frontiers | Frontiers in Artificial Intelligence

TYPE Review
PUBLISHED 03 July 2025
DOI 10.3389/frai.2025.1608014

OPEN ACCESS

EDITED BY
Yuquan Leng,
Southern University of Science and Technology, China

REVIEWED BY
Katya Mkrtchyan,
California State University, Northridge, United States
Wei Gai,
Shandong University, China

*CORRESPONDENCE
Pietro Morasso
✉ pietro.morasso@iit.it

RECEIVED 08 April 2025
ACCEPTED 23 June 2025
PUBLISHED 03 July 2025

CITATION
Sandini G, Sciutti A and Morasso P (2025) Mutual human-robot understanding for a robot-enhanced society: the crucial development of shared embodied cognition. *Front. Artif. Intell.* 8:1608014. doi: 10.3389/frai.2025.1608014

Mutual human-robot understanding for a robot-enhanced society: the crucial development of shared embodied cognition

Giulio Sandini, Alessandra Sciutti and Pietro Morasso*

COgNiTive Architecture for Collaborative Technologies Research Unit, Robotics, Brain and Cognitive Sciences Research Unit – Italian Institute of Technology, Genoa, Italy

The conception of autonomous, intelligent, collaborative robots has been the subject of science fiction rather than science in the second half of the previous century, with practical applications limited to industrial machines without any level of autonomous, intelligent, and collaborative capacity. The new century is facing the challenge of pressing industrial and social revolutions (4, 5, 6, ...) with the prospect of infiltrating robots in every sector of human society; however, this dissemination will be possible if and only if acceptable degrees of autonomy, intelligence, and collaborative capacity can be achieved. Scientific and technological innovations are needed within a highly multidisciplinary framework, with a critical integration strategy and functional characterization that must ask a fundamental question: the design of autonomous, intelligent, collaborative robots should aim at a unified single template to be mass-produced including a standard setup procedure for the functional adaptation of any single prototype, or should the design aim at "baby" robots with a minimal set of sensory-motor-cognitive capabilities as the starting point of a training and educational process in close connection with human companions (masters, partners, final users)? The former alternative is supported by EAI, i.e., the Embodied variant of the Artificial Intelligence family of computational tools based on large foundation models. The latter alternative is bio-inspired; namely, it attempts to replicate the computational structure of Embodied Cognitive Science. Both formulations imply embodiment as a core issue. Still, we think this concept has a markedly different meaning and practical implications in the two cases, although we are still far away from the practical implementations of either roadmap. In this opinion paper, we explain why we think the bio-inspired approach is better than the EAI approach in providing a feasible roadmap for developing autonomous, intelligent, collaborative robots. In particular, we focus on the importance of collaborative human-robot interactions conceived in a general sense, ranging from haptic interactions in joint physical efforts (e.g., loading/unloading) to cognitive interactions for joint strategic planning of complex tasks. We envision this type of collaboration only made possible by a deep human-robot mutual understanding based on a structural equivalence of their embodied cognitive architecture, based on an active, first-person acquisition of experience rather than a passive download of third-person knowledge.

KEYWORDS

embodied artificial intelligence, embodied cognitive science, enaction theory, simulation theory of cognition, developmental psychology, ecological psychology, prospection, extended mind hypothesis

Introduction

The conception of autonomous, intelligent, collaborative robots has been the subject of science fiction rather than science in the second half of the previous century, with practical applications limited to industrial machines without any level of autonomous, intelligent, and collaborative capacity. The new century faces the challenge of pressing industrial and social revolutions (4, 5, 6, …) with the prospect of infiltrating robots in every sector of human society. Still, this dissemination will be possible if and only if acceptable degrees of autonomy, intelligence, and collaborative capacity can be achieved. Scientific and technological innovations are needed, pursued within a highly multidisciplinary framework, with a critical convergent strategy and functional characterization that must ask a fundamental question: the design of autonomous, intelligent, collaborative robots should aim at a unified single template to be mass-produced, including a standard setup/tuning procedure for the functional adaptation of any single prototype, or should the design aim at baby robots with a minimal set of sensory-motor-cognitive capabilities as the starting point of a training and educational process in close connection with human companions (masters, partners, final users)? Considering that both alternatives are decades from an actual implementation/application level, we suggest weighing the pros and cons of the different options and the robustness of their founding bases beyond the strong and/or excessive acclamation of AI technologies.

Knowledge, cognition, intelligence, wisdom

The 2024 Nobel Prizes in Physics and Chemistry awarded to two artificial intelligence scientists have highlighted the social and economic expectations for a scientific methodology based on big data and big computational power, whose scientific substance, in the framework of the modern scientific method established by Galilei and Newton, is still to be understood and still to be proven. There is no doubt that commercially available large language models (LLMs) or vision language models (VLMs) would pass the Turing test, exhibiting the machine's ability to participate in plausible, natural language conversation. In particular, recent experimental results show that GPT-4.5 can pass formalized versions of such test when suitably prompted, with human judges identifying them as human in over 70% of cases—even more frequently than actual human participants (Jones and Bergen, 2025).

We should remember that the test was initially called the imitation game, and as such it could be considered more as a humanlikeness test rather than a direct test of intelligence. Moreover, the only aspect tested is verbal ability, which, by itself, is not the only source for acquiring and organizing knowledge and is certainly not related to the acquisition and organization of goal-driven motor abilities required in robotics. At the same time, there is no doubt that AI models can accelerate the search of complex problem spaces and thus can be used as powerful tools for discovering regularities and hidden properties in large data sets, e.g., for predicting proteins complex structures.

Moreover, the efficiency of the new computational tools does not imply that they characterize a new and superior scientific paradigm, as proposed by AI enthusiasts (Martin and Mani, 2024; Xu et al., 2024) who expect foundation models in generative AI to evolve into a new scientific domain: the crucial innovation would be to overcome the cognitive limitations of human minds, thus achieving reasoning and mental abilities far exceeding those of well-educated humans, including the most distinguished scientists as well as Nobel prize winners. In other words, the evolution of large AI models is expected by AI enthusiasts to quickly achieve a form of general AI (AGI) at the highest level of human intelligence, and then overcome it, ultimately becoming a superhuman omniscient sage, capable of "wisdom," on top of immense amounts of encyclopedic knowledge; without attempting to define in a cogent way such challenging concept, either from the natural or artificial point of view, the jump from intelligence to wisdom is difficult to explain and justify.

In our opinion, the expectation of a future form of "artificial wisdom" on the side and the top of a supreme form of artificial intelligence is somehow illogical and frightening at the same time: in particular, we should consider the well-known ethical problems of AI expressed by many, including one of the two Nobel prize winners (Geoffrey Hinton), interviewed by Heaven (2023). Although the ethical issue is outside the focus of this paper on cognitive robotics, the scientific and technological rationale of an envisaged artificial wisdom needs to be addressed at the beginning of this article, which is based on the crucial role of embodied cognition for human-robot interaction and collaboration principles.

Knowledge, intelligence, and wisdom are related but distinct concepts about human nature, which a standard template cannot capture because each individual is somehow unique, i.e., an exception concerning any conceivable standard, and thus, human nature, in general, is paradoxical, contradictory, and subject to continual change. An inteligent system in the framework of current AI is mainly a system with problem-solving ability that implies the knowledge of large sets of facts and rules. Human wisdom is characterized by several potentially conflicting components (Jeste and Lee, 2019), such as social decision-making, a value system, emotional regulation, prosocial behaviors, self-reflection, acceptance of uncertainty, decisiveness, fusing knowledge with experience, insight, good judgment, etc. Moreover, wisdom has a fundamentally social function, namely, to suggest and induce people to consider the consequences of their actions to themselves and society in the framework of a value system, and there is experimental evidence that the relationships between intelligence and wisdom in individuals is far from linear (Glück and Scherpf, 2022). Thus, moving from the natural to the artificial domain, there is no solid ground to believe that betting on intelligence to achieve wisdom is a promising evolutionary outline. Modern history, marked by the Renaissance and illuminism, which nurtured and consolidated the emergence of the scientific paradigm as an undisputed methodological standard, is a clear witness of the erratic evolution of the shared understanding of the concept of wisdom: in any case, wisdom appears as a changing work in progress, as something to be decided and conquered by a community of individuals, within some democratic framework, not the exceptional individual capability to be evaluated in a competition, as a chess game. Consequently, the expectation of autonomous artificial wisdom as the emerging asymptotic property of higher and higher versions of AI foundation models appears illogical and misleading, scientifically and socially.

