AI & SOCIETY
https://doi.org/10.1007/s00146-024-02165-x

OPEN FORUM

Dehumanizing the human, humanizing the machine: organic consciousness as a hallmark of the persistence of the human against the backdrop of artificial intelligence

Sergio Torres-Martínez[1]

Received: 14 May 2024 / Accepted: 16 December 2024

Abstract
The rise of *generative artificial intelligence* (GenAI), especially *large language models* (LLMs), has fueled debates on whether these technologies genuinely emulate human intelligence. Some view LLM architectures as capturing human language mechanisms, while others, from a posthumanist and a transhumanist perspective, herald the obsolescence of humans as the sole conceptualizers of life and nature. This paper challenges, from a *practical philosophy of science* perspective, such views by arguing that the reasoning behind equating GenAI with human intelligence, or proclaiming the "demise of *the human*," is flawed due to conceptual conflation and reductive definitions of humans as performance-driven semiotic systems deprived of agency. In opposing theories that reduce consciousness to computation or information integration, the present proposal posits that consciousness arises from the holistic integration of perception, conceptualization, intentional agency, and self-modeling within biological systems. Grounded in a model of *Extended Humanism*, I propose that human consciousness and agency are tied to biological embodiment and the agents' experiential interactions with their environment. This underscores the distinction between pre-trained transformers as "posthuman agents" and humans as purposeful biological agents, which emphasizes the human capacity for biological adjustment and optimization. Consequently, proclamations about human obsolescence as conceptualizers are unfounded, as they fail to appreciate intrinsic consciousness-agency-embodiment connections. One of the main conclusions is that the capacity to integrate information does not amount to phenomenal consciousness as argued, for example, by *Information Integration Theory* (ITT).

Keywords Consciousness · Concept formation · Extended humanism · Generative AI · Large language models · Posthumanism

1 Introduction

> Any technology that fundamentally reorients the ways humans exchange and interact with information must be taken seriously as a potential threat, or at least a disruptive variable, in the context of democratic life. It is therefore no surprise that fears over the rise of generative artificial intelligence (AI) and its potential usurpation of human agency and democratic self-rule continue to draw attention from the public and policy-makers around the world. Wihbey 2024, 3

The expansion of *generative artificial intelligence* (GenAI), particularly large language models (LLMs), has spawned a range of theories and approaches aimed at defining whether these deep learning technologies genuinely reflect human intelligence. While some neuroscientists and linguists have suggested that LLM architectures accurately capture the mechanisms of human language production in the brain (and that, as a consequence, LLMs actually understand language in all its purposeful, agentive dimension (e.g. Tuckute et al. 2024), posthumanist and transhumanist thinkers have seized upon the fuzziness of LLM functioning to proclaim the impending obsolescence of humans as the sole conceptualizers of nature and our relationship to it. In this paper, I set out to refute these views, arguing that the reasoning behind such interpretations of GenAI and the projected role of humans is fundamentally flawed, marred by folk theorizing that encompasses "an

✉ Sergio Torres-Martínez
surtr_2000@yahoo.es

[1] Universidad de Antioquia, Cll. 67 #53 - 108, Medellín, Antioquia, Colombia

Published online: 28 January 2025

inarticulate yet influential *science-before-testimony* picture" (Gerken 2022, 1; emphasis in original).

For our purposes, ushering definitions that depict humans as mere performance-driven semiotic systems devoid of agency qualifies as a *before-testimony* conclusion. Specifically, I argue that human intelligence and consciousness cannot be reduced to mere computational processing for information integration, as proposed by theories like the *Information Integration Theory* (*IIT*, Tononi 2004). Instead, I suggest that consciousness arises from the holistic integration of perception, conceptualization, intentional agency, and self-modeling within biological systems. The present proposal seeks to contribute to the growing body of work in the practical philosophy of science, a form of meta-science that focuses on the potential for science-making and theorizing to cause harm in the broadest sense (see also Shrader-Frechette 2014, 6). This inevitably leads me to define humans as an endangered species, thereby avoiding the endorsement of changes that benefit only certain segments, such as the ideological distinction between vulnerable individuals or communities, and privileged individuals, since "the survival of sentient biological life" (Hedlund and Persson 2024, 454) does not concern some individuals but all individuals.

Drawing upon an *Extended Humanist* perspective, I assert that the domain of human consciousness and agency is intrinsically tied to biological embodiment and the experiential interactions of conscious agents with their environment. The emphasis on mere information processing or disembodied existence, as envisioned by transhumanist visions of mind and digital consciousness endorsed by many scientists and philosophers, is viewed in this paper as a potential source of harm that fails to capture the role of human consciousness and agency in the construction of more inclusive and humane societies.

Similarly, posthumanist attributions of agency, consciousness, and meaning-making capabilities to machines and AI language generators neglect the fundamental role of biological embodiment and the holistic integration of perception, conceptualization, intentional agency, and self-modeling.

1.1 The brain-machine myth as an algorithmic ontology of the mind

1.1.1 The ghost's fingerprint

This paper endorses a *weak AI* position by which consciousness in machines can only be possible at the level of simulation of thought and sentience (e.g., Searle 1980). However, the question as to whether machines have reached consciousness or will ever develop a sense of being in the world is useless (Levy 2009), since what we can know about the conscious states of machines is defined by what we know about our own consciousness as a first-person phenomenal experience. In this sense, our comprehension of what a robot could do with thought and reason is mediated by our grasp of reality, our definition of objects and entities, and what affordances thereof (in the Gibsonian sense) are salient enough and useful to us as part of the biological world.

To take a specific example, AI-powered robots need to be cognizant in the ways we define the term, that is, in terms of wakefulness and self-awareness, object identification and conceptualization, or how we deal with our surroundings in an agentive manner, through intention and purpose, through mind reading and making decisions on the basis of a survival imperative (see also Kinouchi and Mackin 2018, 1, Torres-Martínez 2024c). A diluted form of consciousness based on access to inner data and monitoring of processes (Block 1995) as a response to outward stimuli is, at present, the most feasible form of computational consciousness available to machines (Dehaene et al. 2017). However, the problem in defining consciousness pertains inevitably to the possibility of having autonomous machinic agents capable of interacting with humans socially. These agents need to be capable of representing the world to themselves:

> So we see that representation means *to bring to oneself a picture of the world through bodily affordances, biological memory, and incoming experience with that world*. In this sense, a picture of the world is both the result of perception and its further coupling with a chain of actions upon the world that is defined by the application of patterns of action-behavior in order to attain optimization (uncertainty-reduction) (Torres-Martínez 2024c).

More concretely, the prospect of machinic agency and the development of artificial consciousness depends critically on whether representations are connected merely with adaptive capabilities or constitute a fundamentally different type of unconscious agency. This alternative agency would deviate radically from biological systems by focusing on action execution within a performance range, rather than energy minimization for survival.

One of the primary limitations of the AI enterprise in developing conscious artificial agents lies in the constraints of hardware and available resources. These systems are inherently bound to exhaust their bodily and environmental resources while attempting to solve increasingly complex tasks through unsustainable processing schemas. Consequently, the pursuit of artificial consciousness reveals more about the developers' ambitions and misplaced expectations than about potential emergent machine intelligence, potentially leading to catastrophic consequences.

This perspective does not negate the fundamental human motivation behind artificial consciousness research: our

intrinsic need to comprehend consciousness and intelligence. The creation of artificial intelligent agents essentially establishes a dialogue between the Human and the Other—the non-human conceptualizer. This dialogue has increasingly motivated investments in embodied forms of AI. For example, Reinforcement Learning (RL) exemplifies a nuanced approach to computational agency by which systems learn by pursuing goals through extended environmental interactions. As Butlin (2022, 2023) argues, these systems are fundamentally sensitive to their outputs' consequences, developing strategies that maximize cumulative rewards across multiple time steps.

The sophistication of RL becomes particularly evident in complex decision-making scenarios. Researchers have developed computational models capable of simultaneously managing multiple reward functions, mimicking biological systems' ability to balance competing homeostatic drives. Studies by Keramati and Gutkin (2014), Juechems and Summerfield (2019), and Andersson et al. (2019) have explored various computational approaches to this challenge. Notably, Dulberg et al. (2023) proposed a model using multiple independent modules that learn to maximize distinct reward functions, with actions selected based on the highest total score.

The concept of embodiment, explored through output-input models or forward models, represents another critical dimension of artificial intelligence. These models enable systems to distinguish between environmental changes and self-generated sensory inputs—a hallmark of embodied intelligence.

In perceptual contexts, forward models allow systems to predict and differentiate the sensory effects of their own actions from external environmental changes. For motor control, they facilitate rapid and precise adjustments to complex effectors. However, current research suggests that while many AI systems employ forward models, few fully satisfy the embodiment criteria outlined in theoretical frameworks. A notable example is the system described by Friedrich et al. (2021), which utilized Kalman filtering (Todorov and Jordan 2002) to estimate a system's current state by integrating forward models with delayed sensory inputs. This approach demonstrates a sophisticated method of combining predictive models with real-time environmental interactions.

Despite these promising developments, the critical perspective previously articulated exposes fundamental limitations in the embodied artificial intelligence enterprise, particularly regarding resource constraints and representational challenges. As noted, artificial agents are constrained by materiality and potentially unsustainable processing schemas. Unlike biological systems that have evolved energy-efficient survival mechanisms, computational approaches to agency fail to adequately address the enormous energy requirements for creating adaptive systems capable of complex, open-ended problem-solving. This reveals that the computer metaphor has evolved within a conceptual vacuum that privileges specific pragmatist definitions of technology and *information-as-right-and-asset* at the expense of environmental constraints and broader systemic rights. This represents what I have termed *disembodied pragmatism*—a utilitarian mindset that relocates responsibility, accountability, and limitation into the realm of unbounded development of marketable information and informational goods and services, systematically undermining individual rights and the public good. The use of the American *First Amendment* for the protection of corporate speech serves as a quintessential example of this conceptual displacement.

> The justification for covering commercial speech under the First Amendment has shifted from protecting the interests of listeners to protecting the interests of speakers to valuing information for its own sake regardless of authorship or source. (Cortez and Sage 2023, 746)

It is no insight that over the last 60 years, the computer metaphor has nurtured the machinic-brain myth by which specific research designs have provided us with baroquely complex models of the mind based on the positing of arbitrary architectures with a weak *ontological basis*, but possessing *enhanced informational potential*. A paradigmatic set of beliefs backing up this mindset is the AI *transcendence hype* nurtured by Hollywoodian tropes by which "AI [is viewed] as the ultimate technology (and therefore the ultimate solution to all problems, i.e. pollution, the energy crisis, disease), yet at the same time the ultimate threat to humanity; the reduction of the individual to data and computation, and therefore the possibility of digital immortality" (Cave and Dihal 2023, 4).

I will put this analysis on hold for the moment to concentrate on the impact the machinic myth has had on the current view of *language-as-code*. On the one hand, nativists like Chomsky (Chomsky et al. 2019) see the brain as a computer naturally wired to symbolic processes within a black box, which results in the generation of language assembled on the fly by an elusive linguistic competence encoded in our *linguistic genes*. On the other, usage-based cognitive linguists and constructionists (empiricists, e.g. Goldberg 2019; Jackendoff 2012) see the brain as a statistically driven network of learnable symbolic representations fine-tuned to map and produce linguistic content.

