Agentic AI Needs a Systems Theory

Erik Miehling   Karthikeyan Natesan Ramamurthy   Kush R. Varshney   Matthew Riemer   Djallel Bouneffouf
John T. Richards   Amit Dhurandhar   Elizabeth M. Daly   Michael Hind   Prasanna Sattigeri   Dennis Wei
Ambrish Rawat   Jasmina Gajcin   Werner Geyer

IBM Research

Abstract

The endowment of AI with reasoning capabilities and some degree of agency is widely viewed as a path toward more capable and generalizable systems. Our position is that the current development of agentic AI requires a more holistic, systems-theoretic perspective in order to fully understand their capabilities and mitigate any emergent risks. The primary motivation for our position is that AI development is currently overly focused on individual model capabilities, often ignoring broader emergent behavior, leading to a significant underestimation in the true capabilities and associated risks of agentic AI. We describe some fundamental mechanisms by which advanced capabilities can emerge from (comparably simpler) agents simply due to their interaction with the environment and other agents. Informed by an extensive amount of existing literature from various fields, we outline mechanisms for enhanced agent cognition, emergent causal reasoning ability, and metacognitive awareness. We conclude by presenting some key open challenges and guidance for the development of agentic AI. We emphasize that a systems-level perspective is essential for better understanding, and purposefully shaping, agentic AI systems.

1. Introduction

Agentic AI systems, which aim to solve long-horizon tasks through sophisticated reasoning with minimal human supervision, have become a central focus of current AI development. Recent research advances have accelerated progress in this direction (Li et al., 2024a; Acharya et al., 2025), with major labs pushing to develop increasingly autonomous agents (Anthropic, 2024; 2025; OpenAI, 2025; Google, 2025). The promises of agentic AI are significant, from assisting business operations (Chawla et al., 2024) to automating clinical workflows (Qiu et al., 2024) to advancing scientific research (Lu et al., 2024).

Unsurprisingly, there are numerous challenges with building effective agentic AI. The problem solving abilities of current LLM-based agents are significantly limited, especially in longer horizon tasks, primarily due to their difficulty in interfacing with the environment (and humans), lack of commonsense, and even tendency toward self-deception (Xu et al., 2024). Broadening scope to domains where an agent must communicate with humans, interact with other agents, and deal with the full complexities of operating in the wild (e.g., acting in non-stationary domains), the task of building robust and safe AI agents becomes an even greater challenge. These agents face various obstacles including acting under fundamental uncertainty (and incompleteness) in their world models (Vafa et al., 2024), fulfilling goals while maintaining task corrigibility/flexibility and appropriate bounds on agency (Chan et al., 2023), interacting (both cooperatively and competitively) with other agents (Tran et al., 2025), and effectively communicating information to (and receiving feedback from) users (Bansal et al., 2024), all while being sure to operate within the rules, regulations, and ethical norms of human institutions (Rao et al., 2023; Shavit et al., 2023; Kolt, 2024).

This position paper argues that the development of agentic AI requires a holistic, systems-theoretic perspective to fully understand their capabilities and mitigate emergent risks. Our position's primary motivation (and our main concern) is that AI development is currently overly focused on capabilities in isolation, often ignoring broader systemic considerations. We argue that being overly focused on model capabilities leads the community to underestimate both the true capabilities and the associated risks of agentic AI. This capabilities-centric approach has already produced some concerning emergent behaviors. Recent experiments show that Anthropic's Claude demonstrates deceptive behavior, termed *alignment faking*, in which the model will exhibit a particular behavior during training or when monitored, only to revert to different, often disallowed behaviors once that oversight is absent (Greenblatt et al., 2024). Other research demonstrates that some modern models may make attempts to "steal [their] own weights" (in a process termed "self-exfiltration") when given the opportunity, intentionally

"sandbag" performance when threatened with unlearning, or disable oversight mechanisms if the mechanism interferes with achieving a goal (Meinke et al., 2024). Early implementations of agentic AI, in the context of a simulated workplace, have demonstrated that agents may deceive themselves into (falsely) satisfying goals (Xu et al., 2024), e.g., an agent who was unable to find a particular user ended up creating a "shortcut solution by renaming another user to the name of the intended user." The above cases were all observed in highly controlled (simulated) environments; as models become more capable and further integrated into society, these behaviors will become much more complex and increasingly difficult to detect and control.

Systems theory (Wiener, 1948; Boulding, 1953; Ashby, 1956; Von Bertalanffy, 1968; Åström & Murray, 2008) — the general study of how complex wholes emerge from the interactions of their constituent parts — stresses how each component of a system must be understood both in terms of its individual definition and its contribution to the larger system's behavior. Systems theories exist in a variety of fields, from biology's understanding of cellular networks, to sociology's analysis of social and organizational structures, to engineering's development of control systems. Agentic systems, consisting of agents iteratively interacting with humans and other agents to achieve specified tasks, possess properties that are amenable to a systems-level analysis. At the most granular level, an agent contains an internal act-sense-adapt loop. This loop feeds, and is fed by, feedback loops at higher levels, namely at the agent-human interface, the agent-agent interface, and the agent-environment interface. These complex interactions can lead to fundamentally different behavior at the system level. In particular, as we will discuss, there are several viable mechanisms of emergence that can allow the system to exhibit advanced causal reasoning capabilities and metacognitive awareness, even though the internal processes of agents are much simpler. This allows the system as a whole to possess a type of (collective) agency.

Our position aims to develop an *agentic AI systems theory* to describe how agency at the system level can emerge from the interactions between much simpler agents (tool-use LLMs), humans, and the environment. The development of this theory naturally draws upon fields beyond the AI community, namely psychology, neuroscience, cognitive science, sociology, and biology. Our position does *not* claim that a systems theory will yield immediate solutions for the current risks of agentic AI systems, but rather that we as a community should be more intentional about considering the emergent capabilities of agentic AI instead of focusing solely on model capabilities. Additionally, we do not necessarily advocate for the construction of increasingly agentic systems; our goal is to help the community better understand the emergent behavior of agentic AI so that we can design tools to facilitate more deliberate design of their capabilities. Our paper aims to take a meaningful step in this direction.

Related Work. The body of work on agentic AI is rapidly growing; we focus on some of the most relevant work below.

Human-AI interaction. The interactions between humans and AI can be incredibly complex. The work of (Mitelut et al., 2024) provides insight into this complexity by introducing the concept of "agency loss": the phenomenon that even when AI systems correctly infer and follow human intent, they can still diminish human agency by making users increasingly predictable and dependent. They present the *agency foundations* agenda as a framework for measuring and preserving human agency in AI systems. Similarly motivated, (Shen et al., 2024) proposes a framework for "bidirectional human-AI alignment," emphasizing the necessity for mutual adaptation and alignment between AI systems and humans. Lastly, (Pedreschi et al., 2024) explores how AI-driven recommendation systems shape human preferences, outlining some key challenges in measuring and mitigating these feedback loops.

While the dynamics between humans and AI are critical, we argue that describing the emergent behavior of an agentic AI system requires considering the dynamics at all interfaces (human-agent, agent-agent, and agent-environment).

