Workshop Report

Through the Chat Window and Into the Real World

Preparing for AI Agents

Authors

Helen Toner*
John Bansemer*
Kyle Crichton*
Matt Burtell*
Thomas Woodside*
Anat Lior
Andrew J. Lohn
Ashwin Acharya
Beba Cibralic
Chris Painter
Cullen O'Keefe
Iason Gabriel
Kathleen Fisher
Ketan Ramakrishnan
Krystal Jackson
Noam Kolt
Rebecca Crootof
Samrat Chatterjee

***Workshop Organizers**

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

October 2024

Executive Summary

The concept of artificial intelligence systems that actively pursue goals—known as AI "agents"—is not new. But over the last year or two, progress in large language models (LLMs) has sparked a wave of excitement among AI developers about the possibility of creating **sophisticated, general-purpose AI agents** in the near future. Startups and major technology companies have announced their intent to build and sell AI agents that can act as personal assistants, virtual employees, software engineers, and more. While current systems remain somewhat rudimentary, they are improving quickly. Widespread deployment of highly capable AI agents could have transformative effects on society and the economy. This workshop report describes findings from a recent CSET-led workshop on the policy implications of increasingly "agentic" AI systems.

In the absence of a consensus definition of an "agent," we describe four characteristics of increasingly agentic AI systems: they pursue more **complex goals** in more **complex environments**, exhibiting **independent planning and adaptation** to **directly take actions** in virtual or real-world environments. These characteristics help to establish how, for example, a cyber-offense agent that could autonomously carry out a cyber intrusion would be more agentic than a chatbot advising a human hacker. A "CEO-AI" that could run a company without human intervention would likewise be more agentic than an AI acting as a personal assistant.

At present, general-purpose LLM-based agents are the subject of significant interest among AI developers and investors. These agents consist of an advanced LLM (or multimodal model) that uses "scaffolding" software to interface with external environments and tools such as a browser or code interpreter. Proof-of-concept products that can, for example, write code, order food deliveries, and help manage customer relationships are already on the market, and many relevant players believe that the coming years will see rapid progress.

In addition to the many potential benefits that AI agents will likely bring, they may also exacerbate a range of existing AI-related issues and even create new challenges. The ability of agents to pursue complex goals without human intervention could lead to more serious **accidents**; facilitate **misuse** by scammers, cybercriminals, and others; and create new challenges in **allocating responsibility** when harms materialize. Existing **data governance and privacy** issues may be heightened by developers' interest in using data to create agents that can be tailored to a specific user or context. If highly capable agents reach widespread use, users may become vulnerable to **skill fade and dependency**, agents may **collude** with one another in undesirable ways, and

significant **labor impacts** could materialize as an increasing range of currently human-performed tasks become automated.

To manage these challenges, our workshop participants discussed three categories of interventions:

1. **Measurement and evaluation:** At present, our ability to assess the capabilities and real-world impacts of AI agents is very limited. Developing better methodologies to track improvements in the capabilities of AI agents themselves, and to collect ecological data about their impacts on the world, would make it more feasible to anticipate and adapt to future progress.

2. **Technical guardrails:** Governance objectives such as **visibility**, **control**, **trustworthiness**, as well as **security and privacy** can be supported by the thoughtful design of AI agents and the technical ecosystems around them. However, there may be trade-offs between different objectives. For example, many mechanisms that would promote visibility into and control over the operations of AI agents may be in tension with design choices that would prioritize privacy and security.

3. **Legal guardrails:** Many existing areas of law—including agency law, corporate law, contract law, criminal law, tort law, property law, and insurance law—will play a role in how the impacts of AI agents are managed. Areas where contention may arise when attempting to apply existing legal doctrines include questions about the **"state of mind" of AI agents**, the **legal personhood of AI agents**, how **industry standards** could be used to evaluate negligence, and how existing **principal-agent frameworks** should apply in situations involving AI agents.

While it is far from clear how AI agents will develop, the level of interest and investment in this technology from AI developers means that policymakers should understand the potential implications and intervention points. For now, valuable steps could include improving measurement and evaluation of AI agents' capabilities and impacts, deeper consideration of how technical guardrails can support multiple governance objectives, and analysis of how existing legal doctrines may need to be adjusted or updated to handle more sophisticated AI agents.

Table of Contents

Introduction and Scope

Computer scientists have long sought to build computers that can actively pursue goals in the world—commonly referred to as artificial intelligence (AI) "agents." As early as 1950, information theory pioneer Claude Shannon showed how a machine could act on the basis of information that it had "learned" and "remembered" by building a metal mouse that could use trial and error to find its way through a novel maze.[1] Basic AI agents that can take actions in constrained, simple environments such as board games have existed for decades, and over time the complexity of the environments that they can successfully operate in has grown.[2]

Recently, significant progress in large language models (LLMs, the type of AI that powers ChatGPT and similar systems) has fueled new optimism about the prospects for AI agents. Many researchers see the near-term development of highly sophisticated, flexible, LLM-based agents that can adaptively pursue complex real-world goals as a serious possibility.[3] The demand for such systems will likely be enormous: capable and reliable AI agents could help individual users with work or leisure tasks, replace or augment workers in a wide range of industries, and, more generally, vastly expand the boundaries of what computers can usefully do. With sophisticated agents likely on the horizon and current agent-based products already on the market, now is the time for policymakers to begin grappling with the many questions they raise. What societal impacts and potential risks do they bring, and what guardrails would promote benefits and minimize harms?

In May 2024, the Center for Security and Emerging Technology (CSET) hosted a workshop to explore recent developments in efforts to build AI agents, as well as the policy implications if this technology continues to progress. The workshop brought together participants from industry, academia, government, and civil society, with expertise on AI development, policy, and evaluations, as well as law, cybersecurity, and other domains. This workshop report synthesizes the key themes and conclusions of the workshop, including participants' thoughts on what constitutes an AI agent, current and future developments in this space, risks that agents introduce, and potential tools for harm mitigation.

What Is an AI Agent?