In short, we believe that considering advanced forms of AI as a new scientific paradigm is wrong. However, such technologies will be a powerful tool in developing the fourth industrial revolution and

beyond. Moreover, there is a predictable "saturation" of the AI-driven futuristic scenarios determined by the maturation of new computing scenarios such as quantum computing (Morasso, 2023; Gill et al., 2024) and/or wetware/organic computing (Jordan et al., 2024), thus overcoming the limitations of the digital framework that characterizes the current computational formulation of AI. Quantum computing is expected to address the quantum effects that underlie dynamic interaction at the nano-scale in biology, opening the door to an entirely new understanding of the self-organizing processes that characterize the organization and the development of the central nervous system, thus supporting the evolution of natural intelligence. Moreover, a crucial difference between the all-digital scenario and the scenario based on quantum/wetware computing is the energetic consumption, i.e., the energetic frugality intrinsic in the non-digital or only partially digital future scenarios.

On the other hand, the surprising performance of recent foundation models, such as large language models (LLMs) or vision language models (VLMs), hides the fact that these models are essentially passive: they are trained based on vast amounts of data, and thus are intrinsically unable to provide the robot's body with specific inference capabilities, necessary for identifying in any given task the crucial and typically small set of information and combination of actions that make the difference between success and failure or wise versus foolish behavior.

AI vs. embodied cognition

AI is a disembodied computational process conceived to conduct all sorts of imitation games, starting with the Turing test (Turing, 1950). The issue of embodiment is called on stage if an AI agent is requested to act, i.e., the agent is supposed to answer questions coming from the physical world, and the answers should somehow produce effects in the same world. Thus, the cognitive agent must have a body, which includes sensors for detecting and evaluating what happens in the environment and actuators for generating physical effects within a well-defined time framework: in real-time, delayed-time, or intermittent-time. Probably, the first example of a minimal embodied AI agent was proposed by Braitenberg (1984) describing *Vehicles*, namely a class of simple, autonomous moving agents where simple, conceptually analog, wired schemes implemented the process between visual sensors and motor actuators. Depending on the chosen scheme and the structure of the environment, the vehicle could show a variety of complex behaviors that may appear flexible, adaptive, goal-directed, and even intelligent, although in a nutshell scale, without any specific cognitive processes.

This minimalistic approach to what is now known as EAI (Embodied AI) was expanded by Brooks (1991), who proposed that there is no need for complex algorithms or internal representations for producing intelligent behaviors of autonomous agents because the key source of adaptive dynamics is the direct physical interactions of the agent with its environment. Since such interaction is made possible by the agent's body, Brooks concluded that Intelligence must have a body and suggested calling it "embodied intelligence." This *embodiment hypothesis* was further elaborated upon by Smith and Thelen (1994), Pfeifer and Scheier (2001), and Smith (2005), among others, ending up with the current understanding that EAI is a variety of AI that integrates artificial intelligence into physical entities like robots, endowing them with the ability to perceive, learn from, and dynamically interact with their environment: in other words, the explicitly stated goal of EAI is to build General-Purpose Robots via Foundation Models (Hu et al., 2023), namely "robots that operate seamlessly in any environment, with any object, and utilizing various skills to complete diverse tasks." The problem is how to integrate the rationale of data-driven foundation models in the sensory-motor-cognitive structure of a robot, covering the large variety of functions related to perception, prospection, task planning, and action generation: the combined "space" of environments, tasks, and actions is virtually infinite, and attempting to sample it to train a set of foundation models that may encapsulate the cognitive capabilities expected of a generally intelligent robot looks like a hopeless goal, at least in an open environment and in the framework of bounded resources.

On the other hand, we should consider that the same goal of conceiving and designing robots with general intelligence has been investigated well before EAI under the label of Cognitive Architectures for Cognitive Agents. This research field has been very active for several decades, as reported by a recent review (Kotseruba and Tsotsos, 2020), considering tens of projects at different levels of development. However, we are still far away from some standard framework. Among the well-developed prototypes, some architectures focus on modeling human cognition in general as a unified theory of cognition, like SOAR (Laird et al., 1987), ACT-R (Anderson, 1996), CLARION (Sun, 2007), and others aimed explicitly at developmental robotics as iCub (Vernon et al., 2007, 2011) or the cognitive software framework of humanoid robotscognitive robotics, such as ISAC (Kawamura et al., 2008), ArmaX (Vahrenkamp et al., 2015), and CRAM (Beetz et al., 2023). In different manners, such prototypes integrate essential cognitive functions for autonomous cognitive agents (e.g., active perception, purposive action, perceptual inference, learning, adaptation, anticipation, prospection, motivation, attention, action selection, memory, reasoning) with a hybrid combination of computational tools, including symbolic tools (based on logic-based programming and the use of rules and axioms to make inferences and deductions) and sub-symbolic ones, similar to the neural networks of the foundation models of EAI.

However, in both approaches considered above, the embodiment issue plays only a minor role, namely the integration of input–output peripherals with a reasoning/inference machinery of different complexity levels: in the case of the Vehicles model (Braitenberg, 1984), the computational machinery is a simple hand-wired electronic circuit; in the case of CRAM (Beetz et al., 2023) the computational machinery includes self-programmability entailed by physical symbol systems, a plan language for generalized action plans, implicit-to-explicit manipulation, generative models, digital twin knowledge representation, and narrative-enabled episodic memories; in the case of the envisaged innovation toward general-purpose robots via foundation models (Hu et al., 2023) it is expected to apply both existing vision and/or language foundation models already modified for robotic applications (Ahn et al., 2022; Chen et al., 2023) as well as models specifically developed for robotic functions (Brohan et al., 2023a, 2023b) counting on their potential generalization ability across different tasks and even embodiment schemes. In this view, embodiment is somewhat limited and reduced to a one-way flow of information from the sensory periphery toward more remote areas of the brain and then back to the motor periphery. More generally, there is ground to doubt to which extent EAI is really embodied (Hoffmann and Patni, 2025).

In contrast to the minimal utilization of the embodiment concept that characterizes EAI, we should consider an alternative view, namely a form of Embodied Natural Intelligence based on cognitive neuroscience, in particular the subfield known as embodied cognition (Varela et al., 1991; Clark, 1997, 1998) which emerged in the nineties in opposition to the Cartesian dualism and, more recently, to cognitivism and computationalism. Embodied models of cognition are opposed to the disembodied Cartesian model, according to which all mental phenomena are non-physical and thus not influenced by the body, as well as to EAI models where embodiment is limited to one-way interaction between brain, body, and environment. By embodiment, the supporters of embodied cognition refer to the circular, bi-directional interaction where the body allows the brain to physically interact with the environment to accumulate and distill personal experiences, driving the formation and evolution of the agent's cognition. An account of this process is proposed by Varela et al. (1991) as enaction theory, whereby cognitive processes incorporate sensations into a sensorimotor loop, through which active experience of the environment is realized ("enacted"). In this framework, the goal is not to learn the model of the world through interaction but to learn the model of the interaction between the agent and the world. Such a view about embodied cognition, based on the working hypothesis that cognitive processes are deeply rooted in the body's bi-directional interactions with the world, is summarized by Wilson (2002) into six claims about the fundamental features of human embodied cognition which have a direct computational relevance: (1) situated-ness, (2) time-pressured-ness, (3) exploitation of the intrinsic dynamics of the environment, (4) integration of the environment in the cognitive architecture, (5) cognition is mainly an online, action-oriented process; (6) even offline cognition is body based. The experimental baseline to support such claims is diverse and includes the following lines of evidence: ideo-motor theories of perception (Prinz, 1987); the developmental psychology of Piaget (1952) who traced the evolution of cognitive levels from the consolidation of sensorimotor abilities up to higher levels; the ecological psychology of Gibson (1977) who characterized active perception as the discovery of potential interactions with the environment, i.e., affordances; the linguistic decomposition of abstract concepts in terms of qualitative explanations based on bodily metaphors (Lakoff and Johnson, 1999); the sociocultural theory of Vygotsky (1978) emphasizing the role of social interaction in shaping cognitive development.

