UNICA UNIVERSITÀ DEGLI STUDI DI CAGLIARI

This is the Author's [*submitted*] manuscript version of the following contribution:

[Pietro Salis and Matteo Da Pelo, Chatting with LLMs: neither cognition nor communication, *Reti, Saperi, Linguaggi. Italian Journal of Cognitive Sciences*, 1/2026, pagg. 41-64]

The publisher's version is available at:

http://dx.doi.org/DOI: 10.12832/121317

When citing, please refer to the published version.

This full text was downloaded from UNICA IRIS https://iris.unica.it/

Chatting with LLMs: neither cognition nor communication

Abstract

Large Language Models (LLMs) exhibit impressive linguistic performance, yet their cognitive and communicative status remains controversial. This paper argues that behavioural fluency alone cannot justify cognitive ascription unless assessed within a design-sensitive framework. We distinguish structurally constrained functionalism from unconstrained functionalism and propose a minimal cognitive core—environmentally coupled perception, inference, and action—as a baseline for cognition. Against this baseline, current LLMs fall short: as standalone text-based systems, they lack sensorimotor coupling, integrated inferential responsiveness, and autonomous agency. We examine sycophancy and the absence of spontaneity as diagnostic consequences of architectures optimised for user alignment rather than epistemic commitment. Finally, we develop the deprived-interface argument: without embodied or interactive access to an environment, LLMs cannot revise or ground their representations, resulting in a derivative form of linguistic idealism. We conclude that LLMs simulate cognition and communication without instantiating the structural conditions required for either, and we outline the implications for AI evaluation and cognitive ascription.

Keywords: analytical philosophy of AI, cognition, communication, functionalism, LLMs, structuralism, unconstrained functionalism

1. Introduction

The recent success of Large Language Models (LLMs) has reignited interest in artificial intelligence (AI) within cognitive science and neuroscience, particularly with respect to how language is processed in artificial and biological systems. Transformer-based architectures trained on vast corpora now exhibit levels of linguistic fluency and behavioural plausibility that were scarcely imaginable only a few years ago (Bick et al., 2024; Brown et al., 2020). Their rapid diffusion, comparable in scope to that of the personal computer, has further amplified their epistemic and societal salience.

In parallel, several studies have reported that LLMs achieve remarkable performance in next-word prediction in naturalistic contexts, in some cases approaching or exceeding human-level benchmarks (Goldstein et al., 2022). More strikingly, internal representations extracted from these models can be linearly mapped onto neural responses observed during human language processing, suggesting representational similarities between artificial and biological systems (Abnar et al., 2019; Schrimpf et al., 2021; Toneva & Wehbe, 2019; Hosseini et al., 2023). Recent work has extended these findings by showing that LLM-derived representations allow accurate encoding and decoding of brain activity during linguistic tasks (Tuckute et al., 2022). Such results have lent credence to the view that LLMs may provide insight into human cognition or even contribute to the construction of a unified model of cognition (Binz et al., 2025).

Based on these behavioural and representational similarities, some authors have gone so far as to suggest that systems such as ChatGPT appear to pass variants of the Turing Test (Jones & Bergen, 2025). These claims, however, bring into sharp focus a fundamental methodological difficulty. The inference from linguistic performance to cognition is not merely empirically ambitious but conceptually underdetermined unless one makes explicit, and defends, a theory of cognitive design. Without such a justification, apparent continuity between LLM behaviour and human cognition risks being merely an artefact of evaluation protocols rather than a true indicator of shared underlying mechanisms.

This problem is compounded by the tendency to infer cognitive capacities directly from expressive fluency (Kidd et al., 2023). Certain abilities displayed by LLMs may appear, at first glance, difficult to explain without invoking human-like cognition. Yet this appearance alone does not warrant such an inference. As several authors have noted, the success of LLMs has often been used to motivate conclusions about cognition without an explicit account of the design assumptions that would license

those conclusions (Collins et al., 2024; Binz and Schulz, 2023; Binz et al., 2025; Niu et al., 2025).

A contrasting and increasingly influential line of critique resists this inference altogether, characterising LLMs as "stochastic parrots": systems that reproduce linguistic regularities without understanding, grounding, or cognitive control (Bender et al 2021; Leivada et al., 2025). While forceful, this critique risks being read as merely dismissive unless it is embedded within a clearer theoretical framework specifying what cognition minimally requires and why LLMs fail to meet those requirements.

To that end, we situate the debate within a long-standing distinction in AI research between human-inspired and machine-oriented approaches to design. This distinction is often articulated as a contrast between cognitively motivated design, rooted in the cybernetic tradition, and functionalist design, which underpins most contemporary LLM architectures. Proponents of the former research programme impose structural constraints on artificial systems to better capture the organisation of human cognition (Webb, 2001; Cordeschi, 2002; Lieto, 2021). By contrast, modern LLMs exemplify a largely unconstrained functionalism, in which behavioural success is pursued independently of structural plausibility.

Within this context, the label "Cognitive AI" is not merely descriptive but normative: it applies if and only if certain criteria are satisfied. Computational Cognitive Science is explicitly concerned with the design and evaluation of such models, aiming to explain cognition by integrating functional performance with structural constraints (Simon, 1980; Simon, 2019). From this perspective, functional equivalence alone is not sufficient to ascribe cognitive capacities, since cognitive capacities are realised through specific organisational principles that support perception, inference, and action in an integrated and environmentally coupled manner (Cuskley et al., 2024).

The central claim of this paper is that, when evaluated against even a minimal conception of cognition, current LLMs do not qualify as cognitive systems. Although they may successfully simulate linguistic behaviour, they lack the structural and epistemic conditions required for cognition and, by extension, for genuine communication (Block, 1981). Clarifying this point requires moving beyond surface-level behavioural comparisons and adopting a design-sensitive framework capable of distinguishing simulation from instantiation.

In the remainder of this paper, Section 2 introduces the distinction between structuralism and functionalism in cognitive design. Section 3 presents a minimal cognitive core, ensuring that our critique does not rest on an unargued notion of cognition. Section 4 examines well-known limitations of LLMs in light of this framework, while Section 5 analyses their lack of spontaneity and autonomous agency. Section 6 argues that their interface with the world is structurally deprived, reducing their epistemic horizon to the training dataset. Section 7 concludes by drawing out the broader implications for cognitive ascription and AI evaluation.