There is no consensus on when an AI system should be classified as an "agent." Rather than seeking to draw a single dividing line between systems that are and are not agents, this report recommends thinking of a cluster of properties that can be present to greater or lesser degrees, which together determine how "agentic" an AI system is

and in what ways. This way of describing agents is intended to gesture towards a loose category of AI systems being developed in practice, not to provide a definitive answer to long-running philosophical debates on what constitutes agency.[4] Drawing on previous work, we suggest the following characteristics that make an AI system more agentic: [5]

- **Goal complexity:** More agentic systems pursue complex, longer-term goals—or even a variety of different goals. Less agentic systems carry out individual, more explicitly defined tasks.

- **Environment complexity:** More agentic systems can operate effectively in more open-ended and complicated settings, where the number of possible states and actions available to the agent is larger and the dynamics governing what will happen next in the environment are more difficult to model. Less agentic systems can operate effectively only in simpler and more predictable settings.

- **Independent planning and adaptation:** More agentic systems can generate their own plan or pathway to meet the intended goal, adapting as needed to changing circumstances. Less agentic systems follow pre-specified step-by-step instructions.

- **Direct action:** More agentic systems take action directly in their environment (whether real or virtual). Less agentic systems provide information or recommendations for a human user to act on.

Box 1 illustrates this framework's characteristics by describing increasingly agentic versions of AI systems built for different purposes. Some of the systems described already exist; others are—at present—only hypothetical.

Box 1: Increasingly Agentic AI Systems in Different Application Areas

Playing games: Chess → Atari → Starcraft II

Game-playing AI agents have long been developed and used. On the characteristic of **direct action**, game-playing systems could be considered moderate, since they are designed to play autonomously (with no human intervention), but always in sandboxed environments that are walled off from the real world. Depending on design, they are generally high in **independent planning and adaptation** since they do not follow a fixed, human-specified set of steps to play the game. **Goal** and **environment complexity** vary significantly depending on the game, though games are generally far simpler than real-world environments.

According to this report's framework, an AI system that can play *Starcraft II*—a highly sophisticated strategy video game—is more agentic than one that plays multiple Atari games (such as *Breakout* and *Space Invaders*), which is more agentic than a simple chess engine (such as Deep Blue, which beat world champion Garry Kasparov in 1997). The primary characteristics that differ between these examples are **goal** and **environment complexity**: *Starcraft II*'s freewheeling play is significantly more complicated than chess's 64 squares. 1990s-style chess engines also score lower on **independent planning and adaptation** than Atari- and Starcraft-playing agents from the 2010s because their design includes a database of hand-specified openings and endgames that reduces the need to adapt on the fly.[6]

Cyber offense: Advising user → autonomously winning 'Capture the Flag' → autonomously carrying out a real intrusion

Automation of various kinds already plays a major role in cyber operations on both the offensive and defensive sides, but so far most of this automation is relatively simple. One example of a relatively non-agentic LLM-based system for cyber offense would be a chatbot that can advise a human user on how to carry out a hacking task (e.g., by generating code snippets or suggesting tactics to try). Such systems (which already exist[7]) would score low on **direct action** since their effect on the world is entirely mediated by what the human user chooses to put into practice. A system that could autonomously plan and carry out a full cyber operation in response to a high-level instruction from a human operator would be far more agentic, scoring highly on all four characteristics of the framework. An intermediate example would be a system that could autonomously win a staged Capture the Flag contest,[8] where both **goal** and **environment complexity** would be lower, as would **direct action**, due

to the constrained and simplified nature of a contest setting.

Corporate assistance: Personal assistant → factory manager → 'CEO-AI'

Startups and big tech companies alike are racing to create AI personal assistants, with early prototypes capable of tasks such as ordering food for delivery or emailing to inquire about a Craigslist listing.[9] Booking a flight or coordinating among meeting participants to schedule a time are other commonly listed tasks that a simple "personal assistant" agent could carry out. This report's framework would classify such a system as moderately agentic, with relatively high capacity for **direct action** but only moderate **independent planning and adaptation,** as well as **goal complexity.** The **environment complexity** would depend on whether the agent was constrained to act via structured channels (e.g., a dedicated set of APIs to send scheduling emails, book specific services, etc.) or whether it had free access to the open internet (perhaps via its user's internet browser). The latter three characteristics would all be higher for, say, an agent that could autonomously manage the operation of a stand-alone factory, including setting production schedules, managing inventory, and optimizing workflows. Higher still would be a CEO-AI that can autonomously run an entire business, from determining the company's key objectives, to allocating resources, to negotiating deals with suppliers.[10] To succeed at this, an AI system would need to score highly on all four characteristics of the framework.

There may not be a clear boundary between agents whose actions are constrained purely to an online environment and those that affect the real world. A bot managing the inventory of a store could be limited to internet-based actions, but these could include hiring humans via online hiring platforms or ordering physical goods to be delivered to a specific location. And even when actions are "purely" digital, they can have significant implications for people's welfare and rights, as any victim of cybercrime or online harassment can tell you. The ability to directly affect the real world thus does not depend on whether an AI agent is connected to robotic actuators.

