Position: A Call for Embodied AI

Giuseppe Paolo [1] **Jonas Gonzalez-Billandon** [2] **Balázs Kégl** [1]

Abstract

We propose Embodied AI (E-AI) as the next fundamental step in the pursuit of Artificial General Intelligence (AGI), juxtaposing it against current AI advancements, particularly Large Language Models (LLMs). We traverse the evolution of the embodiment concept across diverse fields (philosophy, psychology, neuroscience, and robotics) to highlight how E-AI distinguishes itself from the classical paradigm of static learning. By broadening the scope of E-AI, we introduce a theoretical framework based on cognitive architectures, emphasizing perception, action, memory, and learning as essential components of an embodied agent. This framework is aligned with Friston's active inference principle, offering a comprehensive approach to E-AI development. Despite the progress made in the field of AI, substantial challenges, such as the formulation of a novel AI learning theory and the innovation of advanced hardware, persist. Our discussion lays down a foundational guideline for future E-AI research. Highlighting the importance of creating E-AI agents capable of seamless communication, collaboration, and coexistence with humans and other intelligent entities within real-world environments, we aim to steer the AI community towards addressing the multifaceted challenges and seizing the opportunities that lie ahead in the quest for AGI.

1. Introduction

Over recent years, the field of artificial intelligence (AI) has experienced a significant surge, leading to substantial breakthroughs in areas ranging from computer vision (CV) and natural language processing (NLP) to neuroscience. This journey through AI's development has been marked by a series of significant triumphs interspersed with setbacks, including the well-documented AI winter of the mid-1980s. The ambitious goal that has propelled AI research forward from the beginning was to create intelligence that either parallels or exceeds human abilities. This quest for superhuman intelligence, commonly termed Artificial General Intelligence (AGI), has been seen differently by experts across different disciplines, yet it broadly refers to the ability of a system to understand, learn, and apply knowledge in a wide array of tasks and contexts, mirroring the cognitive flexibility of humans and animals.

The remarkable progress in AI over the past decade can largely be attributed to three pivotal developments: i) advancements in deep learning algorithms, ii) the advent of powerful new hardware, and iii) the availability of extensive datasets for training. A prime illustration of this advancement is the creation of Large Language Models (LLMs) like OpenAI's GPT-4 (Achiam et al., 2023) and Google's Gemini (Team et al., 2023). The surprising abilities of these LLMs have sparked discussions within the AI community, with some pondering whether these models have already achieved nascent forms of AGI. Foundation models (large networks with billions of parameters trained on massive datasets) have found success in varied fields, ranging from predicting 3D protein structures (Cramer, 2021) and robotic control (Brohan et al., 2023), to generating images and audio (Ramesh et al., 2022; Radford et al., 2022). This breadth of achievement supports the hypothesis that continued scaling and refinement of foundation models could be a viable path toward realizing AGI.

In our paper, we argue that despite the significant advances made by current AI technologies, they represent only the initial steps towards truly intelligent agents. Despite their impressive capabilities, these large networks are static and unable to evolve with time and experience. They leverage large datasets and cutting-edge hardware for scaling, but they lack the ability to properly *care* about the truth (Vervaeke & Coyne, 2024), which in turn makes it impossible to dynamically adjust their knowledge and actively search for *valuable* new information. The two primary manifestations of this fundamental shortfall are i) the difficulty in effectively aligning LLMs (Ouyang et al., 2022), and ii) their propensity to generate plausible but inaccurate information, a phenomenon known as confabulation (Huang et al., 2023; Wei et al., 2023). Current strategies to mitigate these issues, such as post-processing, fine-tuning, prompt engineering,

[1]Noah's Ark Lab, Huawei Technologies France, Paris, France [2]London Research Center, London, UK. Correspondence to: Giuseppe Paolo <giuseppe.paolo@huawei.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

and incorporating human feedback, are undeniably valuable. However, we argue that these methods address only the superficial aspects of the problem and fall short in dealing with the core issue at play: the inherent lack of a deeper, grounded sense of *care* in LLMs.

Pursuing the development of AGI, we draw upon the insights of Vervaeke & Coyne (2024) to advocate for designing AI agents that are bound to, observe, interact with, and learn from the real world (including humans) in a continuous and dynamic manner. These Embodied AI (E-AI) agents ought to prioritize their continued existence and our bindings to them, thereby learning the *value* of truth. They should also be capable of adapting to environmental changes and evolving without human intervention.

While Large Language Models play a significant role in the development of AI systems, they fall short of capturing the essence of what constitutes an intelligent agent. Notably, intelligent beings, whether humans or animals, are characterized by three fundamental components: the mind, perception, and action capabilities (Kirchhoff et al., 2018). LLMs, or more broadly, foundation models, may be likened to an aspect of the mind's reasoning function (Xi et al., 2023). Yet, the perceptive and action-oriented dimensions of intelligence, along with the pivotal ability to dynamically revise beliefs and knowledge based on experiences, remain unaddressed. Autoregressive LLMs are not designed to understand the causal relationships between events, but rather to identify proximate context and correlations within sequences (Bariah & Debbah, 2023). In contrast, a fully embodied agent should have the ability to grasp the causality underlying events and actions within its environment, be it digital or physical. By comprehending these causal relationships, such an agent can make informed decisions that consider both the anticipated outcomes and the reasons behind those outcomes.

In short, **this paper argues that the necessary next step in our pursuit for truly intelligent and general AIs is the development and study of Embodied AI. We propose that current LLM-based foundation models could lay the groundwork for designing these agents, but are just one component of a truly embodied agent.** This approach is akin to how neonates come into the world equipped with inherent priors to successfully adapt to the world (Reynolds & Roth, 2018).

In the next section of this paper, we will define the concept of embodiment and what we mean by E-AI in Sec. 2, analyzing the literature and various scientific and philosophical currents. In Sec. 3, we discuss why we believe this is a necessary step towards Artificial General Intelligence (AGI). In Sec. 4 we analyze the main components of a truly embodied agent, and we discuss the major challenges to achieving this ambitious goal in Sec. 5. Our motivation behind the need to develop E-AI and its fundamental role in our path towards AGI is proposed throughout. Finally, Sec. 6 provides a short recap of our proposition.

2. What is embodied AI

Embodied AI (E-AI) is s sub-field of AI, focusing on agents that interact with their physical environment, emphasizing sensorimotor coupling and situated intelligence. As opposed to mere passive observing, E-AI agents act on their environment and learn from the reaction. E-AI is deeply rooted in *embodied cognition* (Shapiro, 2011; McNearney, 2011), a perspective in philosophy and cognitive science that posits a profound coupling between the mind and the body. This idea, challenging Cartesian dualism — the historically dominant view that distinctly separates the mind from the body (Descartes, 2012) — emerged in the early 20th century. Pioneers like Lakoff & Johnson (1979; 1999) have significantly contributed to this paradigm by proposing that reason is not based on abstract laws but is grounded in bodily experiences. Embodied cognition forms a critical part of the 4E cognitive science framework (Varela et al., 1991; Clark, 1997; Clark & Chalmers, 1998), encompassing embodied, enactive, embedded, and extended aspects of cognition. Within E-AI, the focus is predominantly on implementing the 'embodied' and 'enactive' aspects, while the 'embedded' and 'extended' components are more pertinent to situating AI in a social context and as an augmentation of human (individual or collective) cognition.

