Discover Artificial Intelligence

RESEARCH Open Access

4E cognition and the coevolution of human–AI interaction

Jörg Noller[1*]

*Correspondence:
Jörg Noller
joerg.noller@lrz.uni-muenchen.de
[1]LMU Munich, Munich, Germany

Abstract
This paper examines the interaction between humans and large language models (LLMs) through the lens of 4E cognition, which encompasses embodied, embedded, enactive, and extended cognitive processes, to provide a comprehensive understanding of their relational dynamics. It argues that LLMs should not be conceived of in terms of objects but be understood as a processual and relational phenomenon that co-constitutes human agency in a shared sociotechnical environment. As such, the paper reframes AI as an interactive medium of enactive and extended human action. By focusing on the relational entanglement of social databases, dynamic patterns, and algorithmic structures, the paper proposes a 4E-compatible connectionist account of AI—one that understands AI not only as a technological artifact but as a co-evolving component of the extended cognitive ecology of human life, shaping and shaped by enactive practices, intentions, and norms. Finally, the paper discusses the limits of AI, discussing the problem of AI hallucination and collapse from a 4E cognition perspective.

Keywords LLM, 4E cognition, Human–AI interaction, Coevolution, Hallucination

1 Introduction

Large language models (LLMs) such as GPT or LLaMA represent a qualitatively new stage in the history of artificial intelligence (AI). Unlike earlier symbolic AI or recommender systems, they are built on transformer architectures [72] that enable real-time generativity, linguistic responsiveness, and context-sensitive adaptability. Their outputs are not mere retrievals from stored data but dynamically recombined sequences, shaped by attention mechanisms and prompt inputs. This distinctiveness becomes especially evident in phenomena such as "prompt injection" [44], jailbreaks, or hallucinations [39], where the generative structure of LLMs interacts with human sense-making in unpredictable ways. Far from being neutral tools, LLMs enter dialogical practices, scaffold reasoning, and reshape communication norms. These features demand a conceptual framework that can account for their hybrid, relational role in human cognition and agency. As Millière and Buckner [46] have recently emphasized, LLMs occupy a distinctive place in contemporary philosophy of AI, reviving classical debates on intelligence, compositionality, and semantic competence. Their diagnosis of the "Redescription

Fallacy" [[46], 9–10]—i.e., the tendency to dismiss neural models as "a collection of statistical calculations"—highlights the need for frameworks that take seriously their role in reshaping cognition and agency. It is precisely in this spirit that the present paper develops a 4E account of LLMs, treating them not as trivial statistical devices but as constitutive elements within extended human sense-making.

To date, debates about AI have often been trapped in a dichotomy: on the one hand, the anthropomorphizing view that treats AI as a quasi-subject such as "artificial agents" [19], on the other, the reductionist view that regards it as a mere instrument [12]. LLMs challenge both perspectives. They neither "think like us" in any literal sense nor can they be reduced to the logic of hammers or thermometers. Their operation on socially sedimented corpora reflects and extends human purposes [6, 49] while simultaneously introducing new teleological orientations, such as filtering, ranking, and generating content [34]. The main problem, therefore, is not whether LLMs are intelligent but how their integration into human practices transforms action, responsibility, and autonomy. Fazi [17, 18], for example, argues that LLMs manifest a synthetic world-constitution rather than simulating human thought. This complements my account of AI as teleologically constitutive of the lifeworld, though I stress their relational embedding in normative practices.

This paper develops a 4E account of artificial intelligence that situates LLMs as constitutive elements of cognition and agency. The guiding thesis is that cognition is embodied, embedded, enactive, and extended [24, 71], [47]. Against anthropomorphic projections and reductionist trivializations, the paper proposes a relational and processual framework in which humans and AI coevolve within a shared digital lifeworld. According to this view, LLMs are not external forces acting on otherwise self-contained minds, but integrated participants in distributed systems of sense-making, decision-making, and purposive action.

The line of argument unfolds in six steps. Section 2 lays the groundwork by advancing a 4E agency account of LLMs, showing how their computational, generative, and linguistic capacities extend the lifeworld's teleological structure. Section 3 develops the notion of AI-extended agency, reframing classical debates about simulation and intelligence by analyzing LLMs as "extension relations" that redistribute intentionality across human–machine assemblages. Section 4 broadens this to AI-extended action, highlighting language as the medium through which LLMs transform collective intentionality and normativity. Section 5 introduces the paradigm of human–AI coevolution [56], situating LLMs in recursive feedback loops that reshape human preferences and institutions. Section 6 turns to the epistemic limits of LLMs, analyzing hallucination and model collapse as consequences of their associative architecture and disconnection from embodied lifeworlds. Finally, Section 7 concludes by drawing the ethical consequences of this framework, arguing that LLMs must be governed as co-constitutive elements of extended human agency rather than as autonomous moral subjects.

In sum, the distinctive features of LLMs—generativity, linguistic responsiveness, and susceptibility to prompt steering—provide the motivation for adopting a 4E framework. By analyzing them as part of an evolving ecology of human–AI interaction, this paper seeks to illuminate both the promises and risks of our entanglement with generative language technologies.

2 LLMs and 4E cognition

The rise of LLMs has placed human cognition and agency into a novel technological ecology. From a 4E perspective—where cognition is conceived as embodied, embedded, enacted, and extended—this does not come as a surprise. Decades of research have already shown that technologies, artifacts, and social partners can become constitutive elements of human cognitive ecologies [8, 24, 37].

Such a reframing allows us to move beyond the standard dichotomies that dominate current debates: the view of AI as an "artificial agent" [19] and the reductionist view of AI as a mere tool [12]. Consider, for example, how recommender systems or LLMs enter everyday practices. They do not merely simulate intelligence [70], nor are they neutral calculators. Their operation on socially grounded databases reflects and extends human purposes [49], while at the same time introducing new teleological orientations—filtering, ranking, or generating content in ways that guide action [26, 54]. This hybrid entanglement requires a framework that acknowledges how AI extends the horizon of human cognition and agency while remaining dependent on humanly set purposes and the normative structures of the lifeworld [38], [24, 64]. Seen in this light, the central problem is not whether AI is intelligent, but how its integration into human practices transforms the structures of action, responsibility, and autonomy [74], [13].

The task, then, is to develop an account that avoids both anthropomorphization and trivialization, and instead situates AI as part of an evolving ecology of extended human agency [50, 56]. This reframing provides the conceptual basis for the 4E perspective introduced below, in which cognition and agency emerge not from isolated minds but from embodied, embedded, enactive, and extended relations [71], [47].

According to the common 4E approach, cognition is embodied, i.e. shaped by the body's perceptual and motor capacities; embedded, i.e. situated within social and material environments; enactive, i.e. arising through active engagement and sense-making; and extended, i.e. often incorporating external tools and technologies into cognitive processes [7, 28, 45]. Together, these four dimensions highlight that thinking is not confined to the brain but unfolds across bodies, environments, and artifacts [25].

