AI & SOCIETY (2025) 40:5087–5099
https://doi.org/10.1007/s00146-024-02086-9

STUDENT PAPERS

Between world models and model worlds: on generality, agency, and worlding in machine learning

Konstantin Mitrokhov[1]

Received: 16 April 2024 / Accepted: 23 September 2024 / Published online: 7 October 2024

Abstract
The article offers a discursive account of what generality in machine learning research means and how it is constructed in the development of general artificial intelligence from the perspectives of cultural and media studies. I discuss several technical papers that outline novel architectures in machine learning and how they conceive of the "world". The agency to learn and the learning curriculum are modulated through worlding (in the sense of setting up and unfolding of the world for artificial agents) in machine learning engineering. In recent computer science articles, large models trained on Internet-scale datasets are framed as general world simulators—despite their partiality, historicity, finite nature, and cultural specificity. I introduce the notion of "model worlds" to refer to composable interactive environments designed for the purpose of machine learning that partake in legitimising that claim. I discuss how large models are grounded through interaction in model worlds, arguing that model worlds mediate between the sheer scale of language models and their hypothetical capacity to generalise to new tasks and domains, rehashing the empiricist logic of "big data". Further, I show that the emerging capacity of artificial agents to generalise redraws the epistemic boundary between artificial agents and their learning environments. Consequently, superficial statistics of language models and abstract action are made meaningful in distilled model worlds, giving rise to synthetic agency.

Keywords AGI · Artificial cognition · Agency · Worlding · Data

1 Introduction: what even is A(G)I?

Critique of artificial intelligence (AI) in the humanities is often focused on the already existing systems that rely on correlations emerging from "big data" made possible by sensing technologies. It has been shown that such recursive systems are extractive, discriminating, and repressive (Chun 2021; Amaro 2022; Beller 2021; McQuillan 2022; Pasquinelli 2023). AI is described as artificial communication systems that do not have their own intelligence (Esposito 2022) and as a technology centred on banal deception (Natale 2021). Grasping what is theoretically understood as artificial *general* intelligence (AGI), on the other hand, is far from straightforward. Perhaps, the infamous pace of computer-scientific research, its commercial uptake, and the corresponding dilution of concepts are to blame. Merely following the field and making sense of new developments could be a challenge. Some researchers already embrace the "rapid ascent" of AGI and prepare frameworks for studying the phenomenon (Machine + Behaviour Conference 2024), while others either find the terminology misleading and vague at best (Heaven 2020) or attempt to pre-emptively define what is and what is not AGI in pursuit of regulatory or discursive agenda (Morris et al. 2023; Wang 2019; Fitzgerald et al. 2020).

If we consider the technological timeline of the AI branch of computer science from the early 2010s and onwards, it appears that each significant technological breakthrough is followed by the rush of products that capitalise on the public interest in technology. Notable examples of research that spawned new fields and enterprises include deep neural networks cast as a "core technology of the today's fourth industrial revolution" (Sarker 2021), deep reinforcement learning and its positing of intelligence as "subserving the maximisation of reward" (Silver et al. 2021), and more recently conversational systems powered by large language models as implementing AI in the sense of "Agency without

✉ Konstantin Mitrokhov
konstantin.mitrokhov@stud.leuphana.de

[1] Institute of Culture and Aesthetics of Digital Media (ICAM), Leuphana University, Lüneburg, Germany

Intelligence" (Floridi 2023). The cycle underlines the complete conflation of machine learning (ML) as the currently dominant technology in the field with the concept of "artificial intelligence" in the public discourse. This, in its turn, is the often-overlooked consequence of the material capture of the field by the parallel computation infrastructures described by Rella (2024). Considering AI as a resource-intensive process situates the theoretical project of AI within the reality of environmentally unsustainable commercial platforms such as ChatGPT (cf. Bender et al. 2021).

Within the field of AI research itself, "intelligence" is mediated through the variety of benchmarks and milestones and co-constituted through the immense corpus of research spanning fields and disciplines such as cognitive psychology, machine learning engineering, cognitive neuroscience, and philosophy of mind (Guerin 2011; Thórisson 2012; Schmidhuber 2015; Clune 2019; Dindo et al. 2013; LeCun 2022). For the sake of clarity and due to the space constraints of this text, I am staying agnostic on disciplinary frameworks. I do not aim to offer a taxonomy of AI and refer to the manifold concepts of AI and AGI collectively as A(G)I, not distinguishing between the so-called weak and strong, objective and subjective, narrow and general AI. At large, I study the technoscientific project of designing a system with human-like cognitive capacity and the ability to communicate with humans. Rather than contributing to the historical lineage of either delimiting or outright dismissing A(G)I as an impossible project detached from the lifeworld (Fjelland 2020), I consider A(G)I as a discursive performance (Cavia 2023, p. 270) and software agents as its actors.

The standard theory of agency and its critique posits that the minimal agency does not require the possession of mental states and it would suffice to consider agency as a self-maintaining coupling between the agent and its environment (Schlosser 2019). According to this definition, attributing agency solely to an individual agent would be problematic. Furthermore, D'Amato (2024) elaborates on the ethical subject that emerges within the disciplinary setting of AI research, while Shanahan (2024) considers whether it makes sense to think of language-based agents in terms of consciousness. Both authors strive to avoid the dualism of thinking along the lines of mind/body or consciousness/conscious experience divide, respectively. In line with their inquiry, I consider intelligence and agency in relational rather than instrumental terms. With this in mind, the software agent's capacity to acts on behalf of the user is continuously expanding and reconfigured by the means of novel algorithms, experimental development toolkits, and publicly available commercial products. These reconfigurations are blurring the epistemic boundaries of A(G)I. What interests me is the conceptualisation, construction, and (re) configuration of aspects of agency and intelligence as they are figured in the corpus of recent ML research.

My method is strongly aligned with the call elaborated by Amoore et al. (2023) for close and careful reading of computer-scientific texts as sites where contemporary algorithmic models are actively giving accounts of their paradigmatic worldview and normative assumptions. In short, to read such technical articles is to engage with the worldmaking that is taking place there. I read across several technical papers that are exemplary of the emerging research in the A(G)I field from the perspectives of cultural and media studies. I understand agency in A(G)I as a relational capacity of an agent to act autonomously—relational in the sense that this capacity is bound up with the agent's environment. My initial premise is that agency is a necessary constituent of artificial cognition that gives rise to its capacity to generalise. This conceptualisation becomes more differentiated through the exploration of how large models interact with game-like environments. I set out to make sense of the permutations of agency, empirical constitution of the capacity to generalise across domains, and conceptual slippages and borrowings between the subfields and disciplines that contribute to the field of A(G)I.

To make sense of the interwoven technical threads, I turn to two paradigms from computer science, namely the world model and what I provisionally refer to as "model world." *World model* is a computational analogy of the human mental model, a concept that denotes an internal representation of external reality and is borrowed from the field of cognitive psychology. World models are internal to A(G)I agents and implemented with artificial neural networks. Conversely, what I call *model worlds* are simulations of the lifeworld domains designed for the purpose of developing, testing, and training A(G)I. They do not only simulate the complexity of the lifeworld, but also provide an interactive curriculum for the learning agents. Model worlds are external to the agent and implemented as composable software toolkits. Despite their common origins in computer science laboratories, these two notions seem quite dissimilar.