Neuroscientists have taken heed of this theoretical claim in order to provide their research with definitions of language in line with some experimental paradigms (e.g., Kean et al. 2024). However, one problematic assumption of the computer metaphor in neuroscience is that language is conceptualized as consisting exclusively of symbols whose combinatorics can be localized in language-specific brain

areas and identified during ad hoc tasks. This has resulted in the creation of a model of the linguistic mind that is insulated from other cognitive domains.

Moreover, neuroimaging technologies provide machinic representations of processes in the brain that cannot be interpreted as *reliable* renderings of the formation and distribution of concepts in the mind. This, of course, touches upon epistemological considerations that have consequences for the development of both neighboring and not-so-close fields of scientific endeavor. This also compromises the status of the *fact* of the results provided by many studies considered as true and reliable within a single epistemic community: "[S]ome scientific ideas should be called 'facts', and they should be called 'facts' because they are true ideas (…)" (Vickers 2023: 2).

At this juncture, it becomes apparent that aligning with either a *scientific perspective*, as advocated by Massimi (2022), or a *future-proof science* stance, as proposed by Vickers (2023), offers limited assistance in defining true and reliable science. The reason lies in the inherent situatedness of scientific practice, which cannot independently justify the reliability of results accepted by a scientific community. This community operates within a framework of rule-following dynamics, wherein members endorse community-specific knowledge claims (Massimi 2022, 7).

Therefore, the reliability of scientific results, deemed as facts within a community, cannot simply be equated with the means by which these results are integrated into existing knowledge. This process often involves consistent adherence to conventions or institutional *rule-following* (Wittgenstein 1984 [1953]). Henceforth, then, I will stick with the argument that Wittgenstein's conclusion that rule interpretation is akin to rule postulation is flawed since the reasons to do something according to a rule are not dependent on goals existing *prior to* the accomplishment of a "given task". The idea I am advancing here, then, is that an expected consequence of endorsing the rule-following paradox is that any institutional use of language, in the context of a form of life or language game, is always open to a hypothesis-driven reconfiguration of the content of the rule due to the intentional character of actions.

More concretely put, the content of rules is defined, not by *epistemic blindness,* and the systematic adherence to convention (as Wittgenstein suggested), but by the predictive character of human cognition. It is worth pausing to see how this comes about. John McDowell (1992) suggests that to prove a correct reading of convention, a rule has to be postulated that warrants meaning and understanding. This interpretation has a lot to be said for it, for example, after having reached 1000 in a series I need not proceed to add 1002, since I may not want to proceed to add anything at all. In the same vein, the content of the underlying rule should not compel me to repeat the series *that way* to show that what I mean is in accord with previous training. So we may take as proven that the crux here is that both the content of the rule and my grasp thereof are elements of the broader category of reasons which are not reducible to a "correct" or "incorrect" grasp of a rule within an epistemic community.

Seen in this light, both the Wittgensteinian and the McDowellian types of rule are isomorphic in that they establish a normative relation with a context created for it to be meaningful. Another way of putting it is to say that the context provides the norms that make my use of a rule meaningful and acceptable for a community. However, rules cannot be made to contain *behavior*, as reasons do not contain a norm to act in ways that privilege correctness over goals. Perhaps better put, agents engaging in the social use of language will always have the possibility to opt out of conventional language usage to accrue epistemic capital.

Another consequence of this formulation is that the internal structure of rules cannot be framed in terms of a series of steps associated with family resemblance, that is, those stages entertaining an isomorphic relationship with the parts of an event during rule application.

The prevailing ontologies regarding language formation and processing in the brain cannot be simply derived from accepted methodological and epistemic principles, which suggest that highlighted brain regions serve as reliable evidence of ontological reality (e.g., Fedorenko et al. 2024), a mathematical "sketch of a model" (Mougenot & Matheson 2024) that follows an epistemology that can be summarized as: (1) taking a concept from folk psychology (memory, attention, etc.), (2) identifying a suitable information processing theory, (3) correlating performance on a task, and (4) mapping this onto a brain region to explain the concept (Mougenot & Matheson 2024). This approach can be considered folk theorizing, that is, the positing of constructs based on an implicit "science-before-testimony picture" (Gerken 2022, p. 1) that wields significant yet inarticulate influence. A more bizarre assumption consists in applying folk theorizing to experimental outcome prediction, in the belief that neuroscience experiment prediction can be improved by LLMs. According to this assumption, "[i]f LLMs' predictions surpassed human experts, the practice of science and the pace of discovery would radically change." (Luo et al. 2024, 1). Now, is advancing the pace of discovery a realistic goal for a field defined solely by the application of technological metaphors to the study of a single organ through computer scans? Moreover, can the brain become a symbol definable by LLM data averaging? (See subSect. 2.1).

In a similar vein, it is untenable to assert that a linguistic natural kind within the brain, poorly defined as "an ontologically meaningful grouping of brain areas" (Tuckute et al. 2024, 282), can be grounded solely on the traditional encoding–decoding process of generating meaningful symbolic strings. Crucially, the deficiency-fallibility bias backing up

symbolic output-oriented task design characterizes human "performers" as "*fallible* but boundedly *rational* agents who can in principle *eliminate their ignorance* as long as the *task* lies within *cognitively allowed* applications of *inference rules*" (Solaki 2022, 543; my emphasis). As the wording of the previous passage shows, there is an increasingly ubiquitous irony in current algorithmic-driven epistemologies: whereas the Turing test aimed at finding out whether "someone [could] tell accurately whether they [were] communicating with an AI or a human in a decontextualized interaction" (Youssef et al. 2024, 6, citing Nov et al. 2023), the effort is now on determining the extent to which humans can form representations allowing them to communicate efficiently with machines through a reduced set of signs defined a priori by the training of a language model.

As previously indicated, this approach often incorporates unfounded assumptions about the psychological reality of signs-as-action. For instance, defining language as consisting of form-function pairings ("constructions", as proposed by Goldberg 1995, 2019) imposes a particular framework for analyzing linguistic data during the online monitoring of brain activity.

Moreover, the use of non-invasive neuroimaging techniques such as electroencephalography (EEG) and magnetoencephalography (MEG), used to measure the electrical or magnetic activity of the brain, respectively, has provided tools to reinforce a narrative whereby whatever displays either neurochemical electrical or magnetic activity within a response-to-stimuli timeframe during specific tasks attests unequivocally the existence of a physical anchoring for language and the concepts encoded by it. Therefore, the idea of *functional localizers* (Tuckute et al. 2024), that is, a set of stimuli used to detect where a specific linguistic behavior "resides", suffers from a lack of conceptual support other than the formulation of stimulus-matching conditions that make no distinction between signals and signs during language representation (that is, $Language = \Sigma(Symbols, Rules)$).

The ensuing dehumanization of the brain seeks to reduce the complexity of the embodied components of language production and its connection with the identification of conscious states, event recognition-partition, and the agentive character of linguistic actions. According to this reading, the linguistic brain possesses a type of predictable internal variation definable by the amount of observable activity in specific brain areas displaying a distinct pattern of interconnection, which results in the positing of a *ghost fingerprint* (an index) by which a non-reducible natural kind can be identified as an ontological object in its own right. This characterization brings to the fore both an essentialist (the *how*) and an intrinsic (the *what*) definition of the brain as a construct existing on a single layer of reality expressed by one-dimensional propositions, which can be summarized as follows:

$$\text{Brain}\,0_{\text{Language State}} = \text{stimulus} \rightarrow \text{Language} \neq \Sigma(\text{Brain\_Areas}).$$

This, I oppose using the formula:

$$\text{Brain} \neq \text{f(Language\_State} \mid \text{Stimulus} \rightarrow \text{Language} - \text{Specific\_Network}).$$

This formula captures the following points:

1. The brain is not a function of the language state, where the language state is determined by a stimulus leading to the activation of a language-specific network in the brain.
2. The $\text{Language} \neq \Sigma(\text{Brain\_Areas})$ formula reflects the idea that language cannot be reduced to a sum of specific brain areas.
3. The use of the conditional probability symbol "|" emphasizes that the language state is conditionally dependent on the stimulus and the assumed language-specific network, which is rejected in this paper.
4. The negation symbol "$\neq$" indicates that the brain is not adequately represented by this indexical model of language processing, where a stimulus activates a language-specific network to determine the language state.

Since one of the attributes of natural kinds is *iconicity*, that is, they are susceptible to being perceived as distinct signs resembling what they stand for, it is not possible to announce the emergence of a brain-based natural kind based on indexical information only. Hence, what is intrinsic and essential in a natural kind cannot be defined through reaction or contiguity, as the reasons why something happens or is in *that* particular way and not another, remain underrepresented in the mind of the interpreter-conceptualizer. This can be summarized by the following sets and functions:

U = Universal set of all possible signs.

I = Set of iconic signs (resembling the object they represent).

X = Set of indexical signs (indicating the object they represent through a causal or physical connection).

N = Set of natural kinds.

B = Set of brain-based natural kinds.

S = Set of signs used to announce the emergence of a brain-based natural kind.

We can represent the disjoint nature of iconic signs and indexical signs using set operations:

$$I \cap X = \emptyset$$

Additionally, we can define a function $f: U \rightarrow [0, 1]$ that represents the degree of iconicity of a sign, where

0 indicates no iconicity, and 1 indicates a perfect iconic representation. We can define this function as:

$$f(s) = 1, \text{ if } s \in I.$$

$$f(s) = 0, \text{ if } s \in X.$$

To represent the essential nature of iconic signs in representing natural kinds, we can define a function g: $N \rightarrow [0, 1]$ that measures the degree of iconicity required for a sign to represent a natural kind:

$$g(n) = \alpha, \text{ where } \alpha \in (0, 1]$$

The value of α represents the minimum degree of iconicity required for a sign to adequately represent a natural kind. We can then establish the following condition:

$$\forall n \in N, \exists s \in U : f(s) \geq g(n)$$

This condition states that for every natural kind *n*, there exists at least one sign s in the universal set U whose degree of iconicity *f(s)* is greater than or equal to the required degree of iconicity *g(n)* for representing that natural kind. Furthermore, we can represent the limitation of using only indexical signs to announce the emergence of a brain-based natural kind by defining a function *h: S → [0, 1]* that measures the degree of indexicality of the set of signs *S* used for this purpose:

$$h(S) = (1/|S|)\,\Sigma\,(1—f(s)), \textit{ for all } s \in S$$

Here, *|S|* represents the cardinality (number of elements) of the set *S*, and the function *h(S)* calculates the average degree of indexicality of the signs in S by summing up the complement of the iconicity values *(1—f(s))* for each sign s in S and dividing by the total number of signs.

We can then establish the following condition:

$$\forall b \in B, h(S) < \beta$$

This condition states that for every brain-based natural kind *b*, the degree of indexicality *h(S)* of the set of signs *S* used to announce its emergence must be less than a threshold value *β*, where *β ∈ [0, 1)*.

In conclusion, we should resist the temptation to conjure up a Laplacean demon possessing a set of "compact truths" from which all other truths can be inferred (contra Chalmers 2014, xiv) and then impose this framework onto a theory of a linguistic natural kind in the brain.