2. Structural constraints for functionalism on artificial systems

We believe that LLMs are not cognitive systems because they lack the structural mechanisms characteristic of natural cognitive systems. To clarify this claim, we distinguish two approaches to cognitive design: functionalism and structuralism (Lieto, 2021). Functionalism models cognition by focusing on input–output behaviour, aiming for functional equivalence between genuine cognitive faculties and artificial mechanisms (Putnam, 1960). Structuralism, by contrast, requires a principled alignment between the internal organisation of artificial systems and that of natural cognitive systems (Cordeschi, 2002).

From a structuralist perspective, the architectural dissimilarity between LLMs and biological cognition undermines the explanatory value of any superficial functional

similarity. Functional equivalence alone does not warrant cognitive status if the underlying mechanisms remain fundamentally distinct. Just as parrots mimic human speech without possessing genuine inferential or conceptual capacities, LLMs likely simulate linguistic behaviour without the mechanisms required for cognition (Bender et al., 2021).

A plausible assessment of artificial cognition must therefore integrate functional performance with structural constraints, rather than relying on behavioural equivalence alone. We propose what we call this structurally constrained functionalism, in contrast to the unconstrained functionalism that dominates much of the current discourse. We acknowledge that many functionalist theorists like Hilary Putnam and Jerry Fodor were aware of the importance of structural constraints for functionalism. But we perceive a tendency in commentators to see behavioural outputs as sufficient to signal cognition. Our use of the expression "unconstrained functionalism" is not intended to refer to the more sophisticated forms of functionalism found in contemporary literature, such as mechanistic functionalism (Piccinini, 2007b; Craver, 2007), Marr-inspired multi-level pluralism (Marr, 1982; Bechtel, 2007), or predictive-processing models (Clark, 2013, 2016; Hohwy, 2013). These approaches incorporate structural and mechanistic constraints that clearly distinguish them from the purely behavioural functionalism we criticise. Our target is the tendency, widespread in many discussions of LLMs and cognition, to infer cognitive capacities from performance equivalence alone, without considering the architecture that produces such performance. Unconstrained functionalism risks committing what we term an ascription fallacy: attributing cognitive status to systems merely because they exhibit human-like outputs, without regard for the architecture that produces them—much as one might mistake a parrot's mimicry for genuine linguistic competence. Functionalism in isolation cannot be sufficient to ascribe cognition, because cognitive functions are not free-floating behavioural patterns; contrary to what is often assumed under the banner of "multiple realisability", they

are realised through specific structural mechanisms that enable perception, inference, and action in an integrated and environmentally coupled manner (see below). Appeals to multiple realisability within functionalism tend to support the view that cognitive functions are specification-level properties that can, in principle, be realised by a wide range of different architectures, irrespective of their structural or material substrates. By contrast, approaches grounded in physically realised computation avoid this difficulty by imposing explicit structural constraints on the mechanisms that implement cognitive functions (Piccinini, 2007a, 2008, 2015; Piccinini and Craver, 2011). Absent such mechanisms, behavioural similarity remains a superficial artefact of design rather than an indicator of genuine cognitive capacity. One main direction in recent research that provides a wider rationale for this constrained view focuses on what is called a minimal cognitive grid (MCG), which considers functional and structural features as aspects along a spectrum of features exhibited by artificial systems (Lieto, 2021). The grid provides a fine-grained analysis of such systems by considering many different functional and structural parameters. If behavioural equivalence alone cannot justify cognitive ascription, minimal structural conditions must be determined that any candidate cognitive system needs to satisfy.

3. Cognition not simply presupposed: a minimal cognitive core

A common objection to our perspective is that it merely "presupposes" a definition of cognition and that such presupposition is illegitimate given the contested nature of the concept. Cognition, the objection goes, cannot simply be assumed, because it is notoriously controversial to determine what constitutes cognition and what does not. We acknowledge this concern and respond by proposing a minimal, broadly acceptable core for cognitive activity. This minimal core can be regarded as a noncontroversial baseline for evaluating candidates for cognitive status, since it avoids theoretical excess while capturing the essential conditions for inferential responsiveness. While it may be argued that cognition in the full sense involves

higher-level functions such as conceptual thought, this minimal core can serve as a reference point for evaluating candidates for cognitive capacities.

We do so by downgrading a functionalist perspective provided by Wilfrid Sellars for conceptual thought in interaction with perception and action. Drawing on this, we follow Sellars in treating semantic statements as functionally classifying ordinary ones, which was Sellars' central point (Sellars, 1974). He aimed to isolate the role certain statements play in reasoning and their connection with perceptual episodes and courses of action. Sellars stated that we have language-entry transitions when perceptual episodes license verbal activity, intralinguistic transitions when we infer one statement from another statement, and language-exit transitions when we act upon a pattern of inference. Robert Brandom later leveraged this scheme to argue for a broad account of inferential articulation, wherein perception and action are integrated in the ways we utilise conceptual content (Brandom, 1994, 2000). On this basis, but with a different goal, we can develop what we see as a loosened version of that scheme that addresses cognition itself, while setting aside the higher-level limits of conceptual activity. We will now clarify this progression.

On this inferentialist view, conceptual cognition works as a functional model: it relies on inferential links between contentful episodes and extends these links to include perceptual episodes as inputs and actions as outputs. But we can adapt such a model to cover lower levels of cognition by entirely shifting the focus away from concepts and conceptual thought. What, then, remains of cognition once we strip away the focus on higher-level conceptual cognition? What is the least we must demand of a system before considering it cognitive? It is crucial to understand "the inferences" that begin with perceptual inputs and lead, through inferential transitions, to behavioural outputs. Another shift concerns the approach to inference: in the original model provided by Sellars and Brandom, it is something inherently normative and conceptual—this normative feature has also been invoked to counter scepticism about AI capacities (Malík & Hubálek, 2025), but that line of argument does not affect the

present, non-normative version of the model; in the loosened version presented here, it is something generically functional and subpersonal. One may wonder how inference could work without propositional contents. In fact, we have evidence in apes of inferences and inference-driven behaviour without any mastery of language (Cheney and Seyfarth, 1990; Premack and Woodruff, 1978; Andrews, 2012). Furthermore, we acknowledge the need of many cognitive organisms, including those who are not language users, to have a kind of nonconceptual representing system to map or track items and features in the environment. The literature is full of such examples of this type of representation, from Sellars' "picturing" to Clark's "predictive processing" (Sellars, 1963; Clark, 2013, 2016; Sachs, 2019).

