., et al. (2014). A practical and linguistically-motivated approach to compositional distributional semantics. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1, 90–99.

Tiku, N. and De Vynck, G. (2025). Google drops pledge not to use ai for weapons or surveillance. The Washington Post. February 4, accessed February 5, 2025 https://www.washingtonpost.com/technology/2025/02/04/google-ai-policies-weapons-harm/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), 6000–6010. Curran Associates Inc.

Velupillai, V. (2015). Pidgins, creoles and mixed languages: An introduction. John Benjamins Publishing Company.

Weidinger, L., et al. (2021). Ethical and social risks of harm from language models. arXiv preprint. https://arxiv.org/abs/2112.04359
Zhuangzi. (1999). Chuang-Tzu (Chinese-English edition, Vols. 1–2). Hunan Publishing House.

Watch out for precision medicine. Opportunities and dangers of wearable medical devices

Alessandro Demichelis[1]
[1]IMT School for Advanced Studies Lucca

One of the most striking advancements of contemporary medicine is Precision Medicine, an approach to pharmacotherapy that takes into account genetic, environmental, and lifestyle factors to tailor clinical intervention to the specific needs of a given patient. As a scientific endeavour, Precision Medicine is inextricably tied to the new capabilities offered by technological advancement.

One of these technological advancements is the use of digital technologies called medical wearable devices (or simply wearables). Such devices are designed to constantly provide and relay biological and health parameters of a specific individual. The possibility of using these devices for monitoring an individual's physiology, obtaining a continuous, precise, and personalised data collection, and resulting in a continuous stream of clinical data from a single patient without spending time and money on medical visits, is very promising for tailored healthcare, and has been presented and marketed as a new medical revolution. However, we present three reasons for exercising caution.

The first concerns the fact that wearables are simultaneously a consumer item and clinical tool. Indeed, many wearables are nothing but software merged into non-medical wearable devices, such as fitness trackers. This characteristic blurs the lines between what is medically advisable and what is commercially profitable. There is the need to come up with new conceptual and regulatory approaches to understanding what counts as a medical device, as well as greater collaboration between users, data scientists, clinicians, payers, and governments in matters such as device security, user privacy, data standardization, regulatory approval, and clinical validity (Babu et al. 2024).

The second reason to exercise caution is related to increased dangers of overdiagnosis, self-diagnosis, and unwarranted medicalization of life conditions. Wearables monitor a whole array of biomarkers without interruption and this can lead to interpreting them as pathological biological variations, variations that without the device would hardly be noticed and might not carry clinical interest. This over-focus blurs the lines between what counts as abnormal and pathological, and muds the waters on the difference between risk factors and diseases. Moreover, these potential pitfalls compound with doubts over one of the conceptual premises that ground the use of wearables: that measuring will be helpful, and that data will be meaningful. Self-tracking data does not necessarily make sense to users, as a growing body of studies on the concept of data ambivalence and the issues concerning self-tracking shows (Lomborg, Langstrup and Andersen 2020; Wieczorek et al. 2023).

The third reason to question the indiscriminate use of wearables relates to the conceptual difference between Precision and Personalized Medicine or, more precisely, the lack of clear understanding of it (Roden and Tyndale 2013; Ali-Khan et al. 2016; Phillips 2020). Wearables are marketed to users through a conceptual vocabulary closer to the one of old Personalized Medicine, to better cater to the market's needs and requirements; at the same time, they are presented to agencies of medic- and pharmaco-regulation as devices fully in line with the current premises of contemporary Precision Medicine. This ambivalence points to the risks of

unfulfilled expectations and, consequentially, an erosion of trust relationships between the general public and the clinical experts.

References

Ali-Khan S, Kowal S, Luth W, Gold R, Bubela T. Terminology for Personalized Medicine: a systematic collection. PAGEOMICS. Jan 2016

Babu M, Lautman Z, Lin X, Sobota MHB, Snyder MP. Wearable Devices: Implications for Precision Medicine and the Future of Health Care. Annu Rev Med. 2024 Jan 29;75:401-415.

Lomborg S, Langstrup H, Andersen TO. Interpretation as luxury: Heart patients living with data doubt, hope, and anxiety. Big Data & Society. 2020 May;7(1)

Phillips CJ. Precision medicine and its imprecise history. Harvard Data Science Review. 2020 Jan 31;2(1).

Roden DM, Tyndale RF. Genomic medicine, precision medicine, personalized medicine: what's in a name?. Clinical Pharmacology & Therapeutics. 2013 Aug;94(2):169-72.

Wieczorek M, O'Brolchain F, Saghai Y, Gordijn B. The ethics of self-tracking. A comprehensive review of the literature. Ethics & Behavior. 2023 May 19;33(4):239-71.

Limits and Challenges of the Organisational Approach to Health and Disease

Rebecca Riccardo Cuciniello[1], Maria Cristina Amoretti[1]
[1]University of Genova

Within the philosophy of medicine, naturalist theories of health and disease agree in linking health to function and disease to dysfunction, but disagree on how these are defined. The recent organisational approach (OA; Saborido, Moreno 2015; Saborido et al. 2016) departs from biostatistical and selected effects theories, which rely on comparison with normative models (statistical norms or ancestral traits), by relying on value-free, state-descriptions of biological organisation. The OA identifies function with a trait's contribution to an organism's regulated self-maintenance, and dysfunction with its failure to fulfill its 'functional presuppositions', the organism's endogenous norms. While this model aims to naturalise biological normativity, we present two lines of critique.

The first concerns the OA's internal coherence. OA proponents argue that dysfunction ensues when adaptive regulation does not 'fully compensate' for functional loss, but it is unclear whether dysfunction results in different yet suboptimal functioning, or progression towards death. The second concerns the OA's application to the philosophy of medicine. We assume that a robust theory of health and disease should not require radical revisionism of medical nosology. Instead, the OA either fails to map onto the medical use of health and disease concepts—by considering adaptively functioning organisms as functional, and thus healthy, even when they are typically classified as diseased—or reveals itself to be less value-free than it claims to be.

We suggest that while the OA does not fully succeed in naturalising health and disease on exclusively biological basis, it is compatible with a "value conscious naturalism" (Amoretti, Lalumera 2022), wherein non-epistemic values are involved in operationalising function and dysfunction (suboptimal rather than merely different functioning) for medical understanding and practice. In this way, the OA provides a naturalist basis for distinguishing between different levels of functionality, whose medical value can be determined in a context-sensitive and interest-dependent manner.

References

Amoretti, M. C., & Lalumera, E. (2022). Wherein is the concept of disease normative? From weak normativity to value-conscious naturalism.

Saborido, C., & Moreno, A. (2015). Biological pathology from an organizational perspective.

Saborido, C., Moreno, A., González-Moreno, M., & Hernández Clemente, J. C. (2016). Organizational Malfunctions and the Notions of Health and Disease.

Getting Personal in the Philosophy of Psychiatry. Broadening Our Minds ... and More

Benedetta Cogo[1]
[1]University of Wollongong

Philosophical explorations about psychiatry target different topics—among them, issues of classification and characterisations of the various mental disorders, explanations of the different processes intertwined in mental conditions, ethical questions correlated to these accounts, and other aspects that constitute the reality of mental disorder.

In this paper I will show that a proper characterisation of persons is needed in philosophy of psychiatry, if theorising about psychiatry ultimately aims at furnishing insights into the factual, lived reality of psychiatric conditions.

I will justify and develop my argument in three steps.

First, I will motivate the need for pursuing an appropriate philosophy of persons in psychiatry by analysing how three prominent accounts in philosophy of psychiatry fail, in different ways, to reflect the way the person is understood in the psychiatric practice – in their concrete, real-life situations (Gerrans 2014-2024, Fuchs 2021, Gallagher 2024).

Second, I will show one viable direction for developing a philosophy of persons that does not engage with unnecessary and unhelpful metaphysical issues, and that can be applicable to the clinical practice. I will do so by relying on George Graham's (2010) theory of mental disorders as conditions that exhibit a "truncated rationality" (Graham 2010, p. 126). By stressing the relevance of rationality as one important factor in understanding psychiatric conditions, I will show that the three analysed exemplary accounts fail, in different ways, to give congruous consideration to rationality in mental disorders.

Third, I will advance some correctives and clarifications that, starting from Graham's conception of mental disorders, are applicable to a philosophy of psychiatry that aims at being valuable for the clinical practice. I will focus on two main issues identified in Graham's account: the role and extension of the notion of rationality for the understanding of persons, and the appropriateness of the personal/subpersonal distinction in explaining various processes pertaining to the person.

Ultimately, the criterion of success for my argument – and for a coherent characterisation of person in philosophy of psychiatry – is the applicability of this framework to the practice in a way that is of genuine value for the understanding and treatment of psychiatric conditions.

References

Fuchs, T. (2021). In Defense Of The Human Being: Foundational Questions Of An Embodied Anthropology. Oxford: Oxford University Press.

Gallagher, S. (2024). The Self and Its Disorders. Oxford University Press.

Gerrans, P. (2014). The Measure Of Madness: Philosophy Of Mind, Cognitive Neuroscience, And Delusional Thought. Cambridge, Massachusetts: Mit Press.

Gerrans, P. (2024). Anatomy of an Avatar. The virtual self in pathologies of self awareness. Oxford University Press.

Graham, G. (2010). The Disordered Mind. An Introduction to Philosophy of Mind and Mental Illness. Routledge.

Psychedelic introspection: “ego dissolution” and therapeutic alliance

Emiliano Loria[1]
[1]University of Eastern Piedmont

This talk investigates the dialogical structure of the therapeutic alliance within psychedelic-assisted psychotherapy, with particular attention to its epistemic and existential dimensions. Drawing upon phenomenologically rich first-person accounts of psilocybin- and LSD-induced peak experiences in clinical settings, it reflects on the transformative role of psychedelic introspection, especially during episodes of so-called “ego dissolution.” Such altered states appear to suspend habitual ego-bound defenses, allowing for a reconfiguration of the self–other boundary. Within this space of heightened openness, the therapeutic dyad acquires a unique epistemic intimacy: both patient and therapist gain privileged access to the patient’s inner world. This shared access fosters a shift from resistance to intersubjective negotiation, paving the way for psychic repair, integration, and the reconstitution of selfhood. The argument suggests that psychedelic therapy, particularly when grounded in psychodynamic sensibilities, opens up novel philosophical and clinical possibilities for understanding and facilitating human transformation. Central role in thisreconfiguration of the therapeutic relationship is arguably played by the patient’s emotional involvement during the altered state, and by the narratives elicited by the therapist once the psychedelic effect has subsided. These narratives are frequently marked by imagery and emotional states that resist conventional articulation, often expressed through stereotypical phrases emphasizing the ineffability or unspeakable nature of the experience—reducible to the concept of “ego dissolution.”