Despite all this evidence in favor of the view that the mind must be understood only in the context of its relationship with a physical body that interacts with the world in an online manner (the embodiment hypothesis), it has been objected that cognitive activities of various nature can take place as well when the brain is decoupled from any immediate interaction with the environment, i.e., in an offline manner. This coexistence of online and offline operational modes of human cognitive activities contradicts the embodied cognition hypothesis, which claims to characterize the whole of cognition, not only a part of it. However, this is only an apparent paradox if we associate the online vs. offline antinomy (related to the interaction of a cognitive agent with the environment) with two additional antinomies, namely actual vs. imagined activities and overt vs. covert actions: these antinomies reflect the computational and neural equivalence between the sensory-motor-cognitive processes involved in the execution of purposive actions and the processes activated for reasoning about virtual actions, for example in the context of the fundamental cognitive function known as prospection. Prospection is the mental simulation of actions to evaluate their potential sensorimotor, environmental, and social effects in the future, thus supporting an informed (and potentially wise) decision-making process (Gilbert and Wilson, 2007; Seligman et al., 2013; Vernon et al., 2015). The experimental evidence about the equivalence stated above comes from the study of motor imagery (Decety, 1996; O'Shea and Moran, 2017) and different forms of a simulation theory of cognition (Decety and Ingvar, 1990; Jeannerod, 2001; Hesslow, 2002, 2012; Grush, 2004; Ptak et al., 2017); an essential part of this theory is that the simulation is performed by the same neural mechanisms as those typically involved in movement execution and perception, although some researchers suggest that simulation (or emulation) of actions is performed by a neural mechanism that is different and separate from brain areas directly involved in movement and perception (Wolpert et al., 1998). In any case, the equivalence between overt and covert actions does not refer only to the geometry of the involved brain areas but also to the timing of the simulated actions in comparison with the executed ones (Shepard and Metzler, 1971; Decety et al., 1989; Decety and Jeannerod, 1995; Karklinsky and Flash, 2015; Gauthier and van Wassenhove, 2016).

The issue about the apparent online vs. offline paradox, related to the timing of purposive action in support of embodied cognition, is further completed if we consider another facet of human cognition, related to the spatial aspect of purposive actions, i.e., the role of cognitive cortical maps: such brain structures, located in the medial temporal lobe, were proposed by Tolman (1948) for understanding flexible behavior in rodents, e.g., foraging patterns by rats in mazes. In humans, it has become evident that, in addition to their function in spatial navigation, cognitive maps are also the backbone of a systematic organization of knowledge in abstract spaces in such a way as to support the learning of higher-level knowledge (Behrens et al., 2018; Bellmund et al., 2018; Bokeria et al., 2021; Qiu et al., 2024): this means that neurons previously identified in cognitive maps for guiding navigation in the physical environment, such as "place cells," "grid cells" and "head-direction cells" are also likely to support the ability to mentally "navigate" through conceptual spaces for more abstract reasoning tasks.

The double role of cognitive maps, namely the integration in the same brain structure of the "geometrical" aspects of actions at different abstract levels, together with the double role of the brain areas involved in the simulation theory of cognition, for the representation of the "timing," "kinematic," and "haptic" aspects of overt as well as covert actions, explains in which sense and how much embodied human cognition is fundamentally embodied in contrast to the minimal degree of embodiment which characterizes artificial intelligence, in general, and EAI in particular.

From the philosophical standpoint, the strong formulation of the nature and organization of embodied cognition is consistent with the *extended mind hypothesis* (Clark and Chalmers, 1998), namely the belief in the fundamental active role of the environment in driving cognitive processes. Learning, one of the mind's primary functions, emerges from the closed-loop dynamics that link active perception, purposive action, cognition, and dynamically changing environment. In other words, learning and other fundamental cognitive functions as prospection should be understood in the framework of an extended theory of the mind, which includes the changing environment as a

part of the mind dynamical model. In the extended mind hypothesis framework, we should also consider that such extension is naturally articulated in two directions: extension to the physical environment and the social environment. In other words, we should assume that the process of mind extension is not innate or genetically coded, although it is based on genetically based mechanisms, but is mainly the product of different developmental processes, articulated in two main streams: (a) the multi-stage theory of cognitive development, starting with the sensorimotor stage (Piaget, 1952); (b) the sociocultural theory of cognitive development (Vygotsky, 1978).

Figure 1 illustrates in a simplified manner the difference between the two roadmaps examined above for the design of autonomous, intelligent, collaborative robots, namely the roadmap based on AI foundation models and the roadmap based on full embodied cognition. The former alternative, in the top panel, shows that the sensory, motor, control, and reasoning processes that are required for carrying out a given task as a function of a given environment are inferred from a large foundation model, trained by the sampled performance of a population of skilled human agents operating in similar situations, i.e., a large dataset of third-person knowledge: the crucial point is that such data are collected with unnecessary high-resolution but with insufficient filtering of "keyframes." The bottom panel illustrates the main features of the proposed bio-inspired roadmap. In particular, it singles out the crucial features of embodied cognition, based on the accumulation of first-person experience, filtered and organized into a personal episodic/procedural memory, evaluated through a prospective process that combines overt and covert actions using a body model and cooperative interactions with a skilled tutor. Although the body, which is in charge of producing overt actions, and the body-schema, which is supposed to deal with covert (mental) actions, are represented graphically as different blocks, we must remember that, in agreement with the theory on the neural simulation of actions, they incorporate and integrate both functions in the same computational module.

Embodied cognition and computational frugality

Summarizing the analysis of the previous section, we highlight that human cognition substantiates first-person (autobiographical) human experience, characterized as embedded, enactive, and extended, based on the continually changing, embodied, and affective interactions with the world. In contrast, EAI aims at the design of general-purpose robots via foundation models (Hu et al., 2023) based on large amounts of pre-coded, general-purpose, encyclopedic knowledge, providing third-person (impersonal) experience with a

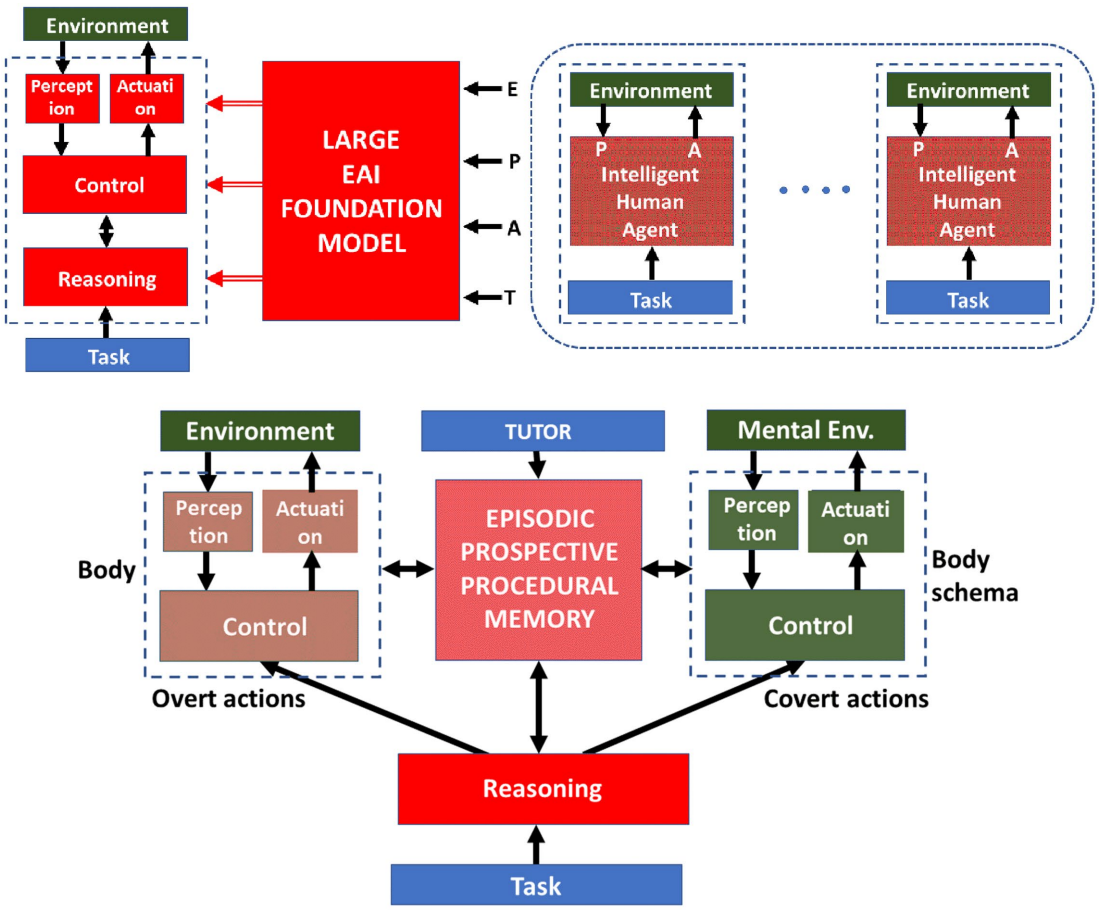

FIGURE 1
Design principles of autonomous, intelligent, collaborative robots. Summary of the difference between the roadmap based on EAI (top section) and the roadmap based on embodied, developmental cognition (bottom section).