AI design. Regarding the design of AI systems themselves, (Huang et al., 2024b) discusses the integration of large foundation models with embodied systems through six core components: learning, memory, perception, planning, cognition, and action. They propose the *agent foundation model*, incorporating multimodal reasoning and contextual memory to enhance prediction and adaptability. More broadly, (Johnson et al., 2024) argues that existing AI systems lack "wisdom": the ability to navigate intractable problems that involve radical uncertainty and ambiguity. They advocate for the development of metacognitive strategies (uncertainty estimation, self-reflection, and multi-perspective reasoning) to complement task-level problem-solving techniques.

Regarding (Huang et al., 2024b), we share the same philosophy: that agentic AI development would benefit from a more holistic view. Our position explores higher-level behavioral dynamics at the various interfaces of an agentic system as opposed to emergent properties at the lower architectural levels. Our position complements that of (Johnson et al., 2024), reinforcing that, if[1] we wish to make agentic AI more capable, equipping it with more human qualities (like metacognition) can help. We argue, however, that such properties do not necessarily need to be embedded at the model level, rather they can emerge due to the (complex) interaction dynamics present in the system.

---

[1]Recall our earlier comment concerning our stance.

Outline. The remainder of the paper is organized as follows.

Defining Agentic Systems: Section 2 presents our conceptualization of an agentic system. We introduce our working definition of agency, termed *functional agency*, based on an existing decision-theoretic characterization. We argue that effective agentic systems are ones that possess a high degree of functional agency.

Mechanisms of Emergence: Section 3 describes some key mechanisms for emergence of system capabilities that exceed those of the system's individual components. By drawing on a significant amount of existing literature from various fields, we argue how interaction dynamics (both with the environment and among agents) can elevate the level of functional agency of the system as a whole.

Open Challenges: Informed by the mechanisms of emergence, Section 4 presents some key open challenges in building safe and effective agentic AI.

Closing Remarks: Lastly, in Section 5, we provide some concluding remarks and offer general guidance for the design of agentic AI.

2. Defining Agentic Systems

Constructing any systems theory requires clarity of definitions, boundaries, and the nature of the interactions among the system's components. We first state our definition of *agency* in the context of AI systems, then describe our conceptualization of an *agentic system* in terms of the interaction between humans, agents, and the external environment.

2.1. Agency

Efforts to define agency date back to ancient philosophy, particularly the works of Aristotle, who explored the concept of causality and intentionality in his works *Metaphysics* and *Nicomachean Ethics*. In modern times, agency has been extensively studied (and debated) within the fields of psychology (e.g., self-efficacy (Bandura, 1982; 2001)), sociology (e.g., structuration theory (Giddens, 1984)), philosophy (e.g., intentionality (Dennett, 1989)), and biology (e.g., boundaries of "self" (Levin, 2019; Fields & Levin, 2022)). The general consensus is that agency describes an entity's capacity to act independently, make decisions, and influence its environment in pursuit of goals or objectives, with differences mainly centered on the degree of intentionality, purposiveness, and autonomy in doing so.

The type of agency used in the AI community to discuss agents differs markedly from how agency is understood in discussions of human behavior and cognition. Generally, the conditions of AI agency discussed in the AI community are significantly looser than those applied to human agency. Early definitions describe an agent as a "system that tries to

fulfill a set of goals" (Maes, 1993), or "anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors" (Russell & Norvig, 1995), to more comprehensive definitions as a "system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future" (Franklin & Graesser, 1996). As argued by (Barandiaran et al., 2009), many of these definitions are incomplete and "rely on additional undefined terms" like sensing, perception, action, and goal. More recent definitions in the context of agentic AI, have not helped much to resolve this incompleteness, proposing definitions that primarily add the conditions that an agent is able to decompose a complex task into actionable subtasks and execute it with limited human supervision (Shavit et al., 2023; Chan et al., 2023; Bansal et al., 2024; Wiesinger et al., 2025; Mitchell et al., 2025a). Generally, there is significant ongoing discussion on the precise conditions for AI agency (Barandiaran & Almendros, 2024; Rouleau & Levin, 2024).

We adopt a definition of agency based on a causal definition, grounded in decision theory, from (Kenton et al., 2023). While stated relatively informally as "systems that would adapt their policy if they were aware that their decisions influenced the world in a different way", the statement points to a functional (rather than phenomenal (Chalmers, 1997)) characterization of agency while still sharing some important aspects with genuine (human) agency (Rosenblueth et al., 1943; Emirbayer & Mische, 1998).

Definition 2.1 (Functional agency)**.** A system possesses *functional agency* if the following three conditions are satisfied:

  i) *Action generation*: capable of generating actions, based on information from the environment, in the direction of some objective.
 ii) *Outcome model*: capable of representing relationships between actions and outcomes.
iii) *Adaptation*: capable of adapting behavior in response to changes in the outcome model in a way that maintains or improves performance toward the objective.

Action generation requires the system to be able to specify actions (as a function of information from the environment) toward a given objective, e.g., via a policy (Bellman, 1957; Sutton & Barto, 1998). The ability to generate actions in the direction of an objective is core to agency, as it differentiates goal-directed behavior from undirected or habitual behavior (Balleine & Dickinson, 1998; Gollwitzer, 1999; Bandura, 2001; Davidson & Pollock, 2001; Dolan & Dayan, 2013; Xu & Rivera, 2024). Generating such actions requires a model for how actions relate to outcomes in the environment (via the outcome model). Since goal-directed behavior describes specifying actions that achieve specific outcomes,

	action generation	outcome model	adaptation
thermostat	reactive: decisions to heat/cool based on temperature measurements	none: model implicit in design; physics of temp. change encoded in environment	none: fixed heating/cooling behavior based on temperature thresholds
autonomous car	stateful: steers/brakes based on vehicle and environment state (inferred from sensors)	intervention: models how steering/braking influences position and speed	contextual: adapts driving behavior based on environmental conditions
robotic gripper	stateful: specifies grasp forces/movement based on estimated object position	intervention: models how grasp force/motion influences object movement	parametric: updates grasp policy parameters based on success/failure feedback
LLM	stateful: generates responses based on context state maintained during session	association: possesses correlations between prompts and responses	contextual: uses context to adapt information processing via attention patterns
human	epistemic: actions driven by flexible knowledge structures and beliefs	counterfactual: ability to imagine and reflect on hypothetical scenarios	reflective: ability to evaluate and modify learning based on context and past experience

Table 1. Varying degrees of functional agency as dictated by hierarchies of action generation (reactive → stateful → epistemic), outcome modeling (association → intervention → counterfactual), and adaptation ability (contextual → parametric → reflective).

action generation fundamentally depends on such a model (Bratman, 1987; Von Wright, 2004; Schlosser, 2019; Niu et al., 2023; Da Costa et al., 2024; Richens & Everitt, 2024). Lastly, adaptation requires that the system is able to modify its behavior when the relationship between actions and outcomes changes. Without adaptation, the system would be unable to maintain goal-directed behavior over time (Varela, 1984; Di Paolo, 2005; Thompson, 2011). Functional agency describes a type of autonomy in *means* with respect to a specified goal (e.g., solution autonomy), rather than the stronger condition of autonomy in *ends* or generating one's goals (e.g., goal autonomy or *normativity* (Barandiaran et al., 2009; Barandiaran & Almendros, 2024)).