In AI, initial explorations into embodiment emerged in the 1980s, driven by a growing recognition of the inherent limitations in disembodied agents. These limitations were primarily attributed to the absence of rich, high-bandwidth interactions with the environment (Pfeifer & Iida, 2004; Pfeifer & Bongard, 2006). An early advocate for this paradigm shift was Brooks (1991), who built walking robots simulating insect-like locomotion. Simultaneously, the field of computer vision was undergoing its own transformation. Researchers and practitioners were increasingly focusing on enabling agents to interact with their surroundings. This emphasis on interaction led to a concentration on the perceptual elements of embodiment, particularly from a first-person point of view (POV) (Shapiro, 2021). This approach aligns with the concept of visual exploration and navigation (Ramakrishnan et al., 2021), where an agent acquires information about a 3D environment through movement and sensory perception, thereby continuously refining its model of the environment (Anderson et al., 2018; Chen et al., 2019). Such exploration techniques empower an agent to discover objects and understand their permanence. As a result of these developments, many contemporary benchmarks in E-AI have emerged predominantly from the domains of vision and robotics (Duan et al., 2022), reflecting the integral role these disciplines have played in advancing the field.

That said, the broader definition of E-AI does not require vision. Sensorimotor coupling may be implemented using *any* physical sense (Pfeifer & Bongard, 2006). In the living world, many organisms survive and thrive without vision, using, for example, chemical or electric sensing (Bargmann, 2006). Levin (2022)'s Technlogical Approach to Mind Everywhere (TAME) framework further explores this idea, suggesting that cognition emerges from the collective intelligence of cell groups, they themselves deeply embodied within their environment (the body they comprise). This framework challenges traditional Cartesian dualism, embedding cognition within the physical and biological makeup of an organism. In the TAME perspective, cognition is not just an attribute of higher-order organisms; it extends throughout the ontological hierarchy of living beings, from individual cells, through tissues and organs, to complex organisms. Each agent demonstrates cognitive capabilities that are inherently connected to its physical structure and the environmental interactions at its proper level. This broadened view of cognition and embodiment goes beyond the conventional focus on vision in robotics and computer vision. It posits that any entity capable of perceiving, interacting with, and learning from its environment, thereby adapting to it and influencing it, qualifies as embodied. A technological instantiation of this concept is an intelligent router in a telecommunication network. This device 'lives' in a realm dominated by electromagnetic sensing. It continuously learns from and adapts to the network traffic, effectively mapping and managing the flow of information. This example underscores the potential of applying the principles of E-AI beyond the traditional domains, embracing a more inclusive and diverse understanding of intelligence and embodiment.

This broadening of the notion of E-AI raises the question: *how close current commercial AI tools are to embodiment?* Here we examine two such tools: Large Language Models (Brown et al., 2020; Devlin et al., 2019) and Social Media Content AI Recommendation Systems (SMAI) (Bakshy et al., 2015; Covington et al., 2016; Eirinaki et al., 2018).

LLMs operate within a linguistic-symbolic domain, representing textual information and generating new text by completing prompts. Their foundational training is essentially static, relying on datasets meticulously compiled and curated by teams of AI engineers. Their goal is supervised: to generate likely tokens following a context. Their secondary training (fine-tuning) may involve both interactions with their symbolic environment (human users) and goals to reach (satisfy their human users), but these interactions are presents some limits due to both technical (e.g., catastrophic forgetting (Kirkpatrick et al., 2017; Parisi et al., 2019)) and business (e.g., managing individuated LLMs (Strubell et al., 2019; Kaplan et al., 2020)) reasons. Looking ahead, we anticipate advancements that might address these limitations, potentially leading to the emergence of "personal assistant" LLMs. These would represent a form of embodied agents within a symbolic realm. However, at present, LLMs largely resemble static *Internet AI (I-AI)* (Duan et al., 2022), differing significantly from the dynamic, interactive nature characteristic of E-AI.

It is intriguing that despite the growing concerns about the risks and alignment challenges of LLMs highlighted in recent research (Bender et al., 2021), SMAIs have attracted comparatively less scrutiny (Huszár et al., 2022; Ribeiro et al., 2020). This is noteworthy considering SMAIs have been around for a longer time and their influence on society is both wider and more profound. We propose that their widespread acceptance and their more integrated, less intrusive presence in our lives are due to their closer alignment with the principles of embodiment, in contrast to LLMs.

What do we mean by SMAIs being closer to embodiment? Firstly, SMAIs are driven by clear objectives: to captivate our attention and maximize our engagement with their respective platforms (Bozdag, 2013; Bodó, 2021). These goals are fundamentally linked to the business models of these platforms, which revolve around advertising. The specifics of these "engagement" objectives are typically proprietary, forming the core of the competitive advantage of these platforms. Although these goals are initially human-designed and not intrinsically generated (Covington et al., 2016), they are subject to evolutionary pressure and adaptation, and thus they are tied to the existence and the survival of the SMAI. Secondly, SMAIs learn almost entirely from the data they collect by interacting with us. This leads to a high level of individuation (adapting to our individual preferences (Nguyen et al., 2014)), and notions of exploration (offering us content not so much to satisfy us but for the sake of learning what we like). This creates a user experience that, when well-executed, resembles interaction with a considerate friend, who wants our best, who connects us to things we like, and who wants to understand us better. The flip side, however, is the potential for these systems to morph into mechanisms that perpetuate addictive behaviors or harmful content (Schüll, 2012). Nevertheless, since SMAIs connect and adapt to us in a more intuitive and deeper manner than LLMs, we often feel a greater sense of control over our interactions with these systems (by, for example, consciously not clicking on content that we know we do not want to see in the long run). This control, albeit limited, is reminiscent of *persuasion* more than mechanical manipulation, aligning with how we interact with other sentient beings rather than machines. This type of relationship with AI systems is a fundamental aspect of Levin (2022)'s TAME proposal . Our stance on E-AI suggests that, while systems akin to SMAIs pose greater risks due to their seamless integration into our social fabric, they also present more natural opportunities for alignment with our values. This alignment process is procedural, perspectival, and evolutionary in nature (Ver-

vaeke et al., 2012; Vervaeke & Coyne, 2024), contrasting with the primarily propositional approaches being applied to LLMs (Shen et al., 2023).

We posit that the potential for more effective and naturally aligned AI systems is, alone, a compelling reason to prioritizing E-AI in the broader AI research agenda.

In the forthcoming section, we further explore the pivotal role that well-executed implementations of E-AI could play in the quest for AGI.

3. Why embodiement?

In the previous section, we examined how contemporary theories of embodiment, particularly the TAME framework (Levin, 2022), challenge the long-standing Cartesian dualism which posits a distinct separation between mind and body (Descartes, 2012). This philosophical stance has significantly influenced the development of current generative AI models, such as LLMs, which primarily rely on static data and lack interaction with the physical or even the symbolic world. It is a prevalent belief that simply scaling up such models, in terms of data volume and computational power, could lead to AGI. We contest this view. We propose that true understanding, not only propositional truth but also the *value* of propositions that guide us how to *act*, is achievable only through E-AI agents that live in the world and learn of it by interacting with it.

The significance of embodiment in cognitive development was demonstrated by Held & Hein (1963)'s carousel experiment with kittens. In this study, one kitten could actively interact with and control a carousel, while the other could only observe it passively. Despite both kittens receiving identical visual input, the one engaged in active interaction exhibited normal visual development, unlike its passively observing counterpart. This seminal experiment underscores the vital role of embodied interaction in shaping cognitive abilities (Shenavarmasouleh et al., 2022). It also reinforces the observation that all known forms of intelligence, including human intelligence, are inherently embodied (Smith & Gasser, 2005), suggesting that embodiment serves as a solid foundation for cognitive learning and development. Current AI learns in a very different way from humans. We humans learn by seeing, moving, interacting with the world and speaking with others. We also learn by collecting sequential experiences, not by passive observation of shuffled and randomized, even if carefully selected, data (Smith & Gasser, 2005; Westho et al., 2020). We advocate for an approach where insights from cognitive science and developmental psychology inform the design of AI systems. Such systems should be designed to learn through active interaction with their surroundings, mirroring the embodied learning processes fundamental to human cognition.