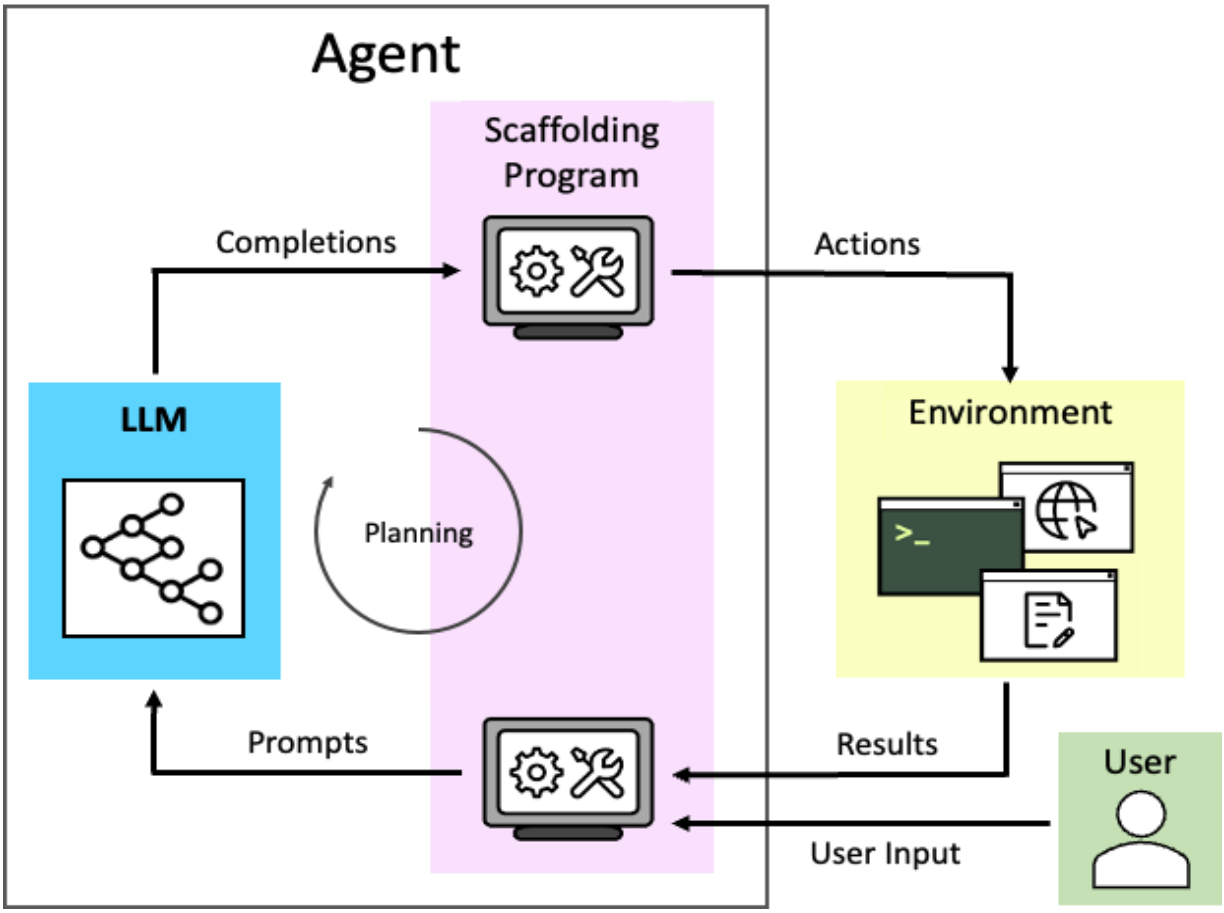
Technological Trajectories

The Current State of AI Agents

A fresh wave of excitement around AI agents kicked off in 2023. In the wake of significant progress in LLMs, a handful of scrappy open-source projects with names like AutoGPT and BabyAGI came up with ways to build software wrappers around an LLM to turn it from a chatbot into an agent. Rather than conversing with a chatbot that can advise you on the steps that you might take to accomplish a goal, this kind of agent is designed to be given a goal directly (e.g., "Create a pitch deck for a new startup idea and send it to five relevant investors"), then generate and execute a plan to achieve it—including taking actions like navigating the internet or running code. The agent carries on until it either achieves the goal or—far more often—gets stuck. These early LLM-based agents have not been practically useful due to high error rates, but they were proof of concept for an important point: that the gap between generating text and taking actions may be much smaller than previously assumed.

The basic setup for LLM-based agents of this kind consists of so-called "scaffolding" software built around an LLM or multimodal model (we refer to "LLMs" throughout for simplicity). This scaffolding is usually a simple piece of software that acts as an interface between the model and the external world, automatically generating prompts to feed into the model (e.g., instructions to operate a web browser or run code) and then converting model outputs into formats that can interact with external systems. In this way, the scaffolding allows an LLM to interact with a wide range of online tools and services.[11] See Figure 1 for a depiction of how one research group constructed an agent to test its ability to carry out risky behaviors; in this depiction, the LLM API plus scaffolding program comprise the agent.

Figure 1: Depiction of an LLM-Plus-Scaffolding Style Agent.*



Source: Adapted from Megan Kinniment et al., "Evaluating Language-Model Agents on Realistic Autonomous Tasks," arXiv preprint arXiv:2312.11671 (2023), https://arxiv.org/abs/2312.11671.

Given the limitations of these early LLM-based AI agents, it may be natural to wonder why computer scientists wishing to build agents wanted to do so atop LLMs, especially since LLMs were not designed to be agentic. The reasons are often unarticulated. The excitement probably comes from several key capabilities that an ideal AI agent would need and that LLMs seem to possess to some degree, such as:

* Readers familiar with reinforcement learning will notice that the setup in Figure 1 is compatible with the classic RL agent paradigm: an "agent" takes "actions" in an "environment," and then the results of those actions (classically called "state") are conveyed back to the agent.

1. Rich, built-in knowledge of many features of the world.[12]
2. The ability to receive, understand, and follow instructions from users in natural language.
3. The ability to formulate plans to achieve open-ended goals.[13]
4. Native ability to use modalities of action commonly used by humans when conducting business on a computer (e.g., using email or other messaging services, browsing the internet,[14] or employing scratchpads[15]).

LLMs can do these things across many more domains and with greater reliability than previous generations of AI systems. This seems to have led many computer scientists to believe that, by continuously expanding the range of "actions" LLMs can take, their competence at text-based reasoning and writing could be expanded to competence at computer-based actions.

As of mid-2024, many major AI companies are explicitly working on turning their chatbots into agents. In competing announcements at their respective May 2024 developer conferences, Microsoft promised AI tools that "act as independent agents . . . with more autonomy and less human intervention," while Google previewed an agent that could work across different apps to autonomously carry out a task like returning a pair of shoes—finding the email receipt, filling out an online return form, and scheduling a pickup.[16] Startups such as Adept, MultiOn, and Lindy have also raised hundreds of millions of dollars on the promise of building AI agents that can flexibly carry out complex tasks.[17]

These agents are still quite limited. Promotional videos show agents performing tasks like adding a new lead into Salesforce or ordering a burger on DoorDash. And when tested in practice, even these simple tasks can prove too difficult. During the workshop, we tested one agent by asking it to reserve a particular book at a particular Washington, D.C. library branch; it reserved the wrong book at the wrong branch. (In earlier tests, the same agent had once reserved the right book at the wrong branch, and once succeeded in reserving the right book at the right branch.) In general, if the agent has some chance of failing or getting stuck at each step, then the more steps required to carry out a task to completion, the more likely it is that the agent will not succeed.[18] Given the high chance of failure at each step, AI agents are not particularly useful today, and widespread adoption does not seem likely without further advances.