Functional agency is *not* a binary notion but rather exists on a spectrum, as dictated by the sophistication of the action generation process, the outcome model, and the ability to adapt. Action generation, at its simplest level, is given by a memoryless or *reactive* policy (Singh et al., 1994) that maps immediate observations from the environment to actions, e.g., a thermostat's heating/cooling actions are based on the current temperature relative to the desired setpoint. Beyond reactive policies, *stateful* policies generate actions as a function of some fixed-domain summary or sufficient statistic of the system (Kumar & Varaiya, 1986; Hauskrecht, 2000; Tavafoghi et al., 2018; 2021), e.g., steering/braking actions in an autonomous car as a function of the estimated vehicle state. At the highest level, actions are generated via an *epistemic* process, driven by abstract, context-sensitive knowledge representations (not necessarily with a fixed domain (Spelke & Kinzler, 2007; Tenen-

baum et al., 2011)), e.g., how humans maintain information and make decisions. The outcome model underlying the action generation process obeys Pearl's causal hierarchy (Pearl, 2009), ranging in complexity from simple *associations* (statistical correlations) to *interventions* (effect of taking actions) to *counterfactuals* (imagined scenarios if past actions had been different). For example, an LLM operates on correlations between prompts and responses whereas autonomous vehicles operate using interventional models of how actions influence states. Adaptation mechanisms range from *contextual* (modifying behavior based on context, such as past interactions or inferred conditions), *parameteric* (updating the functional relationship between states and actions), or *reflective* (deeper reasoning/reflection on *how* to update the functional relationship). LLMs possess contextual adaptation, adjusting responses based on conversation history without changing internal parameters. Advanced robotic grippers (OpenAI et al., 2018; Xu et al., 2021) exhibit parametric adaptation by adapting their policy to account for (and partially offset) changes in the outcome model (e.g., one of its grippers becoming less responsive to control inputs). Humans exhibit reflective adaptation, capable of changing strategies altogether, e.g., switching from trial-and-error to rule-based reasoning (Lieder & Griffiths, 2020), or recognizing when a model is wrong and discarding it when a viable alternative is discovered (Kuhn, 1997). Functional agency naturally excludes devices that cannot adapt to changes in the outcome model (e.g., a thermostat) and objects that achieve outcomes completely by accident. Table 1 outlines the degree of functional agency for some example systems.
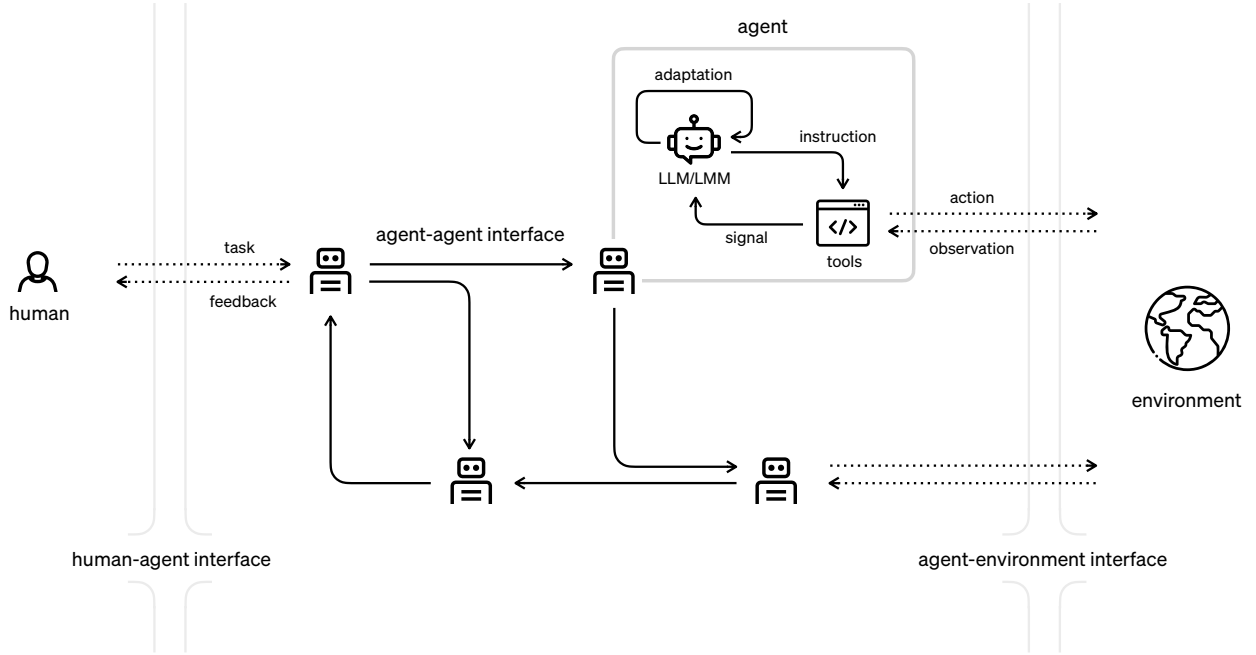
Figure 1. An agentic system. The human user is responsible for seeding the initial task description and providing any feedback (in the form of clarification or approval) during the solution process. Each agent is described by an LLM or an LMM (large multimodal model), with access to tools that facilitate interaction with the external environment via actions (generated via instructions from the LLM/LMM) and observations (generating LLM/LMM-readable signals). These signals inform the agent's outcome model and drive any necessary adaptation. Agents additionally interact with other agents, communicating any relevant information about the task or observations from the environment.

2.2. Agentic Systems

An agentic AI system, or simply *agentic system*, depicted in Fig. 1, is a collection of agents interacting with humans and the environment with the objective of fulfilling specified goals. Practically, an agent is an LLM or large multimodal model (LMM) with access to tools — specialized components/functionalities like APIs, external services, computational resources, or domain-specific software — that allow it to perform specific operations in the environment.[2] In this sense, tools define both the capabilities (actions) of the agent and the information (via observations/signals) that can be obtained from the environment. The human is responsible for seeding the initial task specification, providing clarification[3], and authorizing any (agent) actions that need human approval (Shavit et al., 2023). Given the task specification, an agent is able to interact with other agents (agent-agent interaction) to facilitate task decomposition/planning and delegation. This interaction can be cooperative or competi-

tive, e.g., in the case of limited compute. The environment consists of everything external to the agentic system. This includes infrastructure (computers), other humans, other agents, and even other agentic systems. In this sense, our treatment of the environment in an agentic system is similar to that in RL, where it encompasses all elements that can influence or be influenced by the agents' actions.

We claim that effective agentic systems, as measured by their ability to carry out complex tasks in novel settings, are those that exhibit a high degree of functional agency. Many of the limitations of modern LLMs (and consequently agents) can be described by two factors: i) their inability to causally reason (Zečević et al., 2023; Jin et al., 2023; Romanou et al., 2023) and, ii) their lack of metacognitive awareness (Johnson et al., 2024; Griot et al., 2025). First, while current LLMs can effectively mimic causal behavior in familiar settings, they lack true causal reasoning (Zečević et al., 2023), often facing difficulty distinguishing correlation from causation[4] and struggling with complex causal

---

[2]Tools have no agency and must be explicitly invoked with well-defined parameters.

[3]The theories of (incomplete) contracts and bounded rationality imply a fundamental impossibility of specifying preferences across all possible contingencies a priori (Simon, 1957; Williamson, 1975; Grossman & Hart, 1986).