Even advocates of static learning concede that multimodal learning is the next milestone towards AGI (Fei et al., 2022; Parcalabescu et al., 2021). In I-AI, multimodal data needs to be collected and connected painstakingly. In contrast, E-AI agents, when equipped with multimodal sensors, will inherently collect and correlate multi-modal data by mere co-occurrence. For instance, robots will see (CV), communicate (NLP), reason (general intelligence), navigate and interact with their environment (planning and RL), all simultaneously (Shenavarmasouleh et al., 2022). Intelligent routers will observe requests and traffic (sensing), communicate with other routers, human engineers, absorb news about their surroundings (NLP), reason (general intelligence), and control the traffic (control and RL). Despite the impressive progress in these domains, much of it has relied on the external collection and curation of vast datasets for algorithmic training. This approach has significant drawbacks: i) the collection and preparation of data demands substantial investments; ii) this data can contain biases that are hard to detect and rectify (Li & Deng, 2020; Balayn et al., 2021; Verma et al., 2021). The issue of biases is particularly pertinent in discussions on AI alignment (Shen et al., 2023; Ji et al., 2023). Efforts to align AI through rule-based and procedural methods (such as RLHF (Lambert et al., 2022)) often struggle, producing systems that feel mechanistic and "dumb", rather than an agent which seamlessly *acts* according to values compatible with our society.

An embodied agent, designed to interact with and learn from its environment, fundamentally changes the traditional approach to data collection and curation in AI development. By being inherently integrated with its physical and social contexts, such an agent bypasses the labor-intensive processes previously required. This shift not only simplifies the challenge of aligning AI with human values but also enhances the agent's learning efficiency by utilizing the unique features of its environment. As a result, the focus in AI development transitions from data to *simulators*. These simulators serve a dual purpose: they are both training grounds for E-AI and platforms for testing and refining concepts and algorithms (Duan et al., 2022). Moreover, the process of aligning these agents with human values becomes more intuitive as it involves defining goals reflective of those values. This approach does not claim to fully resolve the alignment challenge, as E-AI systems will still necessitate oversight and guidelines to avert unwanted behaviors. However, the alignment process becomes inherently more natural. Adjusting and defining *goals* is a more straightforward task than the extensive editing and curating of *data*. This methodology draws upon our inherent, non-propositional understanding and instincts about aligning embodied intelligences—whether it is guiding our own actions, nurturing children, or training pets.

Another important characteristic of E-AI, stemming from the coupling between the agent and its environment, is the

agent's capacity for ongoing evolution and adaptation. This adaptability is vital for any agent destined to navigate a world in perpetual change. It underscores the importance of continual learning: the process of assimilating new experiences while retaining previously acquired knowledge (Wang et al., 2023a).

Moreover, Ishiguro & Kawakatsu (2004) have shown, both through theory and practical application in robotics, that a close and effective integration of control mechanisms with body dynamics significantly enhances energy efficiency. Coupled systems lead also to the emergence of intriguing behaviors that can be hard to explicitly program or learn from disembodied datasets (Rosas et al., 2020), an observation aligning with the principles of the TAME framework.

Embodiment is also a prerequisite for learning about *affordances* (Gibson, 1979). Learning, or more precisely *realizing* affordances, according to Vervaeke et al. (2012)'s perspectival learning, is a fundamental capacity of AGI, as affordances are what "fill our world with meaning" (Roli et al., 2022), and are thus necessary for agents that give meaning to their own world. Affordances emerge from the dynamic interplay between an agent's perception, objectives, abilities, and the characteristics of objects and contexts within the environment; for example, a chair affords us to sit, a glass to drink and a hand to grasp and pick up objects. Roli et al. (2022) argue that the capacity to comprehend, utilize, and be influenced by environmental affordances distinguishes biological intelligence from current artificial systems. Besides affordances, E-AI is also indispensable for investigating emergent phenomena such as qualia (Locke, 1847; Korth, 2022), consciousness (Solms, 2019), as well as creativity, empathy (Perez, 2023), and ethical understanding (Lake et al., 2017; Russell, 2021).

Finally, there is the important question of why an intelligent agent would do anything in the first place (Pfeifer & Iida, 2004). What drives it to engage and acquire new knowledge without external prompts? Within well-framed small worlds, such as a chess game, an agent's purpose is straightforward: deciding the next move. However, when navigating large, open worlds, the motivations guiding an agent's decisions grow increasingly ambiguous. The concepts of active inference and the free energy principle (Friston, 2010; Friston et al., 2023) provide a compelling framework for understanding the behaviors of intelligent agents. This principle posits that minimizing surprise and uncertainty is the core objective of the agents. They achieve this through the use of internal models to forecast outcomes, continually updating these models with sensory input, and proactively modifying their surroundings to better match their expectations. This concept resonates within the AI community, particularly in the design of agents equipped with mechanisms for *intrinsic motivation* (Oudeyer & Kaplan, 2007; Pathak et al., 2017), which incentivize agents to explore and acquire new knowledge to reduce uncertainty.

However, what propels an intelligent agent to act, especially beyond mere survival instincts, continues to be a matter of debate. We argue that exploring and developing embodied agents will illuminate this question. Thus, E-AI not only shows potential for significant breakthroughs toward achieving AGI, but also has deep implications for our understanding of cognition in general.

4. Theoretical framework

In previous sections, we have underscored the pivotal role of E-AI in advancing toward AGI. Shifting focus, we now delve into the essential components that, we believe, will comprise E-AIs. We draw heavily on the concept of *cognitive architectures* designed by cognitive scientists aiming to model the human mind (Thagard, 2012). Despite the promise these architectures hold for enhancing modern machine learning methods, progress on this has been notably limited (Kotseruba & Tsotsos, 2020). The slow advancement is largely due to cognitive architectures being the domain of neuroscientists and cognitive scientists, with only a select few within the machine learning community exploring their potential for AGI. We advocate for a synergistic strategy that marries cognitive architectures with machine learning within the E-AI paradigm, proposing it as a viable path toward AGI. The emergence of agent-based LLMs, such as AutoGPT (FIRAT & Kuleli, 2023), which pioneers the generation of autonomous agents, and PanguAgent (Christianos et al., 2023), an agent-focused language model, indicate the potential of this approach.

We identify four essential components of an E-AI system: (i) perception: the ability of the agent to sense its environment; (ii) action: the ability to interact with and change its environment; (iii) memory: the capacity to retain past experiences; and (iv) learning: integrating experiences to form new knowledge and abilities. These components are notably aligned with the active inference framework of Friston (2010). In this framework, the agent models its world through a probabilistic generative model that infers the causes of its sensory observations (perception). This model is hierarchical, forecasting future states in a top-down manner and reconciling these predictions with bottom-up sensory data, with discrepancies or errors being escalated upwards only when they cannot be reconciled at the initial level. The agent acts to minimize the divergence between its anticipations and reality, thus moving towards states of reduced uncertainty (action). Concurrently, it collects and stores new information about its environment (memory) and refines its internal model to minimize predictive errors (learning). In the sections that follow, we will describe in detail these four components and how they comprise the E-AI agent.

4.1. Perception

At the heart of an embodied agent lies the ability to perceive the world in which it exists. Perception is a process by which raw sensory data is transformed into a structured internal representation, enabling the agent to engage in cognitive tasks. The range of inputs that inform perception is vast, encompassing familiar human senses such as vision, hearing, smell, touch, and taste. It extends to any form of stimuli an agent might encounter, be it force sensors in robotics or signal strength indicators in wireless technology. The challenge with sensory data is that it is often not immediately actionable. It typically undergoes a process of transformation, a task where recent advances in machine learning can prove invaluable. The field has seen the development of sophisticated methods for learning feature and embedding spaces, facilitating the conversion of raw data into meaningful information (Golinko & Zhu, 2019; Sivaraman et al., 2022). A particularly effective strategy has been self-supervised learning to learn such representations. Although much of the research has concentrated on single modalities, such as vision (Oquab et al., 2023), the principles underlying these techniques are universally applicable across different sensory inputs (Orhan et al., 2022; Lee et al., 2019).