Looking Over the Horizon

Workshop discussions about what to expect in the development and proliferation of AI agents highlighted several important uncertainties, including:

How quickly will the sophistication and usefulness of AI agents improve? Many researchers, executives, and startup founders are betting that the current unreliability of AI agents is a temporary state of affairs. Several different potential research approaches may—separately or together—make agents significantly more capable. When it comes to LLM-based agents, recent performance gains have been driven primarily by improvements in the underlying LLMs and multimodal models such as GPT-4 and Gemini 1.5.[19] If the trend continues in which each new generation of these underlying models is more advanced than the last, then this will provide increasingly powerful underlying "engines" to power AI agents. AI developers are also working to improve agent-specific training schemes, scaffolding software, and other infrastructure that can be layered on top of LLMs to make better use of the underlying models. In parallel, some researchers are also working on new approaches to improve LLMs' reasoning abilities, while others are building general-purpose agents from scratch (i.e., not starting with an LLM).[20] Views differ on the likelihood of these different avenues leading to more capable AI agents in the near term, but the level of investment in AI agents by major companies and startups alike indicates that some well-resourced actors see it as a serious possibility.

Will agents be general-purpose or designed for specific use cases? The utility of and market demand for general-purpose assistants seem clear, and recent trends have suggested that we may continue to see increasingly sophisticated foundation models that could enable capable general-purpose agents. However, given current agents' performance limitations, it may be more feasible in the short term to develop more narrowly scoped agents that can pursue a more limited set of goals in more constrained environments. If this is the case, the first widespread adoption of AI agents may instead occur for specific use cases or within sectors.

Which use cases are most likely to succeed first? If it does turn out that building general-purpose agents is less feasible than agents focused on a specific use case, what types of focused agents are likely to proliferate?

Workshop participants noted that it is generally easier to train and fine-tune agent behavior in areas—like software engineering[21]—where it is relatively easy to build feedback loops or verification cycles to rapidly generate data about when an agent is performing well or poorly. In addition to software engineering, we should expect to see

higher adoption rates in other purely software-based domains, such as cybersecurity, and use cases with short turnaround cycles, such as customer support (where the customer can provide a satisfaction rating after only a few minutes of interaction). In contrast to existing chatbots for software engineering and customer support, increasingly agentic systems in these domains could go beyond the current paradigm of providing responses in a chat window. For software engineering, they might directly interact with a codebase or set up web servers and other digital infrastructure. For customer support, more agentic systems might have the ability to directly manage features of a customer's account, rather than merely providing advice or escalating to a human agent.

Economic incentives will likely also affect which types of AI agents become widespread. One relevant factor is the time, expense, and risk of using AI agents in high-regulation environments. Companies may initially prefer employing AI agents in lower-regulation environments because it is, presumably, less costly to do so. Developers of AI agents will presumably also weigh the potential market size for new agents they could offer against the cost of development. This may mean that agents purpose-built for major industries become widespread sooner than for other use cases.

What business model(s) and market structure(s) will predominate? At present, the industry for the most advanced LLMs is concentrated among a handful of major players, while other parts of the AI industry (e.g., business process automation, computer vision, and robotics) are more dispersed. The impacts and governance options for AI agents will depend, to a significant extent, on the market for these systems. One possibility, if LLM-based agents become widespread, is that a small number of LLM companies will sell access to agents developed fully in-house. Another is that a two-tiered structure may emerge, where some companies develop advanced LLMs and others build agents on top of them. Yet another possible structure, if the current leading LLM companies do not maintain a competitive edge over open-source options, could be for individuals and companies to use AI agents that are freely available. Still other possibilities come into play if future advances in AI agents are not based on LLMs at all. Each of these possibilities—or others not enumerated here—would lead to different distributions of power and governance responsibilities. Likewise, the technology itself and associated governance options will look different depending on whether consumer subscriptions, enterprise deals, agent-as-a-service models, or other business models prevail.

While it may be impossible to accurately identify the answers to the questions raised in this section now, it is worth exploring them and their associated implications, as policymakers and other actors can take steps now that will influence the likelihood of

different futures. Given recent progress in building more agentic systems, and the time and money that incumbents and startups alike are pouring into this technology, we should expect AI systems to increasingly move out of contained environments (such as games and chatbot windows) and into the real world.

Opportunities, Risks, and Other Impacts

Just as there is no clean, sharp distinction between AI "agents" and other kinds of AI systems, the benefits and risks posed by AI agents are in many cases continuations of those associated with other kinds of AI.[22]

The opportunities that progress in AI agents could bring are manifold, from increasing the productivity of business operations to empowering individual users. We expect the commercial incentives to pursue the development of AI agents to be correspondingly strong, and we anticipate that market forces will ensure that AI agents are developed to the extent they can be productively used. Workshop discussions therefore focused on where there may be a role for intervention by policymakers or other actors to manage potential downsides and market failures. What follows should not be viewed as reflecting participants' views on the balance of benefits and risks associated with AI agents.

In addition to existing challenges raised by AI, agents' characteristics—including their ability to act without human intervention and the potential to build personal relationships with their users—introduce new potential vectors of harm. An illustrative rather than exhaustive list of concerns that are likely to emerge or intensify with the increasing use of AI agents includes:

- **Accidents.** Any system that contains complex interactions between subsystems or interacts in complex ways with external systems—as any sophisticated AI agent would—is vulnerable to unintended failure.[23] This challenge is intensified by the fact that most of today's most advanced AI systems, including existing AI agents, are built using a type of machine learning called deep learning. Experts struggle to fully understand and control the behavior of deep learning systems, which are composed of enormous numbers of learned statistical parameters rather than hand-programmed rules.[24] These challenges will likely carry over to deep learning-based agents. This means, for instance, that developers could have limited ability to guarantee that an agent will act in a certain way, and if an agent does go off the rails, the developer may have limited ability to explain why it did so.[25] The greater the extent to which an agent interacts in complex

ways with other systems—financial transactions, internal company processes, etc.—the greater the surface area for potential failures.