Besides their respective interiority and exteriority, world models and model worlds differ in how they come into being. World models are probabilistic models that the agent acquires through ML *in* and *of* its environment. Model worlds, whether they are based on existing games such as Minecraft or custom-made for specific projects, are interactive environments designed by the researchers themselves. Such is my starting point; however, the meaning and function of both have shifted as the transformer architecture came to dominate the A(G)I industry. I pay particular attention to the ways these paradigms legitimate each other and the difference in how they ascribe and reconfigure what counts as a *world* (and worlding) for an agent. Doing so allows me to explore the machinic agency performed in and through the fundamental computer science research.

2 Generality and scale

In an aesthetically pleasing promotional video published by a start-up called Runway AI on their blog in December 2023, a female voice explains what a "general world model" is. The narrator compares it to a mental map of the environment that her dog builds up from daily interactions in an unnamed city. The analogy feels graspable, even enticing, and the multiple assumptions and challenges that come with building a model like this are solidified into an engineering problem.

The central tenet of the video is that it is possible to train an ML model so that it eventually acquires a detailed understanding of the world's dynamics and a robust capacity for generalisation. To achieve that, a large neural network would have to be trained on extensive datasets that combine different modalities of data such as text, images, sounds, and relations. The probabilistic network would then model the (undifferentiated) world and do so in a general enough manner. Such model would allow for simulations that "closely reflect" our world while being able to "imagine the future based on its knowledge of the world" (Germanidis 2023). Two aspects are important here. First, scaling up both the dataset and the agent's cognitive architecture is framed as building up the generality of the model. World simulation is cast as a radical act of generalisation and prediction from prior knowledge. In other words, "imagining" the future in this paradigm would only be feasible by generalising from the past, at scale.

Furthermore, the model's generality is figured as a capacity for robust predictive control that emerges out of the training data. Put differently, the statistical fact of the large dataset's broad distribution is supposed to be translated into the model's capacity to generalise well and "imagine the future". What this marketing vignette makes tangible—except for reiterating the set of core ML beliefs—is the interest in conceptual uptake of the *world model* notion. As I will discuss further in the text, world model becomes a signpost that a large language indeed figures general abstractions of the world. Just as limitations of generative models come to the fore, general applicability becomes a promise that the technology can outgrow its linguistic confines (cf. Searle 1980)—through persistent scaling that captures ever more of the lifeworld.

2.1 The (re)discovery of world models

World models initially did not have anything to do with scale. As a computational paradigm, world models reappeared on the ML research landscape in 2018 with the publication of eponymous paper by computer scientists David Ha and Jürgen Schmidhuber. The article is building up on Schmidhuber's prior work dating back to the 1990s.

In simple terms, world model can be understood as a computational instance of a mental model, a concept borrowed from cognitive psychology. Mental model is a predictive model of the world dynamics, that is, an abstract representation we develop based on our perception. Predictions generated by our internal mental model can be acted on instinctively and without conscious planning (Ha and Schmidhuber 2018, p. 1). In short, mental models are intuitive and limited by the human senses. Artificial agent's world model, it follows, is its internal representation of the past and current states of the world that integrates them into a predictive model of the future. The novelty of Ha and Schmidhuber's research is the proposed implementation that brings together a large artificial neutral network (the world model) with a relatively small and fairly standard reinforcement learning model (the controller, accordingly). The networks are trained separately and successively. The world model network is first trained on the agent's world, in this case mediated through machine vision. Only after the agent learns the world, it starts training the controller to perform tasks in the external environment in coordination with the already trained internal world model.

Moreover, the authors claim that the agent can learn to perform tasks inside of a "dream", i.e. by rehearsing[1] within an internal simulation that is generated by the world model (ibid., p. 6). Schmidhuber and Ha show that the agent's controller can be trained faster and better by simulating interactions with its environment internally than by training the controller directly in the agent's world. The article is accompanied by an experimental framework that demonstrates the key concepts using a simple car racing game and a first-person shooter game. The joint architecture harnesses the world model's "expressiveness" while allowing the controller to learn tasks in a "small search space" instead of a larger and more contingent space of the agent's actual environment (ibid., p. 2). Learning in a "dream", that is, an internal simulation, decouples the agent from the original training environment, allowing it to learn without direct experience of the agent's world. The agent's directive is to grasp the dynamics of the world around and then train itself within that abstraction. This process goes beyond representing the agent's world and towards the recursive exploration of its internal states.

The notion of world model has since been widely referenced in ML literature. The cognitive ability to grasp and integrate notions such as object permanence, parallax, and material dynamics constitutes a basic onto-epistemology

[1] I am using anthropocentric metaphors in this paragraph for the clarity of explanation.

of the world for babies and animals alike. At large, world models are used as a computational correlate of the human capacity to grasp dynamics, predict, rehearse, and imagine. Sometimes referred to as dynamics model, it is shown to be adaptable for A(G)I tasks as diverse as long-term planning (Ke et al. 2019), continual learning (Ketz et al. 2019), playing Atari games at human-like level (Hafner et al. 2020), indoor navigation (Koh et al. 2021), task space exploration (Mendonca et al. 2021), autonomous control (Hao et al. 2021; Kayalibay et al. 2022; Biza et al. 2022), and causal discovery (Zhu et al. 2022).

Since the publication of Ha and Schmidhuber's article, world models became a prominent paradigm in model-based reinforcement learning community, a subfield of ML usually concerned with domains that are constrained and mediated through camera-based perception. With rare exceptions, researchers utilise the predictive facility of world models to train agents in internal simulations, like the authors have originally proposed. Let us take the third iteration of Google DeepMind's Dreamer algorithm as an example. Hafner et al. (2023) outline an algorithm for reinforcement learning that makes use of the world model paradigm, making it capable of learning in diverse domains without the need for fine-tuning (that is, human expert knowledge) when changing from one domain to another. In other words, DreamerV3 algorithm allows for general learning that makes typically brittle reinforcement learning readily applicable across multiple domains. Empirically tested across several gameworlds, the algorithm was eventually able to learn how to collect diamonds in Minecraft, thus solving a long-standing challenge in reinforcement learning.[2]

Computer scientist Yann LeCun, currently a chief scientist at Meta's AI laboratory, develops the paradigm further by proposing a hierarchy of world models as a cognitive architecture that has the potential to capture a kind of common sense knowledge of the world (LeCun 2022, p. 3). His implementation accounts for the non-deterministic nature of the world and proposes energy cost of actions as a reward metric for the agent's behaviour, rendering reasoning in the broad computational sense as an energy minimization problem (ibid., p. 9). Curiously, towards the end of the paper, LeCun admits that the proposed cognitive architecture does not model autonomous intelligence, reasoning, or learning in the mammalian brain. He goes as far as to suggest that our "illusion of consciousness" could be a "side-effect" of having only one "configurable world model engine" due to the size constraints of the human brain (ibid., p. 44). In other words, LeCun laments that the brain does not scale like artificial networks.

LeCun's cursory speculative account of consciousness highlights the amalgamation of concepts across cognitive science, neuroscience, and computer science. Akin to the brain's neuroplasticity, the notion of intelligence seems plastic as it is figured and re-configured through and by A(G)I research and its omnipresent circular logic. Here, world model implements an efficient processing schema that is analogical to our own cognitive system (such is the claim of the seminal world models paper). Ha and Schmidhuber's joint architecture separates the predictive world model from the mechanical agency of the controller. This separation paradoxically instates in the agent a pseudo-cartesian dualism that neuroscience long sought to collapse. Most importantly, learning in a "dream" blurs the divide between the agent's external world and its internal representation.