4.2. Action

Embodied agents navigate the world by taking actions and observing the outcomes. Acting can be broken down into two steps: (i) choosing *what* action to undertake next, like deciding to relocate to a specific spot, and (ii) determining *how* to execute this action, such as plotting the course to that location. Actions can further be categorized into *reactive* and *goal-directed* types. Reactive actions, akin to human reflexes, occur almost instantaneously in response to stimuli and play a crucial role in an agent's immediate self-preservation by maintaining stability. Goal-directed actions, on the other hand, involve strategic planning and are motivated by high-level objectives. Reactive actions are important for self-preservation, with model-free reinforcement learning methods playing an important role for developing reactive control policies in tasks like robot walking (Rudin et al., 2022). On the other hand, for an agent to achieve more complex, high-level objectives, planning is indispensable, even if efficient planning remains an open area of research (Lin et al., 2022; Shi et al., 2022). Central to the concept of planning is the presence of a "world model" within the agent, which it can use to predict the consequences of its own actions. Model-based RL has made significant strides in developing algorithms that learn these world models and use them for planning (Silver et al., 2016; Kégl et al., 2021; Paolo et al., 2022).

4.3. Memory

Embodied agents learn from their experience, which are stored in memory. Memory encompasses various dimensions, including its duration (short-term or long-term) and its nature (procedural, declarative, semantic, and episodic). Importantly, memory is not necessarily represented as explicit propositional knowledge; it can be implicitly encoded into the weights of a neural network (NN). To navigate cognitive tasks, agents require diverse types of memory systems, each playing a distinct role. Working and short-term memory offer temporary storage to support the agent's immediate objectives. Long-term and episodic memories provide a reservoir for information over longer time. Episodic memory captures and stores unique, perspectival experiences, ready to be accessed when familiar scenarios unfold. Long-term memory, conversely, is the repository for broader propositional knowledge. LLMs, for example, implement long-term memory using Retrieval-Augmented Generation (RAG) (Gao et al., 2024), a technique that reduces hallucinations using an external database. This technique showcases how sophisticated machine learning methods can be synergized with cognitive architectures.

4.4. Learning

A defining trait of intelligent agents is their ability to learn. Yet, how to learn, especially in a continuous and dynamic way, remains a subject of ongoing research and debate (Wang et al., 2023a; Yifan et al., 2023). While recent strides in AI have largely been powered by training on static datasets, the concept of continual learning, essential for adapting over time, faces challenges. These challenges stem primarily from the inherent limitations of deep NNs, such as catastrophic forgetting (Kemker et al., 2018), and the complexities associated with learning from non-stationary data that result from an agent's interaction with its environment (Fahrbach et al., 2023). The embodiment hypothesis suggests that true intelligence is born from such interactions (Smith & Gasser, 2005), underscoring the need for dynamic learning methodologies. In this context, simulators emerge as a vital tool, offering a shift away from the static learning typical of traditional AI. Instead, they enable agents to evolve through ongoing, interactive experiences within simulated environments (Duan et al., 2022).

5. Challenges

E-AIs agents will adopt an egocentric perspective, experiencing their environment from a first-person viewpoint, in contrast to the allocentric perspective prevalent in current AI systems. This shift is not only essential for meaningful interaction with the world but also offers an advantage by allowing the agents to focus on modeling their immediate surroundings rather than the entirety of the world. On the other hand, E-AIs introduces several challenges, includ-

ing extending current learning theories, managing noise in perception and action effectively and safely, and ensuring meaningful communication with humans that adheres to ethical standards. The remainder of this section will cover these challenges, exploring potential pathways and solutions.

5.1. New learning theory

The principles of E-AI challenge us to reevaluate traditional learning theories (Devroye et al., 1996; Vapnik, 1998), bridging a gap between supervised and reinforcement learning. Supervised learning, while foundational in AI, assumes that the data is drawn from an unknown but fixed distribution, collected independently of the learning process. This theory gives rise to the classical notions of generalization, over- and underfitting, bias and variance, and asymptotic or finite-sample statistical consistency. This framing is obviously highly useful: even those who are not explicitly doing theory use it transparently as their lingua technica and cognitive scaffolding when working with algorithms and analyzing results.

When embodied agents interact dynamically with their environment, data collection becomes part of the data science pipeline (Pfeifer & Iida, 2004; Thrun et al., 2005). Classical supervised learning theory is insufficient to analyze these cases and to guide algorithm building. Extensions, like transfer learning (Pan & Yang, 2010), multitask learning (Caruana, 1997), distribution shift (Quiñonero-Candela et al., 2009), domain adaptation (Csurka, 2017) or out-of-distribution generalization, have been proposed to patch basic supervised learning theory, but most of these cling to the original framing, pretending that the data is coming from outside the learning process, encapsulating the value (business or otherwise) of the predictive pipeline. Practically, this is obviously not the case: the data on which we learn a predictor is often collected by the data scientist, responsible for the quality of the pipeline (O'Neil & Schutt, 2013; Provost & Fawcett, 2013). Furthermore, most of the debates around responsible AI turn around the data, not the learning algorithm (O'Neil, 2016; Selbst et al., 2019). Collecting, selecting, and curating data is obviously part of the pipeline. The text we use to train LLMs is created by its writers, rather than drawn from a distribution. In some cases, when collection and model-retraining are automated, the situation may be even worse. For example, in click-through-rate prediction (Bottou et al., 2013; Perlich et al., 2014) or recommendation systems (Deldjoo et al., 2020), the deployed predictor affects the data for the next round of training, generating an often adversarial feedback. A similar phenomenon is happening in the LLM world: as these AIs become the go-to tools for creative and business writing, the data collected for the next round of training will, in large part, be coming from the previous generation of LLMs.

Reinforcement Learning (Sutton & Barto, 2018) and related paradigms (Bayesian optimization (Mockus, 1989) or contextual bandits (Langford & Zhang, 2008)) offer a closer fit for embodied AI, when the prediction is not the end-product, rather part of a predictive pipeline that also includes data collection. RL affords the data scientist to design a higher-level objective, letting the algorithm optimize both the predictor and the data it is trained on. Here, the mismatch between theory and practice is different from supervised learning. The analysis in RL or bandit theory often focuses on the convergence of the agent to a theoretical optimum, given a fixed but often unknown environment. RL theory usually does not offer tools to analyze the data collected during the learning process, especially when the collection is semi-automatic (includes a human curator in the loop). RL agents, in practice, usually do not converge even in a stationary environment, they rather individuate, making, for example, quite perversely, the random seed part of the algorithm (Henderson et al., 2018). This is even more pronounced in non-stationary environments where the agent's actions alter the environment; a situation which AGI will definitely find itself (da Silva et al., 2006; Zhou et al., 2024).

A new learning theory for embodied AI must transcend these limitations. It should account for the dynamic, interactive nature of data in E-AI, where the agent's actions continuously reshape its learning environment. This theory should not just aim for optimal performance in a fixed setting but should embrace a spectrum of behaviors suitable for evolving environments. Moreover, it should provide diagnostics to assess the quality and relevance of data generated through these interactions.

5.2. Noise and uncertainty

E-AI agents are tasked with navigating the real world, rife with noise and uncertainty. These elements can drastically affect both the agent's perception of its surroundings and the quality of its decision-making. For example, elevated noise levels may distort the agent's interpretation of environmental cues, leading to suboptimal decisions. This challenge is accentuated in an egocentric perspective, where agents frequently encounter continuous streams of fluctuating and imprecise data. Sources of noise include the natural imprecision of sensors and actuators, which might lack accuracy due to manufacturing inconsistencies, degradation over time, or external disturbances. Additionally, quantization error, a byproduct of converting analog signals into digital form (Widrow & Kollár, 2008), can further compromise data integrity.