The talk will begin with a definition of the concept of therapeutic alliance. Then, attention will be given to the transformative potential triggered by the psychedelic substance—more precisely, by the dosage required to induce a hallucinogenic alteration of consciousness—which, in this framework, is not the curative agent in itself. Psychedelic therapy does not merely consist in the ingestion of a drug, as is typically the case with standard pharmacological treatments. From a clinical perspective, the psychedelic substance functions instead as a channel-drug, a conduit that enables the patient—and, subsequently and indirectly, the therapist—to gain conscious and reflective access to a deep and significant emotional dimension. Such experiential knowledge takes the form of narrated peak experiences within the therapeutic relationship. In the literature, these accounts are often characterized by recurring metaphors about the inexpressibility of the emotional intensity and vividness of the trip. This ineffability frequently concerns the dissolution of the ego and the consequent sense of fusion with the surrounding world—an experience that may be simultaneously frightening and profoundly welcoming. It would be precisely this vivid experiential core that appears to constitute the vehicle for therapeutic transformation, integration, and the overcoming of trauma.

In conclusion, the distinctive dialogical nature of the patient will be taken into account, emphasizing the use of emotionally charged yet stereotyped language. It will be argued that the emotional resonance of this narrative phase enhances the bond of trust between therapist and patient, helping to avert empathic failures on the part of the analyst. Moreover, this process would contribute to a rebalancing of the epistemic dynamics within the therapeutic alliance, whereby the analyst shifts from the role of interpretive guide to that of listener-witness in the patient’s journey toward reintegration and healing.

Dynamic interactions and the Dark Room problem

Catalin Teoharie[1]
[1]University of Bucharest

The Free Energy Principle (FEP) framework can be successfully applied to model the dynamic interactions between a human agent and its environment. According to this framework, all agents that are in equilibrium with their environment must minimize free energy, which serves as a measure of the agent's uncertainty regarding its environment. Interaction with the environment involves either updating internal models based on sensory input or acting upon the environment to confirm those models. An important objection to the FEP is known as the Dark Room problem. In brief, this problem posits that if agents solely aim to minimize free energy, they could achieve this by withdrawing into parts of the environment that offer minimal or no sensory stimulation—settings that are highly predictable. In such low-stimulation environments, free energy would be very low, seemingly satisfying the principle without necessitating meaningful interaction with the world.

I will introduce the Dark Room problem and the solutions proposed by Karl Friston and others within the framework of the Free Energy Principle. I will then formalize the Dark Room problem, discuss its different versions, and analyze how its premises and conclusions have been treated by Friston and others—ultimately they are rejecting the intermediate conclusion that agents tend to inhabit Dark Rooms.

Next, I will propose my own solution to the Dark Room problem, arguing that agents do, in fact, tend to inhabit Dark Rooms on many occasions, and that this behavior serves as an effective strategy for reducing free energy. I will use depression as a compelling example of human agents inhabiting Dark Rooms in order to minimize free energy. Depression can be understood as an effective strategy to minimize free energy by forming negative beliefs and aligning with them through behaviors such as retreating into dark rooms—environments with reduced sensory input and minimal social interractions that offer greater predictability. Successful treatments for depression include the use of antidepressants or psychotherapy. These approaches aim to alter negative beliefs and help individuals move out of the dark room, thus changing the way they interact with the external environment.

References

American Psychiatric Association. (2013). DIAGNOSTIC AND STATISTICAL MANUAL OF DSM-5 TM.

Andersen, I., Thielen, K., Bech, P., Nygaard, E., & Diderichsen, F. (2011). Increasing prevalence of depression from 2000 to 2006. Scandinavian Journal of Public Health, 39(8), 857–863. https://doi.org/10.1177/1403494811424611

Anderson, M. L., & Chemero, T. (2013). The problem with brain GUTs: Conflation of different senses of "prediction" threatens metaphysical disaster. In Behavioral and Brain Sciences (Vol. 36, Issue 3). https://doi.org/10.1017/S0140525X1200221X

Atkinson, & Hilgard. (2022). Introduction to Psychology (15th ed.). Wadsworth Pub Co.

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

Chekroud, A. M. (2015). Unifying treatments for depression: An application of the Free Energy Principle. Frontiers in Psychology, 6(FEB). https://doi.org/10.3389/fpsyg.2015.00153
Clark, A. (2016). Surfing Uncertainty Prediction, Action, and the Embodied Mind. Oxford University Press (OUP).
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
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, Issue 2, pp. 127–138). https://doi.org/10.1038/nrn2787
Friston, K., Moran, R. J., Nagai, Y., Taniguchi, T., Gomi, H., & Tenenbaum, J. (2021). World model learning and inference. Neural Networks, 144, 573–590. https://doi.org/10.1016/j.neunet.2021.09.011
Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. Frontiers in Psychology, 3(MAY). https://doi.org/10.3389/fpsyg.2012.00130
Froese, T., & Ikegami, T. (2013). The brain is not an isolated “black box,” nor is its goal to become one. Behavioral and Brain Sciences, 36(3), 213–214. https://doi.org/10.1017/S0140525X12002348
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
Huys, Q. J., Pizzagalli, D. A., Bogdan, R., & Dayan, P. (2013). Biology of Mood & Anxiety Disorders Mapping anhedonia onto reinforcement learning: a behavioural meta-analysis. http://www.biolmoodanxietydisord.com/content/3/1/12
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
Schwartenbeck, P., FitzGerald, T., Dolan, R. J., & Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. Frontiers in Psychology, 4(OCT). https://doi.org/10.3389/fpsyg.2013.00710
Sun, Z., & Firestone, C. (2020). The Dark Room Problem. Trends in Cognitive Sciences, 24(5), 346–348. https://doi.org/10.1016/j.tics.2020.02.006
Thornton, C. (2022). An informal reconstruction of the free-energy framework, examining the conceptual problems that arise. In Neuropsychologia (Vol. 173). Elsevier Ltd. https://doi.org/10.1016/j.neuropsychologia.2022.108281
Xu, W., Liang, H. N., Baghaei, N., Ma, X., Yu, K., Meng, X., & Wen, S. (2021). Effects of an immersive virtual reality exergame on university students’ anxiety, depression, and perceived stress: Pilot feasibility and usability study. JMIR Serious Games, 9(4). https://doi.org/10.2196/29330

Decision-Making Processes in Psychopathological-Forensic Assessments: Analysis of 489 Italian Expert Reports

Simona Casale[1], Daria Mikhaylova[2], Folco Panizza[1], Giovanni Parmigiani[3], Pietro Pietrini[1]
[1]IMT School for Advanced Studies Lucca
[2]Università di Pisa
[3]Università degli Studi di Modena e Reggio Emilia

Psychopathological-forensic assessments of criminal responsibility represent a critical stage in the criminal justice process, with significant consequences for the defendant's legal trajectory and the allocation of public resources. Ensuring that such complex decisions are made with care is essential to safeguard defendants' rights and avoid the mismanagement of public funds. Biased evaluations can result in two significant types of error: the conviction of defendants with severe psychiatric conditions - thus denying them appropriate care - or the acquittal of fully accountable individuals who are wrongly assigned to the forensic psychiatric system, consuming limited treatment resources.

This study aims to explore the decision-making dynamics underlying forensic psychiatric evaluations by analysing 489 expert reports produced between 2006 and 2022 by court-appointed psychiatrists across Northern, Central, and Southern Italy. These reports were used as evidence in criminal trials and focused exclusively on judicial assessments of criminal responsibility. The data were manually extracted based on factors identified through a prior systematic review of the relevant literature. They can be categorised into four areas: (1) expert characteristics and methodological approach, (2) defendant's sociodemographic variables, (3) defendant's clinical and diagnostic profile, and (4) defendant's criminal history and offence details.

The reports were analysed using Knowledge Representation and Reasoning (KR&R) methods and association rule learning. KR&R techniques enabled the structured organisation and interpretation of complex forensic knowledge, helping to uncover consistent patterns in expert decision-making.

The analysis revealed three main preliminary findings:

1. More severe psychopathological symptoms are strongly associated with findings of non-responsible or substantially diminished responsibility.
2. Male defendants are more frequently judged criminally responsible compared to female defendants.
3. Substantially diminished responsibility is more likely to be recognised when the victim is female.

These findings offer theoretical and practical insights, enhancing our understanding of expert judgment formation and informing interventions to promote fairer, more consistent, and health-conscious judicial processes. Ultimately, this study contributes to ongoing efforts toward greater standardisation and transparency in forensic psychiatric practice.

Epistemic Vigilance Under Asymmetry: Rethinking Argumentation Norms Between Experts and Laypeople

Piero Avitabile[1]
[1]IMT School for Advanced Studies Lucca

Epistemic vigilance is a crucial cognitive capacity that enables individuals to assess the reliability of information and the trustworthiness of its source. In many everyday contexts, exercising epistemic vigilance is uncontroversial and even expected. However, there are situations where a different epistemic posture – epistemic dependence – seems more appropriate, particularly when laypeople engage with expert claims they cannot independently verify. This tension is especially salient in settings such as courtrooms, medical consultations, public debates, and policy discussions, where laypeople are asked to evaluate expert testimony without access to the relevant technical knowledge. In these cases, they must strike a delicate balance: remaining epistemically vigilant while relying on others' superior expertise. While cognitive science has explored the psychological foundations of epistemic vigilance, much less attention has been paid to the normative frameworks that should guide its application in dialogical contexts marked by epistemic asymmetry.

This talk addresses that gap by developing a normative model of epistemic vigilance in expert–layperson dialogue. It asks: (a) under what conditions can a layperson legitimately challenge an expert? and (b) what dialogical norms should regulate such asymmetrical interactions?

We begin by analyzing Walton's influential framework for the appeal to expert opinion (AEO), which treats appeals to epistemic authority as defeasible arguments assessed through critical questions. While effective in peer-to-peer contexts, this framework faces serious limitations in epistemically asymmetrical dialogues. If AEOs carry presumptive force, then direct expert testimony – on which such appeals rely – must also carry some default credibility, even without expliciting supporting arguments. This challenges a core tenet of traditional argumentation theory: that every claim must be open to scrutiny and defended upon request.

To resolve this, we propose that direct expert testimony be treated as a condensed argument – a stand-in for a more complex inferential structure that cannot always be fully articulated in real time. We then offer a revised set of critical questions suited to expert–layperson interactions. These shift the dialectical burden onto the expert in a way that accommodates the layperson's limited epistemic resources, thereby enabling epistemic vigilance under conditions of dependence.