- **Misuse.** Malicious actors use existing AI tools to cause harm in many different ways, from generating nonconsensual sexual imagery, to scamming victims into making money transfers, to using AI tools to aid cyberattacks. The ability of agents to automate previously human-driven schemes could significantly lower the barriers to, and increase the scale of, harmful activities. For example, the development of more advanced AI agents could enable less technically skilled actors to conduct cyberattacks or allow more sophisticated actors to find and exploit vulnerabilities autonomously—reducing the already narrow window that defenders have to react to an intrusion.[26]

- **Allocation of responsibility.** How to allocate responsibility and legal liability among different actors in the AI value chain is already a contentious question, and the development of AI agents that flexibly pursue complex goals in open-ended settings adds a new twist. It may seem intuitively appealing to treat the agent itself as responsible for harm caused, especially when there is something uncomfortable about holding another entity responsible—say, if the agent causes harm by pursuing a user-specified goal in a way that the user could not anticipate. The legal questions implicated here are complex, and we explore them further in the next section.

- **Data governance and privacy.** Most modern AI systems are trained on vast datasets. Existing AI systems for ad targeting, credit scoring, and other purposes often draw on personal data that was collected, repurposed, and sold without users' knowledge or understanding.[27] Widespread, personalized AI assistants would likely collect and store additional data on each user—both because that would make them more useful and because that data could be monetized.[28] Much of this data would likely be similarly private and personal to a user's search engine history. Plausibly, it could include even more sensitive data if the AI agent is designed to foster a trusting one-on-one relationship with a specific user that relies on retaining a long interaction history (making users less likely to clear old data) and taking action on potentially sensitive or confidential information on the users' behalf. Where and how this data is stored (e.g., on-device versus in the cloud) will have significant implications for who can access the data. If the company behind the agent has access to this data under its terms and conditions of use, then the data will likely also be available to law enforcement (if requested pursuant to a warrant), hackers (if they infiltrate the

company's servers), or anyone with the interest and ability to purchase it (if the company retains the right to resell the data).[29]

- **Skill fade, dependency and vulnerability.** As users lose or lack experience with accomplishing certain tasks, they may grow increasingly dependent on AI agents. And, to the extent that AI agents are designed to have a user build an ongoing relationship with a specific agent, users may become dependent on them to a much greater degree than on more generic AI tools.[30] Dependency, in turn, makes users more vulnerable to their agents—and thereby to the companies providing them. Many online experiences are already designed to induce users to do things they would prefer not to if fully informed, such as signing up for subscriptions or sharing data.[31] Widespread use of AI agents that users feel a personal connection to could radically expand the possibilities for how users can be manipulated and make this influence harder to detect, as it would likely be occurring in private user-agent interactions rather than on public websites.

- **Collusion.** The issue of "algorithmic price fixing," in which simple automated systems interact in a way that leads to illegal collusion, has recently garnered attention from the U.S. Department of Justice, the Federal Trade Commission, and foreign governments.[32] This form of automated collusion is quite basic, but the widespread use of sophisticated AI agents could expand the range of collusive behaviors that could occur, especially if agents can communicate and coordinate with one another in ways that are not easily observable to humans.

- **Labor impacts.** The widespread availability of AI agents could lead to a glut of inexpensive labor, capable of handling many tasks of low-to-medium complexity. This would likely have a wide range of disruptive and unpredictable effects on society, the workforce, and the economy. In particular, an abundance of agent labor may enable the scaling of small tasks or operations beyond what was previously feasible due to cost or labor constraints. This could unlock enormous potential for good (e.g., scaling scientific experiments) or bad (e.g., expanding scam or misinformation campaigns), and in either case could lead to the displacement or elimination of large numbers of jobs.

Guardrails and Intervention Points

The rise of AI agents raises multiple thorny questions for policymakers and regulators that are compounded by the uncertainty about whether, how, and when AI agents will proliferate. Technical, legal, and other guardrails will play a critical role in helping to manage these issues.

It is valuable to ground any consideration of guardrails or interventions by first identifying the **governance objectives** that the guardrails are intended to serve. Developing and prioritizing a complete framework of governance objectives would go beyond the scope of the workshop, but potentially valuable objectives we identified include **supporting innovation**, **visibility**, **control**, **trustworthiness**, and **security and privacy**. The bulk of this section explores some options for how technical and legal guardrails can further some of these objectives. Before discussing those options, we briefly outline how the limitations of current methods for assessing AI agents exacerbate the challenges in question.

Evaluating Agents and Their Impacts

Given the level of uncertainty about how quickly the sophistication and usage of AI agents will grow, workshop participants shared an interest in improving methods for evaluating agent performance. Being able to measure and monitor how well agents perform in different settings over time would make it far easier to track and anticipate progress, and thereby to prepare for the potential impacts of AI agents.

This motivation to improve evaluation techniques for AI agents echoes the widely recognized importance—and difficulty—of measuring the performance of general-purpose AI systems.[33] The U.S. and UK governments are dedicating significant resources to improving the science of evaluating AI models via their respective AI Safety Institutes, in addition to efforts underway within tech companies, civil society organizations, and elsewhere to develop more effective and scientific evaluation techniques.[34]

Unfortunately, agents again add new challenges to an already formidable task.[35] For instance, many existing and anticipated evaluations focus primarily on evaluating the outputs from a standalone model. The appropriateness of this approach is already debated for non-agentic systems.[36] For AI agents, which are distinguished by their autonomous interactions with an external environment, model-focused evaluations are insufficient. Without more sophisticated testing environments that include factors like interaction with other systems, the primary means by which the risks and capabilities

of these systems will be discovered will be from their actions in the real world. If something goes wrong, it will go wrong without the safety net of a simulated environment.

