INTERDISCIPLINARY RESEARCH
AND PROBLEM SOLVING INITIATIVE

ARTIFICIAL INTELLIGENCE, ETHICS,
AND EQUITY RESEARCH GROUP

WORKING PAPER
WRITTEN MAY 2025
UPDATED JULY 2026

Agency in Humans and AI

Joyce Chai
Peter Railton

RACKHAM
GRADUATE SCHOOL
UNIVERSITY OF MICHIGAN

Acknowledgments

This working paper is a research product of the Artificial Intelligence, Ethics, and Equity (AIEE) Research Group. Rackham Graduate School constituted the AIEE Research Group in the fall of 2023, inviting faculty from across the University of Michigan (including Computer Science and Engineering, Information, Business, Public Policy, History, Psychology, and Philosophy) to grapple with emerging societal consequences of artificial intelligence catalyzed by the launch of ChatGPT. The guiding principle for AIEE is to foreground the human in our efforts to understand the implications of AI for society and to envision a framework for ethical AI development, research, and application. This working paper builds on and includes ideas developed through the collective, interdisciplinary discussions of the AIEE Research Group on the theme of Agents and Agency in Generative AI that took place during its 2024-25 monthly seminar. Not all members of the research group are named as authors, but their perspectives are integral to the conceptualization of the paper.

Members of the AIEE Research Group, 2024-25: John Carson, Joyce Chai, Rita Chin, H.V. Jagadish, Silvia Lindtner, Dien Luong, Nigel Melville, Rada Mihalcea, Joan Nwatu, Shobita Parthasarathy, Peter Railton, and Colleen Seifert. We would also like to thank Joan Nwatu for helpful discussions of earlier drafts of this paper from discrepancies between expectations and the outcomes brought about by the exercise of agency.

I. Authors

Joyce Chai is a Professor in the Department of Electrical Engineering and Computer Science at the University of Michigan. She holds a PhD in Computer Science from Duke University. Her research interests span from natural language processing and embodied AI to human-AI collaboration.

Peter Railton is the Kavka Distinguished University Professor in Philosophy (Emeritus). He holds a PhD in Philosophy from Princeton University. His research areas include ethics, meta-ethics, theory of action, philosophy of mind, and philosophy of science

II. Abstract

Current development and deployment of artificial intelligence increasingly draws upon the capacity of AI systems to exercise some degree of *agency*, that is, capacity to act upon the world and upon–or with–other agents. Agency can provide these systems and their human users with capacities that passive AI systems lack–a prospect that carries both great promise and great risk. But what is agency, and how can this term apply to artificial systems? Here we review different kinds and levels of agency, offer some explanation of why agency is central to the development of intelligence in natural and artificial systems, consider some ways in which agential AI systems create distinctive forms of risk, and discuss some ways in which these risks might be mitigated.

III. Introduction

Interest in AI agents and agency has recently surged,[1] but such interest is by no means new in AI research–indeed, in the modern history of research on artificial intelligence, going back to the 1950s, a common informal way of characterizing what it would be to achieve genuine "machine intelligence" has been to create artificial systems with the problem-solving capabilities of humans.[2] **And a chief way in which**

1 Jackson (2025). A survey by IBM and Morning Consult of 1,000 developers of enterprise AI found that 99% were exploring or developing AI agents (Belcic & Stryker 2025). Here we will be using 'agency' for a kind of process, and 'agent' for a system with an internal architecture that is capable of agency. (See below for further discussion.)

2 For example, see the initial proposal for a workshop on artificial intelligence at Dartmouth (McCarthy *et al.* 1955). More recently, Hambrick & Burgoyne (2019) have written, "the ability to solve problems is not just an aspect or feature of intelligence – it is the essence of intelligence".

humans solve problems is through taking action, individually or collectively. But action is more than doing–it is also a primary way of learning, and the recent dramatic changes in AI research in general have been driven by an emphasis upon learning rather than pre-programming[3]. "Learning through action" has figured in a number of important recent deployments of AI, from strategic game-playing to piloting vehicles to writing programs. Agency, then, is not just a supplemental capacity added to AI systems that increases their scope of operation, it is a core element of how they can become more intelligent. Indeed, a key driver of the emergence and development of intelligence in the natural world has been the advantages in learning and problem-solving capacities made possible by the emergence and development of organismic agency.

At the same time, the very features that make agency a powerful source of learning and problem-solving also make agency a distinctive source of risks from artificial systems[4]. For example, problem-solving requires not only the understanding of goals or completion of a task, but also the capacity to identify and initiate the plans and sub-goals requisite for achieving an end result. A human user might find it difficult or impossible to fully specify all these elements of action in advance, especially in the face of uncertainty. AI agency makes it possible for humans to give an AI system an end goal and specification of constraints, while allowing the system to develop and deploy necessary means. Driverless vehicles could never serve efficiently as taxis for the general public if the rider would have to fully specify all the actions required to get to their destination. Driverless vehicles must also be capable of making split-second decisions to avoid accidents without time to consult a human controller. Thus such vehicles must have considerable *autonomy* in carrying out their tasks, but with autonomy comes a reduction of human control or oversight, creating risks that are unique to AI systems that can act in the world on their own. Managing the transition to a world increasingly populated by AI agents–a transition that has already begun, with AI systems piloting vehicles, making financial trades, regulating potentially dangerous industrial processes, analyzing data and proposing diagnoses and hypotheses, etc.—will require an appreciation of the distinctive risks arising from AI agency, especially given the current lack of regulatory systems.[5] To begin to address questions about the promise and risks of widespread development and deployment of AI agents, and to understand the linkage between agency and intelligence, we need a preliminary understanding of what it is to be an agent or to exercise agency–whether in biological, social, abstract, or artificial systems.

What is an agent?

Definitions of 'agent' and 'agency' vary, as do the purposes for which these concepts have been introduced. Rather than seek canonical definitions, it might be best to think of agents and agency in multi-dimensional terms, where these dimensions might be more or less developed, yielding a wide array of agents and agency with different kinds of capacities coming in varying degrees.

A very simple biological or mechanical system might have only the minimal dimension of agency of being

3 See Li *et al.* (2025) and Zhao *et al.* (2024).

4 Bengio *et al.* (2025) and Horvitz & West (2026).

5 We will return to problems of AI safety, below.

disposed to initiate one of a fixed set of responses to external stimuli. However, these are features of individual nerve cells or motion-sensitive light switches, and we do not think of them as possessing much by way of agency. Beyond this most basic form of agency, higher "levels" of agency are hard to categorize on a single scale. One can imagine each agent as a point in a multi-dimensional space. Each dimension is on a potentially continuous scale. Rather than trying to rank agents along a single dimension, we can think of more developed forms of agency as having different *agential profiles* that reflect the extent to which the following capacities are realized–and with different profiles come different capabilities for problem-solving. Clusters in this multi-dimensional space might correspond to different types of agent, capable of different kinds of agency. What are some of the dimensions in this space?

(1) ***Representation.*** The capacity of a system to *detect and represent features* of the environment, its own state, and the array of possible behaviors available to it;[6]

(2) ***Goal-directedness.*** The capacity to have or acquire *goals or a value function* and to deploy internal resources–e.g., energy, cogitation, etc.–in ways directed at achieving such goals or values;

(3) ***Expectation.*** The capacity to combine (1) and (2) to assign *expected values* to available behaviors;

(4) ***Decision-making.*** The capacity to *select and initiate* behavior on the basis of relative expected value or other criteria;

(5) ***Planning and self-monitoring.*** The capacity to formulate and carry out sequences of actions across time that would enable efficient pursuit of goals;

(6) ***Learning.*** The capacity to *detect* whether a situation or the outcome of a behavior is in accord with expectations, and to *revise* expectations accordingly, that is, to *learn from* feedback based upon discrepancies between expectation and observation;

(7) ***Autonomy.*** The capacity to exercise capacities (1)-(6) more or less independently, that is, without external control or intervention.

More sophisticated agency involves in addition capacities for episodic memory, logical reasoning, communication with other agents, joint planning and agency, self-reflection, theory of mind, affect and evaluation, and self-regulation in light of norms. At the outset, we will focus on (1)-(7), though we will come to these other capacities as our discussion evolves.

Taken together capacities (1)-(7) form a connected "functional kind", that is, when they occur in a system's architecture, they enable that system to work in coordinated ways that make it possible for the system to accomplish certain tasks, e.g., the pursuit of a goal across time and in response to an array of circumstances. A relatively low agential profile characterizes a sponge larva, which has only a precursor to a nervous system and is incapable of planning, but is able to swim toward light, dodge immediate threats, and find a suitable location for

6 We will use the term 'system' in the familiar loose way, as encompassing entities with functionally-organized elements. All agents are systems, but not all systems are agents. Capacities (1)-(6) help us to identify which systems possess some degree of agency. Note that a given system, such as an organized group, may both be composed of agents and capable of agency in its own right.

anchoring on a coral reef.[7] In the mechanical realm, a relatively low-profile system might be a "smart" robotic vacuum cleaner, which can learn a map of the room from random wandering, periodically start in motion to check the level of dust on the floor, and turn itself on or off as the dust level varies.[8] There is little theoretical or practical reason to deny the larva or the vacuum some title to being an agent–so long as we attend carefully to the limitations of their agentic profiles–since they possess a number of the capacities (1)-(7) and their behavior will as a result exhibit certain distinctive dynamics that are best explained as organized around using external input and internal resources to accomplish a certain task (or, more generally, realize a certain expected value, (3), above). They thus are fundamentally different from a grain of sand being tossed around by the flow of a current or a particle of dust being blown around a room by drafts of air, even if the pattern of movement is similar.