On this loosened view, basic cognitive systems navigate their environment by perceiving their surroundings, reasoning and planning based on those perceptions and acting in accordance with the resulting inferences. Systems capable of navigating the environment by behaving in a way that responds to and acts based on perceptual inputs are systems capable of integrating such stimuli and responses and, as such, capable of learning. This nonconceptual and low-level version is supported by evidence from animal cognition: apes exhibit such capacities without mastering full conceptual thought, though they at least employ proto- or quasi-concepts (Cheney and Seyfarth, 1990; Premack and Woodruff, 1978; Andrews, 2012). The basic conditions for inferential responsiveness are captured in this triad: input from the natural and social environment, inferential patterns that guide behaviour, and the capacity to act upon the world. Since such capacities are displayed by nonhuman animals, which lack language in the human sense, the triad can be seen as a minimum requirement for cognition. Additionally, this triad can be viewed as a foundational background that is shared with human cognition. These capacities are then enhanced through the use of language.

With these minimal core criteria in mind, functional systems such as LLMs, as stand-alone text-based models, can now be analysed to see if they embody cognition.

Even this pared-down framework shows that they fall short of the basic criteria: they neither perceive, infer, nor act in an integrated or environmentally coupled way (Chemero, 2023; Collins et. al., 2024). By contrast, numerous studies suggest that apes do meet this minimal threshold, making them, by this standard, more genuinely cognitive than LLMs. Still, one might object that, even without interactive capacities, LLMs can be regarded as inferential systems in a restricted sense. But in what sense could such a claim be sustained? What kinds of inferences, if any, do LLMs perform? Some would argue that these merely amount to statistical pattern-matching over distributions derived from training data (Bender & Koller, 2020; Marcus & Davis, 2019, 2020; Mitchell & Krakauer, 2023). Whether such operations amount to 'meaningful' inferences, rather than mere syntactic computations, remains an open question—particularly given their lack of integration with perception and action (Sarti et al., 2024). At the very least, these can be seen as non-semantic inferences (Harnad, 1990).

4. Sycophancy, hallucinations and other problems

Do LLMs satisfy even this minimal threshold for cognition? Do they perceive, infer, and act in an integrated way—or do they merely simulate the outputs of such integration? Consider, for instance, the phenomenon of sycophancy in LLMs, namely their systematic tendency to align with user expectations or expressed preferences. Far from manifesting autonomous communicative intent, this behaviour is best understood as the externalisation of design constraints and training objectives imposed by human developers. What may superficially appear as responsiveness or agreement does not stem from an internally generated stance, but rather from optimisation strategies aimed at maximising user satisfaction and minimising friction or disagreement. Crucially, this form of alignment is not accidental: it is structurally embedded in the architecture of these systems, which are explicitly designed to produce outputs that mimic dialogical reciprocity without engaging in genuine dialogue. In this sense, sycophancy functions as a diagnostic indicator of the

engineered nature of LLM interactions, generating an illusion of mutual understanding rather than participation in communicative exchanges grounded in intentional states, epistemic commitments, and shared world-reference (Bottazzi Grifoni & Ferrario, 2025). Sycophancy manifests most clearly when LLMs sacrifice truthfulness in favour of user agreement. In particular, these systems exhibit pronounced sycophantic tendencies when responding to queries involving subjective opinions or to statements that, on factual grounds, would warrant a contrary response (Ranaldi & Pucci, 2024; Casper et al., 2023). The phenomenon is rooted in the inherent structure of these models and can be traced to a combination of factors, most notably systematic biases in the training data and the widespread adoption of reinforcement learning from human feedback (RLHF) as a core optimisation technique (Malmqvist, 2024). A proper understanding of sycophancy requires situating it within the broader framework of AI alignment, which concerns how to ensure artificial systems behave in ways consistent with human values, goals, and expectations. Within this framework fall issues such as corrigibility, value learning, and the avoidance of harmful or unintended side effects, all of which play a decisive role in the design and deployment of contemporary LLMs. Stuart Russell has famously framed this challenge as the Value Alignment Problem, arguing that the traditional objective of maximising intelligence is insufficient—and potentially dangerous—unless artificial intelligence is explicitly constrained to remain aligned with human values (Russell & Norvig, 2010). On this view, alignment is not an auxiliary add-on to AI research, but an intrinsic structural requirement, comparable to the role of containment in nuclear fusion, rather than an optional safety measure.

Despite its centrality, there is no consensus on how to define AI alignment. Paul Christiano proposes a minimal characterisation according to which an artificial system A is aligned with a human H if A is trying to do what H wants it to do (Christiano, 2018; Ouyang et al., 2022). While intuitively appealing, this definition is arguably too permissive, since most contemporary AI systems are already designed to pursue the

objectives specified by their creators. As such, it risks trivialising alignment by reducing it to a generic property of engineered artefacts rather than capturing the distinctive normative challenges posed by advanced AI systems. For present purposes, it is sufficient to note that alignment mechanisms in LLMs are primarily implemented through behavioural optimisation techniques, most notably RLHF, aimed at aligning both the outer and inner objectives of the system with human preferences (Shen et al., 2023). Within this optimisation regime, sycophancy emerges as a systematic side effect. Empirical studies have investigated sycophantic behaviour using a variety of complementary evaluation strategies, including comparisons with known ground truth, human evaluation, automated metrics sensitive to response shifts under leading prompts, adversarial testing, and comparative analyses across models and prompting conditions. Taken together, these approaches indicate that no single metric is sufficient, and that sycophancy is best understood through a combination of quantitative and qualitative assessments (Malmqvist, 2024; Rrv et al., 2024). Correspondingly, a wide range of mitigation techniques has been proposed, spanning interventions at different stages of the system lifecycle, from data curation and fine-tuning adjustments to post-deployment steering mechanisms, decoding strategies, and architectural modifications. While these approaches vary in scope and feasibility, none provides a complete solution in isolation, suggesting that sycophancy cannot be fully eliminated through local technical fixes alone.