We extend the analysis to two further cases: the expert's appeal to their own expertise (AOE) and the layperson's appeal to a second expert. The former, we argue, is generally fallacious, as it functions as an inquiry-stopper that undermines critical engagement. The latter, on the other hand, offers a viable strategy for challenging expert claims through higher-order evidence. Yet this move, too, must be subject to normative constraints. We therefore articulate a complementary set of obligations for lay participants, including criteria for responsibly invoking second opinions and engaging with expert responses. This dual framework ensures that the burdens of argumentation are distributed in a way that preserves both epistemic vigilance and dependence, allowing for genuinely critical, yet fair, expert-layperson dialogue.

Reclaiming Epistemic Agency in the Age of Deepfakes

Selene Arfini[1]
[1]University of Pavia

Epistemic agency refers to the extent to which individuals can control their belief formation and revision processes (Schlosser, 2019). While the debate over the voluntariness of belief formation has been longstanding, recent concerns highlight how the proliferation of fake news and misinformation, especially carried with the use of AI-generated content — specifically deepfakes — undermines the ability of internet users to exercise their epistemic agency (Rini, 2020; de Ruiter, 2021; Coeckelbergh, 2022). A central question in the philosophical discourse surrounding this issue is: How can we maintain a rational and evidence-based relationship with technology-mediated news when the authenticity of what we see and read is increasingly in doubt?

In this article, I argue that the rise of deepfakes offers an unexpected avenue for embracing a Peircean fallibilist approach to news consumption. Charles Sanders Peirce's doctrine of fallibilism is studied as a theory in three parts (Margolis, 1998): first, it is grounded on the idea that any agent can hold on to a belief that they mistakenly think is true; secondly, that a society of inquirers can, in time, unveil which beliefs are true and false and progress toward a better understanding of the state of things; third, the collective self-correction is tied to inferential empirical and fallible processes(Peirce, 1992).

In this paper I contend that, by acknowledging the existence and inconspicuousness of misinformation and technological manipulation, deepfakes make it easier for individuals to adopt a fallibilist mindset – one that recognizes the inherent uncertainty in belief formation and emphasizes the possibility of error. By defending a deflationary account of epistemic agency – in agreement with Setiya (2013) – I will argue that endorsing the need for fallibilism in news consumption online leads to reclaiming a more active stance over the formation and revision of our beliefs. Indeed, this paper contends that certain forms of ignorance, though seemingly obstructive, can be understood as both a catalyst for the spread of disinformation and a reminder of the need for greater epistemic humility. In this context, adopting a fallibilist perspective not only helps individuals navigate the complexities of modern media but also underscores the importance of collective action toward improving science communication and democratizing the way news is consumed.

Referencees

Coeckelbergh, M. (2022) 'Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence', AI and Ethics, 3, pp. 1341–1350.

Margolis, J. (1998) 'Peirce's fallibilism', Transactions of the Charles S. Peirce Society, 34(3), pp. 535–569.

Peirce, C.S. (1992) 'How to make our ideas clear', in N. Houser and C. Kloesel (eds) The Essential Peirce Selected Philosophical Writing. Indiana: Indiana University Press, pp. 124–141.

Rini, R. (2020) 'Deepfakes and the Epistemic Backstop', Philosopher's Imprint, 20(24), pp. 1–16.

de Ruiter, A. (2021) 'The Distinct Wrong of Deepfakes', Philosophy & Technology, 34(4), pp. 1311–1332.

Schlosser, M. (2019) ‘Agency’, in E.N. Zalta (ed.) The Stanford Encyclopedia of Philosophy. Winter 2019. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/win2019/entries/agency/ (Accessed: 8 August 2023).
Setiya, K. (2013) ‘Epistemic agency: Some doubts’, Philosophical Issues, 23, pp. 179–198.

Psychological Correlates of Pseudoscientific Beliefs

Giorgia Adorno[1], Carlo Martini[1]
[1]Università Vita-Salute San Raffaele

Pseudoscience, characterized by its resistance to falsification and appeal to intuitive reasoning, plays a crucial role in fulfilling the human need for cognitive closure. Unlike scientific inquiry, which is inherently provisional and fragmented (Cartwright 1999), pseudoscience offers a homogeneous explanatory system that delivers immediate and seemingly definitive answers to complex phenomena. This paper examines the psychological and epistemic mechanisms underlying the appeal of pseudoscience, focusing on its function as a closure-inducing belief system. We argue that the homogeneity of pseudoscientific discourse, which provides internally consistent yet empirically unsubstantiated explanations, serves as an epistemic refuge for those seeking certainty.

A central aspect of this analysis is the role of pseudoscience in fostering negationist attitudes—such as climate change denialism, vaccine skepticism, and flat Earth beliefs. These examples show that pseudoscientific frameworks do not merely propose alternative explanations but actively reject mainstream scientific consensus. This deliberate rejection satisfies a cognitive need for a coherent and closed system, in which counterevidence is dismissed or reinterpreted to reinforce prior beliefs (Lewandowsky et al., 2012; Hansson, 2017). The negationist dimension of pseudoscience can thus be seen as a defense mechanism against discomfort caused by scientific uncertainty and complexity (Festinger, 1957; Kruglanski, 2004).

To frame this discussion, we integrate epistemological and psychological perspectives that highlight the contrast between scientific and pseudoscientific reasoning. From an epistemological standpoint, the philosophy of science acknowledges the ‘patchwork’ nature of scientific knowledge—where theories evolve, interact, and sometimes conflict before converging toward refined understanding (Laudan, 1977; Longino, 1990). In contrast, pseudoscience exhibits “epistemic rigidity,” a refusal to revise core assumptions in light of new evidence. This rigidity enhances its persuasiveness by reducing the cognitive effort required to engage with uncertainty (Osman et al., 2022).

Psychologically, research on the need for closure (Kruglanski et al., 1997) shows that individuals with a high desire for certainty are prone to favoring simplified, black-and-white explanations. Pseudoscience caters to this preference, drawing on cognitive biases such as the illusion of explanatory depth (Rozenblit & Keil, 2002), motivated reasoning (Kunda, 1990), and mistrust of scientific authorities (Lewandowsky et al., 2013). These biases reinforce the self-sealing nature of pseudoscientific belief systems, making them resistant to refutation (Boudry & Braeckman, 2011).

Building on these insights, we propose that pseudoscience functions as an epistemic coping strategy with two key roles: (1) providing an illusion of understanding through a coherent, unfalsifiable framework, and (2) shielding individuals from discomfort associated with scientific complexity. This perspective has significant implications for science communication and public policy. Rather than focusing solely on debunking pseudoscienific claims, effective interventions should address the psychological needs that drive their adoption.

In conclusion, the prevalence of pseudoscience is not merely a byproduct of misinformation or lowscientific literacy, but reflects deeper epistemic and psychological needs. Viewing

pseudoscience as a cognitive closure mechanism provides fresh insights into its resilience against counterevidence and scientific progress. This understanding can inform theoretical developments in the philosophy of science and practical efforts in education and policy-making, ultimately fostering greater epistemic openness and critical thinking.

References

Boudry, M., & Braeckman, J. (2011). Immunizing strategies and epistemic defense mechanisms. Philosophia, 39(1), 145–161.

Cartwright, N. (1999). The dappled world: A study of the boundaries of science. Cambridge University Press.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. Science Education, 86(2), 175–218.

Festinger, L. (1957). A theory of cognitive dissonance. Stanford University Press.

Hansson, S. O. (2017). Science denial as a form of pseudoscience. Studies in History and Philosophy of Science, 63, 39–47.

Kruglanski, A. W. (2004). The psychology of closed-mindedness. Psychology Press.

Kruglanski, A. W., Webster, D. M., & Klem, A. (1997). Motivated resistance and openness to persuasion in the presence or absence of prior information. Journal of Personality and Social Psychology, 73(5), 1034–1049.

Kunda, Z. (1990). The case for motivated reasoning. Psychological Bulletin, 108(3), 480–498.

Laudan, L. (1977). Progress and its problems: Towards a theory of scientific growth. University of California Press.

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2013). Misinformation and its correction: Continued influence and successful debiasing. Psychological Science in the Public Interest, 13(3), 106–131.

Lobato, E. J. C., Mendoza, J., Sims, V. K., & Chin, M. G. (2014). Examining the relationship between conspiracy theories, paranormal beliefs, and pseudoscience acceptance. Applied Cognitive Psychology, 28(5), 617–625. Longino, H. (1990). Science as social knowledge: Values and objectivity in scientific inquiry. Princeton University Press.

Osman, M., Heath, H., Evans, J. S. B. T., Stanovich, K. E., & Over, D. E. (2022). The psychological basis of pseudoscientific beliefs: A dual-process perspective. Trends in Cognitive Sciences, 26(3), 213–224.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatorydepth. Cognitive Science, 26(5), 521–562.

Does Embodied Cognition Reshape Our Understanding of Mind Evolution?

Antonella Tramacere[1]
[1]Roma Tre University

How can embodied cognition contribute to our understanding of the evolution of the mind? Does this new cluster of approaches in cognitive science shift or transform how we explain the emergence of mental traits?

This talk addresses these questions by focusing on a core case of content-oriented cognition: mindreading—the ability to track and respond to the mental states of others. Comparative and developmental studies show that children typically acquire mindreading capacities around four years of age. Moreover, infants as young as two, and some non-human animal species, behave in ways suggesting they can act in social contexts by taking others' visual or mental perspectives into account. I examine whether and how embodied approaches might offer a genuinely alternative explanation of this evidence—and what challenges they face in doing so.

Classical modular accounts, rooted in evolutionary psychology, maintain that mindreading is subserved by domain-specific, innately specified cognitive mechanisms. These internal "theory of mind" modules are seen as evolved solutions to social coordination problems, selected for their adaptive value in ancestral environments (Baron Cohen, 1999). Cultural views challenge this innatist perspective by emphasizing the role of external social practices, language, and shared symbolic systems. On these views, content-oriented cognition is not fully hardwired but co-constructed and transmitted through enculturation processes and cooperative learning (Heyes, 2018).

Despite their differences, both approaches rely on an underlying representationalist assumption: that cognitive operations involve manipulating internal representations of other minds. Embodied and enactive accounts break with this assumption. They argue that cognition—including mindreading—is not best understood as inner modeling but as skilled engagement with the world and with others through bodily interaction. Meaning, in this view, is not internally stored but dynamically enacted in situated activity.