Drawing on Froese and Ziemke [23], recent work has extended this insight specifically to digital technologies. AI technologies co-shape human agency within these extended and enactive systems, drawing upon insights from phenomenology, enactivism, and connectionism [49, 50]. In developing a 4E account of AI, it is essential to acknowledge both the promises and the risks highlighted in the recent literature. Aagaard [1] cautions against what he terms the dogma of harmony in 4E cognition, the tendency to portray human–technology relations as inherently collaborative and mutually beneficial. According to Aagaard, such depictions risk ignoring the disruptive or conflictual aspects of our engagements with technology, including phenomena such as "digital akrasia"—i.e., the tendency to act against one's better judgment when drawn in by digital devices—or the deskilling that results from overreliance on automated systems. From this perspective, AI cannot be understood merely as a benign cognitive extension but must also be analyzed as a potential source of friction, dependency, or erosion of human capacities. Clowes [9], by contrast, stresses the ecological dimension of digital technologies, with particular focus on the internet as a pervasive "cognitive ecology". He rejects the "impact thesis," which frames digital artifacts as external forces acting upon otherwise self-contained minds. Instead, drawing on Material Engagement Theory, Clowes argues that

digital infrastructures are constitutive of new forms of human agency. Far from being passive recipients of technological influence, humans actively co-construct cognitive practices with digital artifacts, from cloud-based memory systems to algorithmic recommendation platforms. Importantly, this analysis highlights both the empowering and the constraining roles of digital technologies: they can scaffold planfulness and self-regulation, but they may also inhibit autonomy when their operations remain opaque or externally controlled. Taken together, these contributions offer a more nuanced backdrop for theorizing AI within a 4E framework. Aagaard reminds us that not all couplings between humans and technologies enhance agency—some may disrupt it—while Clowes demonstrates that such couplings are not merely external influences but constitutive elements of cognitive ecologies. A 4E account of AI must therefore attend to both the generative and the problematic dynamics of human–AI interaction, situating artificial intelligence within a lifeworld that is neither wholly harmonious nor purely disruptive, but marked by complex, evolving patterns of co-constitution.

What remains underexplored, however, is how AI reshapes the teleological and normative structure of the human lifeworld. If AIs are not merely tools but integrated elements in the field of embodied, socially embedded action, then **t**heir role is not exhausted by scaffolding cognition. They also transform how purposes, goals, and norms are distributed across human–machine systems. This paper therefore develops a connectionist and phenomenological account of AI-extended agency that centers on coevolution: the mutual shaping of human practices, linguistic patterns, and algorithmic structures within a shared lifeworld [36, 14, 49, 50].

Rather than rehearsing the basic 4E claim that technologies can be parts of cognition, I shall focus on AI's distinctive role: its computational power, pattern-recognition capacities, and linguistic responsiveness allow it to become entangled with human sense-making in ways that differ from earlier technologies. Building on Don Ihde's phenomenology of technics [38] and enactive AI research [23, 63], I will argue that AI functions as an extension relation: a structure that co-constitutes human cognition and action rather than standing merely as an instrument or alterity.

A crucial question is how the proposed 4E agency account differs from existing sociotechnical frameworks on AI. Much recent scholarship has emphasized the sociotechnical embedding of AI, for instance in terms of socio-technical systems [[11], 52], responsible innovation [74], or digital condition [64]. These frameworks highlight the entanglement of technology, institutions, and social norms, and they often frame AI governance as a matter of aligning technical systems with societal values. Similarly, human–AI interaction research frequently employs concepts of affordances [75] or human–computer symbiosis [43] to account for the co-shaping of practices and technologies.

While these approaches provide valuable insights, they tend to remain at the level of external relations between "social" and "technical" domains. By contrast, a 4E agency framework conceptualizes AI as part of the *constitution* of cognition and action. Rather than treating AI as an external factor that merely influences or mediates human behavior, the 4E approach argues that AI can become integrated into the very processes of sense-making, decision-making, and purposive action. This difference matters in that it shifts the focus from governance of external impacts to the analysis of extended human

agency [49], where human purposes, embodied practices, and algorithmic operations are entangled in distributed teleological structures.

In this sense, the novelty of the 4E agency account lies in its ontological commitment: AI is not only situated in socio-technical systems but also participates in the enactment of cognition and action itself. This allows us to capture both the empowering and the disruptive dynamics of AI more precisely, and to frame normative questions not only in terms of external regulation but also in terms of how human autonomy and responsibility are transformed within an AI-extended lifeworld.

This reframing enables us to move beyond the familiar dichotomies—AI as a quasi-subject versus AI as mere tool—and instead situate AI within the evolving lifeworld of human action. Language is crucial here: large language models (LLMs) operate on socially sedimented corpora [6], but once integrated into human practice they enter sensorimotor loops and dialogical interactions that reshape the very norms of communication and agency. As Buckner [6] emphasizes, the performance of contemporary language models depends on their immersion in socially structured corpora and human feedback, more precisely Reinforcement Learning with Human Feedback (RLHF). Rather than being abstract statistical devices, LLMs embody a form of empiricism in which human experiences, purposes, and biases are sedimented into digital corpora. In this sense, their generative capacities are socially grounded from the outset, reflecting the same kind of culturally mediated scaffolding that Buckner associates with developmental trajectories of human social cognition, therefore speaking of "cultural training data" [[6], 341]. Thus, AI should be examined less as a detached computational entity and more as a co-evolving component of extended human agency within the digital lifeworld.

This paper contributes to current debates on AI not only at a theoretical level but also with practical implications. By reframing AI as part of extended human agency within the digital lifeworld, the argument given offers several benefits across multiple domains. For data curation and model training, it highlights the importance of situating datasets within social practices and normative contexts [10], [4], thereby drawing attention to how human purposes and biases are sedimented in training data. For system and interface design, the analysis encourages designers to consider AI not as a neutral instrument but as a participant in human action [2, 51, 65], shaping and being shaped by users' goals and environments. For AI governance, the account provides conceptual resources for design guidelines that go beyond anthropomorphic or instrumental framings [11, 21], [74], instead emphasizing instead the need to regulate the relations between humans, data, and algorithms. For evaluation and benchmarking, it suggests that assessments of AI systems should not only measure performance metrics in isolation, but also account for their teleological integration into human lifeworlds [52, 55]. Finally, for educational programs about AI, the framework supports a more holistic curriculum that integrates technical, ethical, and phenomenological perspectives [40, 60, 77], equipping learners to critically understand and responsibly use AI technologies.

3 Towards a 4E agency account of AI

While the notion that technologies can serve as "prostheses" of the human mind has a long intellectual lineage, stretching from Freud's [22] reflections on artificial supports to contemporary accounts of distributed cognition and sociotechnical entanglement, the framework developed here advances two distinct claims. First, large language models

should be understood not as embodied but as linguistic enactors. Their mode of participation in human sense-making is not grounded in sensorimotor coupling but in dialogical and symbolic responsiveness. This distinguishes them both from classical cognitive artifacts, such as writing or tools, and from embodied human partners. Second, the proposed concept of extension relations is not reducible to external influence or mediation. It introduces an ontological shift: LLMs do not simply affect cognition from the outside but enter into the teleological constitution of the lifeworld, reconfiguring how purposes, goals, and norms are distributed across human–machine interaction. These two moves—the relocation of enactment from the embodied to the linguistic domain, and the shift from external impact to teleological co-constitution—mark the novelty of the present account. They allow us to capture the specificity of LLMs within the broader history of cognitive technologies, and to clarify what is at stake in human–AI coevolution.