low degree of embodied cognitive interaction. In any case, the goal of EAI is still far away, and possibly the expectation of functionally open-ended, general-purpose, super-intelligent robots is an unreasonable dream. Even EAI enthusiasts (Liu and Wu, 2024) recognize that there is still a lot to do because large foundation models (LLMs and VLMs) may support only a part of the essential cognitive capabilities of autonomous intelligent robots, providing efficient inference capabilities. In the framework based on foundation models, what is lacking is related to the crucial first-person cognitive experience that may allow a robot to be truly autonomous, intelligent, and efficient. In particular, it is suggested that EAI robotics must develop several brand-new cognitive models as the following: an evolutionary learning process driven by the agent's physical interactions with open environments; multiverse representations of a virtual environment that can effectively emulate the real world, and interact with the EAI systems (Hall et al., 2022); understanding the physical world, such as the concept of gravity, by using intuitive physics models (Piloto et al., 2022).

In general, at the current level of development and understanding, EAI-based robots are likely to face a kind of deep "personality conflict," namely the conflict between the third-person, disembodied, offline-trained foundation models that should provide the central core of the cognitive capabilities and the first-person, online interaction with the environment and the associated training processes. There is no guarantee that the two coexisting paradigms can avoid conflicting situations and/or may face conflicts without a principle of arbitration and solution in the short and long term. Conversely, findings from developmental psychology demonstrate that in humans abstract cognitive skills—such as the use of abstract verbs or numerical reasoning—are fundamentally grounded in sensorimotor activity, and bodily experience scaffolds symbol formation. Such a process allows a smooth integration of first-person and third-person knowledge, as suggested by computationally modeled formulations (Cangelosi and Stramandinoli, 2018).

The foundational role of the body becomes particularly evident when we contrast it with the limitations of current large-scale learning architectures. Despite their remarkable performance in language and vision tasks, foundational models—whether LLMs or VLMs—struggle to generalize seemingly trivial spatial concepts. Notions such as height, relative position, or reachability—intuitively mastered by humans from early infancy through bodily interaction—remain elusive for these models, especially when they are applied in robotic contexts. For instance, robotic arms trained through behavioral cloning on extensive demonstrations show impressive proficiency when reproducing known actions in familiar settings. However, even minor changes—such as a slight shift in the object's position, a different tablecloth pattern, or a slight variation in height—can cause performance to drop dramatically, as recently pointed out by Dieter Fox in his "Where's RobotGPT" talk (Fox, 2024). These changes, which humans would effortlessly generalize due to their embodied spatial understanding and reasoning, often require retraining or additional examples for the model to adapt, thus illustrating how the lack of embodied grounding and causality severely limits the generalization and adaptability of current AI models in physical environments.

A fundamental difference between passive learning—as performed by LLMs—and the kind of first-person learning we advocate lies in the nature of captured relationships. LLMs, trained on vast corpora of linguistic data, have proven remarkably capable of extracting statistical regularities, many of which reflect deep correlations embedded in language. This correlational power allows them to solve complex tasks with apparent fluency. However, this mechanism radically differs from how humans (and robots with a fundamentally embodied training experience) learn through active engagement with the world. In first-person learning, the agent does not merely observe correlations—it experiences *causation*. By performing an action with a specific goal in mind and perceiving its consequences, the agent can establish a direct link between its behavior and the resulting outcome. This ability to isolate causal mechanisms provides a powerful filter that distinguishes meaningful action-effect relationships from coincidental correlations, a fundamental process well documented in the developmental literature. According to the "interventionist" view of causality (Gopnik and Schulz, 2007), knowing that X causes Y implies that manipulating X leads to a change in Y. Children learn about causation precisely through intentional interventions and the observation of contingent outcomes. By age four, children can actively experiment to infer causal structures (Schulz and Bonawitz, 2007), going beyond early Piagetian learning, where actions are simply associated with their direct outcomes (Piaget, 1930). Even at around 24 months, infants are adept at observational causal learning: they do not merely imitate or detect correlations between events but infer causal relationships from others' actions and use those inferences to plan their own interventions (Meltzoff et al., 2012). These findings suggest that the human capacity for causal learning is rooted in embodied, intentional activity from a very early age. Crucially, this causally grounded understanding—emerging from both action and observation—enables robust generalization. Instead of brute-force pattern matching over all conceivable correlations, the agent can reason about what actions are likely to produce desired effects, even in unfamiliar scenarios. Replicating this capacity in artificial agents requires grounding learning in embodied, interactive experience. Otherwise, the ability to generalize causal knowledge across domains and contexts will likely remain severely limited.

Moreover, it may be observed that the computational model, which forms the philosophical foundation of artificial intelligence, implies intractable problems (Clark, 1999): in particular, an information bottleneck occurs when the mind is requested to construct detailed representations of the external world to produce appropriate purposive actions. The problem is that the world is constantly changing, for its dynamics and as an effect induced by the agent's actions, and thus, the demands on the mental system are likely to preclude the agent from producing appropriate actions just in time. The nature of such information bottleneck, due to the supposed need for a multiverse representation of the environment, is another aspect of the computational prodigality that characterizes AI in general and EAI in particular, based on the "brute-force" assumption that infinite amounts of training data are available and computational resources are vast and free. In contrast, as observed by Clark, humans need relatively little information about the world before they manage to act effectively upon it.

Vision and, in general, the multi-sensory perception of the peripersonal space in the surrounding environment (Di Pellegrino and Làdavas, 2015; de Vignemont et al., 2021) is an active, purposive, attention-driven process not a passive, high-resolution, virtual representation. Although the spatial awareness implicit in everyday life supports the illusion of a stable and fully detailed representation

of the world, this subjective impression (Clark, 1997) obscures the reality of minimal and low-detail environmental information where the constraint of quick action guides the search and acquisition of missing perceptual evidence to extract information "just in time." This concept exemplifies the computational frugality of the bidirectional human embodied cognition. It avoids the computationally expensive reconstruction of a detailed world model on the working assumption that the world is its best model, which only needs to be sampled where and when required.

Another crucial aspect of the computational frugality of the human-embodied cognitive system is related to the role of episodic memory (Dickerson and Eichenbaum, 2010) in the framework of versatile and articulated human memory systems. The episodic memory system is implemented in the brain by an extended circuitry centered around the medial temporal lobe (MTL), interacting with several cortical and subcortical areas: the functions of the cortical components address many aspects of perception and cognition, whereas the MTL system mediates the formation and retrieval of the associative network of memories whose details are stored in the cortical areas. Episodic memories (EMs) are related to specific personal experiences that occur in daily life and, for some reason, are isolated for their "exceptional" relevance and stored in long-term memory. The motivations to single out such episodes from the sensory-motor flow of daily life can be of different types, such as curiosity, novelty detection, emotional drive, social interaction with a teaching master, etc. These memories are structured chunks of information that include spatio-temporal patterns about sequences of actions and the surrounding environment. They include a declarative component, expressed explicitly by direct conscious access to information and communicated by spoken language, and a nondeclarative component, such as a procedural memory about the learned sequence of movements or actions appropriate for the memorized episode.

Episodic memories are unique samples extracted from any embodied cognitive agent's continuous sensorimotor experience flow and coded in some associative storage. Sensorimotor intelligence implies the dual capacity, on one side, to identify and code relevant or crucial episodes and, on the other, to detect the resonance with a stored episode in a given action sequence. Thereafter, the cognitive agent is supposed to quickly retrieve the detailed episode, adapt it to the specific circumstance, and produce the corresponding procedural behavior. In any case, this mnemonic process is not reproductive, as a kind of playback routine of stored information, but reconstructive, namely the activation of an internal simulation model based on the key parameters stored in the episodic memory. Of course, episodic memories are far from detailed digital recordings and do not need to be so. When retrieved from long-term memory to guide an action plan, they are instantiated with slight modifications induced by the situation and the agent's state. However, this kind of flexibility may include margins of failure if the recall occurs in extreme conditions. This issue is well known in forensic psychology (Sarwar et al., 2004) concerning the possible contradictions of eyewitnesses: due to the reconstructive nature of the mnemonic process, it is likely that the event recalled by an eyewitness, in situations with intense emotional stress, is corrupted by unrelated memory fragments that have nothing to do with the truth. Such a problem affects large associative memory systems, such as Hopfield networks (Hopfield, 1982), in case of overloading. However, this does not affect the main issue, namely the fact that the episodic memory system, in association with procedural memories, is a formidable mechanism of computational frugality that allows a cognitive agent to store and retrieve the minimum amount of information together with the minimum amount of computational power.