1.1.2 The "predictive" black box

Embodiment theory is gaining traction in contemporary research (e.g., Lux et al. 2021). The discussion is often framed around whether concepts are represented in modality-specific formats, as demonstrated by effects like the *action-sentence congruency effect* (ACE; Glenberg and Kaschak 2002), or whether perceptual reenactment is unnecessary for conceptualization (e.g., Mahon 2015a; Vannuscorps et al. 2021). A working definition of embodiment for a theory of mind can be formulated as the process by which cognizers integrate external object-entity affordance configurations (external cognition) with the internal processing of somatosensory signals (internal cognition) to form approximate representations of the world. This perspective challenges the notion that perception generates its own images, emphasizing instead that "it is the relevant tract of the environment that is present to consciousness, not an image of it" (McDowell 1992, 455). For example, consider frogs. Neuroscientists might claim that the frog's eyes convey specific information to its brain about an object in motion, but, as McDowell (1992) argues, this "telling" does not originate internally. Instead, the environment itself, through its affordances, provides the frog with actionable information, facilitated by the frog's perceptual apparatus (McDowell 1992, 450). Here, I posit that our understanding of the world is mediated by the body via *multiple modalities*—such as proprioception, interoception, and exteroception—and further structured by somatosensory maps, which deliver sensory feedback to the brain (a process described here as *extended cognition*). This argument counters the view that perceptual information is merely re-accessed in modality-specific systems or brain regions (e.g., Barsalou 1982, cited in Machery 2010). Instead, cognition is framed as a distributed process involving the body and the environment.

On the other hand, the independence of thought from perception, as suggested by Mahon (2015a, 2015b, 2015c), assumes that thought resides solely in the brain. However, this paper supports the thesis that online reenactment of perceptually driven representations arises from action-perception loops. These loops, however, should not be simplistically reduced to limb-action matching (e.g., Mahon 2008). Instead, cognition occurs beyond the brain, extending to the body and the environment. From this perspective, the distinction between concrete and abstract concepts dissolves, as the form and content of representations are accessed as they are. The framework proposed here prioritizes biological processes over psychological ones, viewing cognition as involving entities external to the agent's body within its local environment (Sims and Kiverstein 2021). Consequently, purely symbolic concepts are untenable, as they possess semiotic properties—iconic and indexical (see Torres-Martinez 2022b)—that do not necessitate specific action-related mechanisms to achieve representation or systemic activity (contra Mahon and Hickok 2016).

An even more radical conceptualization of the disembodied brain has been proposed by theoretical neuroscientists, in the context of the *free-energy principle* (Friston 2010), who project mathematical models onto the biological function of the brain to create a functional abstraction that, nonetheless, fails to create an accurate representation of the role of language in the construction of reality. In this context, the notion of linguistic competence has been extended to computational models of the brain consisting of hierarchical, bottom-up processes whereby the "wetware" is deemed to work as an energy minimization machine (Murphy et al. 2024), with language being conceptualized as a code separated from its agentive dimension: "[E]mbracing the traditional distinction between competence and performance, our focus will be on the former (the mental formatting and generation of linguistic structure), and not on the range of complex cognitive processes that enter into the use of language in a specific context" (Murphy et al. 2024, 8).

Clearly, the authors introduce a model of language generation that reduces the cognitive processes involved in producing language to efficiency-led operations mapped onto an idealized brain-kind, which can hardly be made to correspond to actual brain functioning, as defined by neuroscience. Indeed, computational models of language processing do not necessarily overlap with the brain's chemical and electrical reactions associated with identified language regions (Tuckute et al. 2024).

While I would be the last to negate the benefits of proposing a predictive, uncertainty-minimization model of the brain (see Torres-Martínez 2023a; 2024a,b,c), it is also true that human language cannot be defined by merging symbolic syntax with Bayesian models of the brain. What we get is a self-referential mathematical model more concerned with maintaining internal theoretical coherence than providing evidence of its utility as a basis for the identification of ontological objects and natural kinds in the brain-body interface.

As I have stated elsewhere (Torres-Martínez 2024a,c), a more refined reading, informed by active inference (*AIF*, Friston 2009, 2010, see Sect. 3), posits that language, as part of a biological system, serves to link an organism's need to negotiate uncertainty with its goal-directed actions impacting the world. In this sense, a variation of Murphy et al.'s. formalism (2024, 5) states that the expected free energy $G(\pi)$ for a sequence of actions (policy) π over time t is given by:

$$G(\pi) = \int_t G(\pi, \tau).$$

Where $G(\pi, \tau)$ at the current time step τ is:

$$G(\pi, \tau) = E\_Q[\ln Q(\tilde{s}|\pi) - \ln Q(\tilde{s}\_\tau|o\_\tau, \pi)] - E\_Q[\ln P(\tilde{o}\_\tau)].$$

This can be rearranged as:

$$G(\pi, \tau) = E\_Q[lnQ(o\_\tau|\pi) - lnQ(o\_\tau|\tilde{s}\_\tau, \pi)] - E\_Q[lnP(o\_\tau)]$$

The components are:

1. Negative mutual information: E_Q[ln Q($\tilde{s}|\pi$)—ln Q($\tilde{s}$_τ|o_τ, π)] represents the reduction in uncertainty about the state $\tilde{s}$ by taking actions π compared to the expected uncertainty given observations o_τ and actions π.
2. Expected log evidence: E_Q[ln P($\tilde{o}$_τ)] is the expectation of new evidence or observations $\tilde{o}\_\tau$ based on the current actions π and observations o_τ.
3. Negative epistemic value: E_Q[ln Q(o_$\tau|\pi$)—ln Q(o_$\tau|\tilde{s}_\tau, \pi$)] quantifies the reduction in uncertainty about outcomes o_τ through actions π versus the expected uncertainty given the state $\tilde{s}\_\tau$ and actions π.
4. Extrinsic value: E_Q[ln P($\tilde{o}$_τ)] represents the expected benefit or utility of outcomes $\tilde{o}\_\tau$, regardless of the current state and actions.

We can thus assert that the expected free energy balances the extrinsic value (desirability of outcomes) and epistemic value (information gain) when selecting actions. Actions that minimize expected free energy are favored, as they lead to desired outcomes while reducing *uncertainty*.

Seen through the lens of non-artificial modeling of language construction with AIF, an organism can only maintain its internal homeostatic integrity when capable of detecting and reducing the effects of environmental disturbances. In other words, when the organism's internal states are attuned to the external conditions shaping its interactions with the world. A statistical boundary (the Markov blanket) interfaces incoming sensory signals, providing the system with information about external conditions. Perceptual surprise is thus a crucial evolutionary motivator, as the degree of acquaintance with the unknown is what ensures survival.

From this perspective, emergent human linguistic behavior is structured around recurrent hypothesis-driven semiotic models (termed herein as *Constructional Attachment Patterns* or CAPs, Torres-Martínez 2018a,b, 2019, 2020, 2021a,b, Torres-Martínez 2022a, Torres-Martínez 2022b), that is, hypothesis-driven semiotic relations among constructions that facilitate an embodied, perceptually-driven type of reality reconstruction (see Torres-Martínez 2022a). For example, CAPs are integral to broader cognitive processes that enable individuals to evaluate, predict, or hypothesize about events involving others' decisions, attitudes, beliefs, and behaviors. These include mechanisms like *action recognition* (e.g., Ferstl et al. 2017), which involves identifying and interpreting others' actions, and the reflection of specific mental simulations in language. For instance, simulations of speed are mirrored in linguistic structures (see Speed and Vigliocco 2014; Pan et al. 2024), as is the processing of abstract action

sentences (e.g. Balduin-Philipps et al. 2021; Schaller et al. 2016). Additionally, CAPs support cognitive tasks like *spatial perspective taking*, which involves adopting someone else's viewpoint to understand spatial relationships (e.g. Kessler and Thompson 2014; He et al. 2022). These examples illustrate how CAPs function within the complex interplay of cognitive, perceptual, and linguistic processes, emphasizing their role in bridging mental simulations and linguistic expression.

CAP selection is governed by "agentive imperatives" rather than purely statistical mappings, thereby creating meaningful categories. This process is summarized by the formula below (see also Torres-Martínez 2021a, 2021b):

$$(\mathbf{A}sl + \mathbf{B}sn) = \boldsymbol{P}\left(L \rightarrow \frac{l+}{s-}\right)^{MB}$$

According to this formula, when A (the salience of an exogenous or endogenous signal) becomes linked to B (the salience or psychological relevance of the resulting sign), it leads to P (the associative strength of the signal in reconstructing an event). In this framework, language (L) functions as a tool to reduce states of uncertainty (see Torres-Martínez 2020, 2021b, 2022a, 2022c). Ultimately, specific CAP selections are employed to define, shape, and guide an agent's actions in the world. In simpler terms, the association of events—bridging two experiential domains through event segmentation and sequencing—must meet two criteria: (1) it should be retrievable through a coherent and non-conflicting combination of CAPs, and (2) it must be experientially accessible to both the conceptualizing agent and the intended addressee. Relying on CAPs (Constructions as Attachment Patterns) rather than static signs provides significant advantages for understanding language. CAPs align with a dynamic view of the mind, where reality is constructed through the interplay of prior knowledge and new sensory input (see Torres-Martínez 2020, 2021b, 2022a, 2022c). This perspective supports the idea that language is not a fixed code but a flexible, adaptive tool shaped by users to navigate and interpret their experiences. CAPs seamlessly integrate embodied cognition—how physical and perceptual experiences shape thought—with intentionality, offering a multidimensional model of language that emerges from the interaction of speaker goals, contextual factors, and linguistic materiality. Thus, language is treated as a semiotic system actively molded by speakers, rather than as a static repository of symbols. CAPs show that language is a user-driven system, enabling speakers to creatively combine diverse signals—ranging from phonetic and tonal features to syntactic structures—into coherent, context-sensitive utterances. This adaptability reflects the speaker's ability to resolve uncertainty by tailoring language to specific communicative needs. For instance, in Chácobo, a language spoken by approximately 1200 people in the northern Bolivian Amazon (see Tallman and Elías-Ulloa 2020), tonal and stress patterns work in tandem with verb roots, affixes, and clitics to disambiguate events. In one example, stress and tone signal an intransitive event:

a.

pí-	ti[k	i̋ (affix)]	ki
['pí (stress)	ti̋ (tone) /k	i̋ (tone)/	ki̋ (stress)]
wing	break	intransitive affix	declarative predicate clitic

"S/he broke her/his own arm."

b.

pi	tiki̋ k i
['pì (tone)	ti̋ (stress). k i̋] (tone)
eat	AGAIN (affix) declarative predicate clitic

"S/he ate it/him/her again."

In this sense, the idea that language emerges from statistically relevant associations between objects/entities (natural kinds) being noticed by infants, during their interaction with caregivers, and its further coupling with specific argument structure constructions (syntactic templates with slots for lexical content), provides a very poor reading of language as a cognitive tool that is not governed solely by symbolic permutations moving on a continua of entrenchment and weakening associations. This computational reductionism of language has given a weak edge to usage-based linguists to postulate a possible link between human language construction and AI-generated language (e.g., Beuls and Van Eecke 2024).