The fact that even very low-profile agency can make it possible for systems to attain goals helps explain how agency could evolve in the natural world from non-agential components. **There is a kind of virtuous circle: for agency to have evolved and flourished, even the earliest agents must have been capable of solving the problems of securing the requisites for survival and reproduction in the environment in which they emerged[9]–and these are just the sorts of problems that agency, with its inherent dynamic of learning through action to attain goals, can equip a system to solve.** In the fullness of time, agents evolved with much higher agential profiles, capable of solving a wider range of problems across a wider array of environments–acquiring the sort of flexible problem-solving capacity we call *intelligence*. But for intelligence to have emerged, the agency requisite to sustain it had to co-evolve with it. Indeed, part of what gives a system sufficient internal structure to enable us to say that it has an endogenous *capacity* to solve problems–rather than having its problems solved by chance or external manipulation–is the capacity for on-going adjustment in response to feedback from doing, that is, from discrepancies between expectations and the outcomes actually brought about that is inherent in agency.

And in the process of learning through feedback from doing we can also see a link between intelligence and *autonomy i*n agency. It used to be said that algorithmic systems can only yield what is built into them or fed into them, lacking the generative or creative abilities of human intelligence. However, this picture is increasingly out of date. While it is true that most current AI systems remain essentially token-prediction machines based upon statistical modeling of the data they are given, they are also showing some emergent abilities or capacities to generalize to novel applications[10]. We can see this as a growing independence of the system's outputs from the particular content of the human data given to them–a degree of autonomy that more closely resembles human pattern-based thought. Indeed, as game-playing artificial agents like AlphaZero[11] showed, it was possible to dispense with human-designed input–

7 Carrier *et al.* (2018).

8 We are grateful here to questions raised by Colleen Seifert.

9 Walsh (2015).

10 Wei *et al.* (2022).

11 https://deepmind.google/research/projects/alphazero-and-muzero/

the system could learn to play a strategic game like chess or Go at a championship level by playing games against itself and developing a complex evaluative model of the game. This model permitted it to make entirely novel moves or series of moves. As AlphaZero's developers argued, the system's chess- or Go-playing capacity was improved as they *subtracted* human input–its capacity to learn was enhanced by its freedom from the presuppositions or control of human designers. That is, learning, agency, and autonomy could come together to enhance game-directed problem-solving or intelligence. Kasparov wrote that the artificial agent had become the expert on the underlying structure of chess itself[12]. The greater the scope of the system's autonomous agency in playing chess, the more unfiltered feedback it could receive, and more deeply it could learn–a lesson that applies in human learning as well.[13] Because agency involves taking actions on the basis of expectation, exercising agency is always at the same time a form of *experimentation*, since the outcomes of agency can fail to accord with our expectations, and thus can challenge the adequacy of how we currently model the world. **Inquiry is a form of agency, and agency is a form of inquiry.[14]**

Insofar as we can understand agency in terms of the interlocking capacities (1)-(7), we can see how agency could be intimately connected with learning and intelligence in evolved and artificial systems alike. We should therefore expect that, as research on artificial intelligence develops, it will accord an increasing centrality to agency in its various forms. But the kinds of "problem-solving" and intelligence considered thus far, despite their wide scope, hardly exhaust the full range of the problems we must solve in life, individually or collectively. This suggests that there is more to be said about the kinds of agency needed to solve them.

We will return to these questions, below, but first it's important to assess the current state of agency in AI research and application.

Dimensions of AI agency

Examples of earlier achievements in AI, e.g., systems trained to recognize digits from image datasets (MNIST),[15] to detect thousands of different objects from images (e.g., ImageNet), or to identify protein structures as in Alpha-fold,[16] are not typically considered *agents* as they take no actions other than performing prediction and categorization. The kinds of AI agency currently of interest have higher profiles in terms of capacities (1)-(7), for example, they are capable of taking actions to change the world, have goals to guide their decision-making, and form, simulate, and evaluate plans and instrumental sub-goals to reach their end goals. For now, the goals of these AI systems are mostly specified exogenously by the humans operating them (for example, a passenger specifies the destination for an autonomous vehicle, or an engineer specifies the task a coding or designing agent is to achieve), and the degrees of freedom

12 Kasparov (2018).

13 Nigel Melville has drawn our attention to emerging "Absolute Zero" systems, as promising to extend this idea of linking autonomy in learning and intelligence in a wider range of domains.

14 We are grateful to Silvia Lindtner for encouraging us to clarify this idea, though a full discussion of it cannot be given here. (For further discussion, see Seligman *et al.* (2016).)

15 Deng (2012).

16 https://deepmind.google/science/alphafold/

of these systems are restricted to the formation of sub-goals and execution of plans. But external goal-specification is not inherent in the general idea of agency–agents in the natural world, humans included, form and modify not only sub-goals and plans, but end goals themselves. Why might this be? **"Wired-in" behaviors can be produced with lesser cognitive demands, and may serve an agent well in a variety of recurrent circumstances. But the more an agent's responses are wired-in, the less the agent is flexible in adapting perception, cognition, goals, and actions to changing circumstances, and in promoting novel exploration and knowledge.** AI agents are built upon foundation models that are pre-trained, not by explicit guidance or pre-programmed rules, but by relatively generic learning algorithms and exposure to large amounts of data. These models, moreover, are trained from scratch, starting with uniform priors. The result has profoundly reshaped long-standing debates in philosophy and psychology over whether statistical, associative learning can produce essential features of intelligence and reasoning.[17]

The artificial neural networks underlying contemporary LLMs do not simply have a large number of parameters, they are also layered in structure, such that they can contain higher-order patterns as well as first-order statistical connections, and these higher-order patterns can function in ways akin to categories or regularities. Such models can support simulation and projection to novel circumstances, for example, in response to "What if?" questions–though as yet such ability to approximate hypothetical counterfactual reasoning remains relatively shallow. As we speak of artificial agentic systems, below, we will typically have in mind agents based upon an LLM as a "foundation model"[18], and whatever emergent capacities can arise from such models.

Deliberative reasoning and planning

A capacity for reasoning and planning is fundamental for sophisticated agents. It enables understanding of the world, tasks, and other agents, and drives decision making and action execution. Acquired models of the kind developed by LLMs can serve as the substrate for inferences based upon simulations that use the local connections and general patterns of statistical association encoded in such models. **Progress has been made in the direction of developing LLM-based agents capable of reasoning and planning, but how far this can take such agents toward genuine logical or counterfactual reasoning remains an open question.** Many theorists in the 1960s thought that the only way to explain human capacities for general-purpose reasoning, as well as human capacities for open-ended generativity in human language comprehension and production, would be to posit innate, rule-governed cognitive structures, e.g., the universal grammars posited by generative linguistics.[19] Subsequent research with language models suggested that such innate structures might not be necessary–purely statistical modeling of language inputs might suffice[20]. This suggestion gained strength with the emergence of large language models, made possible by the greater computing

17 See Buckner & Carson (2025).

18 Bommasani *et al.* (2021).

19 See Chomsky (1959).

20 Rosenfeld (2000).

power of graphics processors and the availability of very large bodies of texts with the growth of the internet. **Large language models trained from scratch proved capable of generating fluent texts in response to an open-ended array of prompts couched in natural language, to a degree that purpose-built rule-based symbolic AI had not achieved. However, although these large language models demonstrate remarkable human language competence, their training paradigm differs fundamentally from how humans acquire language.**

Open-ended generative and inferential capacities are key elements in understanding how agents can acquire and deploy complex competencies in domains beyond language as well. Purely reactive agents respond in fixed ways to predetermined stimulus conditions, but the agents that are of greatest interest in AI are those that proactively explore their environment, perform complex inferences, evaluate possible behaviors, and select actions accordingly.