Preliminary studies have explored the systematic use of episodic memories in cognitive robotics (Mohan et al., 2014; Vernon et al., 2015). In our opinion, this is one of the crucial research lines to be further investigated in embodied cognitive robotics. The main difference of this approach, in comparison with the EAI approach, based on large foundation models, is that it is based on first-person experience rather than third-person, pre-coded knowledge: computational frugality of the single autonomous agent vs. computational prodigality of the population of super-intelligent agents. However, this does not imply that cognitive robots, educated according to principles of first-person acquisition of experience, cannot take advantage of the consultation of encyclopedic knowledge stored in books, manuals, movies or web browsers using language tools as AI foundation models. This (third-person) knowledge can be used to update or adapt/consolidate first-person know-how obtained through personal experience, for example, by modifying specific parameters of episodic memory or the associated procedural trace. The opposite process, namely integrating first-personal knowledge into a large third-person structure, is unnatural and impractical.

Along the same line of reasoning, we suggest that the issue of computational frugality can be associated with the well-known epistemological, philosophic concept known as Ockham's razor, namely the principle of cognitive parsimony in the search for an explanation of scientific or philosophic problems, i.e., the principle that robust explanations should be constructed with the smallest possible set of elements: entia non sunt multiplicanda praeter necessitate (entities should not be multiplied beyond necessity).

Embodied cognition and the extended mind hypothesis

In a previous section, we briefly discussed the extended mind hypothesis (Clark and Chalmers, 1998), concerning the organization of embodied cognition, and we suggested two interrelated extensions: integration with the physical environment and integration with the social environment. The former issue deals with the need for the embodied cognitive system to incorporate some degree of what is known as commonsense knowledge about the dynamics of the physical world, including causality during physical interaction, the effect of gravity, etc. Commonsense representation and reasoning have been among the central issues addressed by symbolic AI (a.k.a. GOFAI: Good Old-Fashioned AI) in the 70s and 80s, focused on a family of computational models known as expert systems. In particular, the subsets of expert systems oriented to robotic applications were designed to implement *qualitative physics* (Forbus, 1988): the key idea was to find ways to represent continuous properties of the world, traditionally formulated employing differential equations, by discrete systems of symbols, thus allowing different styles of reasoning, like qualitative simulation and envisioning. The success of this approach to commonsense was somewhat limited, with scarce application to autonomous robotics. More recently, with the

expansion of connectionist AI to the level of foundation models, the topic was revisited under the label of intuitive physics (Piloto et al., 2022): it is conceived as a network of concepts focused on the discovery of the hidden principles that explain the interactions of macroscopic objects in the real world. The suggested approach is to use foundation models of the VLMs type, such as the PLATO (Physics Learning through Auto-encoding and Tracking Objects). PLATO is a foundation model that has been trained by a large number of videos depicting objects interacting according to the laws of physics. For simplicity, the videos were generated by simulation experiments, not by observed phenomena. The big dataset was used to train large, deep networks to acquire some commonsense understanding of the dynamics of the physical world, which is useful for reasoning and prospection.

In a different way and with a different sophistication level, qualitative physics, and intuitive physics models fail to achieve the goal of EAI, namely the design of autonomous, intelligent, collaborative robots because are unable to implement an embodied cognitive architecture fully integrated with the first-person experience of a cognitive agent operating in a specific environment and with well-defined functions. The alternative to PLATO is bio-inspired in the sense of exploiting the simulation theory of cognition, which supervises both real and virtual sensorimotor patterns performed in the context of the current world model and of the typical tasks performed by the cognitive agent in real life and cooperation with human or robotic partners. Thus, the training data that are necessary for developing or updating neural models capable of achieving an intuitive understanding of physics are self-generated by the cognitive agent: intuitive physics and/or other intuitive understanding of the dynamics of the environment and the body-environment interaction and are implicitly addressed in a first-person manner. Moreover, this first-person approach combines active synergy formation with data preparation for learning, including the critical labeling step, an Achilles' heel for training foundation models.

The intuitive physics component of the extended mind hypothesis, based on the active physical interaction with the environment, may also be characterized as a computationally frugal strategy, accumulating an amount of data for self-training according to the principle as much as needed, not in general but in a personalized way for the specific cognitive agent.

In a bio-inspired way, we suggest that the extension of the extended mind hypothesis as a result of learning by self-training is the result of a developmental process, organized in layers according to the Piagetian theory with a progressive acquisition of the level of intuitive physics understanding. We may also envisage that, at high levels of cognitive capability, including a sufficient level of linguistic competence, the cognitive agent may be motivated to search for third-person knowledge by interrogating commercially available foundation models trained with massive encyclopedic datasets. For example, the answer provided by the foundation model may allow the cognitive agent to choose one alternative sequence of actions among several possibilities consistent with his previous experiences and incorporate this information in the corresponding episodic memory.

As regards the social extension of the extended mind hypothesis, we should consider the broad research area at the border of motor neuroscience and cognitive neuroscience, namely the large number of experimental studies that emphasize the strong implication of the motor system, specifically responsible for the production of covert and overt actions, also in typical cognitive functions as action observation, imitation and social interaction (Fadiga et al., 1995; Fadiga et al., 2002; Iacoboni et al., 1999; Grezes et al., 2001) as well as activities related to movement in a more abstract way as the observation of manual tools or the use of action verbs (Martin et al., 1996; Grafton et al., 1997). Although these studies were aimed primarily at understanding the interaction of human subjects with the environment and/or the interaction between humans, we believe that they can be naturally extended to the design of autonomous, intelligent, and collaborative robots.

We can view the motor and cognitive systems as forming a pair of equivalent loops, one related to open actions and the other to covert actions: in the former case, motor commands cause muscle contractions with consequent sensory feedback, which in turn influences the control of future motor commands; in the latter case, motor intentions activate an internal body schema with consequent sensory predictions, which in turn affect the ideomotor formulation of future action plans (Mohan et al., 2019). When a person (a naïve performer) interacts with another person (an expert), we can think of an analogous additional loop, namely a social interaction loop in which the "controlled object" is the other person rather than the actual or imagined motion of one's own body. For example, the social interaction loop is instantiated because the naïve person attempts to imitate the expert or because the experts supervise the naïve partner's action, guiding his performance with intermittent intervention. Therefore, in social interactions, by controlling someone else rather than our own body, we can estimate their hidden state, including their mental state, rather than our own body (Wolpert et al., 2003). In other words, the control signals that characterize social interactions may be considered communicative actions, including speech, gestures, and haptic interaction.

One of the primary motivations for adopting this approach for a fully embodied AI roadmap is that it strongly matches the requirement of mutual understanding between humans and robot partners in a very general sense. Equipping robots with a cognitive architecture grounded in first-person experience allows the emergence of a form of cognitive compatibility between human and robot agents. When both share similar developmental principles—such as the incremental accumulation of sensorimotor experiences, episodic memory formation, and action-oriented reasoning—the human partner is more likely to understand the rationale behind the robot's behavior, including its mistakes. This compatibility can play a crucial role in enabling a more natural and effective form of robot education (Matarese et al., 2021). In such a scenario, humans can more easily interpret the robot's errors and provide corrective feedback in ways that the robot can meaningfully incorporate. As a result, robot learning becomes more transparent, interpretable, and ultimately more efficient. Conversely, when it is the robot's turn to provide suggestions or assistance, its behavior is more likely to be perceived as understandable and trustworthy (Matarese et al., 2023). This compatibility not only facilitates the correction of sensorimotor mistakes or the refinement of practical know-how but also opens the door to a richer educational process involving abstract and culturally embedded concepts, including notions of what is considered appropriate or inappropriate behavior. In other words, it becomes feasible to teach a robot not only how to do something, but

also whether it should be done—based on the teacher's values. Much like a child learns through example and imitation what is right or wrong within a given family or culture, a robot equipped with an embodied cognitive architecture may become receptive to similar forms of moral or normative guidance (Sandini et al., 2024). Even in the case of interspecies learning, such as training a dog, the process succeeds despite evident communicative asymmetries because the animal learns through embodied interaction and contextual reinforcement. Similar dynamics can be envisioned in human-robot interaction, provided that the robot cognition is grounded in first-person experience and can structure knowledge accordingly. In contrast, this possibility is largely inaccessible in the case of passive AI systems trained offline on abstract, third-person data, where the encoding of universal ethical rules becomes an ill-defined and arguably unsolvable problem.