As these agents learn and adapt to their environment, they must also grapple with uncertainty. This uncertainty can obscure the agent's understanding of its environment, influencing its performance. This dilemma is especially prevalent in RL scenarios dealing with partial observability, where deci-

sions must be made with incomplete information, leading to uncertainty in predicting the outcomes of its actions (Dulac-Arnold et al., 2021; Hess et al., 2023; Pattanaik et al., 2017). Therefore, managing noise and uncertainty effectively is paramount for the progress of E-AI.

5.3. Simulators

As we pivot towards E-AI, simulators will assume a fundamental role as a key driver of progress, similar to the role data sets play in the training of traditional I-AI models. These simulators offer a controlled, replicable environment where AI systems can be rigorously trained and tested. This setup allows for learning and adapting to diverse scenarios prior to deployment, ensuring both safety and cost-efficiency. A notable advantage of simulators, and requirements, is their speed and ease of parallelization, significantly accelerating training time, making it more feasible to train sophisticated AI models on multiple scenarios simultaneously.

Many advanced simulators have been introduced recently, yet they often demand significant computational resources and are predominantly geared towards robotics applications (Li et al., 2021; Gan et al., 2020; Yan et al., 2018; Puig et al., 2018; Gao et al., 2019). For these simulators to truly serve the needs of E-AI, they must expand their scope to a broader spectrum of environments. A major challenge in the use of simulators is bridging the "reality gap" (Bousmalis & Levine, 2017): the difference between simulated conditions and the agent's eventual real-world or virtual deployment context (Ligot & Birattari, 2020). This gap can lead to a situation where models that excel in simulations fail in actual application, undermining the effectiveness of the training process. Despite numerous strategies being put forward to mitigate the reality gap (Salvato et al., 2021; Daza et al., 2023; Daoudi et al., 2023; Koos et al., 2012; Tobin et al., 2017), it remains an unresolved issue in the field, challenging the applicability of simulated training environments.

5.4. Interaction with humans

A key ambition of E-AI is to seamlessly interact with and learn from humans, enhancing AI's ability to offer personalized and impactful solutions. By improving these interactions, E-AI will also diminish fear and mistrust towards AI technologies, leading to broader acceptance and integration. In this endeavor, LLMs stand out as particularly beneficial, with their ability to comprehend and produce human-like text, facilitating communication in natural language and making engagements with AI more natural and accessible. The domain of Human-Robot Interaction (HRI) offers valuable lessons for enhancing AI-human communication, as researchers in this domain have dedicated efforts to explore innovative methods for robots to better communicate with us (Amirova et al., 2021; Bonarini, 2020). Yet, the challenge of ensuring proper and ethical communication with AI systems persists. The effectiveness of LLMs, for instance, hinges significantly on their training and how well they are aligned with human intentions and values (Wang et al., 2023b). Integrating human oversight directly into the AI development process and establishing comprehensive guidelines and protocols for AI communication are among the proposed strategies to address these challenges, aiming to make AI interactions more meaningful and ethically sound.

5.5. Generalization

An important issue in AI is generalization. There have been many attempts at developing systems capable of quickly generalizing to settings unseen at training time (Pourpanah et al., 2022) in the same fashion living beings do. Nonetheless this is still an open problem that will likely afflict embodied AIs as well, as acting in the real world exposes the agent to situations unseen at training time. For instance, consider a service robot trained in a simulated environment. When placed in a real household, it may encounter novel objects and behaviors not present in its training data, leading to suboptimal or even erroneous actions. This illustrates the critical need for AIs that can adapt and generalize beyond their initial programming. A promising direction in addressing this problem is the leveraging of the enormous amount of internet data. LLMs have demonstrated remarkable zero-shot learning capabilities with minimal fine-tuning (Wei et al., 2021). We can envision that some form of pretraining on internet datasets can kick-start the AI before its embodied phase, enhancing generalization and adaptability.

Recent developments in robotics have started exploring this research direction. Ahn et al. (2024) used a mixed approach between I-AI and E-AI to effectively control multiple robots in different settings. However, only relying on internet data is insufficient. An important aspect is also the ability to accurately identify unknown situations and avoid overconfidence, a common shortcoming of LLMs, that often produce plausible-sounding response that are factually incorrect (Xiong et al., 2024). The ability to assess its own uncertainty is essential, and can prompt the AI to seek human assistance, similarly to how infants ask for help in their early development. We believe that, while the integration of I-AI and E-AI will prove necessary as foundation for the development of the next generation of intelligent systems, the active learning paradigm and precise uncertainty estimation are vital. Active learning, where the AI actively queries for information when uncertain, combined with reliable uncertainty estimation, can enable an E-AI agent to manage novel situations effectively.

Finally, we believe that to properly address the issue of generalization, the community must first clearly define what is the meaning of "generalization". Currently, discussions around this issue often rely on vague terms, referring to an

agent's ability to adapt to unseen settings or data. However, without a formal definition, it is challenging to assess or improve generalization effectively.

Consider the varying degrees of generalization required in different scenarios: transferring skills from driving a car to driving a bus represents generalization within a similar domain, whereas adapting from walking to swimming involves a more profound shift in the type of task. These examples illustrate the spectrum of generalization challenges that embodied AI might face.

To advance this field, it is imperative to develop a precise definition of generalization and establish standardized metrics and benchmarks for measuring an AI's generalization capabilities (Kawaguchi et al., 2017). This necessity ties back to our discussion in Section 5.1, highlighting the urgent need for a new learning theory that can provide a principled approach to developing agents that generalize well. Addressing these questions will hopefully lead to more precise and principled approaches in the development of the field.

5.6. Hardware limitations

A significant challenge to the broad-scale development and integration of E-AI lies in the hardware requirements of these AI systems. Presently, AI technologies largely depend on GPU clusters, which are, while powerful, not ideally suited for embodied agents due to their high cost, energy consumption, and extensive heat output. Additionally, the physical bulk and heft of GPUs pose logistical challenges for mobile agents or those operating within spatial limitations. Addressing these constraints necessitates the innovation of new, energy-efficient hardware solutions that can be embedded within the agents. Promising developments are on the horizon, with Google's Tensor Processing Unit (TPU) (Norrie et al., 2021; Cass, 2019) and Huawei's Ascend chip (Liao et al., 2021) leading the charge. These advancements, coupled with the potential of neuromorphic computing and the strategic synergy of hardware-software co-design, signal a new era of hardware capability. Moreover, the development of energy and data-efficient algorithms is critical. Such breakthroughs in hardware and algorithm efficiency will have a direct and profound effect on an AI's ability to understand, decide, and interact within its environment, enabling E-AI agents to operate more autonomously and effectively in a diverse array of settings.

6. Conclusion

In this paper, we have articulated the critical role Embodied AI plays on the path toward achieving AGI, setting it apart from prevailing AI methodologies, notably LLMs. By integrating insights from a spectrum of research fields, we underscored how E-AI's development benefits from existing knowledge, with LLMs enhancing the potential for intuitive interactions between humans and emerging AI entities. We introduced a comprehensive theoretical framework for the development of E-AI, grounded in the principles of cognitive science, highlighting perception, action, memory, and learning, situating E-AI within the context of Friston's active inference framework, thereby offering a wide-ranging theoretical backdrop for our discussion. Despite the outlook, the journey ahead is fraught with challenges, not least the formulation of a novel learning theory tailored for AI and the creation of sophisticated hardware solutions. This paper aims to serve as a roadmap for ongoing and future research into E-AI, proposing directions that could lead to significant advancements in the field.

Impact Statement

While the development of Embodied AI introduces complexities and challenges, particularly in hardware requirements, ethical considerations, and safety protocols, the potential benefits significantly outweigh these drawbacks. E-AI stands to evolve our interaction with technology, imbuing AI with a deeper understanding of and engagement with both the physical world and human society. This not only paves the way for more natural and effective human-AI interactions but also enhances AI's adaptability and application across a broad spectrum of fields.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ahn, M., Dwibedi, D., Finn, C., Arenas, M. G., Gopalakrishnan, K., Hausman, K., Ichter, B., Irpan, A., Joshi, N., Julian, R., et al. Autort: Embodied foundation models for large scale orchestration of robotic agents. *arXiv preprint arXiv:2401.12963*, 2024.