2.2 Internet-scale data as a world

The technology proposed in the original world models article eventually came to be replaced with generative pre-trained transformers (the *GPT* in ChatGPT). The more powerful and efficient network architecture brought with it the potential to train world models on significantly bigger datasets. Instead, world models subtly shifted from being an instrumental analogy to something akin to an epistemic signpost that has to do with scale.

Transformer networks underlie all recent large language models and the numerous products they spawned. As impressive as transformer-based models appear to be at certain automation tasks, some abilities of large language models are not present in smaller models and only appear at a larger scale. Stanford researchers suggest that these abilities are a "mirage" that emerges due to the choice of testing metrics (Schaeffer et al. 2023). Yet another mirage exhibited by language models, seemingly by design, has to do with their tendency to generate statements that "appear reasonable but are either cognitively irrelevant or factually incorrect" when faced with inputs that are rare or completely outside the training dataset distribution (Ye et al. 2023). This tendency came to be widely known as "hallucinating".

There is an ongoing debate on whether large models acquire interpretable world models, and if so, what aspects of the world are represented in these models. Whether world dynamics is sufficiently mediated through the datasets used to train large models is the fundamental question that the field of ML engineering does not seek to address. Instead, the concern is whether large language models can at all be considered as interpretable models of the world dynamics or they merely represent "surface statistics" (Li et al. 2023).

The epistemic value of world models that are, in theory, acquired by large models is highly contested

[2] Coincidentally, implementing "tool use" is one of the engineering efforts to expand the capabilities of language-based agents. Tool use means that an agent can interact with external software and perform a wide variety of tasks on behalf of the user, cf. for example https://docs.anthropic.com/claude/docs/tool-use.

as well. To give but a few examples, it has been shown that such models lack the structural grasp of geometry (Sarkar et al. 2023), causality (Zečević et al. 2023), or actual understanding (Pezzulo et al. 2024). Nevertheless, some research optimistically suggests that large generative models and their world representations are robust enough to be used as world simulators in scenarios such as autonomous driving (Hu et al. 2023), robotics (Yang et al. 2023) and for training generalist agents (Bruce et al. 2024). Recently, Liu et al. (2024) proposed a large model trained on language and video corpora that acts as a "large world model" in the sense that the model captures both textual knowledge and the physical world dynamics—not unlike the "general world model" teased by Runway AI.

Many discussions in computer science aim to empirically or speculatively demonstrate that language models acquire world models. Same discussions also tend to de-emphasise the lifeworld domains in lieu of their linguistic and visual representations. The immanence of the world in data is taken for granted. Moreover, attributing "worldly", i.e. world-model-based causal knowledge to a large language model admittedly involves "a leap of faith" (Yildirim and Paul 2023). Originally inspired by mental models theorised in cognitive psychology, Ha and Schmidhuber's internal world model are at once exteriorised and re-appropriated in that leap. World model as an agent's representation of its world turns into an external epistemic concept within the context of large neural networks. The demonstrated (or presumed) presence of a world model is now supporting claims that large models do indeed model the world dynamics. In other words, world model came to be a representational benchmark used to frame large models as something more than just surface statistics. A probabilistic model is rendered *causal* in practice.

The appropriation of world model as a signpost for generality is another epistemic slippage and, perhaps, a result of the pressure on researchers to understand how language models can be applied beyond conversational tasks. Internet-scale data appears to approximate representations of at least some complex phenomena. The world model acquired by a large model trained on that data is signalling: the model is suitable for causal reasoning and prediction tasks. In other words, a mere trace of a world model becomes a signpost for general predictive control. This worldly generality turns into an epistemic criterion rather than being a statistical fact pertinent to a dataset. The large models debate hints at the agency that goes beyond the instrumental reach of a software agent. For if an agent can generalise across tasks and domains, it is not tightly bound up with the prescribed environment anymore.

3 A model of the models, not just the training grounds

During the last decade, an implied assumption of ML as *the* technology of most, if not all, data-driven decision systems we now call "AI" became customary in public discourse. Joshua Tenenbaum, computational cognitive science professor at MIT, makes a distinction between AI and its technological means. According to Tenenbaum (2022), we currently have AI technology but no "real" AI, for intelligence is not just function approximation and pattern recognition—two tasks AI technology is already quite good at. From his disciplinary perspective, one of the sensible approaches that could help us to understand human cognition is to "reverse-engineer" the mind, that is, to characterise how the mind works in computational terms and then compare the resulting models to empirical data in order to validate them. Conceiving of the human mind as something that can be (de)constructed is a cognitive science framework rooted in the systems theory and cybernetics of the post-war period.

What piqued my attention is the slogan that Tenenbaum used to describe one of the computational tools for modelling common sense thinking to his audience at a Yale University metaphysics seminar. "The game engine in your head" is a conceptual shorthand for the idea that game engine could be considered as "a model of the model inside the head. Not the training grounds for a learning algorithm, but a *model of the mental models*" (ibid., emphasis mine). Game engine is a type of simulation software that is geared specifically for game design and production. Game engines render graphics and simulate physics in real time—accurate enough to achieve a certain aesthetic quality and responsive enough for playful interaction. In Tenenbaum's slogan, the notion of game engine is put forward as a meta-model for human mental models. "The game engine in your head" is another instrumental analogy for understanding the human ability to learn and predict the world dynamics. To paraphrase, according to this idea from computational cognitive science some aspects of human cognition can be better modelled as a game engine.

Tenenbaum contrasts the "game engine in your head" with the material use of game engines, games, and game-like environments in ML engineering subfields, e.g. autonomous driving and robotics, for transferring models trained in visual simulators to the material world (ibid.). Indeed, since the early days of agent-based cognitive modelling, gameworlds served as knowledge domains that are immediate, visual, complex, and uncertain (cf. Agre and Chapman 1987). Computer games such as Doom, GTA, Starcraft, and Minecraft have paved the way for custom

game-like learning environments that are designed in-house and often implemented with commercial game engines such as Unity and Unreal. Elaborate platforms such as Microsoft's AirSim, DeepMind's XLand, MIT's ThreeDWorld, and others have emerged from computer science laboratories situated in close proximity to the immense venture capital (Ahmed et al. 2023; de Sousa and de Abreu 2024). These platforms are seen first and foremost as software toolkits. From the perspective of cognitive science, game engine as a meta-model of our mental capacity to grasp the world does not have much in common with the engineering use of game engines for simulating the lifeworld. As the algorithmic training grounds are cast aside as a mere technology of AI, what can we derive from exploring this binary?

3.1 Defining model worlds

The game-engine-in-your-head analogy is to human cognition what game engine software is to the lifeworld. Game engines in A(G)I research do not only render graphics, physics, and interactions between agents. They lend the core software, its plugins, and third-party components as composable formal models of the lifeworld's dynamics. Furthermore, game engines assemble these elements and their inherent models as real-time simulations of lifeworld domains. In other words, the software and its components structure the runtime in which artificial agents learn. Contemporary world simulators and idealised microworlds of the past are both part and parcel of the epistemology of A(G)I research and, by extension, the novel cognitive architectures it aims to produce. These software toolkits are not to be dismissed as mere test beds and could be considered as part of the contemporary infrastructural intelligence.