The central claim of this talk is that embodied approaches open new directions for theorizing the evolution of cognitive capacities, but they must go further to fulfill this promise. In particular, if mindreading is to be explained without invoking representational modules (Shapiro, 2019) we need a clearer account of how bodily features—such as movement dynamics, sensory modalities, and motor repertoires—shape the emergence of socially relevant understanding across phylogenetic and developmental scales. Without such an account, the embodied perspective risks falling back on vague appeals to interaction or coordination, without addressing the cognitive flexibility and norm sensitivity that characterize mindreading in humans.

Using comparative and developmental examples, I describe how embodied engagement might explain differences in mindreading abilities. Embodied views suggest that mindreading may not be a matter of inferring mental content, but of participating in structured, embodied social dynamics that grow more complex over evolutionary time (Gallagher, 2008; Varela, Thompson, Rosch, 2017). Yet this is precisely where current embodied theories fall short: in explaining how individuals come to understand unobservable or counterfactual mental states in a way that links

bodily affordances with increasingly abstract forms of social cognition—without reverting to classical modular structures.

References

Baron-Cohen, Simon. *The evolution of a theory of mind*. 1999.
Heyes, Cecilia. *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press, 2018.
Shapiro, Lawrence. *Embodied cognition*. Routledge, 2019.
Gallagher, Shaun. "Direct perception in the intersubjective context." *Consciousness and cognition* 17.2 (2008): 535-543.

4E and technology: towards an enactive mediology

Francesco Parisi[1]
[1]Università di Messina

This paper introduces enactive mediology as a combined, interdisciplinary approach bridging cognitive science and media studies.
Although there is no fully unified theory under the label of “enaction”, all enactive approaches share the view that cognition arises from the ongoing interaction between the agent and their environment. Cognition is thus co-constituted through sensorimotor loops that allow the agent to make sense of the world.
Mediology, on the other hand, investigates the role of media in shaping societies and individuals. Certain disciplinary strands also focus on how media affect our lived experience, exert influence on our perception and cognition, and contribute to a global reconfiguration of how humans relate to the world.
Too often, enactivists underestimate the role of media as constitutive elements of human experience; conversely, media scholars tend to focus on the social dimension while overlooking the cognitive dynamics of the individual agent. This paper proposes a convergence of these two paradigms to develop a more integrated and epistemologically robust understanding of our interactions with others, machines, and environments.
To support this conceptual integration, I present a case study: contemporary electric and optical media dramatically reshape how we engage with the environment by enabling novel sensorimotor conditions—such as telepresence via screens. The paper explores how enactive mediology can account for the complexity of these mediated experiences. This perspective aligns with the view that technology and media are not external tools, but integral components in shaping human experience and social structures—emphasizing the embodied, situated, and dynamic nature of human–technology interactions.

References

Fingerhut, J. (2021). Enacting Media. An Embodied Account of Enculturation Between Neuromediality and New Cognitive Media Theory. Frontiers in Psychology, 12, 1–22.
Grusin, R. (2015) Radical Mediation. Critical Inquiry, 42(1), 124-148.
Kaipainen, M., Ravaja, N., Tikka, P., Vuori, R., Pugliese, R., Rapino, M., & Takala, T. (2011). Enactive Systems and Enactive Media: Embodied Human-Machine Coupling beyond Interfaces. Leonardo, 44(5), 433–438.
Prezioso, E., & Parisi, F. (2025). Extending the extended self: a mediational-constitutional proposal. Synthese, 205(3), 124.
Hansen, M.B.N., & Clarke, B. (eds.) (2009). Emergence and Embodiment. Duke University Press.
Tikka, P. (2008). Enactive Cinema. University of Art and Design, Helsinki.

The extended moral mind

Federico Bina[1]
[1]University of Genoa

This paper discusses the hypothesis of an extended moral mind (EMM), the idea that moral cognition—like other forms of cognition—is not confined into the brains of individual agents but can extend to institutions and technological artifacts that can shape, enable, and constrain it. Building on the extended mind hypothesis (Clark & Chalmers 1998) and its socially extended variants (Gallagher 2013), I analyze the descriptive plausibility of EMM and its implications on four main topics in contemporary ethics: moral testimony, the ethics of AI, individual moral enhancement, societal moral progress.

The first part of the paper situates EMM within the broader literature on cognitive extension and distribution (Heersmink 2017). Verbeek (2011) argued that moral agents often rely on external supports that can become constitutive parts of their deliberative processes. Consider for instance prenatal diagnostic techniques which allow parents' to make decisions about their child's life. In such cases, artifacts play a constitutive role in shaping moral decisions, but the agent retains decisional control. Similarly, Gallagher's hypothesis of the socially extended mind holds that social institutions (e.g., legal norms) can play an analogous role. Cognitive access to what social norms prescribe is often very cheap, and it directly affects people's decisions. If agent A were unaware of, or did not live within a given system of rules, they would think or act differently. By contrast, distributed systems involve interactions among multiple agents, humans and artifacts (Floridi 2013). Floridi's notion of "distributed morality" sees morally relevant outcomes as arising from such interaction in which no single agent bears full decisional control, hence no full moral responsibility. Extended and distributed moral cognition are distinct phenomena, but they both challenge anthropocentric and individualist assumptions about moral agency and raise the question of whether moral responsibility can be ascribed to other entities that are not persons, such as algorithms or institutions.

The second part of the paper addresses some key normative implications of accepting the EMM hypothesis. First, I consider the problem of moral testimony and advice. Is a moral decision based on following a set of rules, or advice from experts or LLMs, more or less autonomous and responsible than one made without such external input?

Second, I discuss the implications of EMM for moral enhancement and progress. Some authors (Macklin 1977; Sauer 2019; 2023) argued that moral progress does not involve individual moral improvement but only institutional change. Others (Buchanan & Powell 2018) defended the value of individual epistemic and motivational moral improvement. I argue that this is a false dichotomy and suggest that agents can improve in the latter sense by relying on external supports, provided they retain critical oversight, openness to learning and revision.

Finally, if we accept that moral agency can be extended, this raises important ethical and political questions about equitable access to the resources that can enable and enhance moral cognition. If such resources become key enablers of responsible moral agency, their unequal distribution may exacerbate epistemic injustices, especially when decision-making systems are opaque or biased.

References

Buchanan, A., Powell, R. (2018). The evolution of moral progress: A biocultural theory. Oxford UP.
Clark, A., & Chalmers, D. (1998). The extended mind. Analysis, 58(1), 7–19.
Floridi, L. (2013). Distributed morality in an information society. Science and Engineering Ethics, 19(3), 727–743.
Gallagher, S. (2013). The socially extended mind. Cognitive Systems Research, 25, 4–12.
Heersmink, R. (2017). Distributed cognition and morality: Agency, artifacts and systems. Science and Engineering Ethics, 23(2), 431–448.
Macklin, R. (1977). Moral progress. Ethics, 87(4), 370–382.
Sauer, H. (2019). Butchering benevolence: Moral progress beyond the expanding circle. Ethical Theory and Moral Practice, 22(1), 153–167.
Sauer, H. (2023). Moral teleology: A theory of progress. Routledge.
Verbeek, P. P. (2011). Moralizing technology: Understanding and designing the morality of things. Univ. of Chicago Press.

Contributed Symposia

Theorising Mixed Affect Across Philosophy, Psychology and Neuroscience

Michele Davide Ombrato[1], Valentina Cesari[2], Luca Cecchetti[2], Giada Lettieri[2], Marco Viola[3]
[1]Università di Bologna
[2]IMT School for Advanced Studies Lucca
[3]Università degli Studi Roma Tre

Mixed emotional affect poses a profound challenge for current theorising about emotions across different disciplines. While experiences such as the enjoyment of horror or tragic drama, the bittersweetness of nostalgia and the felt ambivalence toward events that we perceive as positive in some regards and negative in others show that mixed feelings are a central feature of our conscious emotional life, no consensus has emerged on whether such experiences should be taken at face value and on how to theorise them. Indeed, current models diverge sharply. Some models maintain that affective valence must occupy, at any given moment, a single point on a bipolar dimension of variation from pleasure to displeasure, rejecting the possibility of co-existence of positive and negative emotions. Others, such as bivariate or integrative models, argue that pleasure and displeasure are independent dimensions of variation, allowing for the simultaneous occurrence of discordant emotions. This symposium brings together philosophical investigation, theoretical critique, and empirical synthesis to advance our understanding of mixed emotional affect.

The first paper subjects to scrutiny the view that affective valence must occupy at any moment a single point on a bipolar dimension of variation, highlighting the implications of the view and the theoretical costs of explaining away mixed feelings as the result of rapid vacillation between discordant mental states. It then argues that defending this 'one at a time dogma' undermines the idea that affect is immediately consciously accessible and suggests that the view should be understood as a sub-personal model rather than an account of subjective experience.

The second paper examines the ontological commitments underlying core affect theories, questioning whether the rejection of mixed emotional affect is genuinely mandated by the bipolarity of valence or by the assumption that core affect is a simple, unitary state in conjunction with a constructivist ontology of emotions. It proposes that the 'vacillation hypothesis' to which various core affect theories subscribe rests on different grounds than normally assumed, and that the denial of the existence of mixed affect, once correctly reframed, can be seen to problematically overgeneralize, excluding the coexistence of congruent emotions as well as discordant emotions.

The third and last contribution consists in a systematic review of the neuroscientific literature on mixed affect structured by the introduction of a number of crucial yet much-neglected distinctions such as, the distinction between felt (first-person) and perceived (stimulus-centered) ambivalence, the distinction between unimodal or multimodal stimulation, and the distinction between studies which include a direct assessment of participants' feelings so as to verify that ambivalence was actually elicited, and studies in which such assessments are effected through the application of independent affective norms governing the selection of stimuli deemed capable of evoking contrasting feelings. The review reveals overlapping brain correlates for both

felt and perceived ambivalence but also exposes significant methodological inconsistencies and conceptual ambiguities in the literature, including inconsistent task definitions, lack of self-report assessments of ambivalence, and excessive the reliance on proxy measures, underscoring the need for greater precision in operationalizing ambivalence.
Together, these contributions illuminate the conceptual, empirical and methodological complexities involved in theorizing mixed emotional affect, and chart an interdisciplinary path forward for integrating philosophical insight with psychological and neuroscientific research.