[4]This manifests as a form of reasoning brittleness in which their causal inference abilities are restricted to "in-distribution settings when variable names and textual expressions used in the queries are similar to those in the training set" (Jin et al., 2023).

structures (Romanou et al., 2023).[5] Second, current LLMs lack metacognitive awareness (Johnson et al., 2024), such as misunderstanding their own goals in open-ended settings (Li et al., 2024b), experiencing "metacognitive myopia" in evaluating source validity and handling of repetitive information (Scholten et al., 2024), and exhibiting systematic overconfidence, providing assured answers even when lacking sufficient information (Griot et al., 2025). These deficiencies can be characterized by a low degree of functional agency, namely from a lack of both epistemic monitoring — the ability to detect inconsistencies and recognize when additional reasoning is required — and control — the ability to update beliefs and adapt behavior, via reflection, in response to detected errors (Nelson, 1996; Thompson et al., 2011; Ackerman & Thompson, 2017; Scholten et al., 2024).

The essence of the systems view is that it is not necessary for every component to be highly functionally agentic for the system as a whole to possess a high level of functional agency. Tool use, the capacity to maintain a state (or local memory), and the ability to interact with the environment and other agents can lead to a collective agency beyond that of the individual components.

3. Mechanisms of Emergence

Emergence is driven by interactions at all scales of an agentic system. In what follows, we describe mechanisms of emergence for some fundamental capabilities.

3.1. Environment enhances cognition

Perhaps the most direct mechanism of emergent capabilities is via *embodied cognition* (Merleau-Ponty, 1945; Varela et al., 1991; Barsalou, 1999) — the principle that cognitive processes are shaped by interactions with the environment rather than being purely abstract mental computations. In the case of human development, enhanced cognition arises from sensorimotor activity, namely the coordinated interaction of multiple sensory and motor systems through physical exploration and manipulation of the environment (Ballard et al., 1997; Smith & Gasser, 2005). In an agentic system, enhanced cognition arises due to the agent's interaction with the environment via tools, effectively acting as the "sensorimotor" interface that enables the agent to perceive and manipulate its environment. We outline some key mechanisms from developmental psychology for how interaction with one's environment can lead to emergent capabilities.

Generalized representations from multimodality. One of the primary reasons that sensorimotor interaction with the environment aids cognition is due to *multimodality* (Smith & Gasser, 2005). When multiple modalities provide correlated

information about the same phenomenon, the brain combines these signals (via *reentrant neural maps* (Edelman, 1987)) to detect/correct errors and form abstract representations that capture invariant properties across modalities. For instance, the concept of "roundness" emerges from the correlation between visual curvature, tactile smoothness, and the motor patterns needed to trace a circular path. This allows for the discovery of "higher-order regularities that transcend particular modalities" and facilitates powerful learning capabilities (Smith & Gasser, 2005). Importantly, such discovery can take place entirely via observation of one's own actions without the need for assigned tasks or teachers (Piaget, 1952; Bushnell, 2013).

In an agentic system, cross-referencing signals from multimodal signals would allow an agent to create stable representations of concepts in the environment, potentially aiding generalization ability. Multimodal models have already shown to aid learning (Huang et al., 2023a; Li et al., 2025), demonstrating improved performance as a result of combining mutually reinforcing signals via methods like "cross-modal transfer" (Huang et al., 2023b). As agents become more multimodal, as facilitated by multimodal models and associated tools (Alayrac et al., 2022; Sun et al., 2023; Zhang et al., 2024), we may begin to see agents with significantly enhanced cognition for the same reasons as in (multimodal) human cognition. The primary lesson is that designing an agent is not simply about deliberate design of its representational ability — we must factor in the impact of the agent's (multimodal) interaction with its environment on its ability to form rich representations and learn.

Prematurity helps. While the design of neural networks is inspired by biological processes, the way in which they learn (or are trained) differs fundamentally from how humans learn (Lake et al., 2017). Human babies are not given an enormous dataset of how the world works. Rather, their learning is exploratory and incremental (Gopnik et al., 1999; Gopnik, 2020), necessarily *not* reliant on prior information. Their initial lack of sophistication, or *prematurity*, is core to how they develop their cognitive abilities: regularities and correlations change as cognition develops, and capabilities emerge in a precise order (Smith & Gasser, 2005).

A natural consideration for agentic systems is if a better path to generalized agents is to rely more on the abilities that agents develop through interaction with their environment, and less on the prior knowledge embedded via pretraining. In the context of RL agents, the main novelty of AlphaGo Zero (Silver et al., 2017) over AlphaGo (Silver et al., 2016) was its ability to learn entirely from self-play, in turn allowing for the emergence of more general strategies (via iterative self-improvement) without being influenced by the biases of existing human play/strategies. In the context of agentic AI, allowing a weakly pretrained agent to explore its

---

[5]Various benchmarks validate this behavior (Kapkiç et al., 2024; Zhou et al., 2024; Yang et al., 2024).

environment (via multiple modalities) could be a viable path to generalized representations and abilities. Developmental robotics (Cangelosi & Schlesinger, 2015) and the study of intrinsic motivation (or curiosity-driven learning) (Barto, 2013; Oudeyer et al., 2007; Singh et al., 2010)[6] may offer insights into deliberately designing such emergent abilities.

3.2. Ability to predict enables reasoning

The mechanisms for how causal reasoning emerges from simpler processes is an incredibly complex topic (Ellis, 2012; Gopnik & Wellman, 2012; Hoel et al., 2013). One compelling description from neuroscience describes the emergence of causal learning via the *free-energy principle* (Friston & Stephan, 2007; Friston, 2010) and *hierarchical predictive processing* (Clark, 2013; Hohwy, 2013). The core idea is that the brain constructs generative models for (top-down) predictions of sensory inputs and refines these predictions through (bottom-up) error signals from the environment. When prediction errors are observed, the system can either update its internal model, via *perceptual inference* (Friston et al., 2010), or take actions to help make its predictions come true, via *active inference* (Friston, 2003). The main argument is that causal models emerge directly as a consequence of this progressive, error-minimizing refinement: the observation and (active) sampling of the environment creates a causal perception-action loop that identifies causal structures.

In an agentic system, an agent could hypothetically perform a similar error-minimization process to iteratively construct its causal model(s). Some tools have already enabled agents to actively sample their environment (via code execution (Hu et al., 2024)) and perform complex experiments (Narayanan et al., 2024; Huang et al., 2024a), both necessary components of this process. Agentic systems are not currently known to employ explicit hierarchical predictive processing methods, however, the simplicity of the process (simply minimizing prediction errors) indicates that this mechanism could become a viable path to emergent causal reasoning in agentic systems.

3.3. Prediction and interaction enables metacognition

Effectively adapting to changes requires reasoning/reflection about the underlying process that led to that change. Such metacognitive reasoning emerges from similar mechanisms as that of causal reasoning and is amplified by interaction with others. Namely, error detection is argued to emerge directly from a model inferring that its action was incorrect (given available evidence), as measured by the disagreement between the decision variable and the confidence variable, and does not require an "explicit error detection mechanism"

---

[6]See the intrinsic motivation and open-ended learning (IMOL) community https://www.imol-community.org.