Inspired by the science and technology scholar Bruder's (2021) research on the use of microprocessors as model organisms and arcade video games as test beds for developing AI—allowing for the "mutual contamination of human and artificial intelligence"—I propose to further his analogy. Just as microprocessors provocatively substitute the human brain, world simulators and microworlds act as what I provisionally call *model worlds*. As an early working definition, model worlds are open-ended interactive simulations that are external to artificial learning agents. Model worlds substitute lifeworld domains for the purpose of developing, testing, training, and benchmarking A(G)I architectures and ML algorithms, thus saving costs, time, and addressing safety and regulatory concerns. These sandboxed and composable model worlds are part of the disruptive turn away from the finite historical datasets and towards synthetic data which can be modulated and generated on demand.

Model worlds differ from world simulators and microworlds in that they are neither a scientifically precise simulation nor a highly reduced set of isolated features, occupying a novel middle ground and enabled by the recent technology. The notion stands not so much for mutual contamination as for *distillation* of features of the world and *synthesis* of the learning curriculum. Such curriculum is blending in with the environment, simultaneously grounding it in the simulation of the lifeworld. Model worlds allow for open-ended complexity while being computationally tractable and efficient. Put differently, they integrate cognitive architectures into the material epistemology of computation in yet another round of infrastructural capture (cf. Rella 2024). Game-like model worlds are cultural artefacts as much as technoscientific assemblages, enabling experimental research that recombines concepts from the scientific fields that feed into the modern project of A(G)I.

3.2 The world's runtime

Tenenbaum is one of the principal investigators on a project called ThreeDWorld (TDW), a joint initiative between MIT, Stanford, and IBM. TDW is a model world that simulates environments for the purpose of testing and training artificial agents. It is designed to render fast but accurate physics, photo-realistic graphics, situational sounds, and interactions between agents and objects (Gan et al. 2021). A virtual world is procedurally populated during TDW's runtime by objects from its library. Every runtime of TDW's environment is different but bound by the software's assets. The library is structured by the widely used WordNet database that contains semantic relations between words,—a media-archaeological detail that highlights the continuities within the field of A(G)I research. Agents that populate the environment are virtually embodied as one of the three types: a floating camera; one of the geometric primitives; or a user-defined complex avatar that simulates e.g. a robotic body (ibid., p. 5). TDW enables a physically-based world in which virtually embodied agents can interact between each other and learn, while their agency is delimited by the model world. The runtime recombines environments, objects, materials, and their spatial distribution, thus leading to some degree of contingency in this world. In other words, TDW structures each unique runtime of its environments and populates them with objects and agents.

In a paper published alongside the public release of TDW, the notion of world model is (coincidentally) brought into a model world. The article proposes an agent that can learn an internal model of a dynamic and partially observable world, in this case simulated in TDW.[3] The agent is moved

[3] Gan et al. (2021) references the experiment described in Kim et al. (2020). The latter text curiously does not reference TDW by its name, however the environment is depicted in one of the illustrations.

by its intrinsic curiosity: it is rewarded for actively exploring the environment. The experiment implements a "theory of mind" that allows the learning agent to "understand" and predict the behaviour of other agents (Kim et al. 2020, pp. 15–16). The agent's internal "theory of mind" can be considered as a limited computational form of intuitive psychology. In other words, agents can predict each other's inner states. The model world maintains an external environment that is complex enough to capture the attention of curious agents. In this curiosity-driven learning, novelty becomes a proxy[4] for learning progress (ibid., p. 2). To paraphrase, the agent's intrinsic drive to learn a world model is catered to by the external model world. It seems that TDW's simulation is expressive and efficient enough that there is no need to learn in an internal "dream" anymore. Yet TDW does not represent "unlearnable" noisy dynamics such as falling leaves. By design, TDW simulates a "distillation of a real-world environment" that is grounded in physics (ibid., p. 12). The process of explorative learning is ultimately bound to the action space of the world devoid of leaves. Irrelevant aspects of the lifeworld are left out of the model, and what is considered learnable is instrumentally imbued within the software toolkit.

Custom environments designed for A(G)I research such as TDW are prefigured by the historical use of games as domains that are complex yet bound. In response to the burgeoning use of simple arcade video games in ML research, Microsoft's now-defunct Project Malmo made use of the Minecraft game owned by the company. In 2016, a computer science professor with early access to Project Malmo described its open-ended gameworld as being "very close to the real world in many ways, there are so many possibilities" (Linn 2016). Achieving certain in-game milestones in Minecraft is a challenge for A(G)I researchers to this day. Voyager is an agent based on OpenAI's GPT-4 model and a recent example of work in this area. Voyager explores the Minecraft's gameworld, acquires diverse skills, maintains a "skill library", and incorporates feedback from the learning environment (Wang et al. 2023). Put simply, Voyager learns in Minecraft by playing it. The gameworld effectively acts as a model world that allows for continuous learning by open-ended exploration of the game. The researchers call this an "automatic curriculum" (ibid., p. 2). In this experimental setup, the agent's curriculum, i.e. the list of what has to be learned and the order of learning tasks, is generated by OpenAI's proprietary model as a response to the gameworld's exploration. Overall, the curriculum is based on the overarching goal of seeking novelty, accounting for exploration progress and the agent's internal state. The agent thus "capitalizes on the internet-scale knowledge" contained within the black box of GPT-4 model. This structural dependency is essential to the functioning of Voyager's learning schema (ibid., p. 3).

Project Malmo anticipated the development of Microsoft's own AirSim toolkit and other simulations since then. Particularly interesting are CausalCity and CityLifeSim: two toolkits that implement photorealistic simulations of a Western downtown block populated by cars and pedestrians that act independently from one another (McDuff et al. 2021; Wang et al. 2022). The city block is running on the repurposed AirSim, a plugin for Unreal game engine that Microsoft developed for research on autonomous drones and ground vehicles. While the interaction scenarios can be configured by researchers, the architecture and planning of the area appear fixed, underlining the high-level nature of the agents' instructions. Here, *high-level* means that instructions have to do with navigating the block rather than the physical movement of the agents. CausalCity and CityLifeSim are used to develop algorithms for causal inference and counterfactual reasoning—essential tasks in autonomous driving that also aid the explainability of such systems. Both model worlds structure the causal relations that govern the simulation. These environments act more as conventional test beds and benchmarks for causal learning, a nascent ML approach with the ambitious aim of making inference based on causality and not just correlation. Imperative for the task, CausalCity and CityLifeSim provide causally coherent models that are suitable for developing such high-level agents. Put simply, these model worlds regulate causality and thus rule out spurious correlations that plague "big data" analytics.

DeepMind showed that it is possible to train agents in an open-ended manner on a vast task space to gain general skills, in the sense that they are applicable across multiple tasks and domains (Open-Ended Learning Team et al. 2021). What makes DeepMind's approach feasible is their use of the custom XL and environment that procedurally generates rich 3D worlds and multiagent games. In order to fulfil the curriculum, agents have to acquire capabilities such as navigation, logical reasoning, and theory of mind, among others (ibid., pp. 2–3). Games in XL and are defined solely through rewards, and the gameworld itself is algorithmically variable. Game rules are not directly announced, and the environment is ever changing. The gameworld thus becomes part of the curriculum. In other words, the learning environment is merging with the learning algorithm. This approach implements a task-centric agential worldview: tasks are constituted through the world itself, game rewards, and playing agents. The more recent work from DeepMind offers a technique for training a transformer-based model in XL and (Adaptive Agents Team et al. 2023). In this approach, XLand remains an important part of the learning process as it provides a vast pool of learning tasks, at the same time

[4] A potentially inconsistent proxy, as acknowledged by the article's authors.

grounding the agent's large model in this environmental curriculum. The researchers' goal here is to pave the way for a foundation model for reinforcement learning, that is, a general-purpose action model that can be easily adapted to a broader range of embodied tasks. Open-ended and variable worlds such as XLand appear fundamental for arriving at this kind of generality.