§1. Affective Ambivalence and the One at a Time Dogma

'Affective valence' typically refers to the way an emotion or an affective state feels, that is, the quality of the (dis)pleasure we subjectively experience (Colombetti 2005). In some circumstances, such as when we enjoy tragic drama or riding a roller coaster, we seemingly feel both good and bad at the same time. Yet, several thinkers deny that these Ambivalent Affective States (AAS) are real psychological entities at the level of subjective experience. For instance, James Russell (2003) famously claimed that affective valence at each moment occupies a single point on a bipolar axis going from displeasure to pleasure. I dub this kind of position the "One at a Time Dogma" (OTD). To reconcile OTD with the widespread experience of AAS, ambivalent states are often 'explained away' as some sort of phenomenal illusion. A popular strategy – to which Russell also subscribes to (Russell 2017) – is to reinterpret the putative co-occurrence of pleasure and displeasure as rapid vacillations between positive and negative states, which are then misremembered as co-occurrent. While being viable, this strategy has theoretical costs. In this talk, we emphasize how, in his effort to defend the OTD against a set of well-known experiments where Larsen (Larsen et al. 2003; Larsen & McGraw 2011) provide experimental evidence of putative AAS, Russell ends up giving up on another tenet of his theory, namely that one's subjective affective state is 'consciously accessible as a simple, non-reflective feeling' (2003: 147). Hence, we contend that defending the OTD via the vacillation hypothesis requires giving up any notion of pre-reflexive affective phenomenology. Therefore, Russell's model of affect should no longer be interpreted as an accurate model of spontaneous phenomenology transparent to first-person introspection (cf. Charland 2005), but rather as a psychological construct at the sub-personal level.

§2. Mixed Affect, The Vacillation Hypothesis and Evaluative Switches

Felt emotional ambivalence and other putative cases of coexistence of discordant emotions are normal, and indeed, central aspects of our emotional life. Whether our experience in such cases is correct–i.e., whether discordant emotions literally coexist, or just seem to us to do so due to their rapid alternate occurrence (Russell 2017)–is, however, a much-debated question, with a long history across philosophy and psychology (cf. in philosophy, Hume, [1739/1740]2000, Book 2, "Of the Passions," part 1, "Of Pride and Humility," section 2; Scheler 1973, pp. 330-331; Moran 2001, pp. 180-181. In psychology, see Titchener 1908, pp. 46-47 and Young 1918). The broad aim of this paper is to clarify some of the issues that are at stake in this debate, especially in relation to the shape this debate has taken in psychology and affective sciences in recent years. I will argue that those who deny that discordant emotions may coexist, e.g., Barret & Bliss-Moreau (2009), Condon & Barrett (2013), Russell & Carroll (1999) and Russell (2017), are not actually forced to do so because of an independent theoretical commitment to the view that

valence is ‘bipolar,’ as they claim they are, and as is normally accepted, but because of the role that core affect understood as a simple, unitary state of (un)pleasure plays in their accounts of the ontology of emotions. Anticipating a bit my conclusions, what on core affect views may bar discordant occurrent emotions from coexisting is that the latter would have to be constituted in part out of the very same affective ‘stuff’, namely one’s core affect. Core affect, insofar as it is simple and unitary, cannot in fact bear polarly opposed features, by definition of ‘real polar opposition’ as a species of contrariety (Massin, 2014). In other words, discordant emotions cannot coexist because they would each owe their affective valence to their simultaneous participation to one and the same entity, which, as such, must be either positively or negatively valenced at any given instant. If this is correct, I suggest, the rejection of mixed emotional affect by core affect views overgeneralises to instances of coexistence between distinct but congruent emotions, i.e., it cannot be limited to discordant emotions. I will then suggest that the so-called ‘vacillation hypothesis’ may nonetheless be in some cases an appealing alternative to coexistence, but for different reasons from those provided by core affect theorists.

§3. Brain Correlates of Bittersweetness – A Systematic and Critical Review

Introduction: Mixed affect is characterized by the simultaneous existence of two emotional states, typically opposite in valence (Larsen et al., 2001). To illustrate, remembering a lost love may evoke feelings of warmth and sadness, promoting a sense of bittersweetness. Yet, the exact nature of mixed affect remains theoretically debated. According to Cacioppo’s bivariate model, emotions of opposite valence can occur simultaneously (Cacioppo et al., 1999, 2004). In contrast, Russell’s bipolar model posits that pleasantness and unpleasantness lie on a single continuum, making true simultaneity theoretically incompatible (Russell, 2003), unless one assumes rapid vacillation that mimics co-occurrence (Russell, 2017). More recently, Vaccaro and colleagues attempted to reconcile both views through a neurobiological perspective (Vaccaro et al., 2020). In their model, the brainstem supports Russell’s rapid-switching mechanism, while the insula integrates upstream signals into a genuine experience of mixed affect, as described in Cacioppo’s framework.

Which of these models best account for the existence of mixed affect remains an open question. Theoretical and methodological inconsistencies across studies have rendered it difficult to reach a definitive conclusion (Berrios et al., 2015; Kreibig & Gross, 2017). Notably, many studies fail to clearly distinguish between felt ambivalence (i.e., the subjective co-occurrence of pleasantness and unpleasantness) and perceived ambivalence (e.g., the recognition of conflicting emotional content, such as the mixed response to finishing second in a competition while a close friend wins).

Aim: This systematic review aims to characterize the neuroscientific literature on mixed affect, focusing on the brain correlates of felt (i.e., first-person) versus perceived (i.e., stimulus-centered) ambivalence. To provide an in-depth and comprehensive view of mixed emotions, our review considers several points. For instance, if the study objectives were to address felt versus perceived ambivalence, if ambivalence was evoked through unimodal or multimodal stimulation, and whether a direct assessment of participants’ feelings was included to verify ambivalence was elicited, or if independent affective norms guided the selection of stimuli deemed to evoke contrasting feelings. The objective of structuring the present review

along these core dimensions is to clarify how conceptual and methodological heterogeneity impacts the study of mixed affect and limits theoretical progress.
Methods: in performing the current systematic review, we followed PRISMA guidelines (Moher et al., 2009). Two automated searches were conducted in MEDLINE and Scopus up to December 2024, refined via Google Scholar forward citation tracking. Keywords included “mixed feelings,” “ambivalence,” “emotional conflict,” “emotional complexity,” “emotions,” “incongruence,” “ambiguity,” “MRI,” “fMRI,” “functional magnetic resonance imaging,” “EEG,” “PET,” and “brain mapping.”
Results: The first search strategy yielded 1,864 records (310 from databases, 1,554 from forward search), from which 16 studies were included. The second strategy identified 3,529 records (1,216 from databases, 2,313 from forward search), leading to the inclusion of 53 studies. In total, 69 studies were selected for qualitative synthesis. The distribution of studies based on sensory modality revealed a clear predominance of unimodal paradigms (n = 54) over multimodal ones (n = 15) in the experimental induction of mixed affect. Notably, out of 69 reviewed studies, 11 focused on first-person (i.e., felt) experiences of ambivalence through subjective judgments, while 50 examined stimulus-centered (i.e., perceived) ambivalence by assessing knowledge-based judgments of stimuli’s emotional properties. Two studies addressed both perspectives, and one used a passive viewing task without assessing either. Two studies lacked clarity on their focus, and three involved unrelated tasks. Regarding the assessment of ambivalence, studies on felt ambivalence rely more on explicit measures, including direct ratings of mixed feelings or confidence, while indirect measures infer ambivalence through separate ratings of pleasantness and unpleasantness. Stimulus-centered (i.e., perceived) ambivalence, instead, often relies on judgments of incongruence, affective norms, and implicit assessments.
The brain representation of felt and perceived ambivalence seems to lack specificity, as both first-person and stimulus-centered evaluation recruit overlapping regions, including the anterior and posterior cingulate cortex, and the ventrolateral and ventromedial prefrontal cortex, suggesting shared processing mechanisms regardless of subjective versus stimulus-centered focus.
Conclusions: This systematic review offers two major contributions: it provides a comprehensive neuroscientific synthesis of mixed affect, revealing a convergence of brain regions involved in both felt and perceived ambivalence. Also, it identifies significant methodological and theoretical issues, including inconsistent task definitions, the frequent lack of self-report assessments of ambivalence, and the reliance on proxy measures. In summary, these findings underscore the need for greater precision in defining and operationalizing ambivalence in affective neuroscience. Future research should aim to disentangle felt from perceived ambivalence more rigorously and clarify their distinct contributions to emotional experiences.

References

§1. Affective Ambivalence and the One at a Time Dogma
Charland, L. C. (2005). Emotion Experience and the Indeterminacy of Valence. In L. F. Barrett, P. M. Niedenthal, & P. Winkielman (Eds.), Emotion and consciousness (pp. 231–254). The Guilford Press.
Colombetti, G. (2005). Appraising valence. Journal of consciousness studies,12(8-9), 103-126.

Larsen, J. T., McGraw, A. P., & Cacioppo, J. T. (2001). Can people feel happy and sad at the same time?. Journal of personality and social psychology, 81(4), 684.
Larsen, J. T., & McGraw, A. P. (2014). The case for mixed emotions. Social and Personality Psychology Compass, 8(6), 263-274.
Russell, J. A. (2003). Core affect and the psychological construction of emotion.Psychological review, 110(1), 145.
Russell, J. A. (2017). Mixed emotions viewed from the psychological constructionist perspective. Emotion Review, 9(2), 111-117.

§2. Mixed Affect, The Vacillation Hypothesis and Evaluative Switches

Barrett, L. F., & Bliss‐Moreau, E. (2009). Affect as a psychological primitive. Advances in experimental social psychology, 41, 167-218.
Condon, P., & Feldman Barrett, L. (2013). Conceptualizing and experiencing compassion. Emotion, 13(5), 817.
Hume, D. 2000. A treatise of human nature, D. Norton and M. Norton, eds. Oxford: Oxford University Press.
Massin, O. 2014. “Pleasure and Its Contraries.” Review of Philosophy and Psychology 5 (1): 15-40.
Moran, R. (2001). Authority and estrangement: An essay on self-knowledge. Princeton University Press.
Russell, J. A. (2017). Mixed emotions viewed from the psychological constructionistperspective. Emotion Review, 9(2), 111-117.
Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. Psychological bulletin, 125(1), 3.
Scheler, M. 1973. Formalism in Ethics and Non-Formal Ethics of Value. Evanston: Northwestern University Press, trad. M. Frings, R.L. Funk.
Titchener, E. 1908. Lectures on the elementary psychology of feeling and attention. New York: The Macmillan Company.
Young, P.T. 1918. “An Experimental Study of Mixed Feelings.” The American Journal of Psychology 29: 237–271.