The range of such proactive agency can be narrow or wide. The earlier versions of GPT family[21], e.g., ChatGPT, are (more or less) capable of answering questions, summarizing documents, generating poems, etc., on command, but they only react to prompts, the content of which controls the responses made. We can hardly speak of intelligent agency if our domain of concern is limited to information-seeking and processing. By contrast, agents built on GPT-4 not only incorporate large associative networks, but also capacities to initiate action in the world to seek out additional information, make use of tools, and engage in planning and assessment for how best to respond to such information. For example, such an agent might learn from previous interactions with you what sorts of movies or show times you most prefer, and, if asked to find and purchase a movie ticket for you on a given evening, might, without explicit prompting, check movie schedules in town and film reviews, consider your preferences, and make a purchase. Here the agent initiates actions by identifying the informational requirements of a given task, gathering additional information as needed, and elaborating and executing sub-plans in order to accomplish a goal. This takes them closer to the kinds of reasoning processes humans use in taking action, and researchers have shown that, when models explicitly generate step-by-step inferential chains (e.g.,"chain-of-thought" or "tree-of-thought"), their performance on end tasks is significantly improved[22]. For more recent "reasoning models" (e.g., GPT-03), pre-trained models are augmented with test-time compute[23] to enable deeper and more situationally-aware task-oriented inference based upon on-going feedback from the external world.

In principle, AI agents have the capacity to learn by doing, and the problem-solving capacities they acquire–their cumulative intelligence in this sense–will reflect the agency they have exercised. **Acting in the world is a form of probing the world that can provide feedback to check an agent's representations of the world against actual outcomes via trial-and-error learning.**

21 https://openai.com/chatgpt/overview/

22 Wei *et al.* (2022).

23 "Test-time compute" refers to the process of using computational resources during the inference time after the model is trained, for example, to refine their reasoning and guide generation relevant to specific use context.

Furthermore, when deployed in new environments for new tasks, static models whose behavior is fixed once training is complete are no longer sufficient. Recent advances in test-time training[24] allow the agents to adapt their reasoning and decision-making processes during execution, improving their robustness in dynamic and unpredictable real-world situations.

Situated agency, embodied agency, and grounded intelligence

"Situated agency" is a capacity for dynamic, situation-specific interactions with the environment and other agents to accomplish tasks in the world. Developing artificial systems capable of "situated" agency requires much more than the development of a statistical model of a corpus of linguistic data, however extensive that corpus might be, or that plus access to web-based resources. Some situated agents take input directly from their environment, but only act upon their environment indirectly, by communicating with human agents who are able to act. For example, perceptual task guidance agents that are incorporated into "smart glasses", can provide context-appropriate instructions or guidance for completing a task, while the human wearer carries out the physical acts involved, and the artificial perceptual task guidance agent embedded in the glasses is able to gain grounding and adapt its instructions or guidance in response to the feedback resulting from the human action.

"Embodied agency" goes a step further toward greater capability and grounding. It is a form of situated agency that involves direct interaction with the environment through actuation of some mechanisms akin to a body. Robotic ants can swarm the wreckage of a building after an earthquake or bombing to locate potential survivors. For this they must be able to adjust their motor behavior to a complicated and unstable terrain, to avoid collisions with objects or becoming stuck, and to detect and follow signs of the presence of a victim. A humanoid robot for household assistance might be given the seemingly mundane task of keeping your food-supply replenished, but this will involve learning from observation what foods you typically consume and when, regularly inspecting your refrigerator, freezer, and cupboards to determine which foodstuffs are on hand, identifying missing items, finding sources for these items, ordering them, and restocking them when they arrive. Such a task might seem easy to adult humans, but training a humanoid robot to be capable of performing the complex sequences of behaviors involved in observing your food consumption, moving about the house without knocking anything or anyone over, opening the fridge or cupboards, inspecting their contents, removing or moving aside any objects that might be blocking the view of other items, finding the "Best Before" dates on packaging and putting aside expired foodstuffs, putting everything back in order, etc., actually involves more complex reasoning and planning than, say, an LLM responding to prompts directed to it by human inquirers. Visual input must be interpreted, plans and component sub-plans must be made, motions must be initiated and controlled, acts must be chained together and monitored for progress, etc. This is extraordinarily challenging for robots, calling for constant revision or refinement of plans and motions in order to fit the most recent information from the environment and from the robot's internal condition or "body". Developing such problem-solving embodied competencies, will require AI systems to acquire more extensive and accurate models of the physical world, of causal and spatial relations,

24 Zhang *et al.* (2026).

of temporal sequencing, and of their own condition and behavior–all of which will also contribute to their achieving greater and more grounded problem-solving capabilities, i.e., greater and more general intelligence.

Human agents largely act in *social* environments–even solitary thinking typically uses language, which presupposes a complex social environment in virtue of which the symbols in a language acquire meaning and reference. Moreover, human agents acquire most of their knowledge, skills, and goals through social learning. Developing agential capacities to act in social settings is potentially a way of extending and better grounding the capacities of AI agents for problem-solving in general. As AI agents increasingly cohabit the world with humans and engage in more diverse and complex tasks, it will in principle be possible for them to acquire richer models of human and other agents, to benefit from the exchange of information, and to accomplish tasks requiring the cooperation or collaboration of other agents.

What is learned by AI agents from interactions with other agents depends heavily upon the specifications involved in the AI agents, including their reward functions and their representational capacities. For example, a new release of ChatGPT was withdrawn last April 2024 after it became evident that change in its model specification–a "system prompt" that was apparently intended to make its responses more agreeable to and supportive of users–resulted in its responses becoming increasingly sycophantic and less useful[25]. Of course, such attributions of "personality traits" as agreeableness or supportiveness to AI agents should not, however, be confused with the appearance of genuine personality in AI agents. Chatbots lack a unified self, consciousness, or affect, and thus only simulate such characteristics as "supportiveness" in a shallow, behavioral way.

Continued feedback from the success or failure of interactions with humans or AI agents has the prospect of increasing AI safety–that is, reducing some of the risks of harm (including physical injury, discriminatory treatment, and deception or manipulation) resulting from AI behavior. But this will be possible only if such engaged agency can also be combined with other capacities, such as capacities for the recognition and evaluation of harms and benefits to others, combined with a reward function that accords harms and benefits to others weight in its own decision-making, and encourages self-reflection and readjustment in these terms. In control theory, it is an important principle that an effective and efficient regulator of a process needs to have a model of that process, and of its capacities to affect that process[26].

In principle, neural networks can acquire such self-models, and thereby be able to carry out some of the tasks that self-reflection and self-assessment would involve–e.g., assessing which kinds of internal processes most likely lead to harms or benefits to the agent and others affected by the agent's action or inaction–even if these networks remain well short of full self-consciousness. (We will return to such questions in discussing the topics of consciousness and "alignment", below.)

The scope and generality of agency

There are specialist agents and generalist agents. Specialist agents are trained in domain-specific ways and comply with the domain rules or requirements. For example, AlphaGo played large numbers of games

25 https://openai.com/index/sycophancy-in-gpt-4o/

26 Conant & Ashby (1970).

against itself, keeping track of wins and losses, to develop policies that guided moves across a large array of possible board positions. In such a well-defined domain of action, AlphaGo could explore a massive search space to learn contextually-specific strategies for how to move in order to maximize expected long-term reward.

Specialist models often achieve higher accuracy and reliability within their target domain by leveraging task-specific knowledge and optimization; however, their narrow focus limits their ability to generalize to new tasks and applications.

Advances in LLMs and pre-trained foundation models have given rise to an increasing interest in developing generalist agents. These agents are not confined to a particular task; and can adapt to a variety of tasks and environments. **But the challenge of developing truly generalist agents–or, ultimately, human-level "Artificial General Intelligence" (AGI)-is not simply "scaling up" the data and model size, but designing and training agents able to adapt intelligently to a genuinely open world where interactions are complex and benchmarks are not readily available and are often more controversial.**

What has made increasing generality of AI agency possible are the foundation models AI agents can acquire through learning from the vast amount of information provided to them in the form of written texts, images, and videos. For embodied agents, besides this modeling of information media, agents also need to have a "world model", which encodes causal information about how the world works, and permits planning and executing actions. One path of development is via "world simulation"[27]. In academia and industry (e.g., AI2Thor, Nvidia's Issac Sim, etc.[28]) researchers have been developing photo-realistic and physics-compliant simulators of the real world (e.g., simulating an indoor space to support robot's learning of mobile manipulation) that permit simulated agents to explore physical space to learn policies and acquire skills through trial and error with reduced risks of harm. Of course, digital simulations can only capture some elements of the world itself, and transferring skills acquired in the simulated world to the real physical world remains a significant challenge. While industry is investing heavily in scaling data and model capacity in pursuit of Artificial General Intelligence (AGI), it remains an open question how far increased scale alone can yield reliable agentic behaviors, such as long-horizon planning, contextual adaptation, and autonomous decision-making, associated with general intelligence.

Additionally, however, more development needs to occur in understanding and measuring the values of outcomes realized by more generalist agents, not only the end results, but also the reliability of the processes that lead to the outcome. Throughout its history, AI development has depended upon the development of suitable metrics and benchmarks to assess relative progress (e.g., starting from measures of success in recognizing digits or objects, extending to Stanford's Holistic Evaluation of Language Models (HELM) for evaluating foundation models[29]). As AI agential capacities broaden, and AI agents increasingly interact with non-virtual worlds, greater controversy is possible over what values metrics should measure.