4 On agency and worlding in machine learning

Now, what counts as a world for artificial agents? Ha and Schmidhuber's world model renders the agent's environment as comprehensible and compressible—an internal model that represents the complex world dynamics. As exemplified by the TDW toolkit, the agent's world is typically a bound domain that is devoid of non-deterministic events. Such world could be modelled by a ML algorithm. This procedure is the essence of connectionist neural networks architectures that internally represent the agent's environment by mathematical means. World model as an instrumental analogy is moderately interesting due to the recursive turn inwards, whereby an agent rehearses in a "dream". Somewhat confusingly, in that regard Ha and Schmidhuber's world model can be seen as an internalised learning environment, a kind of probabilistic model world. However, this aspect is becoming increasingly rare as the world model analogy is being exteriorised and re-appropriated in the current discourse on large models.

If world models theoretically acquired by large models are associated with generality, then *world* in large models is a measure of successful scaling operations. World simulation is yet another ability emerging from the resource-intensive process of scaling up the data and compute. Put differently, the world does not have to be captured anymore as it is already immanent in internet-scale datasets. Claiming that a large model acquired a world model means to claim its robust capacity to generalise to new tasks and domains. Jacobsen (2023) writes about synthetic data as a rich source of exposure to variability for the learning agent, considering it as a technology of risk. The epistemic purchase of the large model's generality, then, is that it claims the world as a generally "risk-free zone" (ibid., p. 6) for an agent. In other words, the capacity for world simulation that emerges at scale renders the lifeworld as less risky. World model, initially introduced as a computational analogy of human mental models, becomes an epistemic signpost for general applicability and minimal risk of uncertainty. ML practices speculatively render large models analogous to simulations of the world at large, rehashing the empiricist logic of "big data" and blurring the distinction between language and the world.

4.1 Worlding for A(G)I

Even if we suspend our judgement and take for granted that general aspects of the lifeworld are immanent in large datasets, such dataworlds must be made meaningful for the purposes of A(G)I research. Worldly data needs to be turned into an epistemic ground. As cognitive scientist Harnad (1990) posited, symbol grounding problem is the challenge of making the semantic interpretation of a formal symbol system (such as language) intrinsic to the system and not "just parasitic on the meanings in our heads". In the context of large models, the challenge is to distil language of its situatedness and cultural specificity. Harnad's proposed general approach is to ground symbolic representations in non-symbolic representations in the bottom-up fashion. In spite of Harnad's formulation, in some of the recent work on grounding language this is achieved by grounding models with human feedback (Burns et al. 2023). This move allows for symbolic functions to emerge as the consequence of grounding (Mollo and Millière 2023).

What does worlding, both in its instrumental and new materialist sense,[5] have to do with symbol grounding? As I have shown, claiming that a large model acquired a world model signals a general-purpose representation of the lifeworld. This relation is reversed in model worlds. Model worlds rather avoid any serious claims to generality through their "toy world" designation that distinguishes them from the lifeworld. Despite the label, even commercial games are commonly repurposed as clean and constrained testing grounds for ML algorithms—"microcosms of the real world", as the DeepMind founder Demis Hassabis famously put it (Markoff 2016). While this characterisation certainly frames DeepMind's recent work on multiagent environments, I argue that model worlds generally outgrew their function as development test beds.

Repurposing computer games for technoscientific ends, i.e. as data generation pipelines that eliminate the need for laborious manual annotation, is not a new approach (Richter et al. 2016). Recently, what I provisionally call model worlds are being instrumentalised to ground language in embodied action (Liu et al. 2022; SIMA Team et al. 2024). Here, agent-centric learning environments do not merely generate synthetic data but are grounding language models in their runtime. *World* in model worlds is figured not in the sense of rich material world, but rather as an action space that enables virtual embodiment and warrants control over coherence.

[5] Instrumental worlding is (machine) learning of a world model, while worlding in new materialism is understood as "the setting up of the world, […] a particular blending of the material and the semiotic that removes the boundaries between subject and environment, or perhaps between persona and topos" (Palmer and Hunter 2018).

As designed epistemes external to artificial agents, model worlds provide what is called ground truth in statistics and ML alike. In model worlds, *world* is first and foremost an epistemic ground that stabilises meaning for the learning agent. In short, worlding by the means of model worlds is general symbol grounding.

The growing engineering trend is to build complex, compound AI systems that are integrating multiple sociotechnical agencies instead of designing singular all-encompassing software frameworks (Zaharia et al. 2024). As someone put it bluntly, AI can be seen as a problem of "plumbing" (Sussman 2023). On the practical level, worlding is carried out through the development of research toolkits. Model worlds integrate the topologies of everyday environments, realistic graphics and sound, physically based interactions, causally coherent chains of events, spatially distributed learning curricula, and open-ended games. Some of the software is complex enough to be used for meta-learning research, i.e. developing agents that learn how to learn in an open-ended manner. Each model world is designed with specific applications in mind, yet they can be adapted for a broader set of research tasks. Computer scientist Clune (2019, p. 9) calls such virtual worlds "effective learning environments" and posits that generating them is one of the three pillars that support the development of future AI-generating algorithms, a path towards AGI that he sees as an alternative to the slow "manual" buildup of general artificial cognition.

In this technoscientific worldview, swapping one component for another is a matter of optimisation. But what are the consequences of grounding language models in action instead of relying on human feedback? Model worlds establish in their runtime a material-semiotic relationship between its own grounded action space and an ungrounded language model (and its immanent world model). More importantly, they substitute the "weak" ground provided by human feedback with a ground that is computationally tractable. Model worlds mediate between the worldly generality of large models and their capacity for semantic sense-making. In other words, model worlds as learning environments legitimate the worldly (world model-based) knowledge of language models. Hence, the general capacity for meaning-making is an emerging effect of worlding in the A(G)I project.

4.2 Synthetic agency after model worlds

Agency in the project of A(G)I is clearly bound up with the process of worlding. In world models, an instrumentalised analogy of the human mental model, an artificial agent compresses its environment in the process of abstract representation and then "dreams" in this situated model to learn how to interact with it. This is the primal scene of worlding in the construction of *synthetic* agency. "World" in large models, that is, their worldly knowledge is a measure of scaling operations, as I formulated earlier. Yet agency is not solely an effect of scaling, for the model's robust capacity to generalise of has yet to be made meaningful in model worlds. Synthetic agency then emerges as compound, complex, modulated through software, and more general than before, that is, not bound anymore to the agent's domain. As opposed to its minimal instrumental equivalent, synthetic agency is immanent in the worldly models but only becomes actualised through the agent's action in model worlds.

Worlding blends the understanding of intelligence as distributed across human and technical agencies (Amoore 2019, p. 4) and infrastructural intelligence disguised as cloud-based software (Bruder 2019, p. 10). In the recent A(G)I and ML research efforts, the sociotechnical assemblage of model worlds is interfacing with opaque agents and providing them with a learning environment. The agent's drive to "learn"—and thus to ground what is already learned—is modulated through the process of worlding. This process, in its turn, is blurring the epistemic boundaries between the learning subject and the learning environment. Synthetic agency is negotiated at this boundary, at last, as the difference between the exteriority and interiority of worlding. An agent that has already learned how to use the in-game tools to mine diamonds in Minecraft is one step away from learning how to use other software, granted the necessary interface is in place.