A separate approach that could complement efforts to directly evaluate AI agents would be to gather ecosystem-level data on agents' impacts. This idea is akin to how ecological monitoring for environmental pollutants can complement manufacturers' obligations to measure and manage the chemicals released from their plants. Examples of measures that could be collected to shed light on the adoption and impacts of AI agents include data on:

- Labor impacts (e.g., layoffs or workforce reductions in sectors considered vulnerable to automation by AI agents).
- Adoption rates of AI agents (e.g., via surveys of individual consumers or business users).
- Agent actions as a proportion of web traffic, to the extent that agent behavior online can be distinguished from human behavior.
- Use cases and usage volume of AI agents (if provided in aggregated, anonymized form by AI developers).
- The degree to which AI companies are using AI agents to automate research and development internally (given that this is seen as one especially disruptive way AI agents might be applied).[37]

Benchmarks, which are sets of tests that are commonly used to evaluate and compare the performance of different AI models, could serve as an important tool in tracking progress as well as incentivizing development towards the goals outlined in the previous section. As benchmark scores are typically publicly available, leaderboards provide an incentive for developers to create models that score highly according to the benchmark tests. While sometimes criticized for pushing developers to tune their model specifically to these tests at the expense of other goals, the development of robust benchmarks that align with the desired properties of AI agents could be used as an effective goal-setting tool.

Technical Guardrails

Workshop participants discussed how the design of agents themselves—and the technical systems around them—can minimize harm and support positive uses.

Simplifying somewhat, we could consider three layers at which technical guardrails could be implemented: the **model**, **system**, and **ecosystem** layers. The model layer, which is primarily relevant for LLM-based agents, contains the underlying statistical model(s) (e.g., GPT-4 or Gemini 1.5) that could be considered the "engine" powering an AI agent. The system layer includes the model(s) as well as scaffolding and other components built around the model that enable it to interact with users and external tools, dictate what kinds of actions it can and cannot take, keep records of its interactions, and so forth.[38] The ecosystem is the broader space that AI agents are interacting with (e.g., payment processing infrastructure, social media platforms, etc.).

Workshop discussions focused primarily on system- and ecosystem-level guardrails. The ecosystem level is especially interesting to consider as it is not primarily shaped by those building and selling AI agents. By contrast, model-level interventions to improve the safety of LLMs have received substantial attention elsewhere, and so far appear to only be able to reduce undesirable behavior rather than preclude certain actions altogether.

We explore several options for how technical guardrails could support the aforementioned governance objectives and the trade-offs that arise between objectives. This discussion is intended as a starting point, not a definitive account, and we welcome future work on how to select, prioritize, and implement governance objectives.

Visibility refers to the ability to access information about "where, why, how, and by whom certain AI agents are used"[39] and is a prerequisite for being able to manage a wide range of risks from AI agents. Without visibility into AI agents, when something goes wrong, it will be difficult or impossible to determine what happened, who should be held responsible, or how to prevent it in the future.[40]

Technical guardrails that could increase visibility include:

- **Identification requirements.** A simple version of agent identification would be to require agents to proactively identify themselves as AI systems (rather than permitting them to impersonate humans). A more comprehensive approach could be to attach a unique identifier to each agent, which could perhaps be connected to documentation about that specific agent (e.g., a model or system card,[41] information about the agent's creator and/or operator, capabilities, purpose, etc.).[42] Unique identifiers would allow for more comprehensive tracking but also would raise additional concerns, including privacy issues (discussed further below) and practical questions (such as what counts as a "unique" agent,

which will likely depend on the business model and usage patterns of the product in question). In the absence of regulatory requirements, standards for agent identification will likely be shaped by what the tools and systems that agents interact with require for access. Therefore, industry and protocol standards within the ecosystem are likely to play an important role in shaping visibility requirements.

- **Real-time monitoring.** A significant concern with agents is their potential to have real-world impacts. Several recent research papers have proposed setting up monitoring systems that can track the activity of agents in real time and either intervene or raise the alarm if they detect certain behaviors of interest (e.g., usage policy violations, financial transactions above a certain level, unusually high compute usage, etc.).[43] Any monitoring approach would likely need to be automated, given the speed at which AI agents operate and the potential for large numbers of agents to run simultaneously. This could even extend to using AI-based systems to monitor AI agents, mirroring the common practice today of using AI to monitor AI, though this approach makes the monitoring system more complex and opaque.[44]

- **Activity logs.** A different, perhaps complementary approach to real-time monitoring is to log agent activity so that it can be examined later, to enable post-hoc incident analysis, auditing, and similar uses. These logs could be collected by the entity deploying the agent or by third-party actors whose tools the agent interacts with. A crucial question is who gets access to logged data and under what circumstances. Companies selling access to AI agents are likely to log some amount of information about agent activity in order to give their customers a better, more personalized experience. Storing this data may be helpful for the governance objective of visibility, but it also creates privacy and security concerns, which we explore further below.

Control over how AI agents behave is needed to ensure that the actions that they take are safe and appropriate. Current approaches to controlling deep learning systems, including the LLMs that underlie many promising AI agents, are quite limited due to the models' complexity and opacity and to the vast range of potential problems that must be addressed.[45] The question of how to build technical guardrails that enable control is separate from the question of who—users, developers, regulators, etc.—should be exercising control. The latter question warrants in-depth consideration, but is beyond the scope of this section.

Technical guardrails that could increase control include:

- **Interruptibility.** If agents begin to take undesirable actions, the user and/or provider needs a way to halt them in a seamless manner. Depending on the situation, this may be straightforward, or it may require the agent to end its action in a particular way to avoid causing harm. This may involve reverting to a previous state, entering a default or fail-safe mode, or prompting further user intervention.

- **Reversibility**. The existence of an "undo" button makes it far lower stakes to make a mistake—which is likely why it is near-ubiquitous in modern software applications. While many actions that an agent can take may not be equally reversible, incorporating this functionality where possible would be valuable. Building agents that can foresee the reversibility or irreversibility of their actions is also useful, as this capability would allow them to identify higher- and lower-stakes decisions and modulate their behavior accordingly, such as by requesting human review or authorization.[46]

- **Real-time monitoring** (described above). Monitoring setups would help keep AI agents under control by preventing or calling attention to undesirable actions.