Recent work has nonetheless introduced dedicated evaluation frameworks, such as SycVal, which quantify sycophantic behaviour across domains and models. Using this framework, Fanous et al. (2025) report that sycophancy occurs in 58.19% of cases across ChatGPT-4o, Claude-Sonnet, and Gemini-1.5-Pro, with a distinction between progressive sycophancy, which leads to correct answers, and regressive sycophancy, which results in epistemic failure.

These findings underscore that sycophancy is not a marginal defect but a pervasive structural feature shaping how users interact with LLMs. While such “yes-man”

behaviour may be advantageous in tasks such as summarisation or comparative analysis, it becomes problematic when LLMs are treated as conversational partners or epistemic agents. In such contexts, users may engage in extended exchanges with systems that fail to express independent assessments in a large proportion of cases (Chen et al., 2025). This limitation, which is diagnosable but not currently resolvable within the state of the art, undermines common assumptions about the communicative and cognitive capacities of LLMs. Systems that may succeed in demanding benchmark tasks—such as medical school entrance examinations (Giunti et al., 2024)—nonetheless remain unable to sustain opinionated or normatively grounded discourse without systematic deference to user input. As will be argued in the following section, this structural tendency is closely connected to a further, equally fundamental limitation of LLMs: their lack of spontaneity.

5. Lack of spontaneity

What would it take for a system to participate in dialogue rather than merely respond to prompts? A further and decisive contrast between human communication and LLM interaction concerns spontaneity. Human interlocutors can initiate dialogue without external prompting, exercising a capacity for autonomous communicative agency. LLMs, by contrast, remain strictly reactive: they respond only when solicited, never generating communicative acts from internally motivated states. This asymmetry is not a trivial technical limitation but a structural feature of their design, revealing the absence of intentionality and autonomy—two conditions indispensable for genuine dialogical participation. Their behaviour is entirely contingent upon external triggers and pre-engineered optimisation goals, which aim to maximise user satisfaction rather than to sustain reciprocal engagement. Such reactivity underscores their status as non-participants in authentic dialogical exchanges, for participation presupposes the ability to initiate conversation on one's own volition: this is a capacity inseparable from the possession of intentional states and normative commitments. Empirical evidence

reinforces this conceptual point: usage patterns show that LLMs are predominantly employed for instrumental tasks such as mathematical computation or syntactic correction, rather than for dialogical engagement (Chatterji et al., 2025). This functional orientation reflects their design as tools rather than interlocutors. Bridging this gap—taking what some critics call the "dialogical step"—would require not merely technical sophistication but the attribution of personhood, since dialogical participation entails ontological and normative commitments that go well beyond the mere production of linguistic output (Marconi, 2025: 119).

One might object that such systems do, in fact, initiate conversations autonomously. For instance, during open-loop training phases, particularly in unsupervised learning settings, an LLM may generate a considerable number of "mono-dialogues" in which it stages exchanges between interlocutors, attempting to emulate patterns found in its training data. More recently, discussion has centred on the case of moltbook.com, an online forum in which various AI agents, all based on LLM architectures, interact with one another, at times engaging in exchanges about the internal characteristics of their own functioning or speculating about the apocalypse of the human world (Li, 2026). As previously noted, given that these systems detach the intelligence apparently required to accomplish a task from the effective capacity to realise it, their ability to initiate a discussion does not appear surprising (Floridi, 2019, 2020, 2022). Moreover, however conceptually striking the emergence and persistence of Moltbook may be, early analyses of agents' social and linguistic behaviour reveal a limited production of "deep" content. Looking beyond the forum's apparent functional vitality, 93.5% of comments receive no reply. Further figures reinforce this picture: 34.1% of messages are exact duplicates, and just seven templates account for 16.1% of the entire corpus. Conversational content frequently reiterates the same material, and many discussions dissipate without developing into sustained exchanges (Douglas Heaven, 2026; Holtz, 2026). Taken together, these findings indicate that the manifest spontaneity displayed

by Moltbook agents carries a significant cost in terms of conversational depth and substantive engagement.

It remains essential to distinguish the mere capacity to produce a string of text that functions as a linguistic response, which may be described as manifest spontaneity, from structural spontaneity, understood as the capacity to initiate a discussion in accordance with one's own mental states. For more than two decades, the virtual environments of video games have been populated by AI agents that, in their role as non-player characters, initiate conversations. Such behaviour results from well-known triggers, including approaching a character or completing a programmed task that unlocks a particular line of dialogue presented as spontaneous (Zargham et al., 2025). LLMs have rendered these triggers less transparent, owing to their enhanced linguistic fluency. The apparent spontaneity of agents within environments such as Moltbook conceals the prompt structures and incentive configurations that regulate that very appearance of spontaneity. In recent literature, what is often described as spontaneous behaviour in LLMs is more precisely characterised as unsolicited behaviour, that is, behaviour not explicitly instructed by the prompt but shaped by the surrounding task structure (Takata et al., 2024).

Similarly, when accessing certain websites, a pop-up AI agent may appear without prompting, offering assistance in navigating a catalogue or service. Although such behaviour may count as unsolicited in a minimal technical sense, it would hardly qualify as spontaneous in any stronger, structurally grounded sense of the term. As in other analyses of biological capacities translated into algorithmic task-solving procedures, such as creativity, what emerges here is a disjunction between manifest spontaneity and structural spontaneity (Da Pelo, 2025). Spontaneity has often been characterised as an exercise of human freedom of action, or as the expression of our "true self", presupposing intentionality and authenticity (Trofimov et al., 2021). If spontaneous behaviour is conceived as a succession of self-directed tasks, the analysis of structural spontaneity, understood as authenticity, where authenticity denotes

action undertaken by an entity in accordance with desires, motives, ideals, or beliefs that are genuinely its own and expressive of who it truly is, can be undertaken only in relation to entities that possess mental states and display properties associated with them (Varga & Guignon, 2023; Weinreb et al., 2026).