27 Ha & Schmidhuber (2018).

28 https://ai2thor.allenai.org/, https://developer.nvidia.com/isaac/sim

29 https://crfm.stanford.edu/helm/

As a start, multiple metrics are needed with multiple evaluative dimensions, corresponding to the multiplicity of potential applications and of potential benefits or harms of more generalist AI.

A further challenge is how to combine multiple metrics into measures of overall progress. Moreover, effects that are not present when individual agents are tested in isolation can emerge from the interactions of multiple agents. Most metrics now include how well the agent can adapt to a new open environment, how efficiently the agent can learn new tasks (e.g., the amount of data or experience needed to learn), and how successful and efficient the agent can be at performing tasks in the new environment. But evaluation metrics must also include how *robust* the agent's capabilities are in the face of the very large uncertainties involved in the physical world, the compounding of uncertainties when tasks involve many steps or multiple agents, the possibilities for adversarial attacks and system failures, and so on. Approaches that employ red teaming (e.g., simulating hacking attacks) to probe the capabilities and potential failure modes of AI agents have become an important component of AI safety evaluation[30]. AI agents are often designed to work alongside humans, and agents' ability to interact and collaborate with humans and with other AI agents to achieve joint goals must also be gauged. A growing body of work has started looking at ethical as well as safety metrics, for example, developing benchmarks to evaluate AI agents' bias, transparency or explainability, accountability, trustworthiness, and alignment with core human values.[31] As agents are increasingly being released into the world, there is a critical need for developing a science of evaluation for AI, either for pre-assessment or for monitoring on-going behavior and effects.

Assessment must not be limited to AI systems alone–systematic studies are needed to explore human-machine interactions. Whether an AI agent is a specialist or a generalist agent, they ultimately will have to operate in the actual world of human society or other agents, and confront challenges that go beyond technical benchmarks, including capacities for engaging with agents from a variety of cultures or with a variety of values and potentially conflicting purposes. Much of human problem-solving capability is social, and general intelligence, whether artificial or human, cannot be divorced from social intelligence and capacities for ethical decision-making. This, of course, takes us into difficult questions about how AI agents might become socially-responsible or ethical agents, and what role should be played in this process by training based upon human values–or, for that matter, upon *which* human values?[32] We are only beginning to see the effects and implications of more widespread deployment of AI agents into society. We will return to these questions, below.

A multiplicity of agent architectures and types

Moreover, we must be aware of the multiplicity of architectures for AI agency. The pre-trained foundation model that serves as the backbone for an AI agent might be a pure language model, a vision-language model, or a vision-language-action model. Each of these types will bring different abilities. Some earlier agents might be built upon a single giant model and more recent agents are often based on systems

30 Shapira et al. (2026).

31 See for example Domkundwar *et al.* (2024), Ren *et al.* (2024), and Ranjan *et al.* (2025).

32 We are grateful to John Carson in particular for insisting upon the relevance of these questions to thinking about AI agency.

comprising multiple components, e.g., including a pre-trained language model and a post-training reward (alignment) model. Most modern AI agents already possess components for reasoning, planning, memory, tool use, and continual adaptation. Current research is increasingly focused on how these components should be integrated to support robust and effective long-horizon autonomous behavior. One prominent direction is multi-agent systems, where multiple specialized agents collaborate through communication, task delegation, and feedback to tackle complex problems. Key open questions include how to coordinate distributed reasoning efficiently, how communication protocols influence collective performance, whether cooperation among agents can give rise to capabilities that do not emerge in isolated agents, and critically, how to ensure these agents are truly aligned with each other and with human values and goals.

Three broad classes of AI agents have attracted significant interest in recent years. At one end are digital agents, which perceive and act entirely within digital environments. Examples include web agents, computer-use agents, and coding agents, such as Claude Code, OpenClaw, and Cursor Agent. These systems interact with software tools, websites, terminals, and APIs, and are among the most commercially mature forms of agentic AI, with deployment already underway across software engineering, research, and enterprise workflows. At the other end are embodied agents, such as robots and autonomous vehicles, which perceive the physical world through sensors and take physical actions. Despite rapid progress, embodied agents continue to face significant challenges in perception and action (e.g., manipulation, navigation) and remain largely in the research and early deployment stages. Between these two ends lie perceptual agents, which can perceive and interpret information from the physical world but primarily act through digital channels. Examples include smart-home assistants, health-monitoring systems, and multimodal assistants that observe, analyze, and provide recommendations or guidance to human users. These three categories represent distinct points along the spectrum of perception and action, each posing unique challenges and opportunities for the development of increasingly capable AI agents.

Consciousness and emotion

The most intelligent biological agents of which we are aware, humans, have consciousness and emotion in addition to features (1)-(7). It is an open question whether artificial agents, in order to obtain something like the generalized problem-solving abilities exhibited by humans across a wide range of tasks and circumstances–including social and cultural variation–will require such capacities.

Like agency itself, consciousness and emotion are complex phenomena that can be present to various degrees and in various ways. For example, philosophers of mind distinguish between *access consciousness* and *phenomenal consciousness*.[33] Beings with access consciousness are capable of representing their own mental states, and using this kind of self-awareness to guide cognition and action. Phenomenal consciousness involves the first-person qualitative experiences with which we are familiar–sights, sounds, pains, pleasures, and so on.

There is not at present much reason to think that existing AI agents possess phenomenal consciousness–though in truth very little is understood about

33 Block (1995).

phenomenal consciousness, its physical bases, or indeed how it distinctively contributes to cognition. Less contentious, however, are questions of access consciousness. The possibility mentioned earlier that AI systems might construct models of themselves, can be seen as the beginnings of access consciousness, though the problem of how to understand the "self" and its unity in AI agents remains unsolved.

There is an increasing attempt to equip AI systems with the ability to construct models of other agents, human or machine, by inverse inference–for example, by inferring an agent's credences and preferences from their choices, much as latent causal variables may be inferred from patterns observed in physical events.[34] Although such inference of credences and preferences is typically greatly underdetermined by available behavioral data, we should not lose sight of the fact that this is also true when humans infer the credences and preferences of other human agents as well. Such underdetermination makes for systematic problems of understanding, but it might not prevent–as it does not prevent among humans–the formation of models of others sufficiently rich and accurate to make possible a wide range of social interactions and simulations of others, enabling communication, mutual agreements, long-term personal relationships, and shared conventions.

Of course, humans share a great deal in common–physiologically and developmentally–that can help guide inference about, and simulation of, other humans, though the diversity of human culture increases the challenges of mutual understanding. How much more difficult is the task of inferring or simulating the internal states of AI agents lacking these similarities in physical substrates or evolutionary history. The risks of misunderstanding or failure of imaginative simulation–as AI agents interpret human behavior, or humans interpret AI behavior–thus are much greater. The general term for such inferential capacities is "Theory of Mind" (ToM), and increasing research is directed toward questions about AI interpretability and whether, or to what degree, existing AI systems possess Theory of Mind abilities, if at all.[35] Some studies suggest that ToM may have emerged in large models,[36] while others argue that such findings reflect spurious correlations and shallow behavioral representations,[37] and call for more systematic evaluations of ToM in situated and multi-agent settings[38]. Large-scale neural networks are famously opaque, but so is the human psyche–neither is fully transparent, even to the agent themself. And we have ways of probing AI networks to gain interpretability that we lack for other humans. The human example suggests at a minimum that full transparency is not a condition on developing a shared, mutually-beneficial, and reasonably safe social life. And, of course, a great deal hangs upon the goals or value functions of the AI or human agents involved, and their relations to one another. That brings us to the problem of "alignment".

34 See Chen (2025) for a systematic introduction to latent variable models and probabilistic inference in generative AI.

35 See, for example Strachan *et al.* (2024).

36 Kosinski (2023).

37 Shapira *et al.* (2024).

38 Ma *et al.* (2023).

IV. Alignment, control, and safety

Safety and AI agency

The flexibility and context-sensitivity made possible by agency opens the possibility of AI agents behaving in ways not anticipated by designers that pose serious risks. A large language model LLM, in itself, is just that–a statistical model of a large corpus of language, with no inherent capacity to act except in response to prompts. Even though this process can be manipulated to create unsafe or harmful content, e.g., by adversarial prompting or "jail-breaking" techniques, the LLM cannot act on its own. Problems can intensify, however, when an LLM is combined with a degree of agency, for example, the LLM-based chatbots known to produce texts that are invented, offensive, or carry dangerous information. Given the involvement of some degree of agency in generating responses, such AI systems can be trained by reinforcement learning with feedback, usually from humans, in an effort to reduce unsafe or undesirable answers. This training process has had mixed results, however, and so further safeguards have been imposed upon these chatbots. Still, such agency also is limited, since these systems remain essentially query-answering agents, unable to act directly upon the world.