Rather than contrastingthe digital domain with the analog, I argue—consistent with 4E cognition and agency—that digital structures can become embedded within the embodied, enactive processes that constitute human sense-making. Husserl (1970) describes the lifeworld as teleologically structured, meaning that purposes, goals, and intentions are not external to experience but are the very fabric through which meaning and action unfold. This is echoed in contemporary 4E cognition accounts, which conceive cognition as affordance-responsive and goal-directed interaction between agent and environment. The lifeworld, then, is not simply "the world as experienced," but the pragmatic field of embodied coping and skillful navigation [24, 71]. Therefore, the digital lifeworld, understood in terms of pattern-based, affordance-laden structures, can extend and reshape the analog lifeworld. As LLMs extract and recombine patterns from human-produced data, they operate on socially embedded databases—particularly language—and thus participate in enacted, culturally mediated cognition.

While the 4E framework has established itself as a powerful lens for analyzing human–technology relations, it has also been criticized for being an overly optimistic portrayal of extended cognition. Slaby [62] highlights the limits of the prevailing *user/resource model*, in which individuals are assumed to deliberately and harmoniously integrate external artifacts into their cognitive routines. Against this view, he develops the notion of *mind invasion*: processes in which social and technological infrastructures modulate affectivity and agency from the outside, often contrary to the individual's prior orientations. Rather than simple enhancement, technologies can entrench patterns of habituation and emotional alignment that serve institutional or corporate interests. This perspective underscores that extended cognition is never neutral—it can involve subtle forms of control and subjectification.

Aagaard [1] makes a parallel point with his critique of the *dogma of harmony* in 4E cognition. He warns that scholars too often assume that human–technology couplings are cooperative and beneficial, overlooking conflictual relations such as bad habits, digital distraction, and deskilling. The phenomenon of *digital akrasia*—acting against one's better judgment under the lure of digital devices—illustrates how technological scaffolds can undermine agency as much as they support it. Aagaard therefore calls for a more balanced analysis that admits both antagonistic and generative dimensions of cognitive extension.

Building on these concerns, Haraldsen [27] argues that contemporary digital technologies introduce a qualitatively new factor into extended cognition, which he calls

"economic agency". While historical forms of cognitive engineering (such as the art of memory or literary writing) aligned external supports with user goals, many digital systems are designed to advance corporate interests that often diverge from, or even oppose, those of their users. This means that when we integrate digital technologies into our cognitive routines, we simultaneously allow economic logics to shape attention, memory, and sense-making. Haraldsen's analysis thus extends the critique of optimism in 4E cognition to the political economy of the digital ecology, where external agencies actively reconfigure cognitive processes in line with profit motives.

Taken together, these perspectives complicate the assumption that AI and digital technologies merely enrich human cognitive ecologies. They remind us that the integration of AI into the lifeworld must be analyzed in terms of power, conflict, and embodiment as well as in terms of functional enhancement. Moreover, they point to a key tension for enactivist accounts: while embodied interaction is central to sense-making, much of contemporary AI use is characterized by *low levels of embodiment*, as in text-based interfaces and screen-mediated engagements. The challenge, then, is to reconcile the enactivist emphasis on sensorimotor grounding with the increasingly disembodied modes of interaction that AI affords.

This more critical framing situates the present account within ongoing debates in 4E cognition. It underscores that the significance of AI does not lie only in its role as a novel extension of mind, but also in the ways it can conflict with human purposes, shape affective life, and embed economic and institutional agencies within cognitive systems.

From the perspective of 4E cognition, artificial intelligence must be situated within the broader history of cognitive technologies such as writing, reading, artifacts, and socially distributed practices. Menary's [45] theory of *cognitive integration* emphasizes how writing is not merely an external tool but part of our cognitive system, enabling new forms of reasoning and reflection through the manipulation of external symbolic vehicles. Similarly, Kukkonen [41] demonstrates that reading practices are deeply enactive and scaffolded, involving the embodied engagement of readers with textual structures that extend cognitive capacities in ways that reshape both attention and imagination. Heersmink [28] develops a taxonomy of artifacts—embodied, perceptual, cognitive, and affective—that clarifies how diverse material objects support, transform, and sometimes even constitute human capabilities. Hutchins [37], in his seminal study of ship navigation, shows how cognition in practice is distributed across individuals, tools, and cultural routines, underscoring the irreducible social dimension of cognitive systems. Tribble [69] provides a complementary historical analysis of Shakespeare's theatre as a cognitive ecology, highlighting how memory and attention are scaffolded by material and social arrangements. Sutton [66] similarly stresses that remembering is often scaffolded or co-constituted in socially distributed networks.

Within this broad ecology of cognitive technologies, AI introduces both continuities and novelties. On the one hand, it resembles earlier artifacts insofar as it provides external scaffolds for memory, reasoning, and imagination. Like writing, it stabilizes information across time; like reading practices, it shapes attention and interpretive patterns; and like socially distributed cognition, it embeds individual users in collective systems of knowledge production. On the other hand, AI differs in at least two crucial respects. First, it is highly *responsive*—unlike static artifacts such as notebooks or memory palaces, AI systems adapt in real time to user input, producing outputs that are dialogical

rather than inert. Second, while Williams [76] shows that cognition is always pragmatically oriented toward action rather than passive representation, Haraldsen [27] emphasizes that AI systems often embody *agential forces* that diverge from user intentions. Whereas writing or mnemonic techniques generally aligned with their practitioners' goals, many digital systems are designed to pursue corporate or economic interests, introducing what Haraldsen calls *"economic agency"* into the very fabric of cognition.

Thus, the difference between AI and earlier cognitive technologies is not that the former extends cognition while the latter did not, but that AI participates in extended cognition in ways that are *strategically goal-directed* on behalf of external actors. This raises novel normative concerns: whereas writing, reading, and collaborative memory practices primarily scaffolded human purposes, AI systems frequently embed priorities—optimization for engagement, prediction, or monetization—that may conflict with user agency. Recognizing this both situates AI within the long trajectory of cognitive technologies and clarifies the stakes of its distinctiveness: AI is not just another extension of mind, but a responsive, semi-agentive technology that refracts human agency through powerful institutional and economic logics.

The empirical foundations of artificial intelligence—particularly in the case of machine learning—rest heavily on the availability, quality, and structure of training data. As Buckner [5] argues, contemporary AI systems operate in terms of empiricism by abstracting complex patterns from data bases, in which performance improvements are largely driven not by conceptual breakthroughs in representation or reasoning, but by the scale and character of empirical data.