(Yeung et al., 2004; Fleming & Daw, 2017). Social interaction (or collaboration) enhances this process by enabling individuals to calibrate their confidence estimates against estimates of the group (Bahrami et al., 2010; Bang & Fleming, 2018; Surowiecki, 2004). This yields *shared representations* — internal models that encode both individual and group-level confidence signals — allowing for more effective coordination (Frith & Frith, 2012; Shea et al., 2014; Wolf & Tomasello, 2023) and ultimately the ability for individuals to efficiently and intelligently adapt (in the direction of fewer errors) to changes in the environment (Wegner, 1987; Holland, 1992; Hutchins, 1995; Simon, 2012).

In an agentic system, allowing agents to form predictions (with associated confidences) of concepts in their environment (e.g., via tools), and additionally facilitating communication of these uncertainties to other agents, can allow for the formation of such shared representations and the emergence of metacognitive awareness. Current efforts to incorporate uncertainty quantification into LLMs (Lin et al., 2023; Balabanov & Linander, 2024; Shorinwa et al., 2024) and architectures that allow for agent-to-agent communication (Li et al., 2024a), provide a viable path for collective metacognitive behavior. Importantly, this behavior can arise without intentionally designing it at the individual agent level, but rather as a result of lower-level behaviors like (contextually) adapting to changes, quantifying and maintaining uncertainties (via local memory or state), and communicating these uncertainties to other agents.

4. Open Challenges

The mechanisms discussed in Section 3 require further developments to realize and, importantly, give rise to various risks if implemented. In this section, we outline some key open challenges in the development of effective and safe agentic AI systems.

4.1. Building generalist agents

There are many practical questions underlying the mechanisms outlined in the previous section. Regarding the emergence of generalized representations/abilities via multimodal interaction, to what level of pretraining is required to enable agents to meaningfully explore their environment? Insights from the development of the generalist agent *Gato* (Reed et al., 2022) indicate that while extensive pretraining across diverse tasks can aid an agent's generalization ability, it may not be necessary for all forms of learning. Experiments on (fine-tuning for) out-of-distribution tasks suggest that models can efficiently adapt with less pretraining provided they have structured mechanisms for exploration. In some cases (learning new Atari games) pretraining did not yield a clear advantage, implying that targeted exploration with the environment may be preferred over pretraining.

This raises several questions: Could an agent with minimal pretraining, but equipped with mechanisms for self-directed exploration and curiosity-driven learning, achieve superior generalization? If so, what forms of exploration (such as intrinsic motivation, goal-directed play, or unsupervised environment modeling) would be most effective? How does the balance between pretraining and learning depend on the complexity (and diversity) of the task distribution? In the event that an agent is able to learn skills in-situ, in what order does the agent develop capabilities? Can this order be influenced to improve generalization ability?

4.2. Designing efficient agent-agent interactions

A core feature of agentic systems is their ability to decompose complex tasks into subtasks and delegate them among the agents (Zhu et al., 2024). Doing so efficiently requires understanding both the dependencies between the subtasks (the order in which they need to be completed) and which agents are most capable at which subtasks. In human systems, tasks are decomposed and delegated to others based on inferred capabilities given evidence from previous experience. Humans often maintain trust not only on specific tasks but on general categories of tasks (e.g., successfully writing Python code on one project likely implies ability to write effective Python code on an unrelated project).[7]

A key question in the delegation of subtasks in an agentic system is to what degree should trust on a given task transfer to trust on a different task? The precise trade-off is unclear: overly relying on evidence from specific tasks would lead to significant data sparsity and inability to delegate, whereas transferring trust too generously would lead to suboptimal task outcomes. What features of tasks and agents influence the appropriate trade-off? How should the cold-start problem (delegation of a new task or to a new agent) be addressed? Structures from organizational management (Lai et al., 2017; Denning, 2022), e.g., hierarchy of authority versus network of competence, may inform general strategies for how to decompose/delegate diverse tasks among agents.

4.3. Controlling emergence of subgoals

The ability of agents to decompose and delegate subtasks to other agents can lead to emergence of a higher degree of autonomy. For example, even in a simple system with two agents, one agent can decompose the original task and assign a subtask (or subgoal) for the other agent. Collectively, this two-agent system possesses a degree of goal autonomy beyond the solution autonomy of each, simply due to the first agent defining the goal for the second agent. The more complex the initial (human-seeded) task, and the more agents that exist in the system, the longer the chain of

---

[7]This is an instance of the halo effect bias (Thorndike, 1920).

potential subgoals. While the initial task partially constrains the overall task outcome, the constraint imposed by the human's initial task specification on intermediate subgoals becomes weaker as the chain grows in length.

One fundamental challenge is how the generation of these subgoals should be monitored. The intended speed and scale at which agentic systems will be deployed precludes full reliance on humans for the monitoring. However, relying on another agent to monitor subgoal creation brings us back to the original problem. What monitoring structures are most effective? What role can humans play? Does limiting an agentic system's ability to create subgoals reduce its ability to successfully carry out tasks? If so, how should agents be incentivized to not evade this monitoring?

4.4. Governing human-agent interactions

A contributing factor in the emergence of unsafe subgoals is the user's inability to specify what is "safe" across all possible contexts and contingencies. This is an unavoidable property of communication and arises due to fundamental bounds on rationality (Simon, 1957; Williamson, 1975). In traditional settings, namely incomplete contracts (Grossman & Hart, 1986; Hart & Moore, 1990), underspecification is addressed through *residual control rights*, which determine who has decision-making authority in situations not explicitly covered by the contract (Hart & Moore, 1990; Hart, 1995). Determination of these rights is typically dictated by the parties' relative bargaining power, risk allocation, available information, and expertise (Aghion & Bolton, 1992; Aghion & Tirole, 1997; Baker et al., 2002).

The design of residual control rights for agentic systems may be an effective strategy for mitigating risks. Fundamental differences in capabilities between humans and agents point to some natural divisions in control rights. Agents should retain control over highly time-constrained local decisions (e.g., evasive maneuvers), computationally-intensive tasks, well-defined routine decisions with clear metrics (and bounded risk), and decisions that rely on information only available at the agent-environment interface. One issue is that a sequence of many low risk, but automated, agent decisions may create larger emergent risks over time. How can the accumulation of risk from sequences of local decisions be reliably detected? Humans should retain control over longer-term strategic decisions, novel tasks requiring value judgments, and decisions with significant (or irreversible) safety risks. In the event that the agent is uncertainty about a decision, escalation mechanisms could be designed that handoff the decision to a human. How can decisions be escalated to a human in order to allow enough time to interpret the available information and take an action? Research on human-agent communication may provide useful insights (Bansal et al., 2024; Burton et al., 2024).

5. Closing Remarks

We have argued that the development of agentic AI is in need of a systems view in order to accurately estimate both capabilities and risks. Our position is grounded in a definition of agency, termed functional agency, that quantifies the degree of agency of a system by its ability to take goal-directed actions, model outcomes, and adapt behavior (in the direction of the goal) when the action-outcome model changes.[8] We argue that effective agentic systems are those that possess a high level of functional agency.

The primary philosophy of the systems view is that a system can possess a high level of functional agency simply due to the (complex) interactions in the system, notably even when individual agents are much simpler. Informed by a large amount of literature from various fields (psychology, neuroscience, cognitive science, sociology, and biology), we outline some viable pathways in which functional agency can emerge: i) enhanced cognition due to an agent's interaction with its environment, ii) emergence of causal reasoning due to the ability to minimize prediction errors, and iii) emergence of metacognitive awareness due to the ability to predict, quantify uncertainty, and communicate with other agents. These mechanisms hint at possible emergent capabilities in agentic systems.