Moreover, this is also the roadmap for integrating learning through first-person experience and third-person interaction with various interaction channels, from web-interrogation to human education and tutoring. The relevance of the educational issue of embodied cognition (Hegna and Ørbæk, 2021) is also thoroughly addressed by a theme issue "Minds in movement: embodied cognition in the age of artificial intelligence" of the Philosophical Transactions B (Barrett and Stout, 2024). After having recognized embodiment as a unifying concept in the study of cognition, this study focuses on two key themes, namely the role of language in cognition and its entanglement with the body and the multiple bodily mechanisms of interpersonal perception and alignment across the domains of social affiliation, teaching, and learning: in both themes, AI language models can be valuable tools for robot training.

Conclusion

Summarizing this opinion paper, we may say that the roadmap to the design of autonomous, intelligent, collaborative robots supposed to infiltrate our society for its expected technology-driven reorganization can be characterized according to the following principles:

- Fully embodied cognitive architecture functionally equivalent to the human counterpart
- Learning and training based on prospection capabilities and accumulation of first-person experiences stored in episodic memory
- The crucial role of social interaction for accessing third-person information, defining the intelligence level appropriate for the sought performance target and achieved through a process of sensory-motor-cognitive development
- Human-robot collaboration should adhere to the principle that the limit to a robot's autonomy is the ultimate responsibility of the human partner and/or the social environment at large.

In general, we suggest that it does not make sense (from the scientific and economic sense) to aim at a single design target of a super-intelligent robot to be easily adapted to the variety of application paradigms that should fit social needs. In many situations, this could be too much and a waste of computational resources; in specific situations, it could be too little. We suggest a frugal computational architecture with an initial, minimal configuration to be grown up through learning and training in a well-organized social context. Such minimal architecture could be conceived as an extension of the Vehicle paradigm (Braitenberg, 1984), grown up in a self-organizing and self-training context. Moreover, we fully agree with (Lake et al., 2017) that such machines should be designed to learn and think like people.

An alternative roadmap is pursued by AI companies, counting on the continuously growing progress in developing LLMs to understand human requests and communicate plans of action using natural language. A very recent example is provided by the new model, Gemini Robotics, developed by Google DeepMind (Gemini Robotics Team, Google DeepMind, 2025) that combines its best large language model with robotics. The goal is to give robots the ability to be more dexterous and generalize across tasks, exploiting the generalizing capabilities of LLMs, such as reasoning about which actions to take in a given context. In any case, Gemini robots are trained similarly to most LLMs, namely text, images, and videos from the internet or synthetic data generated by simulation models without any personal accumulation of personal experience and first-person knowledge. The explicitly stated rationale of the Gemini Robotics family is to develop general-purpose robots that realize AI's potential in the physical world: despite the remarkable documented examples of performance, we believe that this is not the appropriate roadmap for the massive diffusion of cooperative cognitive agents in a variety of qualitatively different scenarios as robot teachers, robot helpers, robot companions, and so on. Critical features of such scenarios may not be coded in large collections of text, images, videos, or synthetic datasets but may be hidden in haptic interaction, haptic guidance, and gestural non-verbal communication with human partners, thus allowing the crucial development of shared embodied cognition of robot and human cooperative partners.

The research group that includes the authors of this paper has been working on various building blocks for the design of autonomous, intelligent, collaborative robots according to a bio-inspired roadmap based on embodied cognition for more than a decade. In particular, this research activity has been focused, among other things, on prospection, learning a body schema, embodied simulation of action, imitation learning, episodic memory, understanding physical interaction, social cognition based on embodied communication, and developmental learning (Lungarella et al., 2003; Mohan and Morasso, 2007; Metta et al., 2010; Mohan et al., 2011, 2013, 2014; Vernon et al., 2015; Bhat et al., 2016; Bhat et al., 2017; Sciutti and Sandini, 2017; Sandini et al., 2018, 2024; Pasquali et al., 2025). We are still away from an implementation framework that allows us to integrate the variety of building blocks outlined above in a flexible and self-organizing way. In our opinion, such a framework should be hybrid, combining digital, analog, symbolic, and subsymbolic representations similar to what we know of the human embodied cognitive architecture. In any case, we are confident that the proposed roadmap is naturally suitable for facing ethical issues and social impact because the main design goal is to facilitate a shared embodied cognition between the robot and the human companion as much as possible.

Author contributions

GS: Writing – review & editing, Conceptualization, Funding acquisition. AS: Conceptualization, Funding acquisition, Writing – review & editing. PM: Conceptualization, Writing – review & editing, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work is supported by the Istituto Italiano di Tecnologia (IIT) Genoa Italy throughthe iCog Initiative, coordinated by the RBCS Research Unit, and the "Brain and Machines" Flagship Program. Financial support is also provided by G. A. no. 804388, wHiSPER (Starting Grant from the European Research Council under the European Union's Horizon2020 Research and Innovation Program) and by "Progetto "Future Artificial Intelligence Research (hereafter FAIR)," code PE00000013 funded by the European Union—NextGenerationEU PNRR MUR—M4C2—Investimento 1.3—Avviso Creazione di "Partenariati estesi alleuniversità, ai centri di ricerca, alle aziende per il finanziamento di progetti di ricerca di base" CUP J53C22003010006.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

References

Ahn, M., Brohan, A., and Brown, N., (2022). Do as I can, not as I say: grounding language in robotic affordances.

Anderson, J. R. (1996). ACT: a simple theory of complex cognition. *Am. Psychol.* 51, 355–365. doi: 10.1037/0003-066X.51.4.355

Barrett, L., and Stout, D. (2024). Minds in movement: embodied cognition in the age of artificial intelligence. *Philos. Trans. Royal Soc. B Biol. Sci.* 379:20230144. doi: 10.1098/rstb.2023.0144

Beetz, M., Kazhoyan, G., and Vernon, D. (2023). The CRAM cognitive architecture for robot manipulation in everyday activities.

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., et al. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* 100, 490–509. doi: 10.1016/j.neuron.2018.10.002

Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., and Doeller, C. F. (2018). Navigating cognition: spatial codes for human thinking. *Science* 362:aat6766. doi: 10.1126/science.aat6766

Bhat, A., Akkaladevi, S. C., Mohan, V., Eitzinger, C., and Morasso, P. (2017). Towards a learnt neural body schema for dexterous coordination of action in humanoid and industrial robots. *Autonomous Robot.* 41, 945–966. doi: 10.1007/s10514-016-9563-3

Bhat, A., Mohan, V., Sandini, G., and Morasso, P. (2016). Humanoid infers Archimedes' principle: understanding physical relations and object affordances through cumulative learning experiences. *J. R. Soc. Interface* 13:20160310. doi: 10.1098/rsif.2016.0310

Bokeria, L., Henson, R. N., and Mok, R. M. (2021). Map-like representations of an abstract conceptual space in the human brain. *Front. Hum. Neurosci.* 15:620056. doi: 10.3389/fnhum.2021.620056

Braitenberg, V. (1984). Vehicles: Experiments in synthetic psychology. Cambridge, MA: MIT Press.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., et al (2023b). "Rt-2: vision-language-action models transfer web knowledge to robotic control," in *Proceedings of the 7th Conference on Robot Learning Research, Atlanta, USA*.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., et al. (2023a). "Rt-1: robotics transformer for real-world control at scale," in *Proceedings of the Conference on Robotics: Science and Systems, Daegu, Republic of Korea, July 10–July 14, 2023*.

Brooks, R. A. (1991). Intelligence without representation. *Artif. Intell.* 47, 139–159. doi: 10.1016/0004-3702(91)90053-M

Cangelosi, A., and Stramandinoli, F. (2018). A review of abstract concept learning in embodied agents and robots. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 373:20170131. doi: 10.1098/rstb.2017.0131

Chen, B., Xia, F., Ichter, B., Rao, K., Gopalakrishnan, K., Ryoo, M. S., et al. (2023). "Open-vocabulary queryable scene representations for real world planning," in *2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom*, 11509–11522.