What model worlds enact for synthetic agents, at large, is material-semiotic grounding: they ground language through action in tractable and pseudo-stochastic worlds. As a desirable and therefore rewarded outcome, the epistemic boundary between the learning agent and its environment is eventually stabilised. This stabilisation or, as science and technology scholar Dhaliwal (2022) suggests, addressability is the essence of ML as a cultural technique. ML *is* essentially a material-semiotic worlding practice. Generic minimal agency is expanded into a broader synthetic agency through its grounding in model worlds. Put differently, abstract action and superficial statistics of large models are made meaningful in and by the means of model worlds. Here, agency is entangled with the process of machinic meaning-making.

As I have argued, acquiring a world model is an epistemic signpost for generality in large models. Worldly, that is, world model-based knowledge of an agent distinguishes it from a generic software agent. The minimal agency of the latter is technically inscribed and bound to its domain, while the synthetic agency acquired by the former has the potential to generalise to new tasks and domains. As exemplified by some of the papers I discussed, ML as an epistemic practice is currently dedicated to exploring that potential by finding out what large-model-based agents *can* do. In other words, synthetic agency is an empirical challenge to the critique of AI posited by Dreyfus half a century ago.

5 Conclusion

The "general world model" teased by Runway AI on their blog may be encapsulating more than just the idea of a general-purpose world simulator based on a large model. As of this writing, the company did not publish their research on the topic. The technoscientific and industrial landscape is currently rife with ambitious, yet unfounded propositions such as this one. Currently, there is no consensus on many fundamental epistemic questions pertaining to large models and only a limited number of their reliable use cases (Ye et al. 2023; Juršėnas 2023). This article did not aim to critically assess the technology as it is currently implemented or regulated. Instead, I provided a media-theoretical elaboration of synthetic agency as it is re-configured in and through novel practices in A(G)I research and ML engineering—before the technology is turned into products with the aid of venture capital.

As I have argued, artificial cognition is co-constructed through worlding, a process that is, in its turn, implemented as compound software toolkits. The inert code acquires agency at runtime, thus creating the conditions of possibility for some aspects of A(G)I to emerge. Such process-oriented understanding of A(G)I as software systems—still in development and already in use—embraces their material-semiotic nature and the plurality of concepts of intelligence.

The legitimation of large models by model worlds is mutual. Worlding in ML engineering, in other words, attains the circular logic not unlike the field of A(G)I itself. First, the general capacity for sense-making of large models is ensured by the means of grounding them through interaction in model worlds. Conversely, the very fact of using model worlds for grounding language turns them into something more than mere test beds. To be specific, the use of model worlds for developing and benchmarking A(G)I systems renders them as part of the infrastructure of intelligence. The closed circuit of legitimation is but a means of indirect modelling of emerging properties of A(G)I, all the while moving beyond the shortcomings and criticisms of purely historical datasets.

The simulations I provisionally dubbed model worlds outgrew their epistemic status, but remain relatively understudied as critical AI studies often focus on applied systems and historical datasets. "World simulators", whether it is a probabilistic large model or ML environment manually designed in a laboratory, are also employed as pipelines for synthetic data generation, providing yet another avenue for inquiry. As a technology of risk (Jacobsen 2023), model worlds are already widely used in the automotive industry. Their use in ongoing research is emblematic, however, of the tendency of ML as a technology of A(G)I to become ever more inscrutable. Novel cognitive architectures and engineering techniques are subtly reconfiguring the epistemologies of what we consider as A(G)I and the synthetic agency that these systems acquire when they are deployed (which sometimes happens in an instant, seen historically). Whether the emerging capacities to generalise and simulate aspects of the world is an effect of scale or just an illusion, some company will eventually come up with and release a product that capitalises on that. That is why it is so important for media and cultural studies scholars to stay with the complexity of recent computer-scientific papers, often without access to the datasets and algorithms, simultaneously being attuned to the nuances of discursive work that is captured within these texts.

I focused on the epistemic uptake and permutation of concepts, not attempting to characterise the simulated domains themselves or the language corpora these models are trained on. Another important aspect I have not discussed in this article is the use of ludic elements and technologies in model worlds. The most critical and urgent question of what worlds and whose cultures do language models capture and perpetuate falls within the lineage of classic debates on language (cf. Millière and Buckner 2024). The language model's cultural transfer is further compounded with the model world's cultural transfer through the worlding from the agent's perspective. World-as-language clashes with world-as-model—and the lifeworld.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability No primary dataset was generated for this paper.

Declarations

Conflict of interest The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Adaptive Agents Team, Bauer J, Baumli K, Baveja S, Behbahani F, Bhoopchand A, Bradley-Schmieg N, Chang M, Clay N, Collister A, Dasagi V, Gonzalez L, Gregor K, Hughes E, Kashem S, Loks-Thompson M, Openshaw H, Parker-Holder J, Pathak S, Perez-Nieves N, Rakicevic N, Rocktäschel T, Schroecker Y, Sygnowski J, Tuyls K, York S, Zacherl A, Zhang L (2023) Human-timescale adaptation in an open-ended task space. Pre-print. https://arxiv.org/abs/2301.07608

Agre PE, Chapman D (1987) Pengi: an implementation of a theory of activity. In: AAAI-87 proceedings. https://aaai.org/Library/AAAI/1987/aaai87-048.php

Ahmed N, Wahed M, Thompson NC (2023) The growing influence of industry in AI research. Science 379(6635):884–886. https://doi.org/10.1126/science.ade2420

Amaro R (2022) The black technical object: on machine learning and the aspiration of black being. Sternberg, Berlin

Amoore L (2019) Thinking with algorithms: cognition and computation in the work of N. Katherine Hayles. Theory Cult Soc 36(2):3–16

Amoore L, Campolo A, Jacobsen B, Rella L (2023) Machine learning, meaning making: on reading computer science texts. Big Data Soc 10(1):1–13. https://doi.org/10.1177/20539517231166887

Beller J (2021) The world computer: derivative conditions of racial capitalism. Duke UP, Durham

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp 610–623. https://doi.org/10.1145/3442188.3445922

Biza O, Kipf T, Klee D, Platt R, van de Meent J-W, Wong LLS (2022) Factored world models for zero-shot generalization in robotic manipulation. Pre-print. https://arxiv.org/abs/2202.05333

Bruce J, Dennis M, Edwards A, Parker-Holder J, Shi Y, Hughes E, Lai M, Mavalankar A, Steigerwald R, Apps C, Aytar Y, Bechtle S, Behbahani F, Chan S, Heess N, Gonzalez L, Osindero S, Ozair S, Reed S, Zhang J, Zolna K, Clune J, de Freitas N, Singh S, Rocktäschel T (2024) Genie: generative interactive environments. Pre-print. https://arxiv.org/abs/2402.15391

Bruder J (2019) Cognitive code: post-anthropocentric intelligence and the infrastructural brain. McGill-Queen's UP, Montreal

Bruder J (2021) Donkey Kong's legacy: about microprocessors as model organisms and the behavioral politics of video games in AI. TSANTSA J Swiss Anthropol Assoc 26:71–84. https://doi.org/10.36950/tsantsa.2021.26.6972