While Buckner's [5] analysis of abstraction in neural networks remains valuable, it predates the rapid rise of large language models. Recent discussions by industry players—such as Google's work on transformer architectures [72], Anthropic's reflections on *constitutional AI* [3], or Meta's transparency reports on LLaMA models [68]—demonstrate how the field has shifted toward scale, safety, and alignment. Yet, these accounts remain largely silent on the 4E dimensions of cognition. Training practices emphasize technical optimization while overlooking how models become part of social lifeworlds and human agency. A 4E perspective can therefore complement existing technical discourses by showing that LLMs are not just statistical systems but are already woven into embodied, social, and purposive practices that require critical reflection and governance.

However, the current paradigm often treats data as abstract inputs divorced from the embodied, situated contexts in which human cognition unfolds. From the perspective of 4E cognition, this detachment constitutes a profound limitation. Data in human cognition is not merely informational but is always affectively and pragmatically charged, arising from lived interaction with the world. In contrast, AI systems trained on static datasets lack access to the sensorimotor contingencies and interactive feedback loops that characterize human learning. This epistemic asymmetry points to the need for an alternative approach in AI development—one that acknowledges the empirical importance of data while embedding it within dynamic, embodied engagements, thereby aligning with the principles of 4E cognition.

However, a more precise account of large language models (LLMs) is needed in order to avoid conflating them with earlier forms of AI such as recommender systems. While both can be situated within the digital lifeworld as elements of extended cognition, they operate on different architectures and exhibit distinct interactional dynamics.

Recommender algorithms, for example, primarily function by correlating past user behavior with predefined patterns of similarity across datasets. By contrast, transformer-based LLMs employ the *attention mechanism* to dynamically weight and recombine learned representations during inference. Although trained on static corpora, they do not merely retrieve stored content but generate novel sequences in response to user prompts through real-time contextual recombination.

This distinction matters for 4E analyses. The generative process of LLMs allows for *context-sensitive adaptability* during interaction, including the well-documented phenomena of jailbreaking and prompt injection, where outputs can be steered or manipulated by strategically crafted inputs [44]. While such adaptability does not amount to enactive sensorimotor coupling in the biological sense, it nevertheless exemplifies a form of responsive engagement grounded in linguistic rather than embodied interaction.

Acknowledging this nuance enriches the 4E framework: it shows that LLMs partially instantiate the "enactive" and "interactive" dimensions of cognition, while simultaneously highlighting where they fall short. Their responsiveness is confined to symbolic patterns and lacks the sensorimotor grounding of embodied agents, yet it is precisely this hybrid form of linguistic responsiveness that makes them distinctive within the ecology of cognitive technologies. Rather than treating LLMs as static repositories, we should recognize them as dynamic interlocutors whose generativity reshapes human cognitive practices, even if their "engagement" is circumscribed by the statistical and non-biological nature of their architecture.

As Buckner [6] argues, such databases embody a kind of empiricism in which human experiences, purposes, and biases become embedded and sedimented within digital corpora. These patterns, once processed by artificial neural networks (ANNs), do not remain isolated from human practice. Instead, they enter into sensorimotor loops, practical reasoning, and social interactions. The digital lifeworld is, therefore a relational field of extended cognition and action—a domain in which purposes, symbols, and bodies converge and co-evolve.

Luciano Floridi's concept of the "onlife" [20], understood as a hybrid between online and offline existence, describes the phenomenon of the digital lifeworld. According to this hybrid conception, we no longer operate in distinct "online" and "offline" worlds, but in a merged ecology of analog and digital structures that shape our cognition and action. Within this hybrid space, tools, media, and algorithms actively shape the affordances and trajectories of action. This supports the 4E thesis that cognition is not internal, computational symbol manipulation, but distributed activity across brain, body, tools, and environment [25, 47].

The concept of affordance, originally developed by James J. Gibson and further elaborated in ecological psychology and enactivist philosophy, plays a central role in linking 4E cognition to AI. Affordances refer to "the possibilities for action that are available to agents in their environments" [[32], 3]. As such, they do not only address our cognition but also our action. In the framework of 4E cognition, affordances are not merely objective features of the environment, nor solely internal representations. Rather, they emerge from the dynamic interaction between an agent and its situated world. This relational ontology challenges the traditional subject–object dichotomy by showing that perception and action are co-constituted in practice.

In the digital lifeworld, individuals are not merely consumers of information but also its producers. Alvin Toffler's term prosumer—originally describing people who consumed what they themselves produced [[67], 282]—captures this dual role. Yet the digital prosumer differs from Toffler's original figure: she consumes and produces simultaneously, through her very participation in the digital lifeworld. In 4E terms, this reflects a coupling between humans and digital systems that is not merely unidirectionally causal but potentially constitutive and co-evolutionistic: cognition and action emerge from recurrent sensorimotor interaction with the world, and that world now includes AI systems trained on human data.

Drawing on Don Ihde's phenomenology of technics allows us to further illuminate this structure. His taxonomy of human–technology relations—embodiment, hermeneutic, alterity, and background—can be reinterpreted within the 4E paradigm. While classical AI systems may once have constituted alterity relations, modern ANNs participate in what I propose to call extension relations: they are not simply used or interpreted but are co-constitutive of human cognition and action.

AI systems today not only change what we do, but how we act, what we perceive, and what possibilities we consider actionable. Their influence reaches into domains such as perception (through algorithmic filtering), social interaction (through platform design), and even identity (via recommendation systems). These are not peripheral effects—they indicate that AI systems are now constitutive elements of the digital lifeworld, which in turn extends to the physical lifeworld.

To analyze these dynamics, we must shift our focus from ontological distinctions (machine vs. human) to teleological entanglements: how purposes, intentions, and affordances are redistributed across human-AI systems. This allows us to ask ethical questions not about machine "autonomy" but about how human autonomy is extended, reconfigured, or constrained by its integration with digital technologies.

4 AI extended agency

The distinction between "simulation" and "duplication" of human intelligence—central to classical debates on strong vs. weak AI—rests on the assumption that intelligence is an internal property of a system, whether human or artificial. Yet from a 4E perspective, this distinction can be modified: intelligence is not a static property, but a dynamic process enacted in specific bodily, social, and environmental contexts. Rather than asking whether machines *possess* intelligence, we must ask how cognitive performances emerge from interactions between humans and AI systems, and what roles these systems play in our extended cognitive and agential ecologies.

Alan Turing's famous "imitation game" [70], often interpreted as a simulation benchmark, already hints at this shift. The Turing test does not assess internal mental states but instead evaluates socially situated linguistic performance [57]. It is thus inherently enactive and relational, assessing whether a machine can participate in the dialogical fabric of social cognition. From this angle, the test anticipates the 4E thesis: cognition is not hidden "inside," but enacted in intersubjective exchange.

However, philosophers like John Searle [58] opposed this functional view by emphasizing that machines lack semantic understanding—they only manipulate symbols syntactically. This critique rests on a representationalist framework that the 4E approach explicitly challenges. In the 4E view, meaning arises not from internal representations

alone, but through sensorimotor coupling with the environment, affordance-responsiveness, and norm-governed interaction with others [24, 48].