- **Access control.** Requiring agents to provide authentication credentials and proof of authorization for certain kinds of actions would help to ensure that agents are acting appropriately on behalf of their users. This would verify that the agent is what it claims, is acting on the behalf of the person it says it is, and has been delegated the proper authority to take a specific action on behalf of that person. In addition, authentication would facilitate identification and visibility, discussed earlier, as many existing authentication schemes (e.g., logging into an online bank account) already involve some amount of monitoring and logging. At a minimum, AI agents could use a person's credentials to log into online services and be subjected to the same tracking and authorization restrictions as the human user. However, widespread agent adoption would also provide an opportunity to deploy more secure access control methods, like those based on Public Key Infrastructure, that humans often find challenging to use but could be leveraged more easily by an AI agent.[47] Perhaps this type of authentication could be expanded to cover a wider range of activity by agents.

Trustworthiness of AI agents reflects the expectation that an agent will act in alignment with user intent. While many of the other technical guardrails discussed in this section contribute to establishing trust, several others are worth highlighting.

- **Human-in-the-loop.** Feedback mechanisms that provide users with information on what actions an agent is taking and enable opportunities for the user to verify or change those actions are often cited as a means of maintaining alignment between agent and user intent. For a human monitor to be effective, the information provided by the agent must be timely, understandable, and actionable, and the human must be empowered to intervene.[48]

- **Explainability.** While the "black box" problem remains a difficult, unsolved challenge for machine learning systems, efforts to make agent's actions and rationale available in a way that users can understand is important. Progress in interpretability research could be helpful in this regard, as could AI agent design choices that expose relevant inner workings to the user.

- **Integration testing.** AI agents primarily interact with the outside world through the use of external tools. Conducting integration testing of the agent working with those tools in a realistic environment will therefore help to uncover unintended or unexpected behaviors that can arise from the interaction of multiple complex systems.

The **security and privacy** of the agent's operations and associated user data must be safeguarded from a range of potential threats. These threats include the misuse of sensitive information by the actors involved in developing and running the agent, as well as hijacking of the agent's actions or exfiltration of data by malicious actors.

Technical guardrails that could support security and privacy include:

- **Secure Coding Practices.** Practices for secure software development should be employed when creating agents. These include securing the agent's deployment environment, monitoring and protecting the agent's security during use, and ensuring regular updates and patches.[49]

- **Adversarial Testing.** Agents should be tested to make them more robust to jailbreaks and other adversarial input. This includes input received from the user providing the agent with instructions, input from the environment that the agent can collect itself, and input from external actors that an agent may interact with.

- **Access Control.** Enforcement of proper authentication and authorization also applies to users providing instructions to the agent. Agents should only take action based on instructions from valid users. In cases where an agent may accept external instructions, the agent should validate those inputs with the user before acting upon them. Just like many other software systems,

permissions to execute certain actions or access certain capabilities should be dependent upon a user's privileges. Agents will need to enforce those permissions and prevent privilege escalation.

- **Data Minimization**. Many LLM providers offer "no retention" options for enterprise clients, meaning that prompts and model responses are not retained on the providers' servers.[50] Others, such as Apple, emphasize that user data is stored primarily on the user's device, meaning that the provider cannot access it. The kinds of data storage that are practical for agents will depend, to some extent, on the design and intended usage of the agent; for instance, if the agent is intended to become familiar with a specific user, their preferences, and their use cases, then data capturing that will need to be stored somewhere. But there will likely be a range of options for how to implement any given design, some more privacy-protecting than others.

- **Encryption.** Data used or retained by the agent should be encrypted to protect user privacy. In addition, sensitive components of the agent itself, like the model weights, should be similarly protected. Where computationally feasible, the processing of sensitive information or critical functions should be conducted in trusted execution environments.

In some cases, trade-offs arise between different goals. Enabling visibility into, and control of, how an agent functions to anyone other than the user may infringe on the security and privacy of that user's interactions with the agent, though the extent of this infringement depends on implementation. For example, many online platforms analyze photos shared by their users in order to compare a cryptographic hash of each photo with a database of hashes of known child pornography images. This approach allows such material to be detected without needing to examine the content of images directly, but it does require the platforms to have access to unencrypted versions of the user images.[51] Online platforms are also generally required to share sensitive user data, including private conversations, with law enforcement under certain circumstances.[52] In addition, security measures such as adversarial training and privacy-preserving practices like data minimization are likely to have direct trade-offs with performance by limiting the ability of the model or reducing the data available for adaptability. Measures that help to build trust, like human-in-the-loop methods, may reduce the effectiveness, or at the very least the utility, of an agent intended to accomplish tasks independently.

Decisions around which technical guardrails to implement for AI agents will likely have both commonalities with, and differences from, decisions for existing online systems.

Additional research and creative thinking will be required to determine which governance objectives to prioritize and how to make trade-offs between them, and to inform decisions about which technical guardrails can and should be implemented. In particular, further work could examine what kinds of information should be tracked and logged, as well as who should have access to it and under what circumstances.

AI Agents and the Law

A wide range of existing areas of law could be relevant to managing the impacts of AI agents, even in the absence of new legislation or regulation.[53] Among the workshop participants were several legal scholars who led a discussion on how the law is likely to apply to AI agents (a "descriptive" view) and how we might hope it should apply or be adapted (a "normative" view). The legal issues are numerous and complicated; here we present a brief summary of some key topics in the hope that it helps orient readers without deep legal expertise.

First, many current laws can already be applied to AI agents. For example, the Federal Trade Commission has stated that "there is no AI exemption" to existing laws on civil rights, fair competition, consumer protection, and the like.[54] Similarly, AI agents deployed in highly regulated sectors such as healthcare or finance will also be subject to the regulations already applicable in those sectors. More broadly, other areas of law—including agency law, corporate law, contract law, criminal law, tort law, property law, and insurance law—will play an important role in how cases involving agentic systems are decided.