§3. Brain Correlates of Bittersweetness – A Systematic and Critical Review

Berrios, R., Totterdell, P., & Kellett, S. (2015). Eliciting mixed emotions: A meta-analysis comparing models, types, and measures. Frontiers in Psychology, 6. https://doi.org/10.3389/fpsyg.2015.00428
Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1999). The affect system has parallel and integrative processing components: Form follows function. Journal of Personality and Social Psychology, 76(5), 839–855. https://doi.org/10.1037/0022-3514.76.5.839
Cacioppo, J. T., Larsen, J. T., Smith, N. K., & Berntson, G. G. (2004). The Affect System: What Lurks below the Surface of Feelings? In Feelings and emotions: The Amsterdam symposium (pp. 223–242). Cambridge University Press. https://doi.org/10.1017/CBO9780511806582.014
Kragel, P. A., Sander, D., & LaBar, K. S. (2024). Can Brain Data Be Used to Arbitrate Among Emotion Theories? In Emotion Theory: The Routledge Comprehensive Guide. Routledge.

Kreibig, S. D., & Gross, J. J. (2017). Understanding Mixed Emotions: Paradigms and Measures. Current opinion in behavioral sciences, 15, 62–71. https://doi.org/10.1016/j.cobeha.2017.05.016
Larsen, J. T., McGraw, A. P., & Cacioppo, J. T. (2001). Can people feel happy and sad at the same time? Journal of Personality and Social Psychology, 81(4), 684–696.
Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Medicine, 6(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097
Russell, J. A. (2003). Core affect and the psychological construction of emotion. Psychological Review, 110(1), 145–172. https://doi.org/10.1037/0033-295X.110.1.145
Russell, J. A. (2017). Mixed emotions viewed from the psychological constructionist perspective. Emotion Review, 9(2), 111–117. https://doi.org/10.1177/1754073916639658
Vaccaro, A. G., Kaplan, J. T., & Damasio, A. (2020). Bittersweet: The Neuroscience of Ambivalent Affect. Perspectives on Psychological Science, 15(5), 1187–1199. https://doi.org/10.1177/1745691620927708

Is there hope for deliberative democracy yet? Socio-technical solutions to current political challenges

Fabio Paglieri[1], Carlo Proietti[2], Anna Mikhaylovskaya[3], Andreas Nitsche[4,5]
[1]Istituto di Scienze e Tecnologie della Cognizione, CNR Roma
[2]ILC-CNR Genova
[3]Rijksuniversiteit Groningen
[4]Association for Interactive Democracy Berlin
[5]Project LiquidFeedback

An important way in which human minds are connected is via collective decision-making, and a prominent view on how collective decisions can be best achieved is throughout deliberation. In the political arena, deliberative democracy has been practiced since ancient times, predating representative democracy by a far margin (e.g. in Athenian democracy), whereas it has become the object of theoretical scrutiny only recently: both implicitly, in the seminal work of scholars like Rawls (1971) and Habermas (1984), and more explicitly, with important contributions by Mansbridge (1980), Dryzek (1990), Fishkin (1991), Estlund (2007), Ross (2012), and Landemore (2020), among others.

In spite of the vibrant interest in deliberative democracy, the current political scenario does not invite optimism on the prospects of its success: in a world increasingly dominated by growing international tensions, ideological partisanship, polarization, disinformation, mass abstention from voting polls, widespread distrust towards politics, and the combined effect of weakened democratic institutions and the rise of “illiberal democracies” (Bell et al., 1995), invoking more discussion among citizens as a recipe to revive democracy might seem akin to waltzing on the Titanic while the ship is sinking. On the other hand, some argue that it is precisely in such “diabolical times” (Bächtiger & Dryzek, 2024) that deliberative democracy is most precious, since much of what ails human collectives nowadays stems out of the limitations of traditional decision-making systems, i.e. representative democracy.

This symposium intends to verify the plausibility of renewed hope in deliberation for improving democracy, leveraging recent work in cognitive sciences, agent-based modelling, and political studies. In particular, each presentation emphasizes different facets of the positive impact of deliberation on democratic outcomes: Paglieri discusses the multi-factorial nature of deliberation quality (i.e., what makes an act of deliberation valuable, and in what sense?) and its impact on participatory processes that remain accessible to people with limited resources to invest in politics; Proietti analyzes under what conditions deliberation may promote meta-consensus (i.e., agreeing on the nature of the options we are disagreeing about), instead of fostering polarization, clustering and entrenchment, using multi-agent modelling to capture these dynamics and the factors affecting them; Mikhaylovskaya and Rouméas argue that digital democratic innovations can help individuals develop virtuous qualities such as open-mindedness, critical thinking, empathy, and civic cooperation, and discuss what practical design features could facilitate the cultivation of these virtues; finally, Grossi and Nitsche focus on a specific form of deliberative democracy, i.e. liquid democracy, in which voting rights on specific issues can be delegated to other users, to demonstrate under what conditions such practice does not lead to unrestrained accumulations of power, but rather results in evolving representation patterns.

Deliberation quality and political participation: some musings on a complex socio-cognitive relationship
Fabio Paglieri

At a time when democratic institutions face unprecedented challenges across the globe (Dryzek et al., 2019), proposals for deliberative democracy (i.e., giving discourse and arguments a more central role in collective decision making) appear both more urgent and more fragile, since they need to demonstrate their feasibility under particularly demanding conditions or, as Bächtiger and Dryzek put it, in "diabolical times" (2024). A particularly relevant issue concerns how to measure the quality of deliberation, which is known to include multiple facets: (A) argumentative quality, in terms of correctness, cogency, degree of articulation, etc. of the arguments being exchanged in public debate; (B) epistemic quality, in terms of, e.g., information quality, ease of access; (C) procedural quality, in terms of, e.g., equal voice in the discussion, active participation, perceived democratic agency, tolerance for dissent; (D) outcome quality, both at the collective level (e.g., accuracy of decisions as in 'wisdom-of-crowds' settings, correctness of shared beliefs and their justifications) and at the individual level (e.g., level of understanding of the issue under discussion, critical appreciation of competing standpoints and interests); (E) spill-over effects, e.g. to what extent small-scale deliberative processes may benefit also mass democracy, by influencing the political attitudes of non-participants.
Existing approaches are either based on discourse ethics and focused on contents of individual contributions (e.g., the Discourse Quality Index, Steenbergen et al., 2003), or analyze deliberative quality as a property of relationships between individuals (e.g., the Deliberative Reason Index, Niemeyer et al., 2024): yet these proposals fail to account for some key trade-offs among different aspects of deliberation quality, in particular the fundamental tension between argumentative and epistemic sophistication on the one hand, mass participation and inclusion on the other (Fishkin, 2009; Elliott, 2023).
In this paper, I present a new conceptualization of deliberation quality, characterized by three distinctive features: (i) a multi-factorial structure, with specific predictions on what factors may be incompatible and therefore impossible to jointly maximize; (ii) an effort to turn each factor into one or more metrics, to allow for scalable assessment of complex deliberative processes; (iii) a commitment to develop a notion of deliberation quality compatible with "democracy for busy people" (Elliott, 2023), i.e. in which cognitive demands and efforts are kept to manageable levels, and various forms of participation (speaking, listening, voting, delegating, etc.) and non-participation (being invited, being aware, being ignorant, etc.) are contemplated, under the assumption that effective deliberation should benefit the whole polity and not just self-selected minorities.

Exploring meta-consensus via multi-agent models of opinion dynamics
Carlo Proietti

One major claim of supporters of deliberative democracy is that pre-decision discussion can be beneficial with respect to simple aggregation of preferences among options (voting), insofar as it helps settling differences among views and therefore contributes to more informed and legitimate group decisions. In the most optimistic reading, this would mean aligning individual

preferences among alternatives so to achieve consensual decisions. This type of outcome, however, does not seem realistic to achieve. In fact, experiments on deliberative polling (Luskin et al. 2002) show that, although deliberation increases mutual understanding of positions, it does not seem to induce consensus in such a strong sense. Nonetheless, as stressed by many, deliberation may instead help to achieve a form of meta-consensus that helps participants, if not to align their preferences, at least to find a common dimension on how to situate alternatives. In a nutshell, we may disagree e.g. whether economic growth or environmental preservation is preferable, but we may nonetheless agree that option A favors the first and option B favors the second, thereby "agreeing on how to disagree." This idea of meta-consensus grounds many recent attempts to quantify the effect of deliberation, as it is the case for the Deliberative Reason Index (Niemeyer et al. 2024). One reason why meta-consensus may be a positive social outcome is that it favors the structuring of individual preferences in ways that circumvent well-known problems of preference aggregation (majority voting) as the Condorcet paradox and Arrow's impossibility results (List 2018). This happens, e.g., when deliberation favors so-called single-peaked preferences (Black 1948).
Achieving meta-consensus presupposes however that individuals can reassess the conceptual space in which they locate alternatives via interaction and communication with others. Relevant questions are: how can such mechanism work? Under which conditions can it be effective? This brings us naturally to consider multi-agent models of opinion dynamics as a tool of inquiry. Such models allow to conduct different thought experiments by encoding the conceptual space of individuals as an ordered list of alternatives and investigate how interaction with consequent reshuffling of the order can generate uniformity. One promising tool in this sense is Axelrod's cultural dissemination model (Axelrod 1997) and its extensions. In this model, agents interact with their neighbors by modifying their cultural vector (lists of features expressed as numerical values). Modification occurs by agents reducing the (Hamming) distance of their cultural vectors from those of their neighbors. A similar adaptive mechanism can be applied to implement modifications on the individuals' conceptual spaces to see where meta-consensus arises and where, instead, clustering and entrenchment occurs. The present contribution lies the ground for this investigation providing the conceptual coordinates of this approach.

Nurturing virtues with digital democratic innovations
Anna Mikhaylovskaya & Élise Rouméas

Digital democratic innovations (DDIs) have been widely discussed by both academics and practitioners as a promising way to involve citizens in political decision-making processes on a larger scale. Extensive research has focused on how to design DDIs that are genuinely deliberative, inclusive, and capable of accurately representing people's needs. However, relatively little attention has been paid to how DDIs might nurture individual democratic virtues. We believe that this is a missed opportunity, and our paper aims to address this gap. We argue that the cultivation of democratic virtues should play a more central role when it comes to digital citizen participation. Specifically, we focus on how deliberative DDIs can help individuals develop virtuous qualities such as open-mindedness, critical thinking, empathy, and civic cooperation. Beyond providing a potentially more inclusive and democratic decision-making mechanism, DDIs should be seen as an opportunity for individuals to cultivate these virtues and

become better citizens. We broadly distinguish between two types of virtues—epistemic and moral—and explore how DDIs can nurture them in ways distinct from other forms of democratic engagement. While epistemic virtues enhance one's ability to process and evaluate political information, moral virtues foster respectful, inclusive, and empathetic engagement with others. We argue that both are crucial for the long-term success and legitimacy of democratic systems. To this end, we propose practical design features—such as structured deliberation formats, AI-generated prompts, and inclusive participation tools—that could facilitate the cultivation of these virtues. By conceptualizing DDIs not just as tools for decision-making but as spaces for civic transformation, we offer a novel normative perspective that reimagines the role of digital participation in democracy.