Clark, A. (1997). Being there: Putting brain, body, and world together again. Cambridge, MA: MIT Press.

Clark, A. (1998). "Embodied, situated, and distributed cognition" in A companion to cognitive science. eds. W. Bechtel and G. Graham (Malden, MA, USA: Blackwell), 506–517.

Clark, A. (1999). An embodied cognitive science? *Trends Cogn. Sci.* 3, 345–351. doi: 10.1016/S1364-6613(99)01361-3

Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis* 58, 7–19. doi: 10.1093/analys/58.1.7

de Vignemont, F., Serino, A., Wong, H. W., and Farnè, A. (2021). The world at our fingertips: A multidisciplinary exploration of peripersonal space. Oxford: Oxford University Press.

Decety, J. (1996). The neurophysiological basis of motor imagery. *Behav. Brain Res.* 77, 45–52. doi: 10.1016/0166-4328(95)00225-1

Decety, J., and Ingvar, D. H. (1990). Brain structures participating in mental simulation of motor behaviour: a neuropsychological interpretation. *Acta Psychol.* 73, 13–34. doi: 10.1016/0001-6918(90)90056-l

Decety, J., and Jeannerod, M. (1995). Mentally simulated movements in virtual reality: does Fitts's law hold in motor imagery. *Behav. Brain Res.* 72, 127–134. doi: 10.1016/0166-4328(96)00141-6

Decety, J., Jeannerod, M., and Prablanc, C. (1989). The timing of mentally represented actions. *Behav. Brain Res.* 34, 35–42. doi: 10.1016/s0166-4328(89)80088-9

Di Pellegrino, G., and Làdavas, E. (2015). Peripersonal space in the brain. *Neuropsychologia* 66, 126–133. doi: 10.1016/j.neuropsychologia.2014.11.011

Dickerson, B. C., and Eichenbaum, H. (2010). The episodic memory system: neurocircuitry and disorders. *Neuropsychopharmacology* 35, 86–104. doi: 10.1038/npp.2009.126

Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *Eur. J. Neurosci.* 15, 399–402. doi: 10.1046/j.0953-816x.2001.01874.x

Fadiga, L., Fogassi, L., Pavesi, G., and Rizzolatti, G. (1995). Motor facilitation during action observation: a magnetic stimulation study. *J. Neurophysiol.* 73, 2608–2611. doi: 10.1152/jn.1995.73.6.2608

Forbus, K. D. (1988). "Qualitative physics: past, present, and future" in Exploring artificial intelligence. ed. H. E. Shrobe (Burlington, MA: Morgan Kaufman), 239–296.

Fox, D. (2024). Where's RobotGPT?. Available online at: https://www.youtube.com/watch?v=OAZrBYCLnaA (Accessed April 27, 2024).

Gauthier, B., and van Wassenhove, V. (2016). Cognitive mapping in mental time travel and mental space navigation. *Cognition* 154, 55–68. doi: 10.1016/j.cognition.2016.05.015

Gemini Robotics Team, Google DeepMind (2025) Gemini robotics: bringing AI into the physical world.

Gibson, J. J. (1977). "The theory of affordances" in Perceiving, acting, and knowing: Toward an ecological psychology. eds. R. Shaw and J. Bransford (Hillsdale, NJ: Lawrence Erlbaum), 67–82.

Gilbert, D., and Wilson, T. (2007). Prospection: experiencing the future. *Science* 317, 1351–1354. doi: 10.1126/science.1144161

Gill, S. S., Cetinkaya, O., Marrone, S., Claudino, D., Haunschild, D., and Schlote, L. (2024). Quantum computing: vision and challenges.

Glück, J., and Scherpf, A. I. (2022). Intelligence and wisdom: age-related differences and nonlinear relationships. *Psychol. Aging* 37, 649–666. doi: 10.1037/pag0000692

Gopnik, A., and Schulz, L. E. (2007) in Causal learning: Psychology, philosophy, and computation. eds. A. Gopnik and L. E. Schulz (Oxford: Oxford University Press).

Grafton, S. T., Fadiga, L., Arbib, M. A., and Rizzolatti, G. (1997). Premotor cortex activation during observation and naming of familiar tools. *NeuroImage* 6, 231–236. doi: 10.1006/nimg.1997.0293

Grezes, J., Fonlupt, P., Bertenthal, B., Delon-Martin, C., Segebarth, C., and Decety, J. (2001). Does perception of biological motion rely on specific brain regions? *NeuroImage* 13, 775–785. doi: 10.1006/nimg.2000.0740

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behav. Brain Sci.* 27, 377–396. doi: 10.1017/s0140525x04000093

Hall, B. D., Liu, Y., Jansen, Y., Dragicevic, P., Chevalier, F., and Kay, M. (2022). A survey of tasks and visualizations in multiverse analysis reports. *Comput. Graph. Forum* 41, 402–426. doi: 10.1111/cgf.14443

Heaven, W. D. (2023). Geoffrey Hinton tells us why he's now scared of the tech he helped build. *MIT Technol. Rev.* 2:2023.

Hegna, H. M., and Ørbæk, T. (2021). Traces of embodied teaching and learning: a review of empirical studies in higher education. *Teach. High. Educ.* 29, 420–441. doi: 10.1080/13562517.2021.1989582

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends Cogn. Sci.* 6, 242–247. doi: 10.1016/S1364-6613(02)01913-7

Hesslow, G. (2012). The current status of the simulation theory of cognition. *Brain Res.* 1428, 71–79. doi: 10.1016/j.brainres.2011.06.026

Hoffmann, M., and Patni, S. P. (2025). Embodied AI in machine learning – is it really embodied?

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* 79, 2554–2558. doi: 10.1073/pnas.79.8.2554

Hu, Y., Xie, Q., Jain, V., Francis, J., Patrikar, J., Keetha, N., et al (2023). Toward general-purpose robots via foundation models: a survey and meta-analysis.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., and Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science* 286, 2526–2528. doi: 10.1126/science.286.5449.2526

Jeannerod, M. (2001). Neural simulation of action: a unifying mechanism for motor cognition. *NeuroImage* 14, S103–S109. doi: 10.1006/nimg.2001.0832

Jeste, D. V., and Lee, E. E. (2019). The emerging empirical science of wisdom: definition, measurement, neurobiology, longevity, and interventions. *Harv. Rev. Psychiatry* 27, 127–140. doi: 10.1097/HRP.0000000000000205

Jones, C. R., and Bergen, B. K. (2025). Large language models pass the Turing test.

Jordan, F. D., Kutter, M., Comby, J., Brozzi, F., and Kurtys, E. (2024). Open and remotely accessible neuroplatform for research in wetware computing. *Front. Artif. Intell.* 7:1376042. doi: 10.3389/frai.2024.1376042

Karklinsky, M., and Flash, T. (2015). Timing of continuous motor imagery: the two thirds power law originates in trajectory planning. *J. Neurophysiol.* 113, 2490–2499. doi: 10.1152/jn.00421.2014

Kawamura, K., Gordon, S. M., Ratanaswasd, P., Erdemir, E., and Hall, J. F. (2008). Implementation of cognitive control for a humanoid robot. *Int. J. Humanoid Robot.* 5, 547–586. doi: 10.1142/S0219843608001558

Kotseruba, J., and Tsotsos, J. K. (2020). A review of 40 years in cognitive architecture research core cognitive abilities and practical applications. *Artif. Intell. Rev.* 53, 17–94. doi: 10.1007/s10462-018-9646-y

Laird, J. E., Newell, A., and Rosenbloom, P. S. (1987). Soar: an architecture for general intelligence. *Artif. Intell.* 33, 1–64. doi: 10.1016/0004-3702(87)90050-6

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *BBS.* 40:e253. doi: 10.1017/S0140525X16001837

Lakoff, G., and Johnson, M. (1999). Philosophy in the flesh: The embodied mind and its challenge to western thought. New York, NY, USA: Basic Books.

Liu, S., and Wu, S. (2024). A brief history of embodied artificial intelligence, and its outlook. *Commun. ACM*. (Accessed April 29, 2024).