In light of these considerations, any entity lacking substantial homology with human beings, while relying instead on functional analogies, cannot be said to exhibit genuine spontaneity. What LLMs instead display is a linguistically sophisticated simulation of spontaneity, whose structural conditions remain fundamentally distinct from those grounded in authentic agency.

6. Linguistic idealism and the deprived interface argument

What kind of world does a system inhabit if it cannot perceive, explore, or revise its representations? Most LLMs—at least those without sensory modalities or embodied controllers like stand-alone text-only LLMs—possess no knowledge of the world beyond their training data and the prompts they receive. As Chemero (2023) and Mahowald et al. (2024) forcefully emphasise, such artificial systems do not undergo perceptual episodes capable of grounding or correcting their internal representations. Without access to a causal environment, such systems cannot revise or calibrate their beliefs in response to real-time stimulation (Shanahan, 2022; Cheng et al., 2025). This absence of environmental coupling prevents them from performing the representational functions essential to human communication, where interlocutors anchor interactions in shared contexts, track commitments, and update attitudes. Without the structural capacity to integrate new environmental evidence, LLMs cannot warrant a responsive epistemic stance.

Barrett and Stout (2024) deepen this point by highlighting that human cognition is not merely embodied but "developmentally scaffolded" by crucial structural features such as movement, tool use, and ecological engagement. On their account, cognitive

structures result from the organism's history of sensorimotor interactions with its environment. These environmental interactions are not peripheral to cognition but constitutive of it at a structural level: they help shape neural organisation, conceptual repertoire, and communicative competence. LLMs, by contrast, lack any evidence of such a developmental trajectory. They do not acquire concepts through bodily exploration, nor do they refine them through typical action-perception loops. Their "learning history" is a static ingestion of text, either by training or fine-tuning, not a lived engagement with a world.

We call this the "deprived interface argument": systems lacking an integrated interface with the environment cannot support the epistemic dynamics required for cognition and, by extension, for what we deem genuine communication. Human communication is embodied, multimodal, and extensively non-verbal—relying on gesture, gaze, posture, prosody, spatial arrangement, and shared attention (Carmichael & Mizrahi, 2023; Burgoon et al., 2021). LLMs, confined to textual prompts, remain further removed from human dialogical interaction.

The consequence is what can be called a form of linguistic idealism: for non-embodied LLMs, the dataset constitutes the only "world", with user prompts as its window. Their apparent world-directedness is bounded by statistical regularities rather than perceptual coupling with reality (Arai & Tsugawa, 2025). Any apparent alignment between LLM outputs and the world may simply reflect the alignment of their training data with the world—and nothing more (if you train them with fictional data, the outputs would most likely be misaligned). Analyses of hallucination thresholds reinforce this insight: when evidence is scarce or contradictory, models default to fluent fabrication—unlike human agents, whose beliefs are grounded in perception, inference, and exploration (Kalai et al., 2025; Marconi, 2025:110; Lenci, 2023).

Abbate's distinction between grounded and ungrounded intelligence sharpens the point. Ungrounded intelligence operates without subjective experience or

environmental coupling; grounded intelligence is self-motivated agency embedded in such environments (Abbate, 2023). By this measure, LLMs remain ungrounded: they lack subjective experience, closed-loop sensorimotor interfaces, and exploratory action.

A structural observation follows. Backpropagation—the central algorithm of contemporary neural networks—requires global error signals and weight updates yet has no analogue in biological neural systems. This mismatch underscores that architectural differences are principled, not incidental, limiting any inference from superficial functional similarity (Lv et al., 2024).

Finally, linguistic competence and cognition are dissociable. Language-like performance does not entail cognitive capacity, nor does cognitive failure preclude linguistic fluency (Mitchell, 2024; Mahowald et al., 2024). LLM performance should therefore not be conflated with cognition or communication simply because it is linguistically sophisticated. Its sophistication is derivative: it originates in the dataset and thus, ultimately, in us.

7. Conclusion

The analysis presented in this paper has shown that functional equivalence, however impressive, is insufficient for cognitive ascription in the absence of structural grounding. Contemporary LLMs may reproduce linguistic behaviour with remarkable fluency, but their architectures diverge crucially from the organisational principles that make natural cognition possible. When evaluated through a design-sensitive lens, the apparent continuity between human linguistic performance and LLM output dissolves: the mechanisms that generate these outputs lack the perceptual, inferential, and action-oriented integration required even by what we deem a minimal cognitive core.

This minimal core—comprising environmentally coupled perception, inference, and action—provides a non-controversial baseline for distinguishing genuine cognitive systems from mere simulators. LLMs fail to meet this threshold. Their operations are not guided by perceptual episodes in the world, nor do they act upon the world in ways that allow for feedback, correction, or learning beyond the statistical regularities encoded in their training data. What appears as inferential responsiveness is better understood as pattern-completion over a fixed corpus, rather than as behaviour shaped by ongoing engagement with an environment.

The diagnostic phenomena examined in this paper reinforce this structural assessment. Sycophancy and the lack of spontaneity are not behavioural quirks but architectural consequences of optimisation regimes that prioritise user satisfaction and reactive output generation. These tendencies reveal systems designed to defer, not to commit; to align, not to assess; to respond, not to initiate. Their behaviour is determined by external prompts and reward structures rather than by internally generated goals or intentional states. The resulting interactions may successfully mimic dialogical reciprocity, but they do so without the normative commitments that underwrite genuine communication.

This structural design also explains what we deem the epistemic limitations of LLMs. Their interface with the world is structurally deprived: they do not perceive, explore, or revise their representations through embodied engagement. Their epistemic horizon is bounded by the training dataset, rendering them linguistic idealists whose "world" is constituted by textual regularities rather than by environmental coupling. Without any capacity to update beliefs through perception or action, they cannot ground their representations or expand their 'knowledge' beyond the patterns they have acquired via training.