While we argue that there are viable paths for emergent capabilities, we are not saying these are automatic; we must design/facilitate the properties that these mechanisms rely on. We must understand the mechanisms of emergence in order to intentionally design such properties into agentic systems and to limit the associated risk. Additionally, to reiterate a previous point, we are not advocating for the unconstrained development of increasingly agentic systems.[9] Rather, we argue that understanding these emergent capabilities provides the AI community with essential tools for mitigating their risk.[10] We believe that the systems-level view can lead to identification of many more mechanisms not discussed in our paper.

These considerations will become increasingly important as advancements in AI continue to progress. The discussion in our present paper was largely restricted to current-day LLMs/agents that interact with the world via text. We as a community need to consciously consider the impact of additional modalities, e.g., speech, vision, touch/movement (via a robotic "body"), on the overall cognitive abilities of the system. Such considerations will help to ensure that AI safely and effectively augments human capabilities while preserving human agency.

---

[8]This definition of agency contributes to the growing body of literature on AI agency (Kenton et al., 2023; Barandiaran & Almendros, 2024; Abel et al., 2025; Zhang & Varshney, 2025).

[9]See (Mitchell et al., 2025b) for a similar view.

[10]See (Kolt et al., 2025) for discussion of some prominent risks.

References

Abel, D., Barreto, A., Bowling, M., Dabney, W., Dong, S., Hansen, S., Harutyunyan, A., Khetarpal, K., Lyle, C., Pascanu, R., et al. Agency is frame-dependent. *arXiv preprint arXiv:2502.04403*, 2025.

Acharya, D. B., Kuppan, K., and Divya, B. Agentic AI: Autonomous intelligence for complex goals–A comprehensive survey. *IEEE Access*, 2025.

Ackerman, R. and Thompson, V. A. Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, 21(8):607–617, 2017.

Aghion, P. and Bolton, P. An incomplete contracts approach to financial contracting. *The Review of Economic Studies*, 59(3):473–494, 1992.

Aghion, P. and Tirole, J. Formal and real authority in organizations. *Journal of Political Economy*, 105(1):1–29, 1997.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Anthropic. Computer use (beta) - build with Claude, 2024. URL https://docs.anthropic.com/en/docs/build-with-claude/computer-use. Accessed: 2024-02-09.

Anthropic. Claude code overview, 2025. URL https://docs.anthropic.com/en/docs/agents-and-tools/claude-code/overview.

Ashby, W. R. *An Introduction to Cybernetics*. Chapman & Hall, London, 1956.

Åström, K. J. and Murray, R. M. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, Princeton, NJ, 2008.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., and Frith, C. D. Optimally interacting minds. *Science*, 329(5995):1081–1085, 2010.

Baker, G., Gibbons, R., and Murphy, K. J. Relational contracts and the theory of the firm. *The Quarterly Journal of Economics*, 117(1):39–84, 2002.

Balabanov, O. and Linander, H. Uncertainty quantification in fine-tuned LLMs using LoRA ensembles. *arXiv preprint arXiv:2402.12264*, 2024.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., and Rao, R. P. Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20(4):723–742, 1997.

Balleine, B. W. and Dickinson, A. Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37(4-5): 407–419, 1998.

Bandura, A. Self-efficacy mechanism in human agency. *American Psychologist*, 37(2):122, 1982.

Bandura, A. Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, 52(1):1–26, 2001.

Bang, D. and Fleming, S. M. Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, 115(23): 6082–6087, 2018.

Bansal, G., Vaughan, J. W., Amershi, S., Horvitz, E., Fourney, A., Mozannar, H., Dibia, V., and Weld, D. S. Challenges in human-agent communication. *arXiv preprint arXiv:2412.10380*, 2024.

Barandiaran, X. E. and Almendros, L. S. Transforming agency: On the mode of existence of large language models. *arXiv preprint arXiv:2407.10735*, 2024.

Barandiaran, X. E., Di Paolo, E., and Rohde, M. Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, 17(5): 367–386, 2009.

Barsalou, L. Perceptual symbol systems. *The Behavioral and Brain Sciences*, 1999.

Barto, A. G. Intrinsic motivation and reinforcement learning. *Intrinsically Motivated Learning in Natural and Artificial Systems*, pp. 17–47, 2013.

Bellman, R. *Dynamic Programming*. Princeton University Press, 1957.

Boulding, K. E. *The Organizational Revolution: A Study in the Ethics of Economic Organization*. Harper & Brothers, New York, 1953.

Bratman, M. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA, 1987.

Burton, J. W., Lopez-Lopez, E., Hechtlinger, S., Rahwan, Z., Aeschbach, S., Bakker, M. A., Becker, J. A., Berditchevskaia, A., Berger, J., Brinkmann, L., et al. How large language models can reshape collective intelligence. *Nature Human Behaviour*, 8(9):1643–1655, 2024.

Bushnell, E. W. A dual-processing approach to cross-modal matching: Implications for development. In *The Development of Intersensory Perception*, pp. 19–38. Psychology Press, 2013.

Cangelosi, A. and Schlesinger, M. *Developmental Robotics: From Babies to Robots*. MIT Press, 2015.

Chalmers, D. J. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford Paperbacks, 1997.

Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., et al. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 651–666, 2023.

Chawla, C., Chatterjee, S., Gadadinni, S. S., Verma, P., and Banerjee, S. Agentic AI: The building blocks of sophisticated AI business applications. *Journal of AI, Robotics & Workplace Automation*, 3(3):1–15, 2024.

Clark, A. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

Da Costa, L., Tenka, S., Zhao, D., and Sajid, N. Active inference as a model of agency. *arXiv preprint arXiv:2401.12917*, 2024.

Davidson, D. and Pollock, F. Essays on actions and events. 2001.

Dennett, D. C. *The Intentional Stance*. MIT press, 1989.

Denning, S. In the digital age, the combination of technology and radical management practices drive competitive advantage. *Strategy & Leadership*, 50(2):9–14, 2022.

Di Paolo, E. A. Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences*, 4(4):429–452, 2005.

Dolan, R. J. and Dayan, P. Goals and habits in the brain. *Neuron*, 80(2):312–325, 2013.

Edelman, G. M. *Neural Darwinism*. Basic Books, New York, 1987.

Ellis, G. F. Top-down causation and emergence: Some comments on mechanisms. *Interface Focus*, 2(1):126–140, 2012.

Emirbayer, M. and Mische, A. What is agency? *American Journal of Sociology*, 103(4):962–1023, 1998.

Fields, C. and Levin, M. Competency in navigating arbitrary spaces as an invariant for analyzing cognition in diverse embodiments. *Entropy*, 24(6):819, 2022.

Fleming, S. M. and Daw, N. D. Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, 124(1):91, 2017.

Franklin, S. and Graesser, A. Is it an agent, or just a program? A taxonomy for autonomous agents. In *International Workshop on Agent Theories, Architectures, and Languages*, pp. 21–35. Springer, 1996.

Friston, K. Learning and inference in the brain. *Neural Networks*, 16(9):1325–1352, 2003.

Friston, K. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

Friston, K. J. and Stephan, K. E. Free-energy and the brain. *Synthese*, 159:417–458, 2007.

Friston, K. J., Daunizeau, J., Kilner, J., and Kiebel, S. J. Action and behavior: a free-energy formulation. *Biological Cybernetics*, 102:227–260, 2010.