Therefore, rather than defending or rejecting strong AI, this paper reframes the debate by focusing on AI as an extension of human cognition and agency. A part of human AI-interaction, AI systems do not replicate human minds, but participate in enactive practices that extend and transform human intentionality. This reflects Douglas Engelbart's idea of an "augmented human intellect" [15], although his instrumental focus on "effectiveness" underemphasizes the qualitative shifts in cognition and normativity that occur through such augmentation.

To better capture these qualitative transformations, I propose to understand AI-human interaction not in terms of "simulation" or "duplication," but as an intertwining of affordances, patterns, and actions, forming what Ihde might call an extension relation. This goes beyond the instrumental use of technology (as in mere "tools") and beyond anthropomorphizing (as in "artificial minds"). Instead, AI becomes a co-regulative structure in extended agency—shaping how humans perceive, decide, act, and relate.

The concept of interobjectivity, developed by Latour [42], offers a sociomaterial lens to describe this entanglement. Originally meant to account for how artifacts participate in social life, interobjectivity can be reinterpreted here to describe the material-semiotic coevolution between human purposes, digital artifacts, and algorithmic processes. Importantly, such coupling does not render machines "social agents," but instead highlights how cognitive agency emerges across distributed systems that include humans, tools, languages, and infrastructures.

This is supported by Hui's [35] account of interobjectivity as the structured relationality between digital objects. However, where Hui focuses on the materiality of connections, I emphasize the teleological and agential dimension: how human intentions are extended, delegated, and transformed via algorithmic processes that operate on socially meaningful data. Thus, the key question is not whether AI is intelligent per se, but how it becomes a constitutive part of human practical reasoning, moral responsibility, and cultural meaning-making in coevolution.

From a coevolutional point of view, Searle's distinction between syntax and semantics becomes less decisive. Within a 4E framework, semantic meaning does not reside solely in mental representations but in the dynamically enacted patterns of interaction across humans and their technological environment. By analyzing AI as part of these extended and enacted cognitive systems, we can shift the focus from metaphysical debates to practical concerns about agency, responsibility, and normativity in the digital lifeworld.

Finally, these insights allow us to recast artificial neural networks not as isolated computational systems, but as relational technologies embedded in human teleological fields. The performance of ANNs—pattern recognition, language generation, recommendation—only becomes meaningful when situated within shared intentional contexts, structured by social norms and linguistic practices. As such, AI is not an alien other, but a functionally integrated participant in the ecology of extended human action and coevolution.

5 AI-extended action

Luciano Floridi's concept of "artificial agents" within the infosphere reflects a growing recognition that AI systems are becoming embedded in the normative, social, and moral fabric of everyday life [19]. He distinguishes between *responsibility*, attributed to human designers and users, and *accountability*, which can be ascribed to artificial agents within constrained operational frameworks. While this distinction is insightful, it remains grounded in an informationalist paradigm, which treats AI primarily as a system of symbolic manipulation.

However, from a 4E perspective, such a framing is too narrow. As Gallagher [25] and Varela et al. [71] have argued, cognition—and by extension agency—is best not understood as internal information processing, but as enacted engagement with a meaningful environment, shaped by bodily capacities, social affordances, and temporal dynamics. Consequently, AI systems should not be framed as informational entities operating on abstract levels of description, but as components of extended, embodied, and teleologically structured human action.

The framework of extended mind theory [8] has already shown how tools, symbols, and environments can become constitutive parts of cognitive processes. While the original focus was on belief and memory, this insight applies equally to practical reasoning and intentional action. Artificial neural networks, trained on vast social-linguistic corpora, participate not merely in cognition but in the realization of goals and purposes—a domain traditionally reserved for human agency.

However, 4E theorists have gone beyond Clark and Chalmers by emphasizing that cognition is not just extended, but also embodied, enactive, and embedded. This means that AI cannot be understood simply as a cognitive prosthesis. Rather, its role in the human lifeworld must be seen as co-constitutive of agency: AI systems reshape what actions are possible, desirable, and normatively appropriate. They mediate human sense-making in dynamic and often implicit ways.

Language plays a central role from the perspective of 4E cognition and action. As emphasized in enactivist theories of cognition, language is not merely a code, but a mode of embodied, intersubjective coordination. AI systems like large language models operate on this dimension by transforming linguistic structures and thus enabling or constraining collective intentionality. This does not mean that machines have intentions, but that they participate in the social ecology of meaning, in which human intentions are formed, stabilized, and enacted.

Thus, the notion of AI-extended agency does not imply attributing autonomy to machines, but rather refers to the distributed and hybrid character of intentional action in the digital lifeworld. Drawing on the enactivist concept of human-AI coevolution, we can understand the human-AI relation as one of mutual modulation: human purposes are shaped by technological affordances, while AI outputs are in turn shaped by the social, cultural, and ethical frameworks in which they operate.

Floridi's level-of-abstraction (LoA) model rightly emphasizes that ethical evaluations of AI depend on how we describe and conceptualize them. However, instead of choosing an informational LoA, I argue for a teleological LoA grounded in practical contexts, embodied interactions, and value-laden affordances. This entails not merely a descriptive shift but a normative demand: if AI systems co-determine our field of possible actions, then responsibility must concern the human–AI ensemble, not just isolated agents.

Moreover, as recent theories of cyberbilities [16] suggest, AI does not merely extend cognition, but also modifies emotional, volitional, and social capacities. These cyberbilities are not confined to individuals but extend to institutions and collective practices. In this light, AI becomes part of the distributed infrastructure of practical rationality, shaping how intentions are formed, coordinated, and acted upon.

A normative understanding of AI, then, cannot rest on rules applied to autonomous machines. It must concern the quality of coupling between human purposes and AI performances— i.e., the transparency of affordances, the inclusivity and objectivity of training data, and the intelligibility of algorithmic mediations. Responsibility resides not in the machine, but in the patterns of interaction that sustain, extend or distort human agency.

In summary, AI-extended action reflects a transformation in the structure of human intentionality, one that calls for a 4E-compatible ethical framework. Rather than debating AI's moral status in isolation, we should focus on how human autonomy is transformed through technologically mediated relations. This requires attending not only to representational content, but to the affordances, feedback loops, and power asymmetries that structure human agency.

6 Human–AI coevolution: feedback loops in the digital lifeworld

The emerging paradigm of Human–AI coevolution [[56], 1], which means "a process in which humans and AI algorithms continuously influence each other", offers a compelling framework for understanding the mutual shaping of human practices and artificial intelligence systems. This coevolutionary perspective challenges linear models of technological progress by emphasizing the bidirectional dynamics through which AI systems not only influence but are also shaped by human behavior, institutional norms, and cultural practices. From the standpoint of 4E cognition, this perspective is particularly salient. Rather than viewing AI as an external tool or isolated artifact, coevolution foregrounds its integration into the embodied routines and socio-material environments of human agents. Human–AI interaction thus becomes a site of continuous cognitive and teleological reconfiguration, where agency is not located solely within the human or the machine but emerges relationally through distributed, interactive processes. In this context, 4E agency can be seen as the dynamically extended capacity to act and make sense within a hybrid digital lifeworld of human and artificial agents, technologies, and affordance-rich environments. Recognizing this co-constitutive entanglement is essential for developing normative and design frameworks that support not only effective but also meaningful and ethically sound human–AI relations. The concept of human–AI coevolution, as developed by Pedreschi et al. [56], provides a complementary and empirically grounded perspective on the dynamic interplay between human agency and artificial intelligence. While the 4E approach emphasizes the embodied, embedded, enactive, and extended character of cognition, coevolution theory focuses on the iterative feedback loops through which humans and AI systems mutually transform one another over time.