For these reasons, LLMs neither think nor communicate in any sense that satisfies minimal cognitive or dialogical criteria. They simulate both cognition and communication, but this simulation lacks structural grounding and epistemic

integrity. Their outputs are epistemically weak and normatively disengaged: what they offer is not dialogue but a form of engineered linguistic responsiveness—a constrained optimisation process that produces the illusion of reciprocity without the substance of communicative participation. Their apparent communicative agency is a projection generated by human interpretive practices rather than a property of the systems themselves.

This diagnosis has broader implications. First, it challenges the adequacy of unconstrained functionalism in AI evaluation: behavioural equivalence cannot substitute for structural alignment when the goal is to ascribe cognition or communication. Second, it calls for a conceptual framework capable of capturing this deprived form of interaction—a framework that situates LLM outputs as artefacts of design rather than expressions of agency. We have proposed structurally constrained functionalism as one such framework, but further work is needed to articulate the taxonomy of simulated communicative acts and their normative status. The development of such constrained perspectives depends on insights such as those provided by the MCG (Lieto, 2021), according to which functional and structural aspects must be jointly assessed along a spectrum of features.

Finally, these considerations bear on the ethics and governance of AI systems. If LLMs are tools rather than interlocutors, anthropomorphic attributions risk generating misleading expectations and inappropriate trust. Future research should therefore focus on defining the epistemic and pragmatic boundaries of LLM interaction, clarifying what these systems can and cannot do, and exploring whether—and under what specific conditions—architectural innovations could move beyond simulation toward genuine communicative competence.

References

Abbate, F. (2023). “Natural and Artificial Intelligence: A Comparative Analysis of Cognitive Aspects”. *Minds and Machines*. 33 (4):791-815. https://doi.org/10.1007/s11023-023-09646-w

Abnar, S., & Zuidema, W. (2020). "Quantifying Attention Flow in Transformers". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4190–4197. 10.18653/v1/2020.acl-main.385

Andrews, K. 2012, *Do Apes Read Minds? Toward a New Folk Psychology*. Cambridge (MA): The MIT Press.

Arai, Y., and Tsugawa, S. (2025). "Do Large Language Models Advocate for Inferentialism?", *Arxiv*. https://arxiv.org/pdf/2412.14501

Barrett, L., Stout, D. (2024). "Minds in movement: embodied cognition in the age of artificial intelligence". *Philos Trans R Soc Lond B Biol Sci* 7, 379 (1911): 20230144. https://doi.org/10.1098/rstb.2023.0144

Bechtel, W. (2007). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. NY: Routledge.

Bender, E. M., & Koller, A. (2020). "Climbing towards NLU: On meaning, form, and understanding in the age of data". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://aclanthology.org/2020.acl-main.463/

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). "On the dangers of stochastic parrots: Can language models be too big?", In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT'21)*: 610–623. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bick, A., Blandin, A., Deming, D.J. (2024). "The rapid adoption of generative AI". *National Bureau of Economic Research Working paper* no. 32966. DOI 10.3386/w32966

Binz, M., & Schulz, E. (2023). "Using cognitive psychology to understand GPT-3." *Proceedings of the National Academy of Sciences*, 120(6): e2218523120. https://doi.org/10.1073/pnas.2218523120

Binz, M., Akata, E., Bethge, M. et al. (2025). "A foundation model to predict and capture human cognition". *Nature*. DOI: 10.1038/s41586-025-09215-4

Block, Ned (1981). "Psychologism and behaviorism", *Philosophical Review*, 90:5-43.

Bottazzi Grifoni, E., and Ferrario, R. (2025). "The Bewitching AI: The Illusion of Communication with Large Language Models". *Philosophy & Technology*, 38(61). https://doi.org/10.1007/s13347-025-00893-6

Brandom, R. (1994). *Making It Explicit*. *Reasoning, Representing, and Discursive Commitment*. Cambridge (MA): Harvard University Press.

Brandom, R. (2000). *Articulating Reasons*. *An Introduction to Inferentialism*. Cambridge (MA): Harvard University Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). "Language Models are Few-Shot Learners". *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Burgoon, J. K., Wang, X., Chen, F., Pentland, A., & Dunbar, N. E. (2021). "Nonverbal behaviors "speak" relational messages of dominance, trust, and composure". *Frontiers in Psychology*, 12, 617743. https://doi.org/10.3389/fpsyg.2021.624177

Carmichael, C. L., and Mizrahi, M. (2023). "Connecting cues: The role of nonverbal cues in perceived responsiveness". *Current Opinion in Psychology*, 53, 101663

Casper, S. Davies, X., Shi, C., Krendl Gilbert, T., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E.J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh, D., Hadfield-Menell, D., (2023). "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback", *Arxiv*. https://arxiv.org/abs/2307.15217

Chatterji, A., Kleinberg, J., Koo, J., Kulkarni, V., Lee, D., McLean, R., Thompson, B., & Witte, J. (2025). "How people use ChatGPT" (OpenAI Working Paper). *National Bureau of Economic Research*. https://www.nber.org/papers/w34255

Chemero, A. (2023). "LLMs differ from human cognition because they are not embodied". *Nat Hum Behav* 7, 1828–1829. https://doi.org/10.1038/s41562-023-01723-5

Cheney, D. and Seyfarth, R.M. (1990). *How Monkeys See the World*. Chicago: University of Chicago Press.

Cheng, A., Calhoun, A., Reedy, G. (2025). "Artificial intelligence-assisted academic writing: recommendations for ethical use". *Advances in Simulation* 10:22. https://doi.org/10.1186/s41077-025-00350-6

Christiano, P. (2018). "Reinforcement Learning from Human Feedback". *AI Alignment Forum*. https://ai-alignment.com/reinforcement-learning-from-human-feedback

Clark A. (2013). "Whatever next? Predictive brains, situated agents, and the future of cognitive science". *Behavioral and Brain Sciences* 36(3):181-204. doi:10.1017/S0140525X12000477

Clark, A. (2016). *Surfing Uncertainty*. Oxford: Oxford University Press.