On the contrary, human agents are defined by a type of *bodily awareness*—an umbrella term to refer to the representation of Self in the world through the conjunction of perceptual states, their conceptualization, and the projection of beliefs onto specific events in a context of use. In essence, this viewpoint positions language as arising from an organism's need to reduce uncertainty and sustain itself by accurately modeling its environment, with linguistic constructions emerging from embodied, agent-driven processes rather than purely statistical learning over data. In this sense, it is not possible for current artificial models of language generation to extract meaning from situated interaction with humans simply because human experience is not limited to the confines of an incremental acquisition-verification loop of linguistic data to attain a communicative goal. This, of course, contradicts the argument introduced by Turing and others, and reinforced constantly by computationalists (e.g., Beuls and Van Eecke 2024, 2025; Weissweiler et al. 2023), by which simulation of sentience and consciousness amounts to actual conscious states in a language game:

> We hope, however, to have convinced the reader that better integration of the situated, communicative, and interactional aspects of human linguistic communication constitutes the key to overcoming the limitations of current LLMs, and that more faithfully modeling the situated communicative interactions through which humans acquire their native languages provides a promising path towards more human-like language processing in machines. (Beuls and Van Eecke 2024, 28).

`

2 The risks of folk theorizing: extended humanism or posthumanism?

In this section, I lay out the foundations for the conceptualization of consciousness and first-person experience. No claim is made to completeness or definiteness in this theoretical quest. I am content simply with delimiting the contours of an emerging field of endeavor that, on the one hand, shows how far we have gone in our quest for the destruction of the human and what being human reveals about the *substance of the world*, and, on the other hand, unmasks the lack of consistency in our current knowledge of our very nature and purpose in the universe.

These reflections will revolve around the need to think of the human as a manifestation of a distinct set of properties that separates us from the non-human, especially emerging artificial agents. To dispel possible misunderstandings at the outset, I am compelled to say that my views of humanness and identity are not rooted in a Jungian "development of the I" (Harding 2020[1965]), but in the concept of the *extended human*, that is, an open biological organism bestowed with consciousness and aware of their deep connections with the natural world and other sentient and non-sentient beings. Extended humanism seeks to downplay the underlying tension plaguing discussions of AI/deep learning advancement usually split into views defined by the overconfident tone of nurture-favoring machinic neo-empiricism (privileging an unbound algorithmic narrative of machines "on the verge of achieving escape velocity into world-spanning superintelligence (…) (Buckner 2023, 4), or the *nativist* (also referred as *innatist*) position by which rational thinking can only be achieved by humans thanks to "the human mind's innate startup software" (Buckner 2023, 4).

My project also seeks to dismantle the dehumanizing rhetoric of (empiricist-driven) posthumanism, a program that, as suggested in Sect. 1, extends over several fields through a parsimonious deconstruction of the Self as a means to avoid (nativist) anthropocentrism and, more particularly, monolithic forms of Western-centric "we" narratives (*nos majestatis*, as the Romans called it) (Massimi 2022: 366). As we will see, however, the result has been the emergence of an antihumanist discourse, fueled by beliefs about 'virtuous' and 'vicious' theory building (West 2024) that relativizes the human body and mind (for instance, scientists are often pigeonholed on the basis of their adherence to politically correct discourses, rather than being recognized for the soundness of their contributions), which negates the complex role of humans as a species in need of broader strategies to defeating the "darker side of [their] success" (Desmond and Ramsey 2023: 3). These include, as noted previously, broadening the scope of theoretical philosophy of science so as to include more practical questions such as why humans uncritical overreliance on technological advancement may cause harm to *all of us*, not only particularly *vulnerable groups* (a definition that is in itself biased towards the "sustainable-development" of politically correct narratives through AI ethics, see Curto et al. 2024).

Therefore, in the context of AI-generated language and human–machine interaction, many philosophers and researchers have positioned pretrained transformers and chatbots like ChatGPT as distinct agents capable of creating culture and meaning.[1] This extended *engineering empiricism* "draws oxygen from a simplistic understanding of the relationship between the successes of deep learning systems and the way that humans and animals actually solve problems." (Buckner 2023, 3–4). As a result, meaning is treated in very a loose manner. For example, it has been argued that machines that produce some kind of autonomous meaning

[1] The language generation process of ChatGPT begins with tokenizing input text XX into a sequence of tokens X = {×1, ×2, …, xn}. Each token xi is transformed into a high-dimensional vector representation ei through an embedding function E(xi) = ei.

The model employs a Transformer-based architecture with layers defined by a generalized transformation:

Layeri(input) = LayerNorm(MultiHeadAttention(input) + input).

The attention mechanism, a key component, is calculated using the equation:

Attention(q,k,v) = softmax(q·kT/$\sqrt{}$dk) · v.

Where q, k, and v represent query, key, and value matrices, respectively. The softmax function with scaling by $\sqrt{}$dk helps manage the dot product's magnitude.

During training, the model minimizes the cross-entropy loss:

L(Y,Ŷ) = -∑i yi log(ŷi).

This loss function measures the divergence between predicted and actual token probabilities. The model can be fine-tuned on domain-specific data by updating parameters using gradient descent:

θnew = θold—α · ∇θL.

Where α represents the learning rate and ∇θL is the gradient of the loss with respect to the model parameters. In the decoding phase, the next token xt̂ + 1 is predicted by maximizing the conditional probability:

xt̂ + 1 = argmax_x P(x |× 1, ×2, …, xt).

This means selecting the token that maximizes the probability given the previous tokens.

Finally, the generated text undergoes post-processing steps like detokenization, filtering, and formatting to produce a coherent and polished output.

are *agents* and possess consciousness. The claim goes so far as to assign meaning-making properties to processes that, by virtue of an empiricist heuristic, emerge as embodied cognitive entities with broadened forms of thought, "depending for its specificities on the embodied form enacting it" (Hayles 1999, xiv). This conflation of sign-making processes and output, which does not necessarily contain meaning, has served as a conceptual tool for elevating all sorts of sign users to the status of conscious beings, under the belief that all it takes to be an agent is to process information to complete a task.

Nevertheless, in dealing with signhood as a property of predictive, feed-forward cognition, we must also bear in mind that sign systems—and not signs alone—are the manifestation of purposeful action and not the action per se. The implication of this is that a bedrock condition for consciousness to exist and provide context for meaning is the possibility of an entity to expand its cognitive grasp, offload intentional content, and predict *what-will-be-the-case* within an event that needs to be represented as an actual happening having bearings on a cognizer's grasp of existence. Sign production is thus the result of what is known by the cognizer in terms of purposeful meaning-making and not what is generated in the form of signs through natural or engineered computations. Inevitably, this opposes unconstrained definitions of cognition as "a process of interpreting information in contexts that connect it with meaning" (Hayles 2017, 22; cited in Hayles 2024, 31; Hayles 2021, 6). The reason is that representation is not a homogeneous process geared by "pure" signs directed to and emerging from a species-specific Umwelt (see Torres-Martínez 2023a, b, c), but a nuanced blending of property identification, category assignation, and intentional contextualization for the emergence of a distinct *sign system* providing an organism with identity. Thus, despite the best efforts of posthumanist (new materialistic) views, machines are still far from qualifying as intelligent meaning-making agents, and so are "material processes" such as avalanches and tornadoes, which clearly are not agents (contra Hayles 2024, 33).

The semiotic juggling that conflates information processing with consciousness is also predicated upon the notion of *language-as-performance,* by which LLMs are elevated to the status of "the best predictive models of human language representations at the resolution of data [neuroscientists] have access to" (Tuckute et al. 2024, 285), an idea that has been integrated into ethnographic, antihumanist (posthumanist) research agendas seeking to ascribe an ontological status to pretrained transformers in the belief that these language generators are *authentic* language creators (e.g. Demuro and Gurney 2024).

According to this reading, such models actively contribute to the cultural construction of knowledge through unique "reasoning processes" that, nonetheless, are distinct from those of humans. This assertion stems from a general rejection of any "universalist" definition of language which is associated with an anthropocentric humanist tradition, and instead views language as a manifestation determined by the *performance* of all sorts of agents through *communication-as-action*. Consequently, pretrained language models are seen not just as performers through language, but as active participants in its creation, thereby shaping events where language unfolds through both performance and competence, which "relativize[s] the human by coupling it to some other order of being" (Clarke 2008, 3).

Fittingly, human language is depicted as a process of assembling symbols within a grid of semiotic relations determined by the shifting shapes and affordances of specific cognitive technologies and objects (e.g. Hayles 2021), with speakers functioning as participants bound by the rules of extant signhood preservation practices (language games) within a self-sealing *field event* (Hayles 2018). As a result, human language is portrayed as a residual aspect of interaction, lacking any inherent transcendental qualities (humanness), which redefines LLM language generation as an integral component of authentic "encounters" where participants, both human and non-human, share a sense of experiential adjustment (e.g., Demuro and Gurney 2024; Dynel 2023). This shift obliterates any claim to a construction of the Self-as-symbol through an emphasis on a redefinition of power relations that replace (anthropocentric) universalism with a focus on addressing gender inequality.

> My anti-humanism leads me to object to the unitary subject of Humanism, including its socialist variables, and to replace it with a more complex and relational subject framed by embodiment, sexuality, affectivity, empathy, and desire as core qualities (Braidotti 2013: 26).

And yet, since human agency involves purposeful, goal-oriented action aimed at both reducing uncertain states and constructing preferred models of interaction with the environment, human beings, as distinct biological systems, participate in a continuum of relations with the world and other species that go beyond denaturalized claims to power redistribution through the effacement of the matter-energy boundary. Under this view, it is not possible to posit an essentialist (though undefined) semiotic substance ("language") that can be reduced to underrepresented linguistic practices propelled by language use ("languaging"), definable a priori by a particular set of beliefs in the mind of an interpreter-anthropologist (e.g. Demuro and Gurney 2023).

The primary contradiction in the posthumanist reading lies in the fact that even the most contingent set of relations is driven by biological imperatives that cannot be relativized through the postulation of indefinite intrinsic properties tied to a material network of relations:

For posthuman theory, the subject is a transversal entity, fully immersed in and immanent to a network of non-human (animal, vegetable, viral) relations (Braidotti 2013, 193).

As we can see, the passage offers a chain of rhetorical devices whereby the unsupported negation of humanness, through the medium of "stylistic use of prose to evade stating one's argument" (Rickabaugh and Moreland 2023, 5, citing Searle 1992, 9), is dissolved in a definition that reifies an amorphous *substance of things* that denies the pivotal role of the human in the comprehension of the relations we are part of. Braidotti's wording also points to an underrepresentation of the intrinsicality of the human in the relations she attempts to describe, since, as I have shown elsewhere (Torres-Martínez 2024d,e), the property of being intrinsic (the "what") cannot be separated from the deployment of the essential (the "how").

This new materialist reading of the Human (Coole and Frost 2010), assumes that, since humans and machines are material renditions of broader processes of energy exchange and conservation, consciousness can reside in mathematically feasible recipients other than biological entities. Furthermore, agency is not a property of sentient beings but an affordance of phenomena defining cause-effect events (landslides, storms, hurricanes, and so on). On this reading, proponents of mathematical theories view LLMs as a means to accurately predict how meaning can be extracted from symbols and mapped onto the internal representations and concepts possessed by speakers immersed in an idealized state of *unity*, which is "often thought of as part of a descriptive theory of ideal rational agents, not of real agents" (Kinderman and Onofri 2021, 2). A prime example of this mindset is the attribution of essential properties to chatbots, as Dynel (2023) states:

> Since ChatGPT can adapt its responses to match various discourse styles and registers, users can witness and understand how the model adjusts its language based on the input provided (p. 122).