Pedreschi et al. [56] argue that AI-based systems—particularly recommender systems and digital assistants—are not merely tools but structural components of sociotechnical ecosystems. In such ecosystems, user behavior generates the data AI models are trained on, which in turn shape user preferences and behaviors in subsequent iterations. This recursive interaction creates a nonlinear coevolutionary dynamic that gives rise

to emergent and often unintended systemic effects such as filter bubbles [53], opinion polarization, model degradation, or behavioral conformism.

Viewed through a 4E lens, these dynamics can be interpreted as enactive couplings between human agents and algorithmic structures: cognitive agency does not merely adapt to technological environments—it is co-constituted within them. This resonates with the phenomenological notion of the lifeworld as a teleologically structured horizon of meaning: in the digital lifeworld, intentionality is increasingly mediated and modulated by algorithmic affordances, which themselves are the sedimented products of past human actions.

The coevolutionary perspective foregrounds the distributed nature of responsibility and normativity in human–AI interactions. Since recommender systems adapt based on aggregated user behavior, individual intentions are recursively enfolded into collective algorithmic patterns that then act back upon individuals. In this sense, AI systems become not only epistemic tools but teleological structures—not in the sense of possessing intentions, but in shaping the field of possible intentions available to human agents.

To be sure, the attribution of *teleology* to AI systems requires careful qualification. Large language models, as probabilistic next-token predictors, do not engage in causal reasoning or autonomous goal-setting in the way biological agents do. Their generative process is fundamentally associative, producing outputs by recombining statistical regularities rather than by pursuing ends of their own. In this respect, it may be misleading to speak of AI itself as "teleological." Instead, the teleological dimension emerges relationally, through the way LLMs shape the field of possible intentions available to human agents—by filtering, ranking, or generating patterns that guide subsequent human action.

This point becomes clearer when set against ecological-enactive theories of agency. Segundo-Ortin [59] emphasizes that agency is best understood not as an internal causal sequence of mental states but as a relation between organism and environment, structured by affordances, habits, and sensorimotor schemes. From this perspective, teleology is not an intrinsic property of isolated systems but arises from embodied engagement with meaningful affordances. Extending this insight to AI, we can see that LLMs function less as teleological *agents* in themselves than as artifacts that modulate human affordances and intentions. In this sense, they are comparable to other cultural artifacts—such as architecture or public institutions—that structure the horizon of possible actions without possessing intrinsic purposiveness.

Accordingly, the 4E agency framework does not anthropomorphize AI as an intentional subject. Rather, it analyzes AI as a relational extension of human agency, where the apparent teleology of the system is a reflection of its embedding in human practices, goals, and sociocultural norms.

7 Avoiding AI collapse and hallucination

The phenomenon of hallucinations provides a decisive test case for a 4E agency account of LLMs. Hallucinations show that these systems are not external tools that simply fail, but elements within extended and enactive processes of sense-making, where their breakdowns must be interpreted within the teleological and normative structures of the lifeworld. Recent studies on the epistemology of LLMs converge on the diagnosis of a structural deficit, namely the systematic exclusion of embodied, situated human

engagement in the development and deployment of AI systems. This leads to a form of Human–AI irrationality that manifests in diverse phenomena.

Shumailov et al. [61] describe the phenomenon of "model collapse" as a progressive degeneration that occurs when models are recursively trained on their own outputs rather than on human-authored data. Successive training cycles cause LLMs to "forget" the statistical long-tail of rare but important patterns, resulting in homogenized, impoverished, and epistemically diminished outputs. As they note, the continued availability of genuine human data is crucial if AI systems are to avoid polluting their own training environment. From a 4E perspective, this degeneration exemplifies how disembedded and disextended architectures sever ties to the lifeworld that normally sustain cognitive diversity and meaning.

A similar diagnosis is offered by Hicks et al. [33], who argue that LLMs produce *bullshit* in the Frankfurtian sense: not because they lie, but because they are structurally indifferent to truth. Their design prioritizes linguistic plausibility and surface coherence over epistemic responsibility. In this way, AI systems persuade without convincing, reflecting a deeper disconnection from the embodied and dialogical contexts in which meaning becomes accountable.

Other recent contributions sharpen this insight. Heersmink et al. [29] frame LLMs as "cognitive artifacts" whose opacity and quasi-human phenomenology can foster misplaced trust, especially in cases of hallucination. This aligns with my view that hallucination is not merely a technical failure but an epistemic risk grounded in relational dynamics of trust, responsibility, and integration into the lifeworld. Heimann and Hübener [30, 31] go further by interpreting LLM hallucinations through Lacanian psychoanalysis, likening them to the structural logic of psychosis and "absolute metaphors." Their account resonates with my claim that hallucinations reveal the specifically *linguistic mode of enactment* of LLMs: they generate coherence around gaps in knowledge without embodied grounding.

Kalai et al. [39] add a complementary statistical perspective, showing that hallucinations are predictable outcomes of training regimes that reward confident answers over abstention. Models thereby develop a structural bias toward guessing, which dovetails with my claim that hallucinations are shaped not only by architecture but also by sociotechnical evaluation practices. From the standpoint of extended agency, such failures must be interpreted within the teleological and normative structures of the lifeworld, where incentives and benchmarks co-constitute both model behavior and human expectations.

Taken together, these perspectives clarify why hallucination and collapse cannot be reduced to incomplete datasets or missing embodiment. They are rooted in the generative design of LLMs as next-token predictors, whose associative operations privilege plausibility over truth-tracking. Logical coherence, when it arises, is a secondary effect of pattern learning rather than a built-in objective. This explains why LLMs can generate fluent, contextually appropriate text while producing factually inaccurate or contradictory claims. From a 4E standpoint, hallucinations are therefore not incidental failures but structural features of AI as an associative process. Precisely because of this, human–machine co-regulation is required if LLMs are to function as reliable extensions of cognition and agency.

8 Conclusion

In this paper, I have argued that artificial intelligence, especially in the form of artificial neural networks and LLMs, should not be conceptualized as either a simulation or duplication of human intelligence. Instead, it should be understood as part of an extended, embodied, embedded, and enactive system of human agency that stands in a relationship of human-AI coevolution. Through its interaction with human purposes, language, and practical contexts, AI participates in the ongoing constitution of the digital lifeworld—a space in which cognition, meaning, and action are co-shaped across human and technological domains.