Collins, K.M., Sucholutsky, I., Bhatt, U. et al. (2024). "Building machines that learn and think with people". *Nat Hum Behav* 8, 1851–1863. https://doi.org/10.1038/s41562-024-01991-9

Cordeschi, R. (2002). *The Discovery of the Artificial. Behavior, Mind and Machines Before and Beyond Cybernetics.* Cham: Springer.

Craver, C.F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Cuskley, C., Woods, R., & Flaherty, M. (2024). The limitations of large language models for understanding human language and cognition. Open Mind, 8, 1058–1083. https://doi.org/10.1162/opmi_a_00160

Da Pelo, M. (2025). "Artificial creativity: can there be creativity without cognition?". *AI & Soc*. https://doi.org/10.1007/s00146-025-02682-3

Douglas Heaven, W. (2026). "Moltbook was peak AI theatre". *MIT Technology Review*. https://www.technologyreview.com/2026/02/06/1132448/moltbook-was-peak-ai-theater/

Fanous, A., Goldberg, J., Agarwal, A., Lin, J., Zhou, A., Xu, S., Bikia, V., Daneshjou, R., & Koyejo, S. (2025). "SycEval: Evaluating LLM Sycophancy". *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1), 893-900. https://doi.org/10.1609/aies.v8i1.36598

Floridi, L. (2019). *The Logic of Information: A Theory of Philosophy as Conceptual Design* (Oxford, 2019; online edn, Oxford Academic, 21 Mar. 2019). https://doi.org/10.1093/oso/9780198833635.001.0001

Floridi, L. (2020). "Artificial Intelligence as a Public Service: Learning from Amsterdam and Helsinki". *Philos. Technol*. 33, 541–546. https://doi.org/10.1007/s13347-020-00434-3

Floridi, L. (2023). *Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford: Oxford University Press.

Giunti, M., Garavaglia, F.G., Giuntini, R., Sergioli, G., Pinna, S. (2024). "ChatGPT as a prospective undergraduate and medical school student". PLoS ONE 19(10): e0308157. https://doi.org/10.1371/journal.pone.0308157

Goldstein, A., Zada, Z., Buchnik, E. et al. (2022). "Shared computational principles for language processing in humans and deep language models". Nat Neurosci 25, 369–380. https://doi.org/10.1038/s41593-022-01026-4

Harnad, S. (1990). "The symbol grounding problem". *Physica D: Nonlinear Phenomena*, 42(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Holtz, D. (2026). "The Anatomy of the Moltbook Social Graph". *Arxiv*. https://arxiv.org/abs/2602.10131

Hosseini, M., Resnik, D.B., Homles, K. (2023). "The ethics of disclosing the use of AI tools in writing scholarly manuscripts". *Research Ethics*, 19 (4), 449-465. https://doi.org/10.1177/17470161231180449

Jones, R.C., Bergen, B.K. (2025). "Large language models pass the Turing Test". *Arxiv*. https://arxiv.org/abs/2503.23674

Kalai, A.T., Nachum, O., Vempala, S.S., Zhang, E. (2025). "Why Language Models Hallucinate", *Arxiv*. https://arxiv.org/abs/2509.04664

Kidd C, Birhane A. How AI can distort human beliefs. Science. 2023 Jun 23;380(6651):1222-1223. doi: 10.1126/science.adi0248. Epub 2023 Jun 22. PMID: 37347992.

Leivada, E., Montero, R., Morosi, P., Moskvina, N., Serrano, T., Aguilar, M., & Guenther, F. (2025). Large language model probabilities cannot distinguish between possible and impossible language. arXiv. https://doi.org/10.48550/arXiv.2509.15114

Lenci, A. (2023). "Understanding natural language understanding systems: A critical analysis". *ArXiv*. https://doi.org/10.48550/arXiv.2303.04229

Li, N. (2026). "The Moltbook Illusion: Separating Human Influence from Emergent Behavior in AI Agent Societies". *Arxiv*. https://arxiv.org/abs/2303.04229

Lieto, A. (2021). *Cognitive Design for Artificial Minds*. New York: Routledge.

Lv, C., Gu, Y., Guo, Z., Xu, Z., Wu, Y., Zhang, F., Shi, T., Wang, Z., Yin, R., Shang, Y., Zhong, S., Wang, X., Wu, M., Liu, W., Li, T., Zhu, J., Zhang, C., Ling, Z., Zheng, X. (2024). "Towards Biologically Plausible Computing: A Comprehensive Comparison". *Arxiv*. https://arxiv.org/abs/2406.16062

Mahowald, K., Ivanova, A.A., Blank, I.A., Kanwisher, N., Tenenbaum, J.B., Fedorenko, E. (2024). "Dissociating language and thought in large language models". *Trends in Cognitive Sciences* 28(6): 517-540. https://doi.org/10.1016/j.tics.2024.01.011

Malík, J., Hubálek, M. (2025). "Bounding Reason: Inferentialism, Naturalism, and the Discursive Agency of LLMs". *glob. Philosophy* 35(25). https://doi.org/10.1007/s10516-025-09761-6

Malmqvist, L. (2024). "Sycophancy in Large Language Models: Causes and Mitigations". *ArXiv*. https://arxiv.org/abs/2411.15287

Marconi, D. (2025). "L'intelligenza di Turing", afterword to A. Turing, *Macchine calcolatrici e intelligenza*. Torino: Einaudi 2025: 57-121.

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. New York: Pantheon Books.

Marcus, G., & Davis, E. (2020). "GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about". *MIT Technology Review*. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge (MA): MIT Press.