Burns C, Izmailov P, Kirchner JH, Baker B, Gao L, Aschenbrenner L, Chen Y, Ecoffet A, Joglekar M, Leike J, Sutskever I, Wu J (2023) Weak-to-strong generalization: eliciting strong capabilities with weak supervision. Pre-print. https://arxiv.org/abs/2312.09390

Cavia AA (2023) Interaction grammars: beyond the imitation game. In: Trillo, Poliks (eds) Choreomata: performance and performativity after AI. CRC Press, London, pp 258–277. https://doi.org/10.1201/9781003312338-13

Chun WHK (2021) Discriminating data: correlation, neighborhoods, and the new politics of recognition. MIT, Cambridge

Clune J (2019) AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence. Pre-print. https://arxiv.org/abs/1905.10985

D'Amato K (2024) ChatGPT: towards AI subjectivity. AI Soc. https://doi.org/10.1007/s00146-024-01898-z

de Sousa M, de Abreu A (2024) The shift of artificial intelligence research from academia to industry: implications and possible future directions. AI Soc. https://doi.org/10.1007/s00146-024-01924-0

Dhaliwal RS (2022) On addressability, or what even is computation? Crit Inq 49(1):1–27. https://doi.org/10.1086/721167

Dindo H, Marshall J, Pezzulo G (2013) Conceptual commitments of AGI Systems. J Artif Gen Intell 4(2):23–58. https://doi.org/10.2478/jagi-2013-0004

Esposito E (2022) Artificial communication: how algorithms produce social intelligence. MIT, Cambridge

Fitzgerald M, Boddy A, Baum SD (2020) Survey of artificial general intelligence projects for ethics, risk, and policy. Global Catastrophic Risk Institute Technical Report 20-1. https://gcrinstitute.org/papers/055_agi-2020.pdf. Accessed 15 Mar 2024

Fjelland R (2020) Why general artificial intelligence will not be realized. Humanit Soc Sci Commun 7(10):1–9. https://doi.org/10.1057/s41599-020-0494-4

Floridi L (2023) AI as agency without intelligence: on ChatGPT, large language models, and other generative models. Philos Technol. https://doi.org/10.2139/ssrn.4358789

Gan C, Schwartz J, Alter S, Mrowca D, Schrimpf M, Traer J, De Freitas J, Kubilius J, Bhandwaldar A, Haber N, Sano M, Kim K, Wang E, Lingelbach M, Curtis A, Feigelis K, Bear DM, Gutfreund D, Cox D, Torralba A, DiCarlo JJ, Tenenbaum JB, McDermott JH, Yamins DLK (2021) ThreeDWorld: a platform for interactive multi-modal physical simulation. Pre-print. https://arxiv.org/abs/2007.04954

Germanidis A (2023) Introducing general world models. https://research.runwayml.com/introducing-general-world-models. Accessed 12 Mar 2024

Guerin F (2011) Learning like a baby: a survey of artificial intelligence approaches. Knowl Eng Rev 26(2):209–236. https://doi.org/10.1017/S0269888911000038

Ha D, Schmidhuber J (2018) World models. Pre-print. https://arxiv.org/abs/1803.10122

Hafner D, Lillicrap T, Norouzi M, Ba J (2020) Mastering Atari with discrete world models. Pre-print. https://arxiv.org/abs/2010.02193

Hafner D, Pasukonis J, Ba J, Lillicrap T (2023) Mastering diverse domains through world models. Pre-print. https://arxiv.org/abs/2301.04104

Hao J, Yuan Y, Wang C, Wang Z (2021) ED2: environment dynamics decomposition world models for continuous control. Pre-print. https://arxiv.org/abs/2112.02817

Harnad S (1990) The symbol grounding problem. Physica D 42:335–346

Heaven WD (2020) Artificial general intelligence: are we close, and does it even make sense to try? MIT Technol Rev. https://www.technologyreview.com/2020/10/15/1010461/artificial-general-intelligence-robots-ai-agi-deepmind-google-openai/. Accessed 5 Mar 2024

Hu A, Russell L, Yeo H, Murez Z, Fedoseev G, Kendall A, Shotton J, Corrado G (2023) GAIA-1: a generative world model for autonomous driving. Pre-print. https://arxiv.org/abs/2309.17080

Jacobsen BN (2023) Machine learning and the politics of synthetic data. Big Data Soc 1–12. https://doi.org/10.1177/20539517221145372

Juršėnas J (2023) Can we stop LLMs from hallucinating? https://towardsdatascience.com/can-we-stop-llms-from-hallucinating-17c4ebd652c6. Accessed 11 Apr 2024

Kayalibay B, Mirchev A, van der Smagt P, Bayer J (2022) Tracking and planning with spatial world models. Pre-print. https://arxiv.org/abs/2201.10335

Ke NR, Singh A, Touati A, Goyal A, Bengio Y, Parikh D, Batra D (2019) Learning dynamics model in reinforcement learning by incorporating the long term future. Pre-print. https://arxiv.org/abs/1903.01599

Ketz N, Kolouri S, Pilly P (2019) Continual learning using world models for pseudo-rehearsal. Pre-print. https://arxiv.org/abs/1903.02647

Kim K, Sano M, De Freitas J, Haber N, Yamins D (2020) Active world model learning with progress curiosity. Pre-print https://arxiv.org/abs/2007.07853

Koh JY, Lee H, Yang Y, Baldridge J, Anderson P (2021) Pathdreamer: a world model for indoor navigation. Pre-print. https://arxiv.org/abs/2105.08756

LeCun Y (2022) A path towards autonomous machine intelligence. Pre-print. https://openreview.net/forum?id=BZ5a1r-kVsf

Li K, Hopkins AK, Bau D, Viégas F, Pfister H, Wattenberg M (2023) Emergent world representations—exploring a sequence model trained on a synthetic task. Pre-print. https://arxiv.org/abs/2210.13382

Linn A (2016) Project Malmo, which lets researchers use Minecraft for AI research, makes public debut. https://blogs.microsoft.com/ai/project-malmo-lets-researchers-use-minecraft-ai-research-makes-public-debut/. Accessed 15 Mar 2024

Liu R, Wei J, Gu SS, Wu T-Y, Vosoughi S, Cui C, Zhou D, Dai AM (2022) Mind's eye: grounded language model reasoning through simulation. Pre-print. https://arxiv.org/abs/2210.05359

Liu H, Yan W, Zaharia M, Abbeel P (2024) World model on million-length video and language with blockwise ring attention. Pre-print. https://arxiv.org/abs/2402.08268

Machine+Behaviour Conference (2024) Conference concept. https://machinebehavior.science/concept. Accessed 11 Apr 2024

Markoff J (2016) Alphabet program beats the European human go champion. https://archive.nytimes.com/bits.blogs.nytimes.com/2016/01/27/alphabet-program-beats-the-european-human-go-champion/. Accessed 18 Mar 2024

McDuff D, Song Y, Lee J, Vineet V, Vemprala S, Gyde N, Salman H, Ma S, Sohn K, Kapoor A (2021) CausalCity: complex simulations with agency for causal discovery and reasoning. Pre-print. https://arxiv.org/abs/2106.13364

McQuillan D (2022) Resisting AI: an anti-fascist approach to artificial intelligence. Bristol UP, Bristol

Mendonca R, Rybkin O, Daniilidis K, Hafner D, Pathak D (2021) Discovering and achieving goals via world models. Pre-print. https://arxiv.org/abs/2110.09514