The key to responsible human–AI interaction lies not in attributing subjectivity or moral agency to machines, but in recognizing the distributed structure of intentionality in socio-technical systems. Drawing on the concept of interobjectivity, I have shown that human actions and AI processes are increasingly interwoven in patterns of mutual coordination, mediated by language, databases, and algorithms. These patterns are not neutral, since they reflect and perpetuate human interests, values, and biases, both implicit and explicit.

From a 4E perspective, these interactions are not merely causal but constitutive: the coupling of AI systems with embodied human practices shapes affordances, transforms action spaces, and reconfigures what is perceived as normatively appropriate or meaningful. Responsibility, therefore, cannot be localized solely in the user or designer, but must be understood as emerging from the entire ecology of interaction in which AI systems operate.

Ethical reflection on AI must thus move beyond the framework of rule-based evaluation of machine behavior. Instead, it must ask: how does AI mediate and extend our cognitions and actions? How does it reorient our goals, constrain our options, and modulate our affective or collective responses? The answers to these questions will depend not on abstract ontologies but on concrete, lived patterns of interaction, shaped by institutions, infrastructures, and historical trajectories.

By analyzing AI as part of extended human agency, we open the way for a new form of responsibility that is situated, relational, and sensitive to the dynamics of human-AI coevolution. Rather than moralizing technology in general and AI systems in particular as such [73], we must reflect on the design of interactions that either support or distort human flourishing—a design that allows for productive and objective human-AI coevolution. This includes critically examining the training data, feedback loops, algorithmic structures, and sociocultural practices that form the background of AI systems.

In conclusion, artificial intelligence does not stand beside the human lifeworld as an object, nor does it merge with it as a subject. It is, rather, enacted within it—as a functional extension of our embodied capacities, embedded in our social environments, enacted through our goals and actions, and extended across the tools and technologies we create. Understanding and shaping this relation responsibly requires a 4E-informed ethics, attentive to the ways in which intelligence, action, and meaning emerge not in isolation, but in interaction.

Acknowledgements
The author used OpenAI's ChatGPT to assist in linguistic revision and idea development. All content, arguments, and conclusions are the author's own and have been critically reviewed.

Author contributions
J.N. wrote the entire manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. No funding was received to assist with the preparation of this manuscript. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 541433337.

Data availability
Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Ethics approval
Ethical approval was not required for this study, as it did not involve human participants, animal subjects, or sensitive personal data.

Consent to participate
Consent to participate was not required for this study, as no human participants were involved.

Consent to publish
Consent to publish was not required for this study, as it does not include any personal data, images, or case material from human participants.

Competing interests
The author declares no competing interests.

Received: 30 June 2025 / Accepted: 15 October 2025
Published online: 13 November 2025

References

1. Aagaard J. 4E cognition and the dogma of harmony. Philos Psychol. 2021;34(2):165–81. https://doi.org/10.1080/09515089.2020.1845640.
2. Agre PE. Toward a critical technical practice: lessons learned in trying to reform AI. In: Geoffrey CB, Susan LS, William T, Les G, editors. Social science, technical systems, and cooperative work. New York: Psychology Press; 1997.
3. Bai Y et al. Constitutional AI: harmlessness from AI feedback. In: arxiv:2212.08073. 2022.
4. Bender EM, Gebru T, McMillan-Major A, Shmargaret S. On the dangers of stochastic parrots: can language models be too big?. In: Conference on fairness, accountability, and transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada. ACM, New York. https://doi.org/10.1145/3442188.3445922. 2021.
5. Buckner C. Empiricism without magic: transformational abstraction in deep convolutional neural networks. Synthese. 2018;195:5339–72.
6. Buckner C. From deep learning to rational machines: what the history of philosophy can teach us about the future of artificial intelligence. Oxford: Oxford University Press; 2024.
7. Clark A. Supersizing the mind: embodiment, action, and cognitive extension. Oxford University Press; 2008.
8. Clark A, Chalmers D. The extended mind. Analysis. 1998;58(19):7–19.
9. Clowes RW. Immaterial engagement: human agency and the cognitive ecology of the internet. Phenomenol Cogn Sci. 2019;18(1):259–79. https://doi.org/10.1007/s11097-018-9560-4.
10. Crawford K. Atlas of AI: power, politics, and the planetary costs of artificial intelligence. Yale University Press; 2021.
11. Dignum V. Responsible artificial intelligence: how to develop and use AI in a responsible way. Cham: Springer; 2019.
12. Dreyfus HL. What computers still can't do: a critique of artificial reason. Cambridge: MIT Press; 1992.
13. Dubber MD, Pasquale F, Das S, editors. The oxford handbook of ethics of AI. Oxford: Oxford University Press; 2020.
14. Durt C. Artificial Intelligence and Its Integration into the Human Lifeworld. In: Voenecky S et al., editors. The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives. Cambridge: Cambridge University Press; 2022. p. 67–82.
15. Engelbart D. Augmenting human intellect: a conceptual framework. Menlo Park: Stanford Research Institute; 1962.
16. Essmann B, Mueller O, et al. AI-supported brain-computer interfaces and the emergence of 'cyberbilities.' In: Voenecky S, et al., editors. The Cambridge Handbook of responsible artificial intelligence. Interdisciplinary perspectives. Oxford: Cambridge University Press; 2022. p. 427–43.
17. Fazi MB. The computational search for unity: synthesis in generative AI. J Contin Philos. 2024. https://doi.org/10.5840/jcp202411652.
18. Fazi MB. A transcendental philosophy of large language models. Philos Digit. 2025. https://doi.org/10.18716/pd.v2i1.11665.
19. Floridi L. The ethics of information. Oxford: Oxford University Press; 2013.
20. Floridi L. The 4th revolution: how the infosphere is reshaping human reality. Oxford: Oxford University Press; 2014.
21. Floridi L. The ethics of artificial intelligence: principles, challenges, and opportunities. Oxford: Oxford University Press; 2023.
22. Freud S. Civilization and its discontents (orig. Das Unbehagen in der Kultur), Wien; 1930.
23. Froese T, Ziemke T. Enactive artificial intelligence: investigating the systemic organization of life and mind. Artif Intell. 2009;173:466–500.
24. Gallagher S. The socially extended mind. Cogn Syst Res. 2013;25–26:4–12. https://doi.org/10.1016/j.cogsys.2013.03.008.
25. Gallagher S. Decentering the brain: Embodied cognition and the critique of neurocentrism and narrow-minded philosophy of mind. Constr Found. 2018;14(1):8–21.
26. Gillespie T. Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media. New Haven: Yale University Press; 2018.
27. Haraldsen M. Engineering the mind: the arts of memory, writing literature and economic agency in digital technology. AI Soc. 2025. https://doi.org/10.1007/s00146-025-02253-6.