Lungarella, M., Metta, G., Pfeifer, R., and Sandini, G. (2003). Developmental robotics: a survey. *Connect. Sci.* 15, 151–190. doi: 10.1080/0954009031000165511O

Martin, C. H., and Mani, G. (2024). The recent physics and chemistry Nobel prizes, AI, and the convergence of knowledge fields. *Patterns* 5:101099. doi: 10.1016/j.patter.2024.101099

Martin, A., Wiggs, C. L., Ungerleider, L. G., and Haxby, J. V. (1996). Neural correlates of category-specific knowledge. *Nature* 379, 649–652. doi: 10.1038/379649a0

Matarese, M., Cocchella, F., Rea, F., and Sciutti, A. (2023). "Ex(plainable) machina: how social-implicit XAI affects complex human-robot teaming tasks," in *Proceedings - IEEE International Conference on Robotics and Automation (ICRA), London, UK*, 11986–11993.

Matarese, M., Sciutti, A., Rea, F., and Rossi, S. (2021). Toward robots' behavioral transparency of temporal difference reinforcement learning with a human teacher. *IEEE Trans. Hum.-Mach. Syst.* 51, 578–589. doi: 10.1109/THMS.2021.3116119

Meltzoff, A. N., Waismeyer, A., and Gopnik, A. (2012). Learning about causes from people: observational causal learning in 24-month-old infants. *Dev. Psychol.* 48, 1215–1228. doi: 10.1037/a0027440

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., et al. (2010). iCub humanoid robot: an open-systems platform for research in cognitive development. *Neural Netw.* 23, 1125–1134. doi: 10.1016/j.neunet.2010.08.010

Mohan, V., Bhat, A., and Morasso, P. (2019). Muscleless motor synergies and actions without movements: from motor neuroscience to cognitive robotics. *Phys Life Rev* 30, 89–111. doi: 10.1016/j.plrev.2018.04.005

Mohan, V., and Morasso, P. (2007). Towards reasoning and coordinating action in the mental space. *Int. J. Neural Syst.* 17, 329–341. doi: 10.1142/S0129065707001172

Mohan, V., Morasso, P., Sandini, G., and Kasderidis, S. (2013). Inference through embodied simulation in cognitive robots. *Cogn. Comput.* 5, 355–382. doi: 10.1007/s12559-013-9205-4

Mohan, V., Morasso, P., Zenzeri, J., Metta, G., Chakravarthy, V. S., and Sandini, G. (2011). Teaching a humanoid robot to draw shapes. *Auton. Robot.* 31, 21–53. doi: 10.1007/s10514-011-9229-0

Mohan, V., Sandini, G., and Morasso, P. (2014). A neural framework for organization and flexible utilization of episodic memory in "cumulatively" learning baby humanoids. *Neural Comput.* 26, 2692–2734. doi: 10.1162/NECO_a_00664

Morasso, P. (2023). The quest for cognition in purposive action: from cybernetics to quantum computing. *J. Int. Neurosci.* 2:39. doi: 10.31083/j.jin2202039

O'Shea, H., and Moran, A. (2017). Does motor simulation theory explain the cognitive mechanisms underlying motor imagery? A critical review. *Front. Hum. Neurosci.* 11:72. doi: 10.3389/fnhum.2017.00072

Pasquali, D., Garello, L., Belgiovine, G., Eldardeer, O., Lastrico, L., Rea, F., et al. (2025). "No robot is an island: an always-on cognitive architecture for social context awareness in dynamic environments," in *2025 IEEE International Conference on Development and Learning (ICDL 2025), Prague*. Preprint on Zenodo.

Pfeifer, R., and Scheier, C. (2001). Understanding intelligence. Cambridge, MA: MIT Press.

Piaget, J. (1930). The child's conception of physical causality. New York, NY: Harcourt Brace.

Piaget, J. (1952). The origins of intelligence. New York, NY: W.W. Norton & Co.

Piloto, L. S., Weinstein, A., Battaglia, P., and Botvinick, M. (2022). Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nat. Hum. Behav.* 6, 1257–1267. doi: 10.1038/s41562-022-01394-8

Prinz, W. (1987). "Ideo-motor action" in Perspectives on perception and action. eds. H. Heuer and A. F. Sanders (Londond: Routledge), 47–76.

Ptak, R., Schnider, A., and Fellrath, J. (2017). The dorsal frontoparietal network: a core system for emulated action. *Trends Cogn. Sci.* 21, 589–599. doi: 10.1016/j.tics.2017.05.002

Qiu, Y., Li, H., Liao, J., Chen, K., Wu, X., Liu, B., et al. (2024). Forming cognitive maps for abstract spaces: the roles of the human hippocampus and orbitofrontal cortex. *Commun Biol* 7:517. doi: 10.1038/s42003-024-06214-5

Sandini, G., Mohan, V., Sciutti, A., and Morasso, P. (2018). Social cognition for human-robot symbiosis - challenges and building blocks. *Front. Neurorobot.* 12:34. doi: 10.3389/fnbot.2018.00034

Sandini, G., Sciutti, A., and Morasso, P. (2024). Artificial cognition vs. artificial intelligence for next-generation autonomous robotic agents. *Front. Comput. Neurosci.* 18:1349408. doi: 10.3389/fncom.2024.1349408

Sarwar, F., Allwood, C. M., and Innes-Ker, A. (2004). Effects of different types of forensic information on eyewitness' memory and confidence accuracy. *Eur. J. Psychol. Legal Context* 6, 17–27. doi: 10.5093/ejpalc2014a3

Schulz, L. E., and Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Dev. Psychol.* 43, 1045–1050. doi: 10.1037/0012-1649.43.4.1045

Sciutti, A., and Sandini, G. (2017). Interacting with robots to investigate the bases of social interaction. *IEEE Trans. Neural Syst. Rehabil. Eng.* 25, 2295–2304. doi: 10.1109/TNSRE.2017.2753879

Seligman, M. E. P., Railton, P., Baumeister, R. F., and Sripada, C. (2013). Navigating into the future or driven by the past. *Perspect. Psychol. Sci.* 8, 119–141. doi: 10.1177/1745691612474317

Shepard, R. N., and Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science* 171, 701–703. doi: 10.1126/science.171.3972.701

Smith, L. B. (2005). Cognition as a dynamic system: principles from embodiment. *Dev. Rev.* 25, 278–298. doi: 10.1016/j.dr.2005.11.001

Smith, L. B., and Thelen, E. (1994). A dynamic systems approach to the development of cognition and action. Cambridge, MA: MIT Press/Bradford Books.

Sun, R. (2007). The importance of cognitive architectures: an analysis based on CLARION. *J. Exp. & Theor.l Artif. Intel.* 19, 159–193. doi: 10.1080/09528130701191560

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychol. Rev.* 55, 189–208. doi: 10.1037/h0061626

Turing, A. M. (1950). Computing machinery and intelligence. *Mind* 59, 433–460. doi: 10.1093/mind/LIX.236.433

Vahrenkamp, N., Wachter, M., Krohnert, M., Welke, K., and Asfour, T. (2015). The robot software frameworks ArmarX. *Inf. Technol.* 57, 99–111. doi: 10.1515/itit-2014-1066

Varela, F. J., Thompson, E., and Rosch, E. (1991). The embodied mind: Cognitive science and human experience. Cambridge, MA: MIT Press.

Vernon, D., Beetz, M., and Sandini, G. (2015). Prospection in cognitive robotics: the case for joint episodic-procedural memory. *Front. Robot. AI* 2:19. doi: 10.3389/frobt.2015.00019

Vernon, D., Metta, G., and Sandini, G. (2007). "The iCub cognitive architecture: interactive development in a humanoid robot," in *2007 IEEE 6th international conference on development and learning (ICDL 2007)*. London, UK, 122–127.

Vernon, D., Von Hofsten, C., and Fadiga, L. (2011). A roadmap for cognitive development in humanoid robots. Heidelberg, Berlin, Germany: Springer-Verlag.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Wilson, M. (2002). Six views of embodied cognition. *Psychon. Bull. Rev.* 9, 625–636. doi: 10.3758/bf03196322

Wolpert, D. M., Doya, K., and Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 358, 593–602. doi: 10.1098/rstb.2002.1238

Wolpert, D. M., Miall, R. C., and Kawato, M. (1998). Internal models in the cerebellum. *Trends Cogn. Sci.* 2, 338–347. doi: 10.1016/S1364-6613(98)01221-2

Xu, Y., Wang, F., and Zhang, T. (2024). Artificial intelligence is restructuring a new world. *Innovation* 5:100725. doi: 10.1016/j.xinn.2024.100725