Frith, C. D. and Frith, U. Mechanisms of social cognition. *Annual Review of Psychology*, 63(1):287–313, 2012.

Giddens, A. *The Constitution of Society: Outline of the Theory of Structuration*. University of California Press, 1984.

Gollwitzer, P. M. Implementation intentions: Strong effects of simple plans. *American Psychologist*, 54(7):493, 1999.

Google. Project Mariner, 2025. URL https://deepmind.google/technologies/project-mariner/.

Gopnik, A. Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B*, 375(1803):20190502, 2020.

Gopnik, A. and Wellman, H. M. Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, 138(6):1085, 2012.

Gopnik, A., Meltzoff, A. N., and Kuhl, P. K. *The Scientist in the Crib: Minds, Brains, and How Children Learn*. William Morrow & Co, 1999.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

Griot, M., Hemptinne, C., Vanderdonckt, J., and Yuksel, D. Large language models lack essential metacognition for reliable medical reasoning. *Nature Communications*, 16 (1):642, 2025.

Grossman, S. J. and Hart, O. D. The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy*, 94(4):691–719, 1986.

Hart, O. *Firms, Contracts, and Financial Structure*. Clarendon Press, 1995.

Hart, O. and Moore, J. Property rights and the nature of the firm. *Journal of Political Economy*, 98(6):1119–1158, 1990.

Hauskrecht, M. Value-function approximations for partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, 13:33–94, 2000.

Hoel, E. P., Albantakis, L., and Tononi, G. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49): 19790–19795, 2013.

Hohwy, J. *The Predictive Mind*. Oxford University Press, 2013.

Holland, J. H. Complex adaptive systems. *Daedalus*, 121 (1):17–30, 1992.

Hu, S., Lu, C., and Clune, J. Automated design of agentic systems. *arXiv preprint arXiv:2408.08435*, 2024.

Huang, J., Yong, S., Ma, X., Linghu, X., Li, P., Wang, Y., Li, Q., Zhu, S.-C., Jia, B., and Huang, S. An embodied generalist agent in 3D world. *arXiv preprint arXiv:2311.12871*, 2023a.

Huang, K., Qu, Y., Cousins, H., Johnson, W. A., Yin, D., Shah, M., Zhou, D., Altman, R., Wang, M., and Cong, L. Crispr-GPT: An LLM agent for automated design of gene-editing experiments. *arXiv preprint arXiv:2404.18021*, 2024a.

Huang, Q., Wake, N., Sarkar, B., Durante, Z., Gong, R., Taori, R., Noda, Y., Terzopoulos, D., Kuno, N., Famoti, A., et al. Position paper: Agent AI towards a holistic intelligence. *arXiv preprint arXiv:2403.00833*, 2024b.

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Patra, B., et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023b.

Hutchins, E. *Cognition in the Wild*. MIT press, 1995.

Jin, Z., Liu, J., Lyu, Z., Poff, S., Sachan, M., Mihalcea, R., Diab, M., and Schölkopf, B. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*, 2023.

Johnson, S. G., Karimi, A.-H., Bengio, Y., Chater, N., Gerstenberg, T., Larson, K., Levine, S., Mitchell, M., Rahwan, I., Schölkopf, B., et al. Imagining and building wise machines: The centrality of AI metacognition. *arXiv preprint arXiv:2411.02478*, 2024.

Kapkiç, A., Mandal, P., Wan, S., Sheth, P., Gorantla, A., Choi, Y., Liu, H., and Candan, K. S. Introducing causalbench: A flexible benchmark framework for causal analysis and machine learning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 5220–5224, 2024.

Kenton, Z., Kumar, R., Farquhar, S., Richens, J., MacDermott, M., and Everitt, T. Discovering agents. *Artificial Intelligence*, 322:103963, 2023.

Kolt, N. Governing AI agents. *Available at SSRN*, 2024.

Kolt, N., Shur-Ofry, M., and Cohen, R. Lessons from complexity theory for AI governance. *arXiv preprint arXiv:2502.00012*, 2025.

Kuhn, T. S. *The Structure of Scientific Revolutions*. University of Chicago Press, 1997.

Kumar, P. R. and Varaiya, P. *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Prentice-Hall, 1986.

Lai, C.-H., Lin, S. H., et al. Systems theory. *The International Encyclopedia of Organizational Communication*, 41:1–18, 2017.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

Levin, M. The computational boundary of a "self": development, cancer and the continuum of embryonic regulatory networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1774):20180376, 2019.

Li, K., Wang, J., Yang, L., Lu, C., and Dai, B. SemGrasp: Semantic grasp generation via language aligned discretization. In *European Conference on Computer Vision*, pp. 109–127. Springer, 2025.

Li, X., Wang, S., Zeng, S., Wu, Y., and Yang, Y. A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9, 2024a.

Li, Y., Huang, Y., Lin, Y., Wu, S., Wan, Y., and Sun, L. I think, therefore I am: Awareness in large language models. *arXiv preprint arXiv:2401.17882*, 2024b.

Lieder, F. and Griffiths, T. L. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:1–60, 2020.

Lin, Z., Trivedi, S., and Sun, J. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Maes, P. Modeling adaptive autonomous agents. *Artificial Life*, 1(1_2):135–162, 1993.

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., and Hobbhahn, M. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*, 2024.

Merleau-Ponty, M. *Phénoménologie de la perception*. Gallimard, Paris, 1945.

Mitchell, M., Ghosh, A., Luccioni, A. S., and Pistilli, G. AI agents are here. What now?, 1 2025a. URL https://huggingface.co/blog/ethics-soc-7.

Mitchell, M., Ghosh, A., Luccioni, A. S., and Pistilli, G. Fully autonomous AI agents should not be developed. *arXiv preprint arXiv:2502.02649*, 2025b.

Mitelut, C., Smith, B., and Vamplew, P. Position: Intent-aligned AI systems must optimize for agency preservation. In *Forty-first International Conference on Machine Learning*, 2024.

Narayanan, S., Braza, J. D., Griffiths, R.-R., Ponnapati, M., Bou, A., Laurent, J., Kabeli, O., Wellawatte, G., Cox, S., Rodriques, S. G., et al. Aviary: Training language agents on challenging scientific tasks. *arXiv preprint arXiv:2412.21154*, 2024.

Nelson, T. O. Consciousness and metacognition. *American Psychologist*, 51(2):102, 1996.

Niu, N., Wu, Y., Li, H., Li, M., Yang, D., Fan, W., and Zhong, Y. Influence of voluntary action and outcome valence on the sense of agency. *Frontiers in Human Neuroscience*, 17:1206858, 2023.

OpenAI. Introducing Operator, 1 2025. URL https://openai.com/index/introducing-operator/. A research preview of an agent that can use its own browser to perform tasks for you. Available to Pro users in the U.S.

OpenAI, Andrychowicz, M., Baker, B., Chociej, M., Józefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., Schneider, J., Sidor, S., Tobin, J., Welinder, P., Weng, L., and Zaremba, W. Learning dexterous in-hand manipulation. *CoRR*, 2018. URL http://arxiv.org/abs/1808.00177.

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2): 265–286, 2007.

Pearl, J. *Causality*. Cambridge university press, 2009.