28. Heersmink R. Varieties of Artifacts: Embodied, Perceptual, Cognitive, and Affective. Topics in Cognitive Science. 2021. https://doi.org/10.1111/tops.12549.
29. Heersmink R, et al. A phenomenology and epistemology of large language models: transparency, trust, and trustworthiness. Ethics Inf Technol. 2024;26:41. https://doi.org/10.1007/s10676-024-09777-3.
30. Heimann M, Hübener A-F. The extimate core of understanding: absolute metaphors, psychosis and large language models. AI Soc. 2024;40:1265–76. https://doi.org/10.1007/s00146-024-01971-7.
31. Heimann M, Hübener A-F. Circling the void: using Heidegger and Lacan to think about large language models. Cogn Syst Res. 2025;91:101349. https://doi.org/10.1016/j.cogsys.2025.101349.
32. Heras-Escribano M. The philosophy of affordances. Cham: Springer; 2019.
33. Hicks MT, et al. ChatGPT is bullshit. Ethics Inf Technol. 2024;26:38.
34. Huang T. Content moderation by LLM: from accuracy to legitimacy. Artif Intell Rev. 2025;58:320. https://doi.org/10.1007/s10462-025-11328-1.
35. Hui Y. On the Existence of Digital Objects (= Electronic Meditations 48). Minneapolis: University of Minnesota Press; 2016.
36. Husserl E. The crisis of European sciences and transcendental phenomenology: an introduction to phenomenological philosophy, transl. David Carr, Evanston. 1970.
37. Hutchins E. Cognition in the wild. Cambridge: MIT Press; 1995.
38. Ihde D. A phenomenology of technics. In: Kaplan D, editor. Readings in the philosophy of technology. 2nd ed. Lanham: Bloomsbury Publishing; 2009. p. 76–97.
39. Kalai AT et al. Why language models hallucinate. Preprint at arXiv:2509.04664. 2025.
40. Knox J. Artificial intelligence and education in China. Learn Media Technol. 2020;45(3):298–311. https://doi.org/10.1080/17439884.2020.1754236.
41. Kukkonen K. Literature as "uncertainty practice": an anomaly at the end of literature. Dtsch Vierteljahr Lit Geistesgesch. 2023;97:1143–52. https://doi.org/10.1007/s41245-023-00219-4.
42. Latour B. Une sociologie sans objet? Remarques sur l'interobjectivité sur l'interobjectivité. Sociol Trav. 1994;36:587–607.
43. Licklider JCR. Man-computer symbiosis. IRE Trans Hum Factors Electron. 1960;HFE-1:4–11.
44. McHugh J, Šekrst K, Cefalu J. Prompt injection 2.0: hybrid AI threats. Preprint at https://arxiv.org/abs/2507.13169. 2025.
45. Menary R. Cognitive integration: mind and cognition unbounded. London: Palgrave Macmillan; 2007.
46. Millière R, Buckner C. A philosophical introduction to language models. Part I: continuity with classic debates. Preprint at arXiv:2401.03910 [cs.CL], https://doi.org/10.48550/arXiv.2401.03910. 2024.
47. Newen A, Gallagher S, de Bruin L, editors. The Oxford handbook of 4E cognition. Oxford: Oxford University Press; 2018. https://doi.org/10.1093/oxfordhb/9780198735410.001.0001.
48. Noë A. Action in perception. Cambridge: MIT Press; 2004.
49. Noller J. Extended human agency: towards a teleological account of AI. Humanit Soc Sci Commun. 2024;11:1338.
50. Noller J. Connectionism about human agency: responsible AI and the social lifeworld. AI Soc. 2024. https://doi.org/10.1007/s00146-024-02133-5.
51. Norman DA. The design of everyday things: revised and expanded edition. New York: Basic Books; 2013.
52. O'Neil C. Weapons of math destruction: how big data increases inequality and threatens democracy. New York: Crown Books; 2016.
53. Pariser E. The Filter bubble: what the internet is hiding from you. New York: Penguin Books; 2011.
54. Pasquale FA. The black box society: the secret algorithms that control money and information. Cambridge: Harvard University Press; 2015.
55. Paullada A, Raji ID, Bender EM, Denton EL, Hanna A. Data and its (dis)contents: a survey of dataset development and use in machine learning research. Patterns. 2021. https://doi.org/10.1016/j.patter.2021.100336.
56. Pedreschi D, et al. Human-AI coevolution. Artif Intell. 2025;339:1–13. https://doi.org/10.1016/j.artint.2024.104244.
57. Proudfoot D. Rethinking Turing's Test and the Philosophical Implications. Minds and Machines 2020;30:487–512. https://doi.org/10.1007/s11023-020-09534-7
58. Searle J. Minds, Brains and Science. London: Harvard University Press; 1984.
59. Segundo-Ortin M. Agency From a Radical Embodied Standpoint: An Ecological-Enactive Proposal. Front Psychol. 2020;11:1319.
60. Selwyn N. Should robots replace teachers?: AI and the future of education. Cambridge: Polity Press; 2019.
61. Shumailov I, et al. Ai models collapse when trained on recursively generated data. Nature. 2024;631:755–60.
62. Slaby J. Mind invasion: situated affectivity and the corporate life hack. Front Psychol. 2016;7:266. https://doi.org/10.3389/fpsyg.2016.00266.
63. Smart PK. Human-extended machine cognition. Cogn Syst Res. 2018;49:9–23.
64. Stalder F. The digital condition (trans: Pakis VA). Cambridge: Wiley; 2018.
65. Suchman LA. Human-machine reconfigurations: plans and situated actions. 2nd ed. Cambridge: Cambridge University Press; 2007.
66. Sutton J, Harris CB, Keil PG, Barnier AJ. The psychology of memory, extended cognition, and socially distributed remembering. Phenomenol Cogn Sci. 2010;9(4):521–60. https://doi.org/10.1007/s11097-010-9182-y.
67. Toffler A. The third wave. New York; 1980.
68. Touvron H. et al. LLaMA: open and efficient foundation language models. Preprint at arXiv:2302.13971, https://doi.org/10.48550/arXiv.2302.13971. 2023.
69. Tribble EB. Cognition in the globe: Attention and memory in Shakespeare's theatre. Palgrave Macmillan; 2011.
70. Turing A. Computing machinery and intelligence. Mind. 1950;59:433–60.
71. Varela FJ, Thompson E, Rosch E. The embodied mind: cognitive science and human experience. Cambridge: MIT Press; 1991.
72. Vaswani A et al. Attention is all you need. In: NIPS'17: proceedings of the 31st international conference on neural information processing systems. 2017. p. 6000–6010.
73. Verbeek P-P. Moralizing technology: understanding and designing the morality of things. Chicago: University of Chicago press; 2011.
74. Voenecky S, Kellmeyer P, Mueller O, Burgard W, editors. The Cambridge handbook of responsible artificial intelligence: interdisciplinary perspectives. Cambridge: Cambridge University Press; 2022.

75. Wellner GP. When AI is gender-biased. Humana Mente. 2020;13(37):127.
76. Williams D. Pragmatism and the predictive mind. Phenomenol Cogn Sci. 2018;17(5):835–59. https://doi.org/10.1007/s11097-017-9556-5.
77. Williamson B, Piattoeva N. Objectivity as standardisation in data-scientific educational governance: grasping the global through the local. Res Educ. 2022;114(1):25–46. https://doi.org/10.1080/17439884.2018.1556215.

Publisher's Note