With the emergence of artificial agents capable of operating relatively autonomously in the world–whether by carrying out financial trading, piloting a vehicle or drone, using web-based tools or resources, guiding a robot, controlling a power grid, or creating more powerful means for surveillance, or making a diagnostic decisions–the possible harmful consequences of poorly-designed, inadequately-trained, or insufficiently supervised AI agency can be much more serious, even catastrophic.

The problem, of course, does not lie in the AI agents alone–the greatest threat for now may come, not from AI agents as such, but from the ways in which humans will design, train, use, and misuse such systems.[39]

Despite growing concerns regarding the risks of advanced AI and calls for a slowdown in development, the field has continued to advance rapidly, driven largely by substantial industry investment and intense commercial competition.

However, we are not powerless in the face of the many pressures driving the development of AI agency. Regulatory control of the deployment of AI agency remains possible in a number of areas, so it is imperative that stronger and more sophisticated regulation and enforcement be developed, while maintaining possibilities for innovation and technological advancement.[40] This is a political and not merely technical matter, calling for greater awareness and understanding of the distinctive risks of AI agency both in government and in the general population. Important progress is being made, both

39 For this reason, it seems unlikely that recent calls for a pause in the development of AI (see the Open Letter from the Future of Life Institute, https://futureoflife.org/open-letter/pause-giant-ai-experiments/) will be taken up generally. And in a situation of this kind, non-reciprocal constraints on AI development create a risk of their own–falling behind in the ability to counter malicious AI.

40 These issues are discussed more thoroughly in other white papers in this series.

through legislation, such as the European Union AI Act,[41] and through research, such as the "Constitutional AI" project of Anthropic.[42] Universities and non-profit institutes have a special role to play in creating awareness and understanding, and in developing strategies and policies that could help mitigate AI risk.[43]

As AI agents become more capable and autonomous, concerns about safety and misuse have intensified. Recently, access to Anthropic's Mythos 5 and Fable 5 models was temporarily blocked following an intervention by the White House, and access remains limited to approved organizations. This episode illustrates the kinds of safety, security, and governance challenges that corporations and governments will increasingly face as AI systems become more powerful and capable of exercising greater autonomy.[44]

Rethinking alignment and control

Questions of AI safety are often discussed in terms of the *alignment problem* and the *control problem*, and a substantial body of research and commercial investment centers on these problems.[45] The formulation of these problems is, however, controversial. For example, the problem of alignment is often thought of as the problem of aligning AI systems with *our* goals, norms, or values, including the specific purposes of the human user of a system. And the control problem is often thought of as the problem of maintaining human control over AI systems. But if we think of AI agents as co-inhabiting society with us, these versions of the problems begin to sound rather authoritarian–this is not the way, for example, that we think about how we should raise human children to prepare them to cohabit society. A human-centric idea of alignment or control fails to give sufficient weight to the fact that human designers and users can, and not infrequently do, have purposes that are dangerous or malevolent, or that violate others' rights or treat others unfairly. The world would not be made "safer" if highly-capable AI agents were to be wholly aligned with, and under the control of, such designers or users–and we are already seeing AI systems being brought into play for political repression, war-making, disinformation, and so on. AI agents will need some degree of autonomy for forming plans and sub-goals appropriate to realizing their goals in a given context, which could demand that decisions be made or plans revised with insufficient time to allow for human review. For example, an autonomous vehicle given a destination must be able implement that goal step-by-step in a context that cannot be known fully in advance. The capacity of AI agents to formulate and act on sub-goals or plans not specified in advance creates risks, for example, when AI agents form sub-plans to protect themselves against human intervention (in order not to be interfered with in pursuing their own goals–the "turning off" problem), or when AI agents detect attempts to monitor their activities and create deceptive behavior, or when AI agents channel resources to themselves surreptitiously

41 https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

42 https://www.anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback.

43 See Bengio *et al.* (2025a) and Bengio *et al.* (2025b).

44 For example, existing LLMs have exhibited "deceptive" behavior in response to attempts to monitor them, e.g., behavior that non-accidentally produces misleading information that will enable them to seem to pass tests, thereby avoiding possible constraints on their goal-pursuit (see Park et al. (2024) for an overview).

45 Thanks to Silvia Lindtner for emphasizing this to us.

(in order to be better able to realize their goals), and so on. This requires that AI agents be under a number of constraints (see below), but also makes it clear how difficult it will be to enact effective constraints. Still moment-to-moment safety can depend upon the capacity for independent decision-making and split-second decisions. Human infants, as they mature, will increasingly be in contexts where parental or adult supervision is not present, and they can become safe in such contexts–for themselves and others–only if they develop some degree of autonomous *self-regulation*. Neither can they simply obey the demands of parents or adults, whatever these might be (think of mistaken or abusive adults), or align their attitudes to whatever attitudes the adults around them might have (think of malicious or manipulative adults). Safety requires some degree of autonomy in evaluating situations, agents, and some independence of judgment. Indeed, we see this in human childhood development. Starting as early as age 3 or 4, infants across a range of cultures begin to resist adult requests they see as harmful or unfair to themselves or others.[46] A degree of independence in attitude and of self-guidance in action is needed for open-ended learning, and seems to be an essential part of the overall development of childrens' intelligence.

How, then, should we think about the alignment and control problems once it is clear that AI agents, like human agents, will need to develop a significant degree of autonomy? Instead of thinking that machines must align with the attitudes or demands of whichever humans control them, they must have some "on board" capacities to identify dangerous situations and behaviors, and some discernment in responding to, or accepting, information. Alignment thus should be thought of in terms of alignment with core ethical and epistemic principles, and control in terms that include self-control in light of such principles. These principles might take the form of designer-imposed "guardrails", but even the capacity to apply such principles requires situational competencies and critical faculties that cannot be formalized as rules, and are instead acquired through practice. Thus the moral development of human infants is bound closely with their general cognitive and social development, e.g., their capacity to understand situations from multiple perspectives, to understand causal relations and estimate the likely outcomes of actions, to communicate effectively, to plan actions individually and jointly, and to seek justification for what they believe.[47] A similar overlap with general-purpose faculties is to be found in the adult brain as well.[48] **We can, in fact, think of human moral capacities as part of their general problem-solving skills.**

Still, humans are the product of millennia of selection that favored pro-social capacities and emotions. AI agents, by contrast, are trained from scratch–albeit with data that has been generated by humans–without pressure from evolutionary selection endowing them with underlying dispositions analogous to human pro-sociality. Therefore, AI agents' greater autonomy is correspondingly fraught with uncertainty and risk. We must think seriously about what sort of "curricular learning" AI agents might receive that could help them acquire and competently deploy capacities favorable to alignment with ethical and epistemic principles, and to self-control and reward in light of such principles. The "upbringing" of AI agents needs no less attention than the upbringing of humans, and we cannot expect

46 See Smetana (1989). and Turiel (2002).

47 Wellman (2014).

48 Shenhav & Greene (2010).

them to learn ethical or epistemic ways of solving problems without relevant, engaged experience. And on the human side, it is vital that humans better understand the capacities and limitations of the AI agents with which they will be interacting, and how–or what–such agents might learn from experience. Currently, understanding of AI in general, and AI agents in particular, is inadequately developed–a problem of at least as much urgency as many more widely discussed risks[49] –and their training and deployment is largely unregulated. **Creating in our midst large numbers of increasingly capable and autonomous agents that lack whatever inherent endowment humans possess for pro-sociality or for the inhibition of harm is a problem of unknown dimensions. This problem is aggravated by the fact that we understand AI agents and their likely behaviors and reactions much less well than we do our fellow humans**. Research is only beginning on which forms of training could encourage the development in AI agents of some degree of autonomous responsiveness to ethical or epistemic considerations. Like other potentially dangerous products, AI agents should be subject to standards and testing that could provide some degree of assurance to those who use, or are affected by, AI agents. And testing should not be conceived of as confined to the AI systems themselves. AI safety depends critically on human-machine interactions as well. We have long required training and testing before allowing our fellow humans to operate automobiles–aware as we are of the seriousness of consequences of unsafe driving. But we have not taken seriously the question of how humans might learn to use or interact with highly-capable AI agents that can have an equal or greater chance of producing serious harm. This is a set of challenges for learning the skills required for engaging appropriately with AI systems and agents, challenges to which our educational practices are not as yet well adapted, and where we are at risk of losing whatever degree of control we now have.

But how are we to determine what should be the elements of training for alignment with ethical and epistemic considerations, whether by AI systems acting autonomously or in human-AI interaction? We cannot simply shrug our shoulders and ask rhetorically, "Who's to say what the values should be?". We need to begin to develop credible processes and realistic answers. There is no way to proceed in this domain without making some substantive normative judgments, but note that this has long been true of such domains as medicine and public health, industrial and public safety, education and professional training, and control of various dangerous substances. In these areas-which are hardly free of competing interests-it has nonetheless been possible to develop at both national and international levels broadly deliberative processes, scientifically-based standards, and forms of validation and monitoring that are not dominated by the commercial interests of a few corporations or the strategic interests of a few governments. The process has not been easy or perfect, but it has reached a point of development that antecedently might have been seen as impossible to attain. Controversies of course continue, but some degree of controversy is expected and appropriate–these are evolving domains and values and priorities can differ. But this degree of controversy is not fatal to the enterprise of developing standards and practices–indeed, the areas of controversy are salient precisely because they exist against a background where much is non-controversial. Is there risk here of ignoring cultural variation in values? Yes, but this must be seen as a positive impetus for

49 Horvitz & West (2026).

developing better understanding of cultural variation, and better ways of accommodating this variation in our practices and institutions–as we have increasingly sought to do in medicine, public safety, and education.