Amirova, A., Rakhymbayeva, N., Yadollahi, E., Sandygulova, A., and Johal, W. 10 years of human-nao interaction research: A scoping review. *Frontiers in Robotics and AI*, 8:744526, 2021. URL https://www.frontiersin.org/articles/10.3389/frobt.2021.744526/full.

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., and Van Den Hengel, A. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3674–3683, 2018. URL https://ieeexplore.ieee.org/document/8578485.

Bakshy, E., Messing, S., and Adamic, L. A. Exposure to ideologically diverse news and opinion on facebook. In *Proceedings of the National Academy of Sciences*, volume 112, pp. 5791–5796. National Acad Sciences, 2015. URL https://www.science.org/doi/10.1126/science.aaa1160.

Balayn, A., Lofi, C., and Houben, G.-J. Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, 30(5):739–768, 2021. URL https://link.springer.com/article/10.1007/s00778-021-00671-8.

Bargmann, C. I. Comparative chemosensation from receptors to ecology. *Nature*, 444(7117):295–301, 2006. URL https://www.nature.com/articles/nature05402.

Bariah, L. and Debbah, M. Ai embodiment through 6g: Shaping the future of agi. 2023.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. ACM, 2021. URL https://dl.acm.org/doi/10.1145/3442188.3445922.

Bodó, B. Selling news to audiences–a qualitative inquiry into the emerging logics of algorithmic news personalization in european quality news media. In *Algorithms, Automation, and News*, pp. 75–96. Routledge, 2021. URL https://www.tandfonline.com/doi/full/10.1080/21670811.2019.1624185.

Bonarini, A. Communication in human-robot interaction. *Current Robotics Reports*, 1:279–285, 2020. URL https://link.springer.com/article/10.1007/s43154-020-00026-1.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. In *Journal of Machine Learning Research*, volume 14, pp. 3207–3260, 2013. URL https://jmlr.org/papers/v14/bottou13a.html.

Bousmalis, K. and Levine, S. Closing the simulation-to-reality gap for deep robotic learning. *Google Research Blog*, 1, 2017. URL https://blog.research.google/2017/10/closing-simulation-to-reality-gap-for.html.

Bozdag, E. Bias in algorithmic filtering and personalization. *Ethics and information technology*, 15:209–227, 2013. URL https://link.springer.com/article/10.1007/s10676-013-9321-6.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M. G., Gopalakrishnan, K., Han, K., Hausman, K., Herzog, A., Hsu, J., Ichter, B., Irpan, A., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, L., Lee, T.-W. E., Levine, S., Lu, Y., Michalewski, H., Mordatch, I., Pertsch, K., Rao, K., Reymann, K., Ryoo, M., Salazar, G., Sanketi, P., Sermanet, P., Singh, J., Singh, A., Soricut, R., Tran, H., Vanhoucke, V., Vuong, Q., Wahid, A., Welker, S., Wohlhart, P., Wu, J., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., and Zitkovich, B. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*, 2023. URL https://arxiv.org/abs/2307.15818.

Brooks, R. A. Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159, 1991. URL https://www.sciencedirect.com/science/article/abs/pii/000437029190053M.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020. URL https://arxiv.org/abs/2005.14165.

Caruana, R. Multitask learning. *Machine Learning*, 28 (1):41–75, 1997. URL https://link.springer.com/article/10.1023/A:1007379606734.

Cass, S. Taking ai to the edge: Google's tpu now comes in a maker-friendly package. *IEEE Spectrum*, 56(5): 16–17, 2019. URL https://ieeexplore.ieee.org/document/8701189.

Chen, T., Gupta, S., and Gupta, A. Learning exploration policies for navigation. *arXiv preprint arXiv:1903.01959*, 2019. URL https://arxiv.org/abs/1903.01959.

Christianos, F., Papoudakis, G., Zimmer, M., Coste, T., Wu, Z., Chen, J., Khandelwal, K., Doran, J., Feng, X., Liu, J., et al. Pangu-agent: A fine-tunable generalist agent with structured reasoning. *arXiv preprint arXiv:2312.14878*, 2023. URL https://arxiv.org/2312.14878.

Clark, A. *Being There: Putting Brain, Body, and World Together Again*. MIT Press, 1997. URL https://mitpress.mit.edu/9780262531566/being-there/.

Clark, A. and Chalmers, D. The extended mind. *Analysis*, 58(1):7–19, 1998. URL https://era.ed.ac.uk/bitstream/handle/1842/1312/TheExtendedMind.pdf?sequence=1&isAllowed=y.

Covington, P., Adams, J., and Sargin, E. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 191–198, 2016. URL https://dl.acm.org/doi/10.1145/2959100.2959190.

Cramer, P. Alphafold2 and the future of structural biology. *Nature structural & molecular biology*, 28(9): 704–705, 2021. URL https://www.nature.com/articles/s41594-021-00650-1.

Csurka, G. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017. URL https://link.springer.com/chapter/10.1007/978-3-319-58347-1_1.

da Silva, B. C., Basso, E. W., Bazzan, A. L. C., and Engel, P. M. Dealing with non-stationary environments using context detection. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 217–224, 2006. URL https://dl.acm.org/doi/10.1145/1143844.1143872.

Daoudi, P., Prieur, C., Robu, B., Barlier, M., and Santos, L. D. A trust region approach for few-shot sim-to-real reinforcement learning, 2023. URL https://arxiv.org/abs/2312.15474.

Daza, I. G., Izquierdo, R., Martínez, L. M., Benderius, O., and Llorca, D. F. Sim-to-real transfer and reality gap modeling in model predictive control for autonomous driving. *Applied Intelligence*, 53(10):12719–12735, 2023. URL https://link.springer.com/article/10.1007/s10489-022-04148-1.

Deldjoo, Y., Di Noia, T., and Merra, F. A. Adversarial machine learning in recommender systems (aml-recsys). In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 869–872, 2020. URL https://dl.acm.org/doi/abs/10.1145/3336191.3371877.

Descartes, R. *Discourse on method*. Hackett Publishing, 2012. URL https://hackettpublishing.com/discourse-on-method.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Devroye, L., Györfi, L., and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996. URL https://link.springer.com/book/10.1007/978-1-4612-0711-5.

Duan, J., Yu, S., Tan, H. L., Zhu, H., and Tan, C. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022. URL https://arxiv.org/abs/2103.04918.

Dulac-Arnold, G., Levine, N., Mankowitz, D. J., Li, J., Paduraru, C., Gowal, S., and Hester, T. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, 2021. URL https://link.springer.com/article/10.1007/s10994-021-05961-4.

Eirinaki, M., Gao, J., Varlamis, I., and Tserpes, K. Recommender systems for large-scale social networks: A review of challenges and solutions. *Future Generation Computer Systems*, 78:413–418, 2018. URL https://www.sciencedirect.com/science/article/pii/S0167739X17319684.

Fahrbach, M., Javanmard, A., Mirrokni, V., and Worah, P. Learning rate schedules in the presence of distribution shift. *arXiv preprint arXiv:2303.15634*, 2023. URL https://arxiv.org/abs/2303.15634.

Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., Lu, H., Song, R., Gao, X., Xiang, T., et al. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):3094, 2022. URL https://arxiv.org/abs/2110.14378.

FIRAT, M. and Kuleli, S. What if gpt4 became autonomous: The auto-gpt project and use cases. *Journal of Emerging Computer Technologies*, 3(1):1–6, 2023. URL https://github.com/Significant-Gravitas/AutoGPT.

Friston, K. The free energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. URL https://www.nature.com/articles/nrn2787.