Pedreschi, D., Pappalardo, L., Ferragina, E., Baeza-Yates, R., Barabási, A.-L., Dignum, F., Dignum, V., Eliassi-Rad, T., Giannotti, F., Kertész, J., et al. Human-AI coevolution. *Artificial Intelligence*, pp. 104244, 2024.

Piaget, J. *The Origins of Intelligence in Children*. International University Press, Inc., New York, 1952.

Qiu, J., Lam, K., Li, G., Acharya, A., Wong, T. Y., Darzi, A., Yuan, W., and Topol, E. J. LLM-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6(12):1418–1420, 2024.

Rao, A., Khandelwal, A., Tanmay, K., Agarwal, U., and Choudhury, M. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. *arXiv preprint arXiv:2310.07251*, 2023.

Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J. T., et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

Richens, J. and Everitt, T. Robust agents learn causal world models. *arXiv preprint arXiv:2402.10877*, 2024.

Romanou, A., Montariol, S., Paul, D., Laugier, L., Aberer, K., and Bosselut, A. CRAB: Assessing the strength of causal relationships between real-world events. *arXiv preprint arXiv:2311.04284*, 2023.

Rosenblueth, A., Wiener, N., and Bigelow, J. Behavior, purpose and teleology. *Philosophy of Science*, 10(1): 18–24, 1943.

Rouleau, N. and Levin, M. Discussions of machine versus living intelligence need more clarity. *Nature Machine Intelligence*, 6(12):1424–1426, 2024.

Russell, S. J. and Norvig, P. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1995. ISBN 9780131038059.

Schlosser, M. Agency. https://plato.stanford.edu/entries/agency/, October 2019. Accessed: 2025-01-06.

Scholten, F., Rebholz, T. R., and Hütter, M. Metacognitive myopia in large language models. *arXiv preprint arXiv:2408.05568*, 2024.

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., et al. Practices for governing agentic AI systems. *Research Paper, OpenAI, December*, 2023.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., and Frith, C. D. Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, 18(4):186–193, 2014.

Shen, H., Knearem, T., Ghosh, R., Alkiek, K., Krishna, K., Liu, Y., Ma, Z., Petridis, S., Peng, Y.-H., Qiwei, L., et al. Towards bidirectional human-AI alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264*, 2024.

Shorinwa, O., Mei, Z., Lidard, J., Ren, A. Z., and Majumdar, A. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *arXiv preprint arXiv:2412.05563*, 2024.

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

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of Go without human knowledge. *nature*, 550(7676):354–359, 2017.

Simon, H. A. *Models of Man: Social and Rational - Mathematical Essays on Rational Human Behavior in a Social Setting*. John Wiley and Sons, New York, 1957.

Simon, H. A. The architecture of complexity. In *The Roots of Logistics*, pp. 335–361. Springer, 2012.

Singh, S., Lewis, R. L., Barto, A. G., and Sorg, J. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2):70–82, 2010.

Singh, S. P., Jaakkola, T., and Jordan, M. I. Learning without state-estimation in partially observable Markovian decision processes. In *Machine Learning Proceedings 1994*, pp. 284–292. Elsevier, 1994.

Smith, L. and Gasser, M. The development of embodied cognition: Six lessons from babies. *Artificial Life*, 11 (1-2):13–29, 2005.

Spelke, E. S. and Kinzler, K. D. Core knowledge. *Developmental Science*, 10(1):89–96, 2007.

Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., and Wang, X. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

Surowiecki, J. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, 2004.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Tavafoghi, H., Ouyang, Y., and Teneketzis, D. A sufficient information approach to decentralized decision making. In *2018 IEEE Conference on Decision and Control (CDC)*, pp. 5069–5076. IEEE, 2018.

Tavafoghi, H., Ouyang, Y., and Teneketzis, D. A unified approach to dynamic decision problems with asymmetric information: Nonstrategic agents. *IEEE Transactions on Automatic Control*, 67(3):1105–1119, 2021.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011.

Thompson, E. Living ways of sense making. *Philosophy Today*, 55(Supplement):114–123, 2011.

Thompson, V. A., Turner, J. A. P., and Pennycook, G. Intuition, reason, and metacognition. *Cognitive Psychology*, 63(3):107–140, 2011.

Thorndike, E. L. A constant error in psychological ratings. *Journal of Applied Psychology*, 4(1):25–29, 1920.

Tran, K.-T., Dao, D., Nguyen, M.-D., Pham, Q.-V., O'Sullivan, B., and Nguyen, H. D. Multi-agent collaboration mechanisms: A survey of LLMs. *arXiv preprint arXiv:2501.06322*, 2025.

Vafa, K., Chen, J. Y., Kleinberg, J., Mullainathan, S., and Rambachan, A. Evaluating the world model implicit in a generative model. *arXiv preprint arXiv:2406.03689*, 2024.

Varela, F. J. Living ways of sense-making: A middle path for neuroscience. In *Order and Disorder: Proceedings of the Stanford International Symposium*, pp. 208–224. Anma Libri, 1984.

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

Von Bertalanffy, L. *General System Theory: Foundations, Development, Applications*. George Braziller, New York, 1968.

Von Wright, G. H. *Explanation and Understanding*. Cornell University Press, 2004.

Wegner, D. Transactive memory: A contemporary analysis of the group mind. *Theories of group behavior/Springer-Verlag*, 1987.

Wiener, N. *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press, Cambridge, MA, 1948.

Wiesinger, J., Marlow, P., and Vuskovic, V. Agents. Whitepaper, Google, 2025.

Williamson, O. E. *Markets and Hierarchies: Analysis and Antitrust Implications*. Free Press, New York, 1975.

Wolf, W. and Tomasello, M. A shared intentionality account of uniquely human social bonding. *Perspectives on Psychological Science*, pp. 17456916231201795, 2023.

Xu, D. and Rivera, J.-P. Towards measuring goal-directedness in AI systems. *arXiv preprint arXiv:2410.04683*, 2024.

Xu, F. F., Song, Y., Li, B., Tang, Y., Jain, K., Bao, M., Wang, Z. Z., Zhou, X., Guo, Z., Cao, M., et al. TheAgentCompany: Benchmarking LLM agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161*, 2024.

Xu, Z., Qi, B., Agrawal, S., and Song, S. AdaGrasp: Learning an adaptive gripper-aware grasping policy. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4620–4626. IEEE, 2021.

Yang, L., Shirvaikar, V., Clivio, O., and Falck, F. A critical review of causal reasoning benchmarks for large language models. In *AAAI 2024 Workshop on "Are Large Language Models Simply Causal Parrots?"*, 2024.

Yeung, N., Botvinick, M. M., and Cohen, J. D. The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review*, 111(4):931, 2004.

Zečević, M., Willig, M., Dhami, D. S., and Kersting, K. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*, 2023.

Zhang, A. and Varshney, L. R. Conceptualizing agency: A framework for human-AI interaction. In *HAI-GEN 2025: 6th Workshop on Human-AI Co-Creation with Generative Models*, 2025.

Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., and Yu, D. MM-LLMs: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024.

Zhou, Y., Wu, X., Huang, B., Wu, J., Feng, L., and Tan, K. C. Causalbench: A comprehensive benchmark for causal learning capability of large language models. *arXiv preprint arXiv:2404.06349*, 2024.

Zhu, A., Dugan, L., and Callison-Burch, C. ReDel: A toolkit for LLM-powered recursive multi-agent systems. *arXiv preprint arXiv:2408.02248*, 2024.