Rethinking influence in liquid democracy: delegations as evolving representation patterns

Davide Grossi & Andreas Nitsche

Liquid democracy is a mechanism for the division of labor in decision-making through the transitive delegation of influence. In essence, all individuals possess the autonomy to determine the issues with which they will engage directly, while for other matters, they may appoint a representative of their choosing.

We propose a novel approach to assessing the influence of voting nodes in a transitive delegation graph, taking into account the process nature of real-world liquid democracy in which delegation and voting are distinct and increasingly independent activities. The focal point of this analysis is transitivity, a foundational component of liquid democracy. Transitivity, defined as the capacity to transfer acquired authority to another entity, has been identified as a catalyst for concerns regarding an unrestrained accumulation of power.

However, when examining liquid democracy as a process, rather than a static delegation graph, it can be demonstrated that transitivity may in fact contribute to the regulation of power.

The manner in which a liquid democracy system functions, and consequently its dynamic characteristics, are contingent upon the delegation model and its implementation. Consequently, the present analysis pertains to a prototypical reference model of liquid democracy and its implementation in LiquidFeedback, a widely adopted online platforms for liquid democracy (https://liquidfeedback.com/).

The following argument will be put forward regarding liquid democracy as implemented in LiquidFeedback: While maintaining the one-person, one-vote paradigm for all votes cast, the anticipated influence of an agent, to the extent it is stemming from transitivity, experiences a precipitous decline following an exponential trajectory. In this context, the emergence of representation patterns will be examined, the topic of delegation cycles will be revisited, and the notion of gurus will be challenged.

While this approach does not claim universal validity, it introduces a crucial perspective that is indispensable for evaluating the ramifications of implementation decisions, measures intended to curtail power, and alternative delegation models. It will be exemplified how interventions with commendable intentions can yield unintended consequences.

It is our objective to establish a foundation for an in-depth examination of the dynamic nature of liquid democracy and, moreover, the temporal dimension of liquid democracy as a learning system. These aspects have not yet been explored within the existing academic literature.

References

Axelrod, R. (1997). The dissemination of culture: a model with local convergence and global polarization. Journal of Conflict Resolution, 41(2), 203–226.

Bächtiger, A., & Dryzek, J. (2024). Deliberative democracy for diabolical times. OUP.

Bell, D. A., Brown, D., Jayasuriya, K., & Jones, D. M. (1995). Towards illiberal democracy in Pacific Asia. Palgrave Macmillan UK.

Black, D. (1948). On the rationale of group decision-making. Journal of Political Economy, 56(1), 23-34.

Dryzek, J. (1990). Discursive democracy: politics, policy, and political science. Cambridge University Press

Dryzek, J., Bächtiger, A., Chambers, S., Cohen, J., Druckman, J. N., Felicetti, A., ... & Warren, M. (2019). The crisis of democracy and the science of deliberation. Science, 363(6432), 1144-1146.

Elliott, K. (2023). Democracy for busy people. The University of Chicago Press.

Estlund, D. (2007). Democratic authority: a philosophical framework. Princeton University Press.

Fishkin, J. (1991). Democracy and deliberation: new directions for democratic reform. Yale University Press.

Fishkin, J. (2009). When the people speak. Deliberative democracy and public consultation. OUP.

Habermas, J. (1984). Theory of communicative action, volume I: reason and the rationalization of society. Beacon Press.

Landemore, H. (2020). Open democracy: reinventing popular rule for the twenty-first century. Princeton University Press.

List, C. (2018). Democratic deliberation and social choice. The Oxford handbook of deliberative democracy, 463-489.

Luskin, R. C., Fishkin, J. S., & Jowell, R. (2002). Considered opinions: Deliberative polling in Britain. British Journal of Political Science, 32(3), 455-487.

Mansbridge, J. (1980). Beyond adversary democracy. University of Chicago Press.

Niemeyer, S., Veri, F., Dryzek, J. S., & Bächtiger, A. (2024). How deliberation happens: enabling deliberative reason. American Political Science Review, 118(1), 345-362.

Rawls, J. (1971). A theory of justice. Belknap Press.

Ross, C. (2012). The leaderless revolution: how ordinary people will take power and change politics in the twenty-first century. Blue Rider Press.

Schelling, T. C. (1978). Micromotives and macrobehavior. WW Norton & Company.

Steenbergen, M., Bächtiger, A., Spörndli, M., & Steiner, J. (2003). Measuring political deliberation: a discourse quality index. Comparative European Politics, 1, 21-48.

Cognitive network science for modeling the mind: theories, tools, and insights

Katherine Abramski[1], Massimo Stella[2], Alexis Carrillo Ramirez[2], Salvatore Citraro[3]
[1]University of Pisa
[2]University of Trento
[3]Institute of Information Science and Technologies , National Research Council CNR Pisa

Overview

Cognitive network science is an emerging field at the intersection of computer science, cognitive science, psychology, and artificial intelligence. It is a multidisciplinary approach that applies the tools of network science for understanding cognitive phenomena. Under this framework, cognition is modeled as a complex system of nodes and edges, such that nodes represent entities – like words or concepts – and edges represent relationships between them. Modeling cognition in this way enables investigations of the complex interactions that lead to the emergence of cognitive phenomena. In this symposium, we present recent works that apply cognitive network science for understanding cognition primarily in humans, but also in LLMs. These works are a collection of theoretical frameworks, methodological tools, and empirical investigations that provide insights on a wide variety of topics, including lexical processing, syntax and semantics, creativity, emotion, mental health, and implicit biases. We hope this symposium will be ofinterest to researchers across different fields who are interested in cognitive modeling, or to anyone interested in insights about the complex nature of language and cognition.

Multilayer cognitive networks: The case study of EmoAtlas and other applications in cognitive science and artificial intelligence

Massimo Stella

Cognitive multilayer networks can map multiple types of information at once, thus capturing how different layers of associations might co-exist within the mental lexicon and influence cognitive processing. This review starts with a gentle introduction to the structure and formalism of multilayer networks. We then discuss quantitative mechanisms of psychological phenomena that could not be observed in single-layer networks and were only unveiled by combining multiple layers of the lexicon [1]: (i) multiplex viability highlights language kernels and facilitative effects of knowledge processing in healthy and clinical populations; (ii) multilayer community detection enables contextual meaning reconstruction depending on psycholinguistic features; (iii) layer analysis can mediate latent interactions of mediation, suppression, and facilitation for lexical access. We give emphasis to EmoAtlas [2], a Python package for building cognitive multilayer networks of syntactic, semantic and emotional associations between words in texts. EmoAtlas combines interpretable artificial intelligence (AI) for syntactic parsing in 18 languages and psychologically validated lexicons for detecting the eight emotions in Plutchik's theory. We show that EmoAtlas can match or surpass transformer-based natural language processing techniques, BERT or large language models like ChatGPT 3.5 or LLaMAntino, in detecting emotions from Italian and English online posts and news articles (e.g., achieving 85.6% accuracy in detecting anger in posts vs the 68.8% value of ChatGPT and 89.9% value for

BERT). Combining BERT's semantic features with EmoAtlas' emotional/syntactic networks of words gets substantially better at estimating creativity rates of 1071 stories. This indicates an interplay between the creativity of narratives and their semantic, emotional, and syntactic structure. By outlining novel quantitative perspectives where multilayer networks can shed light on cognitive knowledge representations, including in next-generation brain/mind models, we discuss key limitations and promising directions for cutting-edge future research.

Textual forma mentis networks bridge language structure, emotional content and psychopathology levels in adolescents
Alexis Carrillo Ramirez

This research introduces a novel AI framework that uses language to predict psychopathology dimensions in adolescents. We performed natural language processing to analyze structured interviews from 232 adolescents within the Healthy Brain Network project. In these interviews, adolescents discussed emotionally charged topics, providing rich linguistic data for our analysis. Our methodology centers on textual forma mentis networks (TFMNs), a cognitive-inspired AI approach. TFMNs model the relationships between words, capturing syntactic, semantic, and emotional associations. These networks act as a proxy for the associative knowledge within the mental lexicon. By extracting network features and
emotional profiles from TFMNs, we created predictors for psychometric scores related to social maladjustment, internalizing behaviors, and neurodevelopmental risk. We employed Random Forest and XGBoost regression models to identify significant links between language features and clinical assessments. Our findings revealed that linguistic patterns can be used to predict psychopathology levels. For instance, higher modularity and a core-periphery structure in language networks were associated with increased social maladjustment. We also observed a connection between internalizing symptoms and linguistic expressions of disgust, potentially indicating rumination. Conversely, reduced efficiency in conceptual integration within language was linked to neurodevelopmental risk. This study demonstrates the potential of cognitive network science and AI to extract meaningful psychological information from language. This approach offers a new avenue for assessing adolescent mental health, with potential applications in early detection and intervention within clinical and educational settings.

SpreadPy: Simulating diffusion and superdiffusion in the mental lexicon with cognitive networks
Salvatore Citraro

We introduce SpreadPy as a Python library for simulating activation spreading in cognitive single- and multi-layer networks. Our tool is designed to perform numerical simulations testing structure-function relationships in cognitive processes. By comparing simulation results with grounded knowledge modelling theories, SpreadPy enables systematic investigations of how activation dynamics reflect cognitive, psychological and clinical phenomena. We demonstrate the library's utility through three case studies: (i) Simulations of creativity tasks show that activation trajectories vary with task difficulty, exposing how cognitive load modulates lexical access; (ii) Activation spreading on associative knowledge networks distinguishes students with

high versus low math anxiety, revealing anxiety-related structural differences in conceptual organization; (iii) In individuals with aphasia, simulated activation patterns on lexical networks correlate with empirical error types (semantic vs. phonological) during picture-naming tasks, linking network structure to clinical impairments. SpreadPy's flexible framework allows researchers to model these processes using empirically derived or theoretical networks, providing mechanistic insights into individual differences and cognitive impairments. The library is openly available with documentation for reproducibility in psychology, neuroscience, and education research.