In this passage, the author ascribes essential properties—qualities that allow something to behave or act towards a goal—to the chatbot. This suggests that its adjustments through training and enhanced processing capabilities create a form of *ontological dependence* between the user's needs and the prompt-driven event governing the generation of "polite" language. According to Dynel, this ontological dependence actually mirrors the inherent naturalistic linguistic interactions between human agents. The point to notice, however, is that the human perception of AI-generated language requires an omission of the layers of abstraction inherent in computational artifacts. These layers include, among others, the designer's intentions for the solution formulated to address a given problem. Moreover, it is crucial to recognize that displaying polite behavior (the manner in which someone behaves politely) does not inherently equate to possessing an *intrinsic quality* of being polite (the inherent quality that makes someone behave politely). Consequently, the suggestion that chatbot-human interaction entails a degree of naturalistic alignment of intentions, involving meaningful negotiation and pragmatic adjustment, becomes implausible when only one participant, the human agent, possesses an awareness of their conscious states enabling them to plan beyond the immediate chatting event (the AI's behavior is guided solely by the prompt imposed by the human). Clearly, [t]here is no *one* at the other end for the human to communicate with. Human–computer interaction is not inter*personal* communication" (Duncker 2020, 97).

The type of conceptual relativism and attitude revisionism proposed by posthumanism had been previously framed by Turkle (2005, 59) in terms of an "adulthood" bias by which grown-up humans attribute consciousness only to organic entities:

> The children take a different view. The idea of an artificial consciousness is not what impresses them. They may be the first generation to grow up with such a radical split between the concepts of consciousness and life, the first generation to believe that human beings are not alone as aware intelligences. The child's splitting of consciousness and life may be a case where instead of thinking in terms of adult ideas "filtering down" to children, it makes more sense to think of children's resolutions prefiguring new positions for the computer culture to come.

Indeed, the postulation of humans, "a strange product of evolution that can perceive itself and its fellow humans as paradoxical" (Rosendahl 2013, 224), as self-symbols striving for biological priority and persistence, provides a solid basis for our definition of language as a cognitive tool that cannot be fully emulated by artificial systems of language generation.

Such a radical embodied position goes against the grain of a growing realization among chatbot users that sentience, and even a soul-like otherness, is a by-product of our interaction and our conditioned reactions to chatbot's apparent empathy and well-calibrated responses. Sentience attribution is thus the result of a psychological projection onto an artificial entity that "enlivens the robots, even as the people in their presence are enlivened, sensing themselves in a relationship" (Turkle 2010, 85).

2.1 Consciousness attribution leads to output averageness

The problem of sentience attribution in the study of human perception, regarding both AI-generated and

human-constructed content, is that it is not evident that creativity attribution to LLMs as intelligent tools could eventually be used as a means to postulate a human flawed heuristic (e.g., Porter and Machery 2024). In other words, it is not the case that human heuristic judgments could be compacted in a single-layered experiential proposition that excludes other experiences (e.g. *Information Integration Theory* (*ITT) 3.0*, Oizumi et al. 2014, 3l see Sect. 3 below) and, hence, susceptible to being subsumed to the affordances of LLM performance, an idea that is central to ITT 4.0's formulation:

> By the intrinsicality postulate, the Φ-structure [the structure of the amount of consciousness] of a complex depends on the causal interactions between system subsets, not on the system's interaction with its environment (except for the role of the environment in triggering specific system states). In general, different physical systems with different internal causal structure may perform the same input–output functions (Albantakis et al. 2023, 36).

The reason is that, as a form of inference, heuristic (clue-driven) decisions do not depend on a bias towards *mathematical simplicity*, that is, towards mechanized, meaningless habits within a closed physical system, since the nature of pattern finding for complex biological systems is reliant on a bottom-up identification of those elements (extracted from environmental signaling) that can provide a framework for the postulation of *emergent minimal working systems* (conceptual systems providing agents with a sense of consciousness-as-action). In terms of creativeness, the implication is that AI-generated output provides a minimal working system that accommodates clue-based inference without there being a need to *reduce world complexity*.

Thus, the mixing of *creativity perception*, *content generation, and machinic representation* as a graded dimension of attribution (the assignation of properties to a tool), has led some to conclude, for example, that "people prefer AI-generated poetry to human-authored poetry, consistently rating AI-generated poems more highly than the poems of well-known poets across a variety of qualitative factors" (Porter and Machery 2024, 9). Notwithstanding, the preference for AI-generated content among chatbot users does not imply poor heuristics, but a reduced investment in context analysis through enhanced attribution of precision and accuracy to AI technologies for precision leveraging and human outperformance (Jablonka et al. 2023). In other words, the use of empty signs results in blind sign deployment: "By using signs emptily, we do not even terminate in the consciousness of whatever it signifies anymore, because we use the habitual sign itself habitually, we do not invest mental activity into bringing the signified into focus, since we'know' what it means." (Arnold 2022, 216–217).

Clearly, the conflation of human inference with AI content generation efficiency introduces a major risk, namely the elevation of *output averageness* through data—AI output tends to a mediocre flattening of reality, as is the case in medicine (Clusmann et al. 2023; Schmidgall et al. 2024)—as the new direction in all conceivable fields of human action, science included (Tuckute 2024). For example, enthusiasts of *enhanced* science-making through AI processing, assert, quite unashamedly, that the human mind should eventually become an outgrowth of AI technologies as a means to attain Noble-prize level influence and recognition through "top-level science" pattern definition, thereby positioning AI technologies as a solvent for human error and bias by means of a "Turing grand challenge" (Kitano 2021, 1): "a grand challenge to develop AI systems that can make significant scientific discoveries that can outperform the best human scientist (…)".

However, the reduction of science to technology and "innovation" by way of perplexity values (that is, AI performance values leading to "efficient algorithmic discipline"), announces a world where bad science will become instrumental to a halt in human problem solving and insight. The stated goal of endorsing AI-powered automation as a standard for experiment design, data analysis, information collection, and results interpretation (e.g. Boiko et al. 2023), is thus to increase output optimization, rather than true scientific development. But such advancements come with a cost, namely that LLM deployment is operated on the premise that users "will do and fill the gaps", thereby resulting in unsupervised applications with weak LLM uncertainty quantification (see Catak and Kuzlu 2024): "There remains a possibility for users to elicit fundamentally novel behaviors from LLMs, unforeseen by developers, through causal reasoning processes akin to those observed in GPT-3, defying current evaluation or analytical methodologies." (Kumar 2024, 39).

As we will see, however, the emergence of "disembodied subjectivities", including not only the artificial but the projection and ultimate extension of human cognitive states onto devices and non-human *others*, reinforces the idea that biological agency, driven by phylogenetic rules for survival and semiotic persistence, can guarantee a relevant place for our species in a world dominated by performative actants. To note, my use of "persistence" does not seek to position a code-driven idea of biological reality through heredity only.

As it will become evident in the next pages, my intention is to delineate the conditions for the emergence of an extended organic species evolving within a context of predictive semiosis not restricted to the mere passing of information through DNA-based heredity. That said, the idea of humans transcending their mortal organic bodies (transhumanism) as a means to outpace non-organic performers is replaced in this paper by a type of human that does not

surrender their symbolic self to the promise of enhanced signhood through a machinic turn.

Nor is it suggested that the road ahead for "the Human" implies overcoming the old self to become more tethered to a shifting parcel of reality defined by an endless loop of efficacy and loose sense of progression through disruptive "encounters". Biological signhood is thus the result of pushing aside both the organic and the technological myth—cherished by some as a means "to compose enactment" (Hayles 1999, 196; Hayles 2012)—that separates us from the world of the living and non-living. In this sense, the idea of language as a semiotic tool for self-preservation and the construction of preferred futures entails an irreducible sense of being in the world that cannot be emulated by machines, since "life and sentience are coterminous. Life is based on cells, biomolecular entities of immense complexity. The *living* element, the biological component, is the critical factor." (Reber et al. 2023: xiv; emphasis in original).

The present conceptualization also highlights the idea that being embodied does not simply mean to act through and with a physical body but entails the existence of a sense of boundary that provides us with bodily identity.

This is precisely one of the problems of non-embodied approaches that attempt to fuse the human body with the objects profiled by our extended cognitive affordances (e.g. Hayles 2012). Therefore, in postulating markers of cognitive utility (but, paradoxically, of semiotic dispersion and discontinuity) and usability as internalized "nonconscious" extensions of the cognizer, "the nonconscious cognitive assemblages through which these distributed cognitive systems work" (Hayles 2017, 2), the author adheres to a definition of consciousness governed by a utilitarian performance-driven loop whereby our acts in/on the world become a (contingent) measure of our relevance therein. The relations entertained by entities and objects are hence made to depend on a discrete definition of identity as an atomic component of a "self-referential field" (Hayles 2018, 10) where language is reduced to a code connecting a network of statements referring to themselves through a fuzzily defined continuity governed by family resemblance.

The question that arises is: How can Hayles' field metaphor account for the emergence and further deployment of eventful states to an Interpretant? For consciousness does not emerge as a result of a metalinguistic reconstruction of the process, but the definition of the meaningfulness of *its becoming* as part of an agent's (re)construction of identity; which, of necessity, is and will always be a measure of the mind defining the boundaries of the actions that guarantee its persistence in time and space. It is here that the notion of consciousness (and a definition thereof) becomes handy.

3 The uniqueness of human consciousness

Where, then, is consciousness to be found? What amounts to a conscious state, and what kind of consciousness can be said to provide the access of a system, biological or not, to a coupling of inner-outward experience? *According to Information Integration Theory* (*IIT 1.0*, Tononi 2004; *ITT 3.0*, Oizumi et al. 2014; *ITT 4.0*, Albantakis et al. 2023), the question of consciousness needs to be defined from a first-person perspective. In this vein, IIT considers two problems: 1) the extent to which a system can be considered as having conscious experience, and 2) the definition of the conditions that determine the type of consciousness a system has. This phenomenological reading departs from the assumption that, for a system to be conscious of itself and of what is going on around it, it must, first, be able to deal with large amounts of information (differentiation), and second, be capable of integrating that information into a coherent semiotic picture of the world (integration). Moreover, the whole process is said to be private and occurs in specific areas of the brain.

Expectedly, the most significant weakness of IIT lies in its epistemological limitations, since it posits that systems such as photodiodes may possess consciousness, yet there is no independent method to verify such claims. Furthermore, the theory's central metric, Φ (integrated information), faces significant issues. Calculations of Φ often yield non-unique values that depend on arbitrary modeling choices, making it unclear how such results should be interpreted. Additionally, the range of Φ values lacks specificity, thereby preventing clear conclusions about the relationship between Φ and conscious experience. Philosophically, IIT introduces problematic and counterintuitive claims. For example, the theory implies that consciousness can be quantified, leading to assertions that a 2D grid might possess "10 times the amount of consciousness" as the human brain. Such statements are not only counterintuitive but also conceptually incoherent, as they fail to clarify what "amount of consciousness" means. Moreover, although, as it will become evident further below, I reject the idea of a Hard problem of consciousness (Chalmers 1996), IIT does not adequately address such a problem (its stated goal), that is, the question of why and how physical processes give rise to subjective experiences. This leaves the theory vulnerable to the "zombie argument," wherein one can imagine a universe populated by high-Φ systems that exhibit no inner experience. These scenarios expose IIT's inability to account for the qualitative dimensions of consciousness, which remain essential to any comprehensive explanation.