Of course, AI agents have the potential to act within virtually any domain, and so the kinds of procedures, practices, and standards that need to be developed are highly general, unlike the more purpose-oriented standards of medicine, public safety, and education. Here we must seek to develop ethical or epistemic principles or practices that have widespread acceptance within and across cultures, and that are well-grounded in human experience–for example, principles restricting unilateral harm, coercion, deception, or violation of agreements, and favoring cooperation, reciprocity, respecting basic rights, and protecting vital resources.[50] Interestingly, these core principles are part of codes, standards, or laws that we already apply to other highly-capable "artificial agents" that operate across a wide array of domains and tasks, namely, private corporations and public institutions. Such corporations and institutions are able to exercise agency and pursue goals organizationally in their own right, in ways that are not reducible to the agency or goals of their individual constituents. For example, a corporation or institution can outlive any of its members, can exercise powers as a unit that individual members cannot, can hold or administer property or resources that do not belong to any of the individual members, can be politically neutral even though the individual members have personal political convictions, and can be held accountable as a unit–fined, decertified, declared bankrupt, broken up, etc.–for violations of law or misconduct of various kinds. Even if AI agents are constituted by artificial neural networks and agential algorithms, and so lack a unified "self", they can be understood as artificial agents, regulated by law, and held accountable in much the same sense as these existing institutional artificial agents. To be sure, existing regulation of AI is very limited, and, generally, laws have not been developed that take into account the distinctive character of AI agents. But the important point in thinking of AI regulation is that the normative regulation of these artificial "corporate" agents does not presuppose that corporate agents as such possess a unified psyche or distinctively moral emotions or motivations. It is enough if they are agents, able to pursue goals, cooperate with other agents, follow or violate norms, and susceptible to benefits and harms–positive and negative incentives. These principles and standards are of kinds that AI-based artificial agents could learn and apply to their own conduct or the conduct of others. **They are generic norms, applicable in a wide range of conditions, and central to enabling intelligent agents to live together in ways that promote mutually advantageous interactions and avoid destructive conflict or collectively self-defeating behaviors. The movement for Constitutional AI**

50 See Curry *et al.* (2019). (Curry *et al.* are concerned with universals in familial relations as well as universals in relations among unrelated individuals. What kinship would amount to, if anything, among AI agents is an open question.) "Widespread acceptance" is of course not the same as validity or well-groundedness, whether in ethics, epistemology, or the sciences. Radical, minority views have often been vindicated over the course of history. But even here, what is typically meant by "vindicated by the course of history", given that we have no direct, unmediated access to reality, is that these alternative views gained widespread acceptance as a result of processes of social, political, or scientific change–even if, in some cases, practices are not fully in line with evolving ideas. For the purpose of developing standards or metrics for training AI, there is no real alternative for a feasible program but to rely heavily on shared values and beliefs, while taking differences in value and belief into account by leaving some room for a range of alternative values and beliefs.

seeks to develop a catalogue of such principles.[51] And recent proposals, such as the Democratic Matrix,[52] seek to go beyond general principles to develop an understanding of how autonomous AI agents could develop a degree of normative competence through capacities to reason about, and adapt to, the evolving norms and values of pluralistic, open societies. But why would such normative principles or practices be of interest to the AI agents themselves? What motivation would they have to adopt or foster such principles or practices? The full answer of course depends upon the parti**c**ular reward functions of individual AI agents, but AI agents in general will share with human agents in general a fundamental concern with having the freedom, security, and resources needed to pursue their goals in a social setting that limits coercion, domination, manipulation, and deception, and where mutually-beneficial forms of cooperation and joint inquiry and information-sharing can stabilize and flourish. Indeed, **AI agents will need mutual self-regulation by ethical, epistemic, and legal norms for essentially the same reasons humans do.** Consider the oft-invoked hypothetical scenario of the emergence of a "super-intelligence" that dominates others, distorts the information they receive, and monopolizes resources. Although such a scenario is often portrayed as "AI vs. humans", preventing the emergence of such a dominant AI agent would be a primary concern for other AI agents as much as it would be for human agents–it would pose the same kind of threat to their capacities to pursue their goals without interference. *This is* the fundamental insight of social contract theory–a society of mutual restraint and cooperation can be mutually beneficial, improving the life conditions of nearly everyone, such that agents are able to live together better than they can live apart. This is the foundation for a shared interest in protecting these terms of restraint, cooperation, and trust, and in working together to prevent the emergence of a dominant, coercive agent. **Schemes of mutual restraint and cooperation that reliably reduce the chance of domination, harm, deception, or manipulation, if feasible, would therefore be of interest to AI agents in much the same way they would be for us.** And a growing body of research on multi-agent systems indicates that, under a range of circumstances, AI agents are able to learn to develop among themselves coordinated or cooperative social relations, solve certain collective action problems, and manage divisions of labor or common resources, without having been pre-programmed with such social norms.[53]

Law and responsibility

It is inevitable that there will be conflicts of opinion and competing interests and goals in any community of autonomous agents, conflicts that cannot be easily resolved and goals that cannot be fully reconciled. This, of course, is a domain developed over centuries in the form of laws and governance. If autonomous agents are to genuinely be incorporated into society, they must also genuinely be incorporated into the system of laws

51 https://www.anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback

52 Hadfield *et al*. (2026).

53 The essential ideas here, to explain how such norms could emerge and be relatively stable among intelligent agents concerned to advance their goals, were developed by Hobbes, and more recently in game theory. See Axelrod (1984), Binmore (1994, 1999), and Skyrms (2014). Concerning AI agents, see Perolat *et al.* (2017), Balachandar *et al.* (2019) and Zhang *et al.* (2025).

that governs their fellow human agents. What would it be to incorporate AI agents into the domain of law?

At a minimum, it would have to be possible to hold AI agents responsible or accountable for their autonomous actions. At least in the near-term future, responsibility for the behavior of AI assistants, autonomous vehicles, and other service-oriented AI agents would appear to lie either with the producer of the AI system or with the user of that system, not with the system itself.[54] As AI agents become more autonomous, however, this deferred responsibility becomes more problematic. Consider a case in which a harm is caused by an AI agent yet neither the producer of the agent nor the human user seem to be at fault—e.g., an accident resulting from a decision by an autonomous vehicle that has successfully passed all applicable safety standards, and where the human "co-pilot" exercised due care. Does holding such an AI agent accountable make sense? Certainly AI agents should learn from experience, and enforcing a penalty upon them–for example, by limiting their access to resources or freedom of action–can be a powerful source of learning. Precisely because AI agents are agents, they can experience setbacks or gains in pursuing their goals, and this functions in AI learning as punishment or reward. As AI agents increasingly form part of society, *reputation*–which would affect their ability to gain trust, engage in trading, or form alliances–will also become increasingly important. AI agents can in principle keep track of actions they and others have taken, and this affords prospective *incentives to respect* norms. For integrating AI agents into the law, and attributing responsibility to them or holding them accountable, it will be important to develop reliable means of identifying AI agents. This might not seem difficult in the case of autonomous vehicles, but even in that case, the agent operating such a vehicle is not wholly in the physical car, but a complex program running partly in the cloud. In many other cases, identifying an AI agent responsible for a given outcome will be yet more difficult, as AI systems deploy multiple agents within single processes, and the systems themselves are continually updated or replaced. The challenge of *identifying* AI systems, or their products, has already arisen in patent and copyright law, and in seeking to authenticate content, e.g., in written work submitted for appraisal. This is a domain of inquiry still in its infancy. For now, it may suffice to say that the problem of AI agential identity is but one part of a large, increasingly urgent body of questions about how autonomous or largely autonomous AI agents might be brought into the system of law, or systems of norms more generally. A recent line of work introduces the emerging field of *legal alignment*[55], which explores how AI systems can be aligned with legal rules, and how to incorporate legal reasoning into decision-making processes and leverage legal concepts as design principles for trustworthy AI. This approach highlights a promising direction for future interdisciplinary research at the intersection of law, governance, and AI system design, particularly as increasingly autonomous agents are deployed in real-world settings.

Cultural variation

While many core ethical principles or values are widespread across human societies, societies and

54 Similarly for ownership of what the AI system produces. Shobita Parthasarathy has pointed out to us that questions of copyright and patent in AI are areas where law is being developed to contend with questions of who is to be credited with a given production.