The influence of semantic memory structure on implicit biases in humans and LLMs
Katherine Abramski

Both humans and large language models (LLMs) exhibit harmful implicit biases, so understanding the cognitive mechanisms that underlie such biases can be beneficial to society. According to the dual process theory of cognition, implicit biases in humans emerge largely from our associative "System 1" thinking, while our deliberative "System 2" thinking can reduce biases by adopting a more objective point of view [3]. LLMs on the other hand are "System 1" thinkers by design, so they lack the capacity to mitigate their own biases the same way humans do. Some theories suggest that divergences between "System 1" and "System 2" thinking can be attributed to structural differences in semantic memory [4], but
the relationship between semantic memory structure and implicit biases remains relatively unexplored. In this paper, we investigate this relationship in both humans and LLMs to better understand the cognitive mechanisms that underlie biased reasoning. We use data generated by humans and LLMs to build multilayer semantic network models that represent three different types of semantic memory structures, based on 1) associations, 2) definitions, and 3) categorical relations. We investigate implicit biases by simulating semantic priming within the multilayer networks. Our results highlight important differences between humans and LLMs. First, we find that the three different types of semantic memory structures are irreducible in humans but not in LLMs, suggesting that LLMs lack the human reasoning capabilities necessary to exhibit a richer, more varied understanding of concepts. Second, in humans but not in LLMs
we find a clear and consistent relationship between implicit biases and semantic memory structure. This suggests that structural differences in semantic memory influence biased reasoning in humans but not in LLMs, providing evidence for the absence of "System 2" reasoning in LLMs.

References
1. Stella, M., Citraro, S., Rossetti, G., Marinazzo, D., Kenett, Y. N., & Vitevitch, M. S. (2024). Cognitive modelling of concepts in the mental lexicon with multilayer networks: Insights, advancements, and future challenges. Psychonomic Bulletin & Review, 1-24.
2. Semeraro, A., Vilella, S., Improta, R., De Duro, E. S., Mohammad, S. M., Ruffo, G., & Stella, M. (2025). EmoAtlas: An emotional network analyzer of texts that merges psychological lexicons, artificial intelligence, and network science. Behavior Research Methods, 57(2), 77.
3. Kahneman, D. (2011). Thinking, fast and slow. Macmillan.

4. Strack, F., Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. Personality and social psychology review, vol. 8, no. 3, pp. 220–247.

Contributed Posters

More Than Just 'Good': Modeling the Politeness of Costly Intensifiers in Compliments

Carlotta Marianna Cascino[1]
[1]École Normale Supérieure - Paris

Why do speakers sometimes use needlessly strong words to praise something only mildly good? We propose that costly overstatement—via longer or rarer intensifiers—serves as a rational politeness strategy. Building on Bennett & Goodman's (2018) data on intensifiers' costs and meanings and on the Rational Speech Acts framework for politeness (Yoon et al., 2020), we model how speakers strategically select intensifiers to balance informativeness, social intent, and communicative effort. Our preliminary simulations show that costlier intensifiers ("incredibly good" vs "very good") are favored when speakers aim to be both kind and informative, suggesting that overstatement functions as a politeness strategy.

Sex Differences in Intelligence Quotient Across Developmental Stages: A pre-registered Meta-Analysis of WPPSI, WISC, WAIS

Leognano Ceraudo[1], Eva Tolomeo[1], Marco Tullio Liuzza[2]
[1]Department of Health Sciences, University "Magna Graecia", Catanzaro, Italy
[2]Department of Developmental Psychology and Socialization, University of Padova, Padova, Italy

Sex differences in Intelligence Quotient (IQ) have been extensively examined, but results appear inconsistent. The Developmental Theory of Sex Differences in Intelligence (DT-SDI) might be one explanation of these inconsistencies. This pre-registered meta-analysis aims to evaluate the evidence for this theory by quantitatively synthesizing studies that assess sex differences in Full-Scale Intelligence Quotient (FSIQ) of Wechsler scales. This meta-analysis will adhere to the PRISMA guidelines, and eligible studies will be identified through systematic searches across three databases: PubMed, Scopus, and APA PsycInfo.The findings of this meta-analysis will provide a more comprehensive understanding of sex-related differences in IQ across developmental stages.

Towards Understanding in Large Language Models: Investigating Reasoning Abilities and Behavior using Cognitive Reflection Tests

Dimana Orlinova Anastassova[1]
[1]Centro di Ricerca E.T.h.O.S., Università degli Studi di Verona

This study explores Large Language Models' (LLMs) reasoning skills, a crucial aspect of their understanding. Existing arguments suggest LLMs' understanding limits arise from functional linguistic competence deficits, including reasoning and world knowledge. We employed

Cognitive Reflection Tests (CRTs) to assess LLM reasoning abilities, comparing their performance and error patterns to human subjects. Results indicate LLMs achieve high CRT accuracy, showing some human-like reasoning. However, distinct error patterns highlight current LLM limitations in mimicking human cognitive processes. These insights guide future LLM design for improved functional linguistic competence.

SDF-FuzzIA: a DSS based on LLMs and Fuzzy Ontologies

Marco Elio Tabacchi[1], Lydia Castronovo[2], Giuseppe Filippone[1], Giuseppe Giacopelli[1], Gianmarco La Rosa[1], Manuel Mancini[1], Arianna Maria Pavone[1]
[1]DMI e SCo^2, Università degli Studi di Palermo
[2]MIFT, Università degli Studi di Messina

The proliferation of AI algorithms based on LLMs has highlighted, alongside considerable potential, questions regarding their adaptability to the generality of solvable problems. Among these are explainability, introspectiveness, and precision when applied to information that is not strictly linguistic in nature.
This work contains a description of SDF-FuzzIA, the AI module of the SDF project for consolidation and exploration of geothematic data. The purpose of SDF-FuzzIA is to integrate a GPLMM with a Fuzzy ontology, ensuring that the suggestions which constitute the core of its DSS are grounded in concrete facts.
As SDF is a work in progress, results will not be presented; rather, design and implementation choices will be discussed.

Cognitive Artifacts and their influence on Human Cognition and Learning

Martina Barbera[1], Anna Re[2]
[1]Department of Cognitive Science, Psychology, Educational and Cultural Studies (COSPECS), University of Messina
[2]Institute for Educational Technology, National Research Council, via Ugo La Malfa 153, 90146, Palermo, Italy

According to the extended mind theory, cognition can extend beyond the brain to tools that influence human cognitive processes, such as technological artifacts. Closely linked to Artificial Intelligence, these tools not only emulate but enhance reasoning, learning, and problem-solving. More than supporting human activity, they reshape how cognition interacts with the world, challenging the boundary between mind and environment. The central question is whether such tools can lead to the creation of a “system 0,” marked by minimal cognitive effort, or, if effectively integrated, become true extensions of human thought and learning.

Embodied Time: Toward the Establishment of a New Paradigm for the Cognitive Sciences

Emilio Leone[1]
[1]Università degli studi di Messina

This work proposes a new paradigm of embodied time in cognitive science, emphasizing temporality as dynamically shaped by the interaction between mind, body, and environment. Drawing on phenomenology, it explores Husserl's notion of time as a continuous flow (retention, impression, protention) and Merleau-Ponty's ontological levels—matter, the living, and the human. It also examines recent studies on the sense of agency, showing how time perception is co-constituted through embodied action. The proposal integrates theoretical and empirical approaches to reconceptualize time as lived and enacted, offering new directions for research on temporality, subjectivity, and cognition.

Emotional Evaluation, "For-Me"-Ness and A Sense of Self

Rituparna Roy[1]
[1]Ruhr University Bochum

Emotion has a unique sui generis phenomenology that is fundamentally evaluative. Cognitive theories fail to understand the unique phenomenology (what-it-is-likeness) of emotional experience. First, I will focus on the value-laden features of emotion. I will cover the evaluative-attitudinal account of emotion. Then I shall criticize the cognitive theories of emotion as they are not consistent with the notion of the subject of experience. Then I will offer an alternative take on emotional evaluation that focuses on the felt aspect over epistemology and will grasp the underlying similarities between will, emotion, and pleasure or pain have been ignored in the contemporary philosophy of mind. Finally, I will shed light on the fact why do various individuals have different emotional reactions to the same incident? The brief response is that they do not possess the same as the subjective aspect, even though the phenomenal aspect is analogous. The "self" I am referring to is situated in the body and as a whole, it belongs to a given contextual background such as geographical location, time, community, culture, and philosophy.

Investigating Neural Responses to Narrative Voice Perspective (First vs. Third Person): An ERP Study

Ines Adornetti[1], Valentina Deriu[1], Alessandra Chiera[1], Daniela Altavilla[1], Federica Palumbo, Elisa Valiante[1], Francesco Ferretti[1]
[1]Roma Tre University

This ERP study investigates how narrative voice (who told the story) influences neural processing during story comprehension. 36 participants read short narratives written in either the first or third person, with target verbs varying in semantic congruence. Two primary results emerged. First, incongruent condition elicited a greater N400 response, indicating increased cognitive effort in processing regardless of narrative voice. Second, a hemispheric difference in the P600 component was found in the third-person condition, suggesting differential reanalysis or integration processes. Overall, the study confirms the role of semantic congruence in narrative comprehension but does not support a functional distinction between first- and third-person narratives at the neural level.

Satellite event

Embodied and enactive approaches to cognition

Shaun Gallagher – *September 19, 15:00, Cappella Guinigi*

Che cos'è la cognizione umana? Quanto conta il corpo? E quanto è cruciale la nostra interazione con l'ambiente?
Questo saggio offre un approccio accessibile e pragmatico a tali domande attraverso la lente di una forma originale di enattivismo radicale — capace di integrare molte delle proposte provenienti dal vasto campo degli studi sulla cognizione incarnata. Qui, il sistema cognitivo è concepito come un tutto inscindibile di cervello, corpo e ambiente, dove il corpo comprende l'intero organismo non neurale in relazione dinamica con il proprio contesto e orientato all'azione. Radicata nelle tradizioni filosofiche della fenomenologia e del pragmatismo, l'analisi si confronta con le più recenti sperimentazioni e riflessioni, spaziando dalla psicologia alle neuroscienze fino alla robotica.

What is human cognition? How much does the body matter? And how crucial is our interaction with the environment? This essay offers an accessible and pragmatic approach to these questions through the lens of a distinctive form of radical enactivism—one that integrates many of the proposals from the wide field of embodied cognition studies. Here, the cognitive system is conceived as an inseparable whole of brain, body, and environment, where the body includes the entire non-neural organism in dynamic relation with its surroundings and oriented toward action. Grounded in the philosophical roots of phenomenology and pragmatism, the analysis engages with the most recent experiments and reflections, spanning psychology, neuroscience, and robotics.