Millière R, Buckner C (2024) A philosophical introduction to language models—Part I: continuity with classic debates. Pre-print. https://arxiv.org/abs/2401.03910

Mollo DC, Millière R (2023) The vector grounding problem. Pre-print. https://arxiv.org/abs/2304.01481

Morris MR, Sohl-dickstein J, Fiedel N, Warkentin T, Dafoe A, Faust A, Farabet C, Legg S (2023) Levels of AGI: operationalizing progress on the path to AGI. Pre-print. https://arxiv.org/abs/2311.02462

Natale S (2021) Deceitful media: artificial intelligence and social life after the turing test. Oxford UP, New York

Open-Ended Learning Team, Stooke A, Mahajan A, Barros C, Deck C, Bauer J, Sygnowski J, Trebacz M, Jaderberg M, Mathieu M, McAleese N, Bradley-Schmieg N, Wong N, Porcel N, Raileanu R, Hughes-Fitt S, Dalibard V, Czarnecki WM (2021) Open-ended learning leads to generally capable agents. Pre-print. https://arxiv.org/pdf/2107.12808

Palmer H, Hunter V (2018) Worlding. New Materialism Almanac. https://newmaterialism.eu/almanac/w/worlding.html. Accessed 5 Mar 2024

Pasquinelli M (2023) The eye of the master: a social history of artificial intelligence. Verso, London

Pezzulo G, Parr T, Cisek P, Clark A, Friston K (2024) Generating meaning: active inference and the scope and limits of passive AI. Trends Cognit Sci 28(2):97–112. https://doi.org/10.1016/j.tics.2023.10.002

Rella L (2024) Close to the metal: towards a material political economy of the epistemology of computation. Soc Stud Sci 54(1):3–29. https://doi.org/10.1177/03063127231185095

Richter SR, Vineet V, Roth S, Koltun V (2016) Playing for data: ground truth from computer games. Pre-print. https://arxiv.org/abs/1608.02192

Sarkar A, Mai H, Mahapatra A, Lazebnik S, Forsyth DA, Bhattad A (2023) Shadows don't lie and lines can't bend! Generative models don't know projective geometry…for now. Pre-print. https://arxiv.org/abs/2311.17138

Sarker IH (2021) Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Comput Sci 2:420. https://doi.org/10.1007/s42979-021-00815-1

Schaeffer R, Miranda B, Koyejo S (2023) Are emergent abilities of large language models a mirage? Pre-print. https://arxiv.org/abs/2304.15004

Schlosser M (2019) Agency. In: Zalta (ed) Stanford encyclopedia of philosophy. https://plato.stanford.edu/archives/win2019/entries/agency/. Accessed 5 Mar 2024

Schmidhuber J (2015) On learning to think—algorithmic information theory for novel combinations of RL controllers and recurrent neural world models. Pre-print. https://arxiv.org/abs/1511.09249

Searle JR (1980) Minds, brains, and programs. Behav Brain Sci 3(3):417–424. https://doi.org/10.1017/S0140525X00005756

Shanahan M (2024) Simulacra as conscious exotica. Pre-print. https://arxiv.org/abs/2402.12422v1

Silver D, Singh S, Precup D, Sutton RS (2021) Reward is enough. Artif Intell. https://doi.org/10.1016/j.artint.2021.103535

SIMA Team, Raad MA, Ahuja A, Barros C, Besse F, Bolt A, Bolton A, Brownfield B, Buttimore G, Cant M, Chakera S, Chan SCY, Clune J, Collister A, Copeman V, Cullum A, Dasgupta I, de Cesare D, Di Trapani J, Donchev Y, Dunleavy E, Engelcke M, Faulkner R, Garcia F, Gbadamosi C, Gong Z, Gonzales L, Gregor K, Hallingstad AO, Harley T, Haves S, Hill F, Hirst E, Hudson DA, Hughes-Fitt S, Rezende DJ, Jasarevic M, Kampis L, Ke R, Keck T, Kim J, Knagg O, Kopparapu K, Lampinen A, Legg S, Lerchner A, Limont M, Liu Y, Loks-Thompson M, Marino J, Cussons KM, Matthey L, Mcloughlin S, Mendolicchio P, Merzic H, Mitenkova A, Moufarek A, Oliveira V, Oliveira Y, Openshaw H, Pan R, Pappu A, Platonov A, Purkiss O, Reichert D, Reid J, Richemond PH, Roberts T, Ruscoe G, Elias JS, Sandars T, Sawyer DP, Scholtes T, Simmons G, Slater D, Soyer H, Strathmann H, Stys P, Tam AC, Teplyashin D, Terzi T, Vercelli D, Vujatovic B, Wainwright M, Wang JX, Wang Z, Wierstra D, Williams D, Wong N, York S, Young N (2024) Scaling instructable agents across many simulated worlds. https://deepmind.google/discover/blog/sima-generalist-ai-agent-for-3d-virtual-environments/. Accessed 20 Mar 2024

Sussman GJ (2023) Artificial intelligence: a problem of plumbing? https://www.youtube.com/watch?v=CGxbRJPCoAQ. Accessed 5 Mar 2024

Tenenbaum J (2022) What kind of computation is cognition? https://www.youtube.com/watch?v=NsID1iM8gRw. Accessed 5 Mar 2024

Thórisson KR (2012) A new constructivist ai: from manual methods to self-constructive systems. In: Wang, Goertzel (eds) Theoretical foundations of artificial general intelligence. Atlantis Thinking Machines (4). Atlantis Press, Paris. https://doi.org/10.2991/978-94-91216-62-6_9

Wang P (2019) On defining artificial intelligence. J Artif General Intell 10(2):1–37. https://doi.org/10.2478/jagi-2019-0002

Wang CY, Nir O, Vemprala S, Kapoor A, Ofek E, McDuff D, Gonzalez-Franco M (2022) CityLifeSim: a high-fidelity pedestrian and vehicle simulation with complex behaviors. In: IEEE 2nd international conference on intelligent reality (ICIR), pp 11–16. https://doi.org/10.1109/ICIR55739.2022.00018

Wang G, Xie Y, Jiang Y, Mandlekar A, Xiao C, Zhu Y, Fan L, Anandkumar A (2023) Voyager: an open-ended embodied

agent with large language models. Pre-print. https://arxiv.org/abs/2305.16291
Yang M, Du Y, Ghasemipour K, Tompson J, Kaelbling L, Schuurmans D, Abbeel P (2023) Learning interactive real-world simulators. Pre-print. https://arxiv.org/abs/2310.06114
Ye H, Liu T, Zhang A, Hua W, Jia W (2023) Cognitive mirage: a review of hallucinations in large language models. Pre-print. https://arxiv.org/abs/2309.06794
Yildirim I, Paul LA (2023) From task structures to world models: what do LLMs know? Pre-print. https://arxiv.org/abs/2310.04276
Zaharia M, Khattab O, Chen L, Davis JQ, Miller H, Potts C, Zou J, Carbin M, Frankle J, Rao N, Ghodsi A (2024) The shift from models to compound AI systems. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/. Accessed 25 Mar 2024
Zečević M, Willig M, Dhami DS, Kersting K (2023) Causal parrots: large language models may talk causality but are not causal. Pre-print. https://arxiv.org/abs/2308.13067
Zhu Z-M, Chen X-H, Tian H-L, Zhang K, Yu Y (2022) Offline reinforcement learning with causal structured world models. Pre-print. https://arxiv.org/abs/2206.01474