Friston, K., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G. A., and Parr, T. The free energy principle made simpler but not too simple. *Physics Reports*, 1024:1–29, 2023. URL https://arxiv.org/abs/2201.06387.

Gan, C., Schwartz, J., Alter, S., Mrowca, D., Schrimpf, M., Traer, J., De Freitas, J., Kubilius, J., Bhandwaldar, A., Haber, N., et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv*

preprint arXiv:2007.04954, 2020. URL https://arxiv.org/abs/2007.04954.

Gao, X., Gong, R., Shu, T., Xie, X., Wang, S., and Zhu, S.-C. Vrkitchen: an interactive 3d virtual environment for task-oriented learning. *arXiv preprint arXiv:1903.05757*, 2019. URL https://arxiv.org/abs/1903.05757.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., and Wang, H. Retrieval-augmented generation for large language models: A survey, 2024. URL https://arxiv.org/abs/2312.10997.

Gibson, J. J. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

Golinko, E. and Zhu, X. Generalized feature embedding for supervised, unsupervised, and online learning tasks. *Information Systems Frontiers*, 21:125–142, 2019. URL https://link.springer.com/article/10.1007/s10796-018-9850-y.

Held, R. and Hein, A. Movement-produced stimulation in the development of visually guided behavior. *Journal of comparative and physiological psychology*, 56 (5):872, 1963. URL https://psycnet.apa.org/record/1964-03855-001.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. URL https://arxiv.org/abs/1709.06560.

Hess, F., Monfared, Z., Brenner, M., and Durstewitz, D. Generalized Teacher Forcing for Learning Chaotic Dynamics, October 2023. URL http://arxiv.org/abs/2306.04406. arXiv:2306.04406 [nlin].

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023. URL https://arxiv.org/abs/2311.05232.

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., and Hardt, M. Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences*, 119(1):e2025334119, 2022. doi: 10.1073/pnas.2025334119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2025334119.

Ishiguro, A. and Kawakatsu, T. How should control and body systems be coupled? a robotic case study. In *Embodied Artificial Intelligence: International Seminar, Dagstuhl Castle, Germany, July 7-11, 2003. Revised Papers*, pp. 107–118. Springer, 2004. URL https://link.springer.com/chapter/10.1007/978-3-540-27833-7_8.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023. URL https://arxiv.org/abs/2310.19852.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL https://arxiv.org/abs/2001.08361.

Kawaguchi, K., Kaelbling, L. P., and Bengio, Y. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 1(8), 2017.

Kégl, B., Hurtado, G., and Thomas, A. Model-based micro-data reinforcement learning: what are the crucial model properties and which model to choose? *arXiv preprint arXiv:2107.11587*, 2021. URL https://arxiv.org/2107.11587.

Kemker, R., McClure, M., Abitino, A., Hayes, T., and Kanan, C. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. URL https://arxiv.org/abs/1708.02072.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., and Kiverstein, J. The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The Royal Society interface*, 15(138):20170792, 2018. URL https://royalsocietypublishing.org/doi/10.1098/rsif.2017.0792.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. URL https://arxiv.org/abs/1612.00796.

Koos, S., Mouret, J.-B., and Doncieux, S. The transferability approach: Crossing the reality gap in evolutionary robotics. *IEEE Transactions on Evolutionary Computation*, 17(1):122–145, 2012. URL https://ieeexplore.ieee.org/document/6151107.

Korth, M. The purpose of qualia: What if human thinking is not (only) information processing? *arXiv preprint arXiv:2212.00800*, 2022. URL https://arxiv.org/abs/2212.00800.

Kotseruba, I. and Tsotsos, J. K. *40 Years of Cognitive Architectures: Core Cognitive Abilities and Practical Applications*, volume 53. Springer Netherlands, 2020. ISBN 9550141039. doi: 10.1007/s10462-018-9646-y. URL https://doi.org/10.1007/s10462-018-9646-y.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017. URL https://pubmed.ncbi.nlm.nih.gov/27881212/.

Lakoff, G. and Johnson, M. *Metaphors we live by*. University of Chicago press, 1979. URL https://press.uchicago.edu/ucp/books/book/chicago/M/bo3637992.html.

Lakoff, G. and Johnson, M. L. Philosophy in the flesh : the embodied mind and its challenge to western thought. 1999. URL https://api.semanticscholar.org/CorpusID:16103621.

Lambert, N., Castricato, L., von Werra, L., and Havrilla, A. Illustrating reinforcement learning from human feedback (rlhf). *Hugging Face Blog*, 2022. URL https://huggingface.co/blog/rlhf.

Langford, J. and Zhang, T. The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems*, volume 20, 2008. URL https://proceedings.neurips.cc/paper/2007/file/4b04a686b0ad13dce35fa99fa4161c65-Paper.pdf.

Lee, M. A., Zhu, Y., Srinivasan, K., Shah, P., Savarese, S., Fei-Fei, L., Garg, A., and Bohg, J. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8943–8950. IEEE, 2019. URL https://arxiv.org/1810.10191v2.

Levin, M. Technological approach to mind everywhere: an experimentally-grounded framework for understanding diverse bodies and minds. *Frontiers in systems neuroscience*, 16:768201, 2022. URL https://www.frontiersin.org/articles/10.3389/fnsys.2022.768201.

Li, C., Xia, F., Martín-Martín, R., Lingelbach, M., Srivastava, S., Shen, B., Vainio, K., Gokmen, C., Dharan, G., Jain, T., et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*, 2021. URL https://arxiv.org/abs/2108.03272.

Li, S. and Deng, W. A deeper look at facial expression dataset bias. *IEEE Transactions on Affective Computing*, 13(2):881–893, 2020. URL https://arxiv.org/1904.11150.

Liao, H., Tu, J., Xia, J., Liu, H., Zhou, X., Yuan, H., and Hu, Y. Ascend: a scalable and unified architecture for ubiquitous deep neural network computing: Industry track paper. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pp. 789–801. IEEE, 2021. URL https://ieeexplore.ieee.org/document/9407221.

Ligot, A. and Birattari, M. Simulation-only experiments to mimic the effects of the reality gap in the automatic design of robot swarms. *Swarm Intelligence*, 14(1):1–24, 2020. URL https://link.springer.com/article/10.1007/s11721-019-00175-w.

Lin, H., Sun, Y., Zhang, J., and Yu, Y. Model-based reinforcement learning with multi-step plan value estimation. *arXiv preprint arXiv:2209.05530*, 2022. URL https://arxiv.org/abs/2209.05530.

Locke, J. *An essay concerning human understanding*. Kay & Troutman, 1847. URL https://www.gutenberg.org/files/10615/10615-h/10615-h.htm.

McNearney, S. A Brief Guide to Embodied Cognition: Why You Are Not Your Brain, 2011.

Mockus, J. *Bayesian Approach to Global Optimization: Theory and Applications*. Kluwer Academic Publishers, 1989. URL https://link.springer.com/book/10.1007/978-94-009-0909-0.

Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., and Konstan, J. A. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, pp. 677–686, 2014. URL https://dl.acm.org/doi/10.1145/2566486.2568012.

Norrie, T., Patil, N., Yoon, D. H., Kurian, G., Li, S., Laudon, J., Young, C., Jouppi, N., and Patterson, D. The design process for google's training chips: Tpuv2 and tpuv3. *IEEE Micro*, 41(2):56–63, 2021. URL https://ieeexplore.ieee.org/document/9351692.

O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, 2016.

O'Neil, C. and Schutt, R. *Doing Data Science*. O'Reilly Media, Inc., 2013.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D.,

Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision, 2023.

Orhan, P., Boubenec, Y., and King, J.-R. Don't stop the training: continuously-updating self-supervised algorithms best account for auditory responses in the cortex. *arXiv preprint arXiv:2202.07290*, 2022. URL https://arxiv.org/abs/2202.07290.