It is no coincidence then that my selection of this specific theory of consciousness targets the consequences of reducing self-awareness to a series of symbolic operations subsumed to a computational bias mysteriously encoded in

our brains, usually termed functional consciousness, that is, based on "mere functions and abilities" (see Hassel Mørch 2023, 3).

The paradox here is that ITT sells itself as a phenomenal theory of consciousness by which "[p]henomenally conscious states are characterized by the fact that there is *something that it's like* for a creature or entity to be in them, or that they are subjectively experienced or felt" (Hassel Mørch 2023, 3; emphasis in original). And yet, on closer inspection, ITT's focus is primarily on the physical aspect of sentience, having the brain as a recipient of seeming subjectivity and self-awareness.

Thus, in a manner akin to the posthumanist musings reviewed previously, many studies have seen in ITT the opportunity to dissolve human singularity into the alleged materiality of a "self-conscious" wetware producing signal-driven informational interactions as a response to a stimulus (e.g., Montoya 2023). Central to this claim is the idea that conscious responses are fired by a stimulus that can be further incorporated into the system to produce an experiential response. This leads us to a second objection to ITT: its reductionism of conscious states to an impoverished version of the brain as a physical support of experience (a quantity).

In contrast, the present proposal defines consciousness as a combination of four main elements: (1) The identification of objects in the world associated with events of perception (experience is intrinsic, though modeled on shared experiential structures in common with other organisms). (2) The construction of pictures of the world through the integration of modal and amodal information processes in the brain (information is provided by different modalities and then organized in the brain in specific areas. The nature of experience is defined by degrees of perceptual access to complexes (composite concepts) and not by simulations of experience in the form of "shadows of perception". (3) The content of experiential states becomes integrated into the agentic projection of intentional states onto the world. (4) The alignment of subjective conscious states with other conscious entities is facilitated by the construction of a model of the self both as an individual and as part of a system of agentive relations and hierarchies. Agentive hierarchies are the instances by which decisions and non-decisions are made to reveal to an agent or an observer the purpose and reason for engaging in a particular action given a state of affairs within an event. As previously suggested, such delimitation requires a sense of boundary that separates, though not insulates, the feel of being in possession of a body that is distinct from the flux of energy surrounding it.

As we saw in *sub*Sect. 1.1.2, the notion of bodily awareness as an instance of agency is partly indebted to the AIF theory (Friston 2009, 2010), according to which organisms need to keep a balance between environmental and inner-body information to avoid entropy (systemic collapse). This defines the role of biological systems as predictive entities moving toward specific attracting states, that is, points of equilibrium that minimize entropy to attain homeostatic integrity. Homeostasis (inner-system equilibrium) is accomplished through the identification of sensory states by means of ad hoc receptors placed in a statistical boundary (Markov Blanket) providing a link between system-external and system-internal conditions.

To assess this in the context of human language production, we need to explain what perceptual experience is supposed to be. In the first place, language is constructed on concepts, and the existence of concepts point to a need to recognize the substance of objects and entities in the world. So, a definition of concepts is required at this point.

3.1 Defining concepts

In the context of intelligent agency, concepts can be defined as the sum of animate (natural) and inanimate (natural or constructed, aka. "extended") entities that provide agent-conceptualizers with perceptual maps for the reconstruction of reality. I classify concepts into three main categories: Prototypes, exemplars, and theories.

3.1.1 Prototypes

Prototypes have their origin in a basic set of idealized tendencies that are entrenched as the result of beliefs about intrinsic properties, that is, the *what* making up the substance of objects and entities in the world. These provide conceptualizers with sensory-driven, *part-to-whole* relations leading to semiotic integration. For example, bird, fish, or feline essences consist of particular features that are ascribed to a prototypical natural kind (NK):

Prototype = EssenceN → *prototypicalNK*

Examples:

Feline glance:

Essence_N: The idealized tendency associated with feline features (e.g., eye shape, expression).

Prototypical NK: The general concept of a cat, including shared characteristics.

Example: The way a cat looks at something, characterized by the specific shape and expression of its eyes, can be considered as a prototype that encapsulates the essence of feline visual characteristics.

Cat-like gait:

Essence_N: The idealized tendency associated with the way a cat moves (e.g., posture, movement pattern).

Prototypical NK: The general concept of a cat's movement.

Example: The distinctive walking style of a cat, with its characteristic posture and movement, serves as a prototype representing the essence of feline locomotion.

3.1.2 Exemplars

Exemplars are extended sets of stored prototypes representing *whole-to-part* associations of both natural and inanimate kinds (concepts). In other words, an exemplar is the best sample of a set of prototypes. Exemplars integrate more formally somatosensory experience with higher-order conceptualization.

3.1.3 Theories

Theories, or beliefs, represent the combination of perceptual mappings and experiences aimed at swiftly accessing categories. The assignment of categories involves a process of reality reconstruction, where the actions of the world upon natural kinds (including entities like humans) produce perceptually accessible effects. The essence of natural kinds resides at the intersection of cognition and patterns of biological adaptation to the environment, encompassing phylogenetic extension and ecomorphological convergence. Categories emerge from a synthesis of perception, beliefs, and our interactions with the world and other organisms. Natural kinds, being mind-independent, originate from our phenomenological grasp of ways of being in the world, facilitated by our sensitivity to various natural effects shaping the essence of organisms (confer Torres-Martínez 2024c).

The identification of the essence of things goes beyond the simple categorization of observable features along a continuum of prototypical characteristics that entities possess. Indeed, unobservable essences are not homogeneous, linearly arranged properties of identifiable natural kinds, but rather fragments of experience retrievable across various layers of perceptual modeling in a bottom-up manner. The idea is that language connects these essences to observable surface traits through specific affordances, represented as a vector E = [e1, e2, …, ek]. So the essential properties of a natural kind like a tiger are not just symbolic representations of intrinsic properties in a conceptualizer's mind but reflections of a network of potentialities inherent in events where these essences are active (Torres-Martínez 2024c).

It is important to see how much this objection militates against the possibility of postulating single-layered propositions for the description of intrinsic features in natural kinds. It is perhaps not surprising that defining the existence of an entity or object requires positing a multilayered set of propositions following a model like this: "If condition S becomes true upon the lower scale length ΔL, then condition S* will become true on the higher scale length ΔH" (Wilson 2023, xiv-xv).

On my account, then, a property essential at one scale does not prevent it from being intrinsic at another level (see Torres-Martínez 2024b). For tigers, lacking stripes may not inherently negate core "tigerness," but could hamper camouflage and hunting success existentially. Therefore, the fur property cannot be separated from its functional value in the natural environment. As a result, statements like "Redness and squareness are intrinsic properties. Being next to a red object is extrinsic" (Vallentyne 2014, 31) become nonsensical given the multilayered nature of linguistic reconstruction.

This reasoning has limits since stripes and orange fur are intrinsic for successful hunting but potentially extrinsic from the perspective of prey like Chital that could be deceived by the camouflage. So it is inaccurate to say "[w]hether something is red, or 3 kg, or round is a matter of how it itself is, regardless of anything else" (Denby 2014, 91).

4 Conclusion

Throughout this paper, I have challenged perspectives that attribute genuine intelligence, consciousness, or the capacity to construct meaning and culture to generative artificial intelligence, particularly large language models. This becomes clearer when considering the epistemic blindness pervading the dehumanizing tenor of current linguistic and neuroscientific research, which can only be interpreted as a strategy to amplify these fields' influence on other epistemic communities through the imposition of their own methods, ontological formulations, and agendas (e.g. Tuckute et al. 2024). This is compounded by narratives that obscure the human experience framed within a posthumanist discourse that erroneously posits machinic information integration and performance as a prerequisite for sentience and agency. This has been interpreted as a threat to science in many respects, requiring analysis from philosophers of science who warn against the proliferation of *science-before-testimony* that reduces the Human to data and perplexity values (see Torres-Martínez 2024c). I have argued that a means to correct this trend is to place the focus on the definition of foundational concepts such as consciousness, natural kinds, and concept formation. To underscore the central argument, the thesis has been that human consciousness and agency are intrinsically tied to our biological architecture through the holistic integration of perception, conceptualization, intentional action, and self-awareness.

It is instructive to remember that, in the absence of a clear distinction between the human and the machinic, it becomes painfully easy to fall into the conceptual trap of unfounded sentience attribution. Evidently, advocates of information integration as a purported sign of consciousness quickly run out of defensible ground, since, although computational models may simulate aspects of language

processing, they fundamentally lack the grounded, embodied experience and drive that shape human linguistic behavior. This lack of alignment, that is, the requirement on the part of (human) *choice architects* (see Mills and Skaug Sætra 2024) to align with human values and preferences, departs from the premise that "it is logically necessary that AI does not itself substantially shape the very cognitive resources with which it aims to match; if it does, alignment, for example, would become recursive and incoherent, i.e., AI aligning with itself." (Wihbey 2024, 4).

AI-self-alignment, as supported by some cognitive linguists and neuroscientists poses a threat to the persistence of human knowledge construction and the values providing meaningfulness to the process. This advances us toward an emerging realization that language arises from an organism's need to reduce uncertainty about its environment and sustain itself through accurate predictive modeling that reflects the speakers' worldview and agentivity. Alternatively stated, linguistic constructions emerge from an embodied, agent-driven process rather than purely statistical operations over data (contra Goldberg 2019). This rejects both the empiricist notion that language and meaning can be entirely captured by information processing machines, as well as posthumanist views that conflate sign production with conscious meaning-making. Thus, we may take it as substantiated that the crux here is that sign systems manifest purposeful action but are not equivalent to the action itself, which contends the idea that "the post-humanist movement aims to question human exceptionalism and the role that modernity has given to human beings." (Gómez Redondo et al. 2024, 12). This interpretation has considerable merit, for example, to reveal that the attribution of agency and genuine semiosis to artificial systems is a conceptual overreach stemming from folk, before-testimony theorizing that fails to grasp the nuances of biological consciousness.

Ultimately, this paper has characterized *the Human* as a manifold of qualities and properties associated with an intrinsically embodied species whose linguistic capacities are profoundly intertwined with their biological identities as conscious, sense-making agents tuned to the world. As has been remarked before, the extended humanist position embraced herein stands in stark contrast to reductive perspectives championing the deconstruction of "binarism" and "normalized selves" as a means to impose a blind neo-*triadism* wherein the "Other" (the permanently undefined *Third*, including machines) is solely construed in terms of its capacity to introduce ambiguity and disruption, while failing to recognize other subjectivities as inherent rather than obstructive to a strategy "to disrupt a humanist version of being" (Toffoletti 2007, 84). While generative AI can simulate surface behaviors and smooth the introduction of "a model of the self as either entirely resisting or complying with particular aspects of culture" (Toffoletti 2007, 91), it cannot replicate the holistic experiences and organic self-modeling that fundamentally underlie human language and meaning construction, irrespective of how the Human is conceptualized.