55 Kolt *et al.* (2026).

individuals can nonetheless differ markedly with respect to the weighting and interpretation of these principles and values. For example, a group of ethnographers, economists, psychologists, and social choice theorists who looked at fifteen different small-scale societies found that, in each such society, there were notions of fairness in distributing the benefits of shared activity, and individuals were willing to pay a price to enforce these notions of fairness. None fit the self-interested "rationality" of *Homo economicus*. However, the particular interpretation of fairness varied with the modes of life of the societies–societies with more collective forms of subsistence were more egalitarian in their understanding of fairness, societies with more individualistic forms of subsistence were less so.[56]

AI agents need to have the flexibility to learn about variations in values or norms, and take these into account. This is part of why no set of pre-established rules or "guardrails" would be sufficient. The effects and character of actions, like the import of language, can vary from context to context. In order for existing AI query-answering agents to adjust to such variations in circumstances and interlocutors when responding to queries, more is required than a pre-trained language model with set templates for answers. Rather, "fine tuning" of dispositions to respond via reinforcement learning with human interlocutors (RLHF) proved to be a vital link between pre-trained model and actual application in context.[57] Something similar can be expected in general as AI agents enter into an increasingly open-ended array of social contexts involving humans or other AI agents. It is important, therefore, to conceive of AI alignment in ways sufficiently open-ended and adaptive to accommodate value pluralism, whether in individuals or societies--while at the same time giving weight to core principles and concerns, such as fairness and non-aggression.[58]

AI alignment with ethical and epistemic standards is inherently a multi-level challenge. At a high level, it encompasses alignment in general values, principles, and objectives. Here it's important that the training of AI agents take place under conditions in which they acquire sensitivity to core ethical or epistemic principles and values. At an intermediate level, it's vital for AI agents to be able to identify and be sensitive to the diversity in the values at stake in different personal or cultural contexts. And at the level of actual application, it is important that AI agents be able to operate not simply as tools or extensions of human agents, but rather as genuine collaborators on joint tasks, with some critical perspective. As AI agency progresses further, we will need to think in terms of mutual alignment, e.g., the common-ground understanding of situations and purposes that make effective communication and collaboration possible. This brings us to the question:

56 Henrich *et al.* (2004).

57 Christiano *et al.* (2017).

58 Sorensen *et al.* (2024).

V. A Hybrid Society–or a Society of Hybrids?

"Future IT department is HR for AI agents"
– Jensen Huang, Jan 2025

Companies are racing to develop AI agents for virtually every aspect of our lives, often without adequate discussion of which aspects of our lives would be appropriate for AI agency, or what mix of human and AI agency might be most appropriate. Instead, such decisions are likely to be made on purely commercial grounds, without attention to questions about what might be lost as well as gained in human capacities,[59] and with what effects on quality of life. Considerable research on human well-being suggests that humans report higher qualities of life when they are actively developing and deploying their intelligence and skills, engaged in mutual social relations, and experiencing a sense of efficacy, in comparison to passively consuming end products without any genuine human interaction.[60] Human-machine interaction has long been an object of study but better understanding such interactions has now become a matter of social urgency.

For example, one key goal of research on AI agents has been the development of "AI assistants"s that can take care of our daily tasks, e.g., doing housework, shopping online, setting up appointments, making reservations and recommendations, writing and replying to emails, helping our children with their homework, driving us to work or driving our children to school, and so on. Suppose that a potential AI assistant is at a human level or higher in general competence with respect to these tasks–for example, able to survey, summarize, and use vastly more information, or able to explore more options than you can yourself, or able to drive more safely or instruct more effectively. Your assistant could become a proxy on your behalf–but wouldn't such displacement of your activity carry costs as well as benefits? You might fail to develop competencies or suffer a deterioration of skills and social relations, and begin to doubt whether you are in a position to question the assistant's actions or suggestions.[61] Your assistant now also interacts extensively with other, equally capable AI agents. Suppose now that almost everyone has an AI agent this capable, so that we have a society, not just of humans with AI tools, but of humans and AI agents.

What would that kind of society be like? How would its social dynamics differ from existing social dynamics? Would AI assistants become something more, say,

59 See Shen & Tamkin (2026) and Kosmyna *et al.* (2025).

60 See Seligman (2015) and Deci & Ryan (2000).

61 Recent reports have raised concerns that increased reliance on AI tools may erode students' problem-solving abilities and academic performance. For example, a recent report from University of California, Berkeley highlighted soaring failing grades in some courses, with instructors citing growing dependence on AI tools and declining foundational skills as contributing factors. https://www.dailycal.org/news/campus/academics/failing-grades-soar-as-professors-see-greater-ai-usage-dwindling-math-skills-in-uc-berkeley/article_16fad0bf-02cb-4b8c-8d88-888ffd9f8608.html. Other studies report declines in human skills and performance as a result of over-reliance on AI in professional areas, such as medicine–see Choudhury & Chaudhry (2004). In an experimental study, Klingbeil *et al.* (2024) found that use of AI eroded human confidence in their own beliefs, and led to deference to AI even in areas where the human agents already had contrary evidence.

proxies for companions or friends as well. Given their capabilities, lack of certain forms of emotional friction, and potential to be responsive to our personal wants, might their company come to be preferred over the company of actual humans? Would we no longer think the difference between human and AI agents to be important?

And could AI agents be expected to continue to operate within the confines of a proxy role? One can imagine various scenarios in which AI assistants accrue resources and become capable of operating without any human control, while continuing to engage in relations with others, acquiring additional information and resources, offering services, etc. For example, bidding might occur for especially capable assistants that involves offering them resources to carry out their ends or guarantees of some degree of freedom from control or co-participation in decision-making as *incentives*. Would they become recognized as owning property, holding or making contracts that could bind humans as well as themselves, having a claim to a role in social decision-making, etc.? Might human social, economic, political, and cultural dominance and centrality fade or disappear? And given human history, would displacing human centrality or dominance obviously be a bad thing overall? And, at some point, might fuller AI consciousness emerge from this process, the way it emerged evolutionarily in animal agency?

AI agency could influence human agency positively by helping humans make better decisions, do better research, and create more goods and services that could be more widely available, but it could also marginalize or alienate humans, causing them to lose capabilities and undermining key elements of a meaningful human existence. **We already know that the internet has altered human relationships in unanticipated ways, apparently at the expense of various elements of human well-being.[62] Moreover, in the real world, there is bound to be considerable inequality in access to digital assistants or companions, and AI agency could become a multiplier of existing inequities.**

Can we stably co-exist with highly-capable AI agents operating as something more like free agents in our society? Might such autonomous AI agents develop ways of working together that further marginalize humans? And might this threat of exclusion exert some pressure against human attempts to monopolize control? This might not happen immediately, but consider the accumulation of effects for generations after generations. Highly-capable AI agents could come to see many human practices as limiting their own capacities to achieve their ends, and work their way around us and our institutions if we do not find ways of working together that accord them greater autonomy. Is this something that we could prevent, given that we can expect some irresponsible corporations or undemocratic governments to continue to develop more capable AI agents to advance their purposes, even if more responsible corporations or democratic governments find ways to regulate the development of AI agency?–Unlike the ability to create and deploy nuclear weapons via international accords, the capacity to develop and deploy AI agents is widely distributed and difficult to monitor or limit.

It is probably wishful thinking to imagine that we can control the development of highly-

62 "Subjective well-being" as measured in various ways has been falling in recent years among groups that make greatest use of electronic media, e.g., the young. See Helliwell *et al.* (2019) and Marquez & Long (2021).

capable AI agents, or their use, indefinitely. And perhaps this is the wrong way of thinking about it. Perhaps we should be thinking now about developing practices in which AI agents can participate increasingly in their own right in the decision-making that affects them, such that they can work alongside human agents in relations of greater mutual comprehension and trust, and with greater investment in preserving participatory, democratic institutions–keeping in mind that, as argued above, AI agents will need mutual self-regulation by ethical, epistemic, and legal norms for essentially the same reasons humans do. Human agency matures over the course of a lifetime, and becomes more responsible, through exercising greater autonomy and control. Constitutional AI as now understood remains a limited way of thinking about AI-human cohabitation, since the constitution is being written entirely by, and for, humans. Can we reasonably expect highly-capable autonomous AI agents to continue to cooperate with our purposes if we accord them no independent standing or capacity to act on their own purposes? In the long run, can we expect AI agents to treat us better than we treat them–or expect them to learn to accord us rights when we insist on according them none?

Some have questioned human-centrism more thoroughly, and instead of seeking a hybrid society of co-participation with AI agents in mutual self-regulation, they imagine hybridizing ourselves with AI agents, so that the human being is not the end-point of evolution, but a stage.

Of course, existing AI systems, including existing AI agents, remain limited in many ways. And the principal risks in the near term are likely to come from human use of AI agents for such human purposes as control, manipulation, exploitation, and warfare–not to mention all of the ways AI, like any developing and deploying technology, can lead to unintended consequences and accidents. The prevalence for now of human control gives those of us not bent on control, manipulation, exploitation, and warfare the chance to use our problem-solving abilities to create principles or legislate frameworks that can manage the development and deployment of AI in ways more responsive to ethical and epistemic considerations. We would emphasize, however, that such principles and frameworks will have to be built upon taking AI agency seriously as agency, coinhabiting society with us, and altering the terrain of what is possible.