Oudeyer, P.-Y. and Kaplan, F. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 2007. URL https://www.frontiersin.org/articles/10.3389/neuro.12.006.2007.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359, 2010.

Paolo, G., Gonzalez-Billandon, J., Thomas, A., and Kégl, B. Guided safe shooting: model based reinforcement learning with safety constraints. *arXiv preprint arXiv:2206.09743*, 2022. URL https://arxiv.org/2206.09743.

Parcalabescu, L., Trost, N., and Frank, A. What is multimodality? *arXiv preprint arXiv:2103.06304*, 2021. URL https://arxiv.org/abs/2103.06304.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017. URL https://arxiv.org/1705.05363.

Pattanaik, A., Tang, Z., Liu, S., Bommannan, G., and Chowdhary, G. Robust Deep Reinforcement Learning with Adversarial Attacks, December 2017. URL http://arxiv.org/abs/1712.03632. arXiv:1712.03632 [cs].

Perez, C. *Artificial Empathy - A Roadmap for Human Aligned Artificial General Intelligence*. Publisher, 2023.

Perlich, C., Dalessandro, B., Raeder, T., Stitelman, O., and Provost, F. Machine learning for targeted display advertising: transfer learning in action. In *Machine learning*, volume 95, pp. 103–127. Springer, 2014.

Pfeifer, R. and Bongard, J. *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press, 2006.

Pfeifer, R. and Iida, F. Embodied artificial intelligence: Trends and challenges. *Lecture notes in computer science*, pp. 1–26, 2004. URL https://link.springer.com/chapter/10.1007/978-3-540-27833-7_1.

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., Wang, X.-Z., and Wu, Q. J. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4051–4070, 2022.

Provost, F. and Fawcett, T. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media, Inc., 2013.

Puig, X., Ra, K., Boben, M., Li, J., Wang, T., Fidler, S., and Torralba, A. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8494–8502, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Puig_VirtualHome_Simulating_Household_CVPR_2018_paper.html.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (eds.). *Dataset Shift in Machine Learning*. The MIT Press, 2009.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Ramakrishnan, S. K., Jayaraman, D., and Grauman, K. An exploration of embodied visual exploration. *International Journal of Computer Vision*, 129:1616–1649, 2021. URL https://arxiv.org/abs/2001.02192.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Reynolds, G. D. and Roth, K. C. The development of attentional biases for faces in infancy: A developmental systems perspective. *Frontiers in psychology*, 9:222,

2018. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00222.

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A., and Meira Jr, W. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 131–141, 2020.

Roli, A., Jaeger, J., and Kauffman, S. A. How organisms come to know the world: fundamental limits on artificial general intelligence. *Frontiers in Ecology and Evolution*, 9:1035, 2022. URL https://www.frontiersin.org/articles/10.3389/fevo.2021.806283.

Rosas, F. E., Mediano, P. A., Jensen, H. J., Seth, A. K., Barrett, A. B., Carhart-Harris, R. L., and Bor, D. Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLoS computational biology*, 16(12): e1008289, 2020. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008289.

Rudin, N., Hoeller, D., Reist, P., and Hutter, M. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Conference on Robot Learning*, pp. 91–100. PMLR, 2022. URL https://arxiv.org/abs/2109.11978.

Russell, S. Human-compatible artificial intelligence. *Human-like machine intelligence*, pp. 3–23, 2021. URL http://aima.cs.berkeley.edu/~russell/papers/mi19book-hcai.pdf.

Salvato, E., Fenu, G., Medvet, E., and Pellegrino, F. A. Crossing the reality gap: A survey on sim-to-real transferability of robot controllers in reinforcement learning. *IEEE Access*, 9:153171–153187, 2021.

Schüll, N. D. *Addiction by Design: Machine Gambling in Las Vegas*. Princeton University Press, 2012.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. Fairness and abstraction in sociotechnical systems. *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 59–68, 2019.

Shapiro, L. *Embodied Cognition*. Routledge, 2011.

Shapiro, L. G. Computer vision: the last fifty years. *University of Washington, Last access*, 7(05), 2021.

Shen, T., Jin, R., Huang, Y., Liu, C., Dong, W., Guo, Z., Wu, X., Liu, Y., and Xiong, D. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023. URL https://arxiv.org/abs/2309.15025.

Shenavarmasouleh, F., Mohammadi, F. G., Amini, M. H., and Reza Arabnia, H. Embodied ai-driven operation of smart cities: A concise review. *Cyberphysical Smart Cities Infrastructures: Optimal Operation and Intelligent Decision Making*, pp. 29–45, 2022. URL https://arxiv.org/abs/2108.09823.

Shi, L. X., Lim, J. J., and Lee, Y. Skill-based model-based reinforcement learning. *arXiv preprint arXiv:2207.07560*, 2022. URL https://arxiv.org/abs/2207.07560.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Sivaraman, V., Wu, Y., and Perer, A. Emblaze: Illuminating machine learning representations through interactive comparison of embedding spaces. In *27th International Conference on Intelligent User Interfaces*, pp. 418–432, 2022. URL https://arxiv.org/abs/2202.02641.

Smith, L. and Gasser, M. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11 (1-2):13–29, 2005. URL https://pubmed.ncbi.nlm.nih.gov/15811218/.

Solms, M. The hard problem of consciousness and the free energy principle. *Frontiers in Psychology*, 2019.

Strubell, E., Ganesh, A., and McCallum, A. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645–3650, 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT press, 2018.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Thagard, P. Cognitive architectures. *The Cambridge handbook of cognitive science*, 3:50–70, 2012.

Thrun, S., Burgard, W., and Fox, D. *Probabilistic Robotics*. MIT Press, 2005.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Vapnik, V. *Statistical Learning Theory*. Wiley, 1998.

Varela, F. J., Thompson, E., and Rosch, E. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, 1991.

Verma, S., Ernst, M., and Just, R. Removing biased data to improve fairness and accuracy. *arXiv preprint arXiv:2102.03054*, 2021. URL https://arxiv.org/2102.03054.

Vervaeke, J. and Coyne, S. *Mentoring the Machines*. Hackett Publishing, 2024. URL https://www.mentoringthemachines.com.

Vervaeke, J., Lillicrap, T. P., and Richards, B. A. Relevance realization and the emerging framework in cognitive science. *Journal of Logic and Computation*, 22(1):79–99, 2012.

Wang, L., Zhang, X., Su, H., and Zhu, J. A comprehensive survey of continual learning: Theory, method and application. *arXiv preprint arXiv:2302.00487*, 2023a. URL https://arxiv.org/abs/2302.00487.

Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X., and Liu, Q. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*, 2023b.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Westho, B., Koele, I. J., and van de Groep, I. H. Social learning and the brain: How do we learn from and about other people? *Everything You and Your Teachers Need to Know About the Learning Brain*, pp. 42, 2020. URL https://kids.frontiersin.org/articles/10.3389/frym.2020.00095.

Widrow, B. and Kollár, I. *Quantization noise: roundoff error in digital computation, signal processing, control, and communications*. Cambridge University Press, 2008.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huang, X., and Gui, T. The rise and potential of large language model based agents: A survey, 2023. URL https://arxiv.org/abs/2309.07864.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms, 2024.

Yan, C., Misra, D., Bennnett, A., Walsman, A., Bisk, Y., and Artzi, Y. Chalet: Cornell house agent learning environment. *arXiv preprint arXiv:1801.07357*, 2018. URL https://arxiv.org/abs/1801.07357.

Yifan, C., Yulu, C., Yadan, Z., and Wenbo, L. Continual learning in an easy-to-hard manner. *Applied Intelligence*, pp. 1–21, 2023. URL https://link.springer.com/article/10.1007/s10489-023-04454-2.

Zhou, Q., Chen, S., Wang, Y., Xu, H., Du, W., Zhang, H., Du, Y., Tenenbaum, J. B., and Gan, C. HAZARD challenge: Embodied decision making in dynamically changing environments, 2024.