Author contributions Sergio Torres-Martínez: writing—review and editing, writing—original draft, investigation, formal analysis, conceptualization.

Funding This research was not funded by any institution or agency.

Data availability No data has been used in this paper.

Declarations

Conflict of interest The author declares no conflict of interest.

Data There is no data in this manuscript.

Use of AI tools No AI tools were used in the construction of the paper.

Code No code has been created for the study in the paper.

Ethics approval Not applicable.

References

Albantakis L, Barbosa L, Findlay G, Grasso M, Haun AM, Marshall W et al (2023) Integrated information theory (IIT) 40: Formulating the properties of phenomenal existence in physical terms. PLoS Comput Biol 19(10):e1011465. https://doi.org/10.1371/journal.pcbi.1011465

Andersson P, Strandman A, Strannegård C (2019) Exploration strategies for homeostatic agents. In Artificial General Intelligence: 12th International Conference, pp. 178–187.

Arnold T (2022) the tragedy of scientific culture: Husserl on inauthentic habits, technisation and mechanisation. Hum Stud 45:209–222. https://doi.org/10.1007/s10746-022-09621-x

Balduin-Philipps LS, Weiss S, Schaller F, Müller HM (2021) Abstract action language processing in eleven-year-old children: Influence of upper limb movement on sentence. Compr Behav Sci 11(12):162. https://doi.org/10.3390/bs11120162

Barsalou LW (1982) Context-independent and context-dependent information in concepts. Mem Cognit 10:82–93. https://doi.org/10.3758/BF03197629

Beuls K, Van Eecke P (2025) Construction grammar and artificial intelligence. In Mirjam Fried and Kiki Nikiforidou (eds.), Preprint. To appear in the Cambridge Handbook of construction grammar. Available at: arXiv.2309.00135

Beuls K, Van Eecke P (2024) Humans learn language from situated communicative interactions what about machines? Comput Ling 50(4):1–35

Block N (1995) On a confusion about the function of consciousness. Behav Brain Sci 18:227–247. https://doi.org/10.1017/S0140525X00038188

Boiko DA, MacKnight R, Kline B, Gomes G (2023) Autonomous chemical research with large language models. Nature 624:570–582. https://doi.org/10.1038/s41586-023-06792-0

Braidotti R (2013) The posthuman. Polity Press, Cambridge, UK

Buckner CJ (2023) From deep learning to rational machines: What the history of philosophy can teach us about the future of artificial intelligence. Oxford University Press, Oxford

Butlin P (2022) Machine learning, functions and goals. Croat J Philos 22(66):351–370. https://doi.org/10.52685/cjp.22.66.5

Catak FO, Kuzlu M (2024) Uncertainty quantification in large language models through convex hull analysis. Discov Artif Intell 4(90):1–14. https://doi.org/10.1007/s44163-024-00200-w

Cave S, Dihal K. (eds) (2023) How the world sees intelligent machines: Introduction. In Imagining AI: How the world sees intelligent machines. Oxford University Press, Oxford, pp 3-15.

Chalmers DJ (1996) The conscious mind: In search of a fundamental theory. Oxford University Press, New York, NY

Chalmers DJ (2014) Constructing the world. Oxford University Press, Oxford

Chomsky N, Gallego Á J, Ott D (2019) Generative grammar and the faculty of language: Insights, questions, and challenges. Catalan Journal of Linguistics, Special Issue: 229–261. https://doi.org/10.3390/e25091328

Clarke B (2008) Posthuman metamorphosis: Narrative and systems. Fordham University Press, New York

Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J-N, Laleh NJ, Löffler CML, Schwarzkopf SC, Unger M, Veldhuizen GP, Wagner SJ, Kather JN (2023) The future landscape of large language models in medicine. Commun Med 3(141):1–8. https://doi.org/10.1038/s43856-023-00370-1

Coole D, Frost S (eds) (2010) Introducing new materialisms, New materialism: Ontology, agency and politics, 1–43. Duke University Press, Durham

Cortez N, Sage WM (2023) The disembodied First Amendment. Washington University Law Review 707. https://ssrn.com/abstract=4406190

Curto G, Jojoa Acosta MF, Comim F, Garcia-Zapirain B (2024) Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings. AI & Soc 39:617–632. https://doi.org/10.1007/s00146-022-01494-z

Denby D (2014) Essence and Intrinsicality. In Francescotti, R. (Ed.), Companion to intrinsic properties. Walter DeGruyter, Berlin/Boston pp. 87–110.

Dehaene S, Lau H, Kouider S (2017) What is consciousness, and could machines have it? Science 358:486–492. https://doi.org/10.1126/science.aan8871

Demuro E, Gurney L (2023) Can nonhumans speak? Languaging and worlds in posthumanist applied linguistics. Ling Front 6(2):94–105. https://doi.org/10.1007/s11229-024-04566-3

Demuro E, Gurney L (2024) Artificial intelligence and the ethnographic encounter: Transhuman language ontologies, or what it means "to write like a human, think like a machine." Lang Comm 96:1–12. https://doi.org/10.1146/annurev-neuro-120623-101142

Desmond H, Ramsey G (eds) (2023) Introduction: The manifold challenges to understanding human success. In Human success: evolutionary origins and ethical implications. University Press, Oxford, pp. 1-14.

Dulberg Z, Dubey R, Berwian IM, Cohen JD (2023) Having multiple selves helps learning agents explore and adapt in complex changing worlds. Proc Natl Acad Sci 120(28):e2221180120

Duncker D (2020) Chatting with chatbots: Sign making in text-based human-computer interaction. Sign Syst Stud 48(1):79–100

Dynel M (2023) Lessons in linguistics with ChatGPT: Metapragmatics, metacommunication, metadiscourse and metalanguage in human-AI interactions. Lang Commun 93:107–124. https://doi.org/10.1007/s001090000086

Fedorenko E, Ivanova A, Regev T (2024) The language network as a natural kind within the broader landscape of the human brain. Nat Rev Neurosci 5:289–312. https://doi.org/10.1038/s41583-024-00802-4

Ferstl Y, Bülthoff H, de la Rosa S (2017) Action recognition is sensitive to the identity of the actor. Cognition 166:201–206

Friedrich J, Golkar S, Farashahi S, Genkin A, Sengupta A, Chklovskii D (2021) Neural optimal feedback control with local learning rules. Adv Neural Inform Process Syst 34:16358–16370. https://doi.org/10.5555/3540261.3541512

Friston K (2009) The free-energy principle: A rough guide to the brain? Trends Cogn Sci 13:293–301. https://doi.org/10.1016/j.tics.2009.04.005

Friston K (2010) The Free-Energy Principle: A unified brain theory? Nat Rev Neurosci 11:127–138. https://doi.org/10.1038/nrn2787

Gerken M (2022) Scientific realism: Its roles in science and society. Oxford University Press, Oxford

Glenberg AM, Kaschak MP (2002) Grounding language in action. Psychon Bull Rev 9:558–565. https://doi.org/10.3758/BF03196313

Goldberg AE (1995) Constructions: A construction grammar approach to argument structure. University Of Chicago Press, Chicago

Goldberg AE (2019) Explain me this: Creativity, competition, and the partial productivity of constructions. Princeton University Press, Princeton, NJ

Gómez Redondo S, Rodríguez Higuera CJ, Coca CR, Olteanu A (2024) Transhumanism, society and education: An edusemiotic approach. Studies in Philosophy and Education. https://doi.org/10.1007/s11217-024-09927-6

Harding ME (2020) [1965]) The 'I' and the 'Not-I': A study in the development of consciousness (Bollingen Series LXXIX). Princeton, NJ, Princeton University Press

Hassel Mørch H (2023) Non-physicalist theories of consciousness. Cambridge University Press, Cambridge

Hayles NK (1999) How we became posthuman: Virtual bodies in cybernetics, literature, and informatics. University of Chicago Press, Chicago and London

Hayles NK (2012) How we think: Digital media and contemporary technogenesis. University of Chicago Press, Chicago and London

Hayles NK (2017) Unthought: The power of the cognitive nonconscious. University of Chicago Press, Chicago and London

Hayles NK (2018) The cosmic web: Scientific field models and literary strategies in the twentieth century. Cornell University Press, Ithaca and London

Hayles NK (2021) Postprint: Books and becoming computational. Columbia University Press, New York

Hayles NK (2024) Posthuman bodies: Why they (still) matter. In: Hamilton G, Lau C (eds) Mapping the posthuman. Routledge, New York and London, pp 29–48

He C, Chrastil ER, Hegarty M (2022) A new psychometric task measuring spatial perspective taking in ambulatory virtual reality. Front Virt Real 3:971502. https://doi.org/10.3389/frvir.2022.971502

Hedlund M, Persson E (2024) Expert responsibility in AI development. AI & Soc 39:453–464. https://doi.org/10.1007/s00146-022-01498-9

Jablonka KM, Schwaller P, Ortega-Guerrero A, Smit B (2023) Leveraging large language models for predictive chemistry. Nat Mach Intell 6:161–169. https://doi.org/10.1038/s42256-023-00788-1

Jackendoff R (2012) A user's guide to thought and meaning. Oxford University Press, Oxford

Juechems K, Summerfield C (2019) Where does value come from? Trends Cogn Sci 23(10):836–850. https://doi.org/10.1016/j.tics.2019.07.012

Kean H, Fung A, Pramod RT, Chomik-Morales J, Kanwisher N, Fedorenko E (2024) Intuitive physical reasoning is not mediated by linguistic nor exclusively domain-general abstract