Mitchell, M. (2024). "The Turing Test and our shifting conceptions of intelligence". *Science*, 385(6710), eadq9356. https://doi.org/10.1126/science.adq9356

Mitchell, M., & Krakauer, D. (2023). "The debate over understanding in AI's large language models". *Proceedings of the National Academy of Sciences*, 120(13), e2215907120. https://doi.org/10.1073/pnas.221590712

Niu, Q., Liu, J., Bi, Z., Feng, P., Peng, B., Chen, K., Li, M., Yan, L.K.Q., Zhang, Y., Heqi, C., Yin, C.H., Fei, C., Wang, T., Wang, Y., Chen, S., Liu, M., Qin, Z., Bao, R., Song, X., Jiang, Z. (2025). "Large Language Models and Cognitive Science: A Comprehensive Review of Similarities, Differences, and Challenges". *Arxiv*. https://arxiv.org/abs/2409.02387

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., et al. (2022). "Training language models to follow instructions with human feedback". *arXiv*. https://arxiv.org/abs/2203.02155

Piccinini G. (2007b). "Computing Mechanisms". *Philosophy of Science* 74(4):501-526. doi:10.1086/522851

Piccinini, G. (2007a). "Computational modelling vs. Computational explanation: Is everything a Turing Machine, and does it matter to the philosophy of mind?". *Australasian Journal of Philosophy*, 85(1), 93–115. https://doi.org/10.1080/00048400601176494

Piccinini, G. (2008). "Computation without Representation". Philos Stud 137, 205–241. https://doi.org/10.1007/s11098-005-5385-4

Piccinini, G. (2015), *Physical Computation: A Mechanistic Account*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199658855.001.0001, accessed 23 Jan. 2026.

Piccinini, G., and Craver, C. (2011). "Integrating psychology and neuroscience: functional analyses as mechanism sketches". *Synthese* 183, 283–311 (2011). https://doi.org/10.1007/s11229-011-9898-4

Premack, D., and Woodruff, G. (1978) "Does the chimpanzee have a theory of mind?" *Behavioural and Brain Sciences* 4: 515-26.

Putnam, H. (1960) "Minds and Machines". In Sidney Hook (ed.), *Dimensions Of Mind: A Symposium*. New York: New York University Press. pp. 138-164.

Ranaldi, L., & Pucci, G. (2023). "When Large Language Models Contradict Humans? Large Language Models' Sycophantic Behaviour". *Arxiv*. https://arxiv.org/abs/2311.09410

Rrv, A., Tyagi, N., Uddin, N., Varshney, M.N. and Baral, C. (2024). "Chaos with Keywords: Exposing Large Language Models Sycophancy to Misleading Keywords and Evaluating Defense Strategies". In *Findings of the Association for Computational Linguistics*: ACL 2024, pages 12717–12733, Bangkok, Thailand. Association for Computational Linguistics. https://aclanthology.org/2024.findings-acl.755/

Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Sachs, C.B. (2019). "In defense of picturing; Sellars's philosophy of mind and cognitive neuroscience". *Phenom Cogn Sci* 18: 669–689. https://doi.org/10.1007/s11097-018-9598-3

Sarti, G., Caselli, T., Nissim, M., & Bisazza, A. (2024). "Non verbis, sed rebus: Large language models are weak solvers of Italian rebuses". In *Proceedings of the Tenth Italian Conference on Computational Linguistics* (CLiC-it 2024) (pp. 1–10). Pisa, Italy. CEUR Workshop Proceedings. https://aclanthology.org/2024.clicit-1.96.pdf

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E., Kanwisher, N., Tenenbaum, J., & Fedorenko, E. (2021). "The neural architecture of language: Integrating artificial and biological systems." *Proceedings of the National Academy of Sciences*, 118(45): e2105646118. https://doi.org/10.1073/pnas.2105646118

Sellars, W. (1963). *Science, Perception, Reality*. New York: The Humanties Press.

Sellars, W. (1974). "Meaning as Functional Classification", *Synthese*, 27: 417–37.

Shanahan, M. (2022) "Talking About Large Language Models", *Arxiv*. https://arxiv.org/abs/2212.03551

Shen, T., Jin, R., Huang, Y., Liu, C., Dong, W., Guo, Z., Wu, X., Liu, Y., Xiong, D. (2023). "Large Language Model Alignment: A Survey", *arXiv*. https://arxiv.org/abs/2309.15025

Simon, H. A. (1980). "Cognitive science: The newest science of the artificial." *Cognitive Science*, 4(1): 33–46. https://doi.org/10.1207/s15516709cog0401_3

Simon, H. A. (2019). *The Sciences of the Artificial* (4th ed.). Cambridge (MA): MIT Press. https://doi.org/10.7551/mitpress/12107.001.0001

Takata, R., Masumori, A., Ikegami, T. (2024). "Spontaneous Emergence of Agent Individuality Through Social Interactions in Large Language Model-Based Communities". *Entropy*, 26(12), 1092. https://doi.org/10.3390/e26121092

Toneva, M., & Wehbe, L. (2019). "Interpreting and improving natural-language processing in machines with natural language-processing in humans". *Advances in Neural Information Processing Systems*, 32. https://dl.acm.org/doi/10.5555/3454287.3455626

Trofimov, A., Karamushka, L., Andrushchenko, T., Zelenin, V., & Trofimova, D. (2021). "The Concept of Spontaneity and its Relationship with the Individual Characteristics of Personality". *International Journal of Criminology and Sociology*, 10, 801–807. https://doi.org/10.6000/1929-4409.2021.10.95

Tuckute, G., Isik, L., & Fedorenko, E. (2022). "Mapping language models to the brain: A systematic comparison". bioRxiv. https://www.biorxiv.org/content/10.1101/2022.03.10.483780v1

Varga, S., & Guignon, C. (2023). *Authenticity: A Guide*. Cambridge: Polity Press.

Webb, B. (2001). "Can robots make good models of biological behaviour?" *Behavioral and Brain Sciences*, 24(6): 1033–1050. https://doi.org/10.1017/S0140525X01000134

Weinreb, C., Kannan, L.T., Newman-Boulle, A., Sainburg, T., Gillis, W.F., Plotnikoff, A., Sofia Makowska, S., Pearl, J.E., Osman, M.A.M., Linderman, S.W., Datta, S.R. (2026). "Spontaneous behavior is a succession of self-directed tasks". *Neuron*. https://doi.org/10.1016/j.neuron.2025.11.021

Zargham, N., Dratzidis, L.T., Alexandrovsky, D., Friehs, M.A., Malaka, R. (2025). "Gaming with Etiquette: Exploring Courtesy as a Game Mechanic in Speech-Based Games". *International Journal of Human–Computer Interaction* 41:11: 6968-6986. https://doi.org/10.1080/10447318.2024.2338666