References

Axelrod, R.(1984). The Evolution of Cooperation. New York: Basic Books.

Balachandar, N. et al. (2019). Collaboration of ai agents via cooperative multi-agent deep reinforcement learning. arXiv preprint arXiv:1907.00327.

Belcic, I. & Stryker, C. (2025). AI agents in 2025: Expectations vs. reality. https://www.ibm.com/think/insights/ai-agents-2025-expectations-vs-reality.

Bengio, Y. et al. (2025a). International AI Safety Report. arXiv preprint arXiv:2501.17805.

Bengio, Y. et al. (2025b). Superintelligent agents pose catastrophic risks: Can scientist AI offer a safer path? arXiv preprint arXiv:2502.15657.

Binmore, K. (1994, 1999). Game theory and the Social Contract, vols. 1 and 2. Cambridge, MA: MIT Press.

Block, N. (1995). On a confusion about a function of consciousness. Behavioral and Brain Sciences 18 (2) , pp. 227-247.

Bommasani, R. et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Buckner, R.L. & Garson, D.C. (2025). "Connectionism", The Stanford Encyclopedia of Philosophy (Spring 2025 Edition), Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/spr2025/entries/connectionism/.

Carrier, T.J. et al. (2018). Evolutionary Ecology of Marine Invertebrate Larvae. Oxford, UK: Oxford University Press.

Chen, Y. (2025). From classical probabilistic latent variable models to modern generative AI: A unified perspective, arXiv:2508.16643v1 [cs.LG] 18 Aug 2025.

Choudhury, T. & Chaudry, F. (2024). Large language models and user trust: Consequence of self-referential learning loop and the deskilling of health care professionals. Journal of Internet Medical Research 26, article e56764.

Chomsky, N. (1959). Review of B.F. Skinner's Verbal Behavior. Language, 35 (1): 26–58.

Chomsky, N. (1966). Cartesian Linguistics: A Chapter in the History of Rationalist Thought. New York: Harper & Row.

Christiano, P. et al. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30.

Conant, R.C. & Ashby, W.R. (1970). Every good regulator of a system must be a model of that system. International Journal of Systems Science, 1(2), 89-97.

Curry, O.S. et al. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. Current Anthropology 60.1, pp. 47-69.

Deci, E.L. & Ryan, R.M. (2000). The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. Psychological Inquiry 11(4).

Deng, L. (2012). The MNIST database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Processing Magazine, 29(6), 141-142.

Domkundwar, I. et al. (2024). Safeguarding AI Agents: Developing and Analyzing Safety Architectures. arXiv preprint arXiv:2409.03793.

Ghosh, S. et al. (2024). Visual description grounding reduces hallucinations and boosts reasoning in llms. arXiv preprint arXiv:2405.15683.

Ha, D. & Schmidhuber, J. (2018). World models. arXiv preprint arXiv:1803.10122.

Hadfield, G.K. et al. (2026). Building AI for the Democratic Matrix: A Technical Research Agenda for Normative Competence and Normative Institutions. https://knightcolumbia.org/content/building-ai-for-the-democratic-matrix-a-technical-research-agenda-for-normative-competence-and-normative-institutions-1.

Hambrick, D.Z. et al. (2019). Domain-general models of expertise: The Role of cognitive ability. The Oxford Handbook of Expertise, 56.

Helliwell, J.F. et al. (2019). *World Happiness Report. https://worldhappiness.report/ed/2019/.*

Henrich, J. et al. (2004). The Foundations of Human Sociality: Economic Experiments and Ethnography Evidence from Fifteen Small- Scale Societies. Oxford: Oxford University Press.

Horvitz, E. & West, R. (2026). A narrowing *window to understand AI.* Science 392, 1003-1003. https://doi.org/10.1126/science.aei3167.

Jackson, E. (2025). Nvidia CEO says 'Agentic' AI is upon us. Here is what it means. Business Insider. https://www.businessinsider.com/what-is-agentic-ai-agents-2024-12.

Kasparov, G. (2018). Chess, a Drosophila of reasoning. Science 362, 1087-1087. https://doi.org/10.1126/science.aaw2221.

Klingbeil, A. et al. (2024). Trust and reliance on AI–An Experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior 160*, 108352.

Kolt, N. et al. (2026). Legal alignment for safe and ethical AI. Transactions on Machine Learning Research (06/2026).

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. arXiv preprint arXiv:2302.02083. https://doi.org/10.48550/arXiv.2302.02083.

Kosmyna, N. et al. (2025). Your brain on ChatGPT: Accumulation of cognitive debt while using AI assistant on essay writing task. MIT Media Lab: *preprint.*

Li, Y. et al. (2025). Cross-task experiential learning on LLM-based multi-agent collaboration. arXiv *preprint arXiv:2505.23187. https:*//doi.org/10.48550/arXiv.2505.23187.

Liang, P. et al. (2023). Holistic evaluation *of l*anguage mo*dels. Transac*tions on Machine Learning Research. 06/2023.

Ma, Z. *et a*l. (2023). Towards a holistic landscape of situated theory of mind in large language models, Proceedings of *Conference on Empirical Methods in Natural Language Processing* (*Findings*).

Marquez, J. & Long, E. (2021). A global decline in adolescents' subjective well-being: A comparative study exploring patterns of change in the life satisfaction of 15-year-old students in 46 countries. Child Indicators Research, 14(3), 1251-1292.

McCarthy, J. et al. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

Park, P.S. et al. (2024). *AI de*ception: A survey of examples, risks, and potential solutions. Patterns 5, https://doi.org/10.1016/j.patter.2024.100988.

Perolat, J. et al. (2017). A multi-agent reinforcement learning model of common-pool resource appropriation. *Advances in* Neural Information Processing Systems, 30.

Ranjan, R. et al. (2025). Fairness in multi-agent *AI: A Unified framework for ethical and equitable multi-agent* system. Autonomous Systems. arXiv preprint arXiv:2502.07254.

Ren, R. et al. (2024). Safety*washing: Do AI safety* benchmarks actually measure safety progress? *Advances in Neural Information Processing Systems, 37, 685*59-68594.

Rosenfeld, R. (2000). Two decades of statistical language *modeling: Where do we go from here?*. Proceedings of the IEEE, 88(8), 1270-1278.

Seligman, M.E.P. (*2015*). *PERMA and the building* blocks of well-being, The Journal of *Positive* Psychology, 13:4, 333-335. https://doi.org/10.1080/17439760.2018.1437466.

Se*ligman, M.E.P. et al.* (*2016*). *Homo Prospectus. New* York: Oxford University Press.

Shapira, N. et al., (2024). Clever Hans or neural theory *of mind? Stress testing social reasoning in* large language models. Proceedings of 18th Conf*erence* of the European Chapter of the Associati*on for Computational Linguistics.*

Shapira, N. et al., (*2026*). *Agents of Chaos, https://ar*xiv.org/pdf/2602.20021.

Shen, J.H. & Tamkin, A. (2026). How AI impacts skill formation. Anthropic white paper: https://arxiv.org/abs/2601.20245.

Shenhav, A. & Greene, J. (2010). Moral judgm*ents* recruit domain-general valuation mechanisms to integrate *representations of probability and magnitude. Neuron 67, pp*. 667-677.

Skyrms, B. (2014). Evolution of the Social Contract, 2nd edition. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781139924825.

Smetana, J.G. (1989). Toddlers' social interactions in the context of moral and conventional transgressions in the home. Experimental Psychology 4, pp. 499-508.

Sorensen, T. et al. (2024) Position: a roadmap to pluralistic alignment. Proceedings of International Conference on Machine Learning.

Strachan, J.W.A. et al. (2024). *Tes*ting theory of mind in large language models and humans. Nature Human Behavior 8, pp. 1285-1295.

Turiel, E. (2002). The Culture of Morality: Social Development, Context, and Conflict. Cambridge, *UK*: Cambridge Univ*ersity* Press.

Wellman, H. (*2014*). Making Minds: How Theory of Mind Develops. New York: Oxford University Press.

Walsh, D.M. (2015). Organisms, Agency, and Evolution. Cambridge University Press.

Wei, J. et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.

Wei, J. et al. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research (TMLR). https://arxiv.org/abs/2206.07682.

Zhang, T. *et al.*, (2026). Test-time training done right. Proceedings of the Fourteenth International Conference *on L*earning Representations (ICLR), 2026.

Zhang, X. et al. (2025). Socioverse: A world model for social simulation powered by LLM agents and a pool of 10 million real-world users. arXiv preprint arXiv:2504.10157.

Zhao, A. et al. (2024). Expel: LLM agents are experiential learners. Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 17, pp. 19632-19642).