representations. Available at: https://doi.org/10.1101/2024.11.25.625212\
Keramati M, Gutkin B (2014) Homeostatic reinforcement learning for integrating reward collection and physiological stability. Elife 3:e04811. https://doi.org/10.7554/eLife.04811
Kessler K, Thompson LA (2014) The embodied nature of spatial perspective taking: embodied transformation versus sensorimotor interference. Cognition 114:72–88. https://doi.org/10.1016/j.cognition.2009.08.015
Kinderman D, Onofri A (2021) The fragmented mind: an introduction. In: Borgoni C, Kindermann D, Onofri A (eds) The fragmented mind. Oxford University Pres, Oxford, pp 1–33
Kinouchi Y, Mackin KJ (2018) A basic architecture of an autonomous adaptive system with conscious-like function for a humanoid robot. Front Robot 5:30. https://doi.org/10.3389/frobt.2018.00030
Kitano H (2021) Nobel Turing Challenge: creating the engine for scientific discovery. NPJ Syst Biol Appl 7(29):1–12. https://doi.org/10.1038/s41540-021-00189-3
Kumar P (2024) Large language models (LLMs): survey, technical frameworks, and future challenges. Artif Intell Rev 57(260):1–51. https://doi.org/10.1007/s10462-024-10888-y
Levy D (2009) The ethical treatment of artificially conscious robots. Int J Soc Robot 1:209–216. https://doi.org/10.1007/s12369-009-0022-6
Luo X , Rechardt A, Sun G, Nejad KK, Yáñez F, Yilmaz B, Lee K, Cohen AO, Borghesani V, Pashkov A, Marinazzo D, Nicholas J, Salatiello A, Sucholutsky I, Minervini P, Razavi S, Rocca R, Yusifov E, Okalova T, Gu N, Ferianc M, Khona M, Patil KR, Lee P-S, Mata R, Myers NE, Bizley JK, Musslick S, Bilgin IP, Niso G, Ales JM, Gaebler M, Murty N, Loued-Khenissi L, Behler A, Hall CM, Dafflon J, Dongqi Bao S, Love BC (2024) Large language models surpass human experts in predicting neuroscience results. Nature Human Behaviour. https://doi.org/10.1038/s41562-024-02046-9
Lux V, Non AL, Pexman PM, Stadler W, Weber LAE, Krüger M (2021) A developmental framework for embodiment research: the next step toward integrating concepts and methods. Front Syst Neurosci 15:672740. https://doi.org/10.3389/fnsys.2021.672740
Machery E (2010) Reply to Barbara Malt and Jesse Prinz. Mind Lang 25(5):634–646
Mahon BZ (2008) Action recognition: Is it a motor process? Curr Biol 18(22):R1068–R1069. https://doi.org/10.1016/j.cub.2008.10.00
Mahon BZ (2015a) The burden of embodied cognition. Can J Exp Psychol 69(2):172–178. https://doi.org/10.1037/cep0000060
Mahon BZ (2015b) Response to Glenberg: Conceptual content does not constrain the representational format of concepts. Can J Exp Psychol 69(2):179–180. https://doi.org/10.1037/cep0000059
Mahon BZ (2015c) What is embodied about cognition? Lang Cogn Neurosci 30(4):420–429. https://doi.org/10.1080/23273798.2014.987791
Mahon BZ, Hickok G (2016) Arguments about the nature of concepts: Symbols, embodiment, and beyond. Psychon Bull Rev 23:941–958. https://doi.org/10.3758/s13423-016-1045-2
Massimi M (2022) Perspectival realism. Oxford University Press, Oxford
McDowell J (1992) Meaning and intentionality in Wittgenstein's later philosophy. In: French P, Uehling T, Wettstein H (eds) Midwest studies in philosophy, vol 17. University of Notre Dame Press, Notre Dame, IN, pp 40–52
Mills S, Skaug Sætra H (2024) The autonomous choice architect. AI & Soc 39:583–595. https://doi.org/10.1007/s00146-022-01486-z
Montoya I (2023) What is it like to be a brain organoid? Phenomenal consciousness in a biological neural network. Entropy 25:1328. https://doi.org/10.3390/e25091328
Mougenot, D, Matheson H (2024) Theoretical strategies for an embodied cognitive neuroscience: Mechanistic explanations of brain-body-environment systems. Cognitive Neuroscience. https://doi.org/10.1080/17588928.2024.2349546
Murphy E, Holmes E, Friston K (2024) Natural language syntax complies with the free energy principle. Synthese 203(154):1–35. https://doi.org/10.1007/s11229-024-04566-3
Nov O, Singh N, Mann D (2023) Putting ChatGPT's medical advice to the (Turing) test. *arXiv*. https://doi.org/10.48550/arXiv.2301.10035
Oizumi M, Albantakis L, Tononi G (2014) From the phenomenology to the mechanisms of consciousness: integrated Information Theory 30. PLoS Compuational Biology 10(5):1003588. https://doi.org/10.1371/journal.pcbi.1003588
Pan X, Liang B, Li X (2024) Flexible and fine-grained simulation of speed in language processing. Front Psychol 15:1333598. https://doi.org/10.3389/fpsyg.2024.1333598
Porter B, Machery E (2024) AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. Sci Rep 14:26133. https://doi.org/10.1038/s41598-024-76900-1
Reber AS, Baluška F, Miller WB Jr (2023) The sentient cell: The cellular foundations of consciousness. Oxford University Press, Oxford
Rickabaugh B, Moreland JP (2023) The substance of consciousness: A comprehensive defense of contemporary substance dualism. Wiley-Blackwell, Hoboken, N.J.
Rosendahl Thomsen M (2013) The new human in literature: Posthuman visions of changes in body, mind and society after 1900. Bloomsbury Academic, London/New York
Schaller F, Sabine Weiss S, Müller HM (2016) Pushing the button while pushing the argument: Motor priming of abstract action language. Cogn Sci 41(5):1328–1349
Schmidgall S, Harris C, Essien I, Olshvang D, Rahman T, Kim JW, Ziaei R, Eshraghian J, Abadir P, Chellappa R (2024) Evaluation and mitigation of cognitive biases in medical language models. NPJ Dig Med 7(295):1–9. https://doi.org/10.1038/s41746-024-01283-6
Searle JR (1980) Minds, brains and programs. Behavioral Brain Science 3:417–424. https://doi.org/10.1017/S0140525X00005756
Searle JR (1992) The rediscovery of the mind. MIT Press, Cambridge, MA
Shrader-Frechette K (2014) Tainted: How philosophy of science can expose bad science. Oxford University Press, Oxford
Sims M, Kiverstein J (2021) Externalized memory in slime mould and the extended (nonneuronal) mind. Cogn Syst Res 73:26–35. https://doi.org/10.1016/j.cogsys.2021.12.001
Solaki A (2022) The effort of reasoning: Modelling the inference steps of boundedly rational agents. J Logic Lang Inform 31:529–553. https://doi.org/10.1007/s10849-022-09367-w
Speed LJ, Vigliocco G (2014) Eye movements reveal the dynamic simulation of speed in language. Cogn Sci 38:367–382. https://doi.org/10.1111/cogs.12096
Tallman A, Elías-Ulloa J (2020) The acoustic correlates of stress and tone in Chácobo (Pano): a production study. J Acoust Soc Am 147(1):3028–3042. https://doi.org/10.1121/10.0001014
Todorov E, Jordan MI (2002) Optimal feedback control as a theory of motor coordination. Nat Neurosci 5(11):1226–1235. https://doi.org/10.1038/nn963
Toffoletti K (2007) Cyborgs and Barbie dolls: Feminism, popular culture and the posthuman body. I.B. Tauris, London/New York.
Tononi G (2004) An information integration theory of consciousness. BMC Neurosci 5(42):1–22
Torres-Martínez S (2018a) Constructions as triads of form, function and agency: an agentive cognitive construction grammar analysis of English modals. Cognitive Semantics 4(1):1–38. https://doi.org/10.1163/23526416-00401001

Torres-Martínez S (2018b) Exploring attachment patterns between multi-word verbs and argument structure constructions. Lingua 209:21–43. https://doi.org/10.1016/j.lingua.2018.04.001
Torres-Martínez S (2019) Taming English modals: how a construction grammar approach helps to understand modal verbs. English Today 35(2):50–57. https://doi.org/10.1017/S0266078418000081
Torres-Martínez S (2020) On English modals, embodiment and argument structure: Response to Fong. English Today 38(2):105–113. https://doi.org/10.1017/S0266078420000437
Torres-Martínez S (2021a) The cognition of caused-motion events in Spanish and German: an agentive cognitive construction grammar analysis. Austr J Linguist 41(1):33–65. https://doi.org/10.1080/07268602.2021.1888279
Torres-Martínez S (2021b) Complexes, rule-following, and language games: Wittgenstein's philosophical method and its relevance to semiotics. Semiotica 242:63–100. https://doi.org/10.1515/sem-2019-0113
Torres-Martínez S (2022a) Metaphors are embodied otherwise they would not be metaphors. Linguistics Vanguard 8(1):185–196. https://doi.org/10.1515/lingvan-2019-0083
Torres-Martínez S (2022b) The role of semiotics in the unification of Langue and Parole: An agentive cognitive construction grammar approach to English modals. Semiotica 244(1/4):195–225. https://doi.org/10.1515/sem-2018-0046
Torres-Martínez S (2023a) A radical embodied characterization of German Modals. Cognitive Semantics 9(1):132–168. https://doi.org/10.1163/23526416-bja10035
Torres-Martínez S (2023b) The semiotics of motion encoding in early English: A cognitive semiotic analysis of phrasal verbs in old and middle English. Semiotica 251:55–91. https://doi.org/10.1515/sem-2019-0104
Torres-Martínez S (2024a) Embodied human language models vs. large language models, or why Artificial Intelligence cannot explain the modal be able to. Biosemiotics 17:185–209. https://doi.org/10.1007/s12304-024-09553-2
Torres-Martínez S (2024b) Embodied essentialism in the reconstruction of the animal sign in robot animal design. Biosystems 238:105178. https://doi.org/10.1016/j.biosystems.2024.105178
Torres-Martínez S (2024c) Semiosic translation: A Bayesian-heuristic theory of translation and translating. Language and Semiotic Studies 10(2):167–202. https://doi.org/10.1515/lass-2023-0042
Torres-Martínez, S (2022c). On the cognitive dimension of metaphors and their role in education: A response to Molina Rodelo (2021). Revista Senderos Pedagógicos 13, 113–123. https://doi.org/10.53995/rsp.v13i13.1128
Torres-Martínez, S (2023c) Grammaire agentielle cognitive de constructions : Explorations sémioticolinguistiques des origines de la représentation incarnée. Signata, Annales de Sémiotique 14. https://doi.org/10.4000/signata.4551.
Torres-Martínez S (2024d) A predictive human model of language challenges traditional views in linguistics and pretrained transformer research. Language and Semiotic Studies10(4): 562-592. https://doi.org/10.1515/lass-2024-0018
Tuckute, G. [@GretaTuckute]. (2024). Interested in supporting research at the intersection of cognitive science, neuroscience, and AI? [Tweet]. X. https://twitter.com/GretaTuckute/status/1790045528912199847
Tuckute G, Kanwisher N, Fedorenko E (2024) Language in brains, minds and machines. Annual Review of Neuroscience 47. https://doi.org/10.1146/annurev-neuro-120623-101142
Turkle S (2005) The second self: Computers and the human spirit. The MIT Press, Cambridge, Massachusetts
Turkle S (2010) Alone together: Why we expect more from technology and less from each other. Basic Books, New York
Vallentyne P (2014) Intrinsic properties defined. In Francescotti, R. (Ed.), Companion to intrinsic properties. Walter DeGruyter, Berlin/Boston pp. 31–40.
Vannuscorps G, Rombaux E, Andres M, Pereira Carneiro S, Caramazza A (2021) Typically efficient lipreading without motor simulation. J Cogn Neurosci 33(4):611–621. https://doi.org/10.1162/jocn_a_01666
Vickers P (2023) Identifying future-proof science. Oxford University Press, Oxford
Weissweiler L, Köksal A, Schütze H (2023). Hybrid Human-LLM corpus construction and LLM evaluation for rare linguistic phenomena. arXiv:2403. 06965.
West O (2024) What makes a good theory, and how do we make a theory good? Computational Brain and Behavior. https://doi.org/10.1007/s42113-023-00193-2
Wihbey JP (2024) AI and epistemic risk for democracy: A coming crisis of public knowledge? Ethics Institute Working Paper. https://ssrn.com/abstract=4805026
Wilson M (2023) Imitation of rigor: An alternative history of analytic philosophy. Oxford University Press, Oxford
Wittgenstein L (1984[1953]) Werkausgabe Band 1. Tractatus Logico-Philosophicus/Tagebücher/Philosophische Untersuchungen. Suhrkamp Verlag, Frankfurt am Main.
Youssef A, Stein S, Clapp J, Magnus D (2023) The importance of understanding language in large language models. Am J Bioeth 23(10):6–7. https://doi.org/10.1080/15265161.2023.2256614