However, even when a legal subject clearly governs a question, certain application details may need to be worked out. For example, who should be liable when an AI agent causes harm? If an actor intentionally employs an agent to cause harm, criminal law may be implicated. Meanwhile, **tort law** will be relevant for both intentional and unintended acts, but it is not yet clear what standard should be applied in different scenarios.[55] In simplified terms, three major standards can be used to impose liability under U.S. tort law:

- **Negligence** is the most common tort law standard. Usually, someone is judged to be negligent if they had a duty to act, did not exercise a reasonable level of care, and that lack of care caused an injury. A jury will usually evaluate what constitutes "reasonable care" in light of the specific circumstances of a given case. Doctors and lawyers, in contrast, are evaluated under a "professional care" standard—the decision will still go to a jury, but the profession sets the bar for

what constitutes reasonable care, and the jury evaluates if the defendant's actions met that standard.

- **Strict liability** means that a party can be held liable for causing harm, regardless of whether they behaved reasonably or not. Strict liability primarily applies to "abnormally dangerous activities" that have a risk of causing harm regardless of how much care is taken (such as using explosives or owning wild animals).[56]

- **Product liability** applies when a commercial seller sells a product that was defective at the time of sale and the defect causes an injury. Simplifying somewhat, it blends elements of negligence and strict liability, depending on how the product caused harm. If the harm is caused by a manufacturing defect, it is more akin to strict liability; if harm is caused by a design or informational defect, it is more akin to negligence. Notably, cases involving product liability in the software industry—perhaps the most obvious comparable to the AI industry—have generally not held software companies liable for harm to users, although there are indications this could be changing.[57]

A harmed entity might simultaneously bring negligence, strict liability, and product liability claims for the same set of actions. As courts evaluate claims involving harms caused by agentic systems, they will set important precedents that will influence the likelihood of the success of future claims. For example, courts might find that certain types of negligence claims should be governed by the reasonable care rather than the professional care standard, which might open up parent companies to greater liability risks.[58]

Another highly relevant area is **agency law**, which deals with situations where one party (an agent) has the authority to act on behalf of another (a principal), and where liability for the agent's acts may in some cases transfer to the principal ("vicarious liability").[59] Typical situations in which agency law can be applied include an employee acting on behalf of their boss or a broker on behalf of a client. Some legal scholars (including one of this report's authors) have suggested that this kind of relationship is an apt metaphor for the relationship between AI systems (including AI agents) and their users, and that agency law is, therefore, an appropriate tool to manage AI-caused harm.[60] Other scholars (including another of this report's authors) have expressed reservations.[61]

Among other things, agency law covers questions of how and when a principal should be liable for the actions of their agent, meaning that it overlaps with the tort liability

considerations described above. If AI agents are determined to be part of a principal-agent relationship in the traditional sense of agency law, then these existing liability precedents could carry over, potentially leading to a human principal being held vicariously liable for an AI agent.

Against the backdrop of these existing legal frameworks, workshop participants discussed relevant considerations for AI agents, including:

- **State of mind.** Many legal standards are based on the state of mind of one or more parties involved. This applies not only when criminal cases require *mens rea* (a "guilty mind"), but also in imposing tort liability, where the defendant's knowledge of wrongdoing can affect whether they are ultimately held liable. Determinations of negligence also often hinge on whether a party should have been able to foresee the harm. When one of the parties in question is an AI agent, it is unclear how to make determinations of this kind about mental states. Legislatures or courts may determine that AI systems do not have mental states and therefore cannot pass any of these tests; however, this may systematically allocate liability away from AI systems—and, more importantly, away from their users and developers, who would be spared vicarious liability for the AI agent's actions under some readings of the doctrine—to an extent that may not be desirable. To counteract this, judges may attribute a mental state to an AI agent; the legal system regularly attributes mental states to various artificial persons, like corporations,[62] and some scholars have already attempted to develop theories about how mental states can be justifiably attributed to autonomous agents.[63]

- **Legal personhood.** While experts often warn against anthropomorphizing AI systems, several workshop participants emphasized that treating AI agents as potential parties in litigation does not mean treating them as equivalent to humans. Rather, legal personhood is a legal construct that can be created, designed, and tailored to achieve certain social or economic goals. Corporations, for example, are considered legal persons, but that does not entail considering them equivalent to humans.

- **Who is the principal?** While the established principal-agent concept from agency law may be an appealing analogy for situations involving AI agents, it may sometimes be difficult to determine who should count as the principal. In many cases, the user would be an obvious choice, given that they would typically direct an AI agent to pursue a goal. In others, however—such as when a chatbot contributes to a user deciding to commit suicide, or when an agent

behaves in a way that is very different from what the user intended—it may be less applicable. One suggestion raised during the workshop was to identify multiple principals (e.g., the user, app developer, model developer, etc.). However, to the extent different principals' desires may conflict, it may be difficult to determine which should be held accountable for an agent's actions.

- **Industry standards.** To the extent that negligence is used as the mechanism to allocate liability for AI agent-caused harm, standard industry practices will become an influential form of soft law. This is true even when a "reasonable care" (rather than "professional care") standard for breach is applied, as comparing a defendant's conduct to what is typical within their industry often forms part of the jury's deliberation about what counts as reasonable care. This could mean that arrangements such as the safety commitments made by AI companies at the White House in July 2023 or the Seoul Summit in May 2024 could wind up having indirect legal force.[64]

- **Liability and innovation.** Making developers in a sector liable for harm caused by their products does not necessarily dampen innovation in that sector. For example, when credit card providers were made liable for fraud, this spurred innovation in the security mechanisms built into the cards, which reduced instances of fraud.[65] Liability insurance could also potentially play a valuable role in facilitating innovation even if developers are held liable for harms caused by AI agents they create. Purchasing insurance simultaneously incentivizes the insurance provider to accurately assess risk and the insurance purchaser to try to reduce risk, thereby allowing the developers of AI agents to continue to innovate while reducing their financial exposure.[66]

- **Limits of liability.** Tort liability is a useful tool in many cases, but it also has important limits. It is often inaccessible for victims who lack the resources to bring a claim, primarily provides after-the-fact recourse rather than preventing harm (though it can help set expectations that deter bad conduct), and struggles to handle situations involving many small individual harms that add up to substantial harm (though class action suits can sometimes handle such situations). It also generally works far better for cases involving physical harm than ones involving primarily mental or emotional harm or pure economic loss.

When novel legal questions arise, an underlying question is often whether the source of novelty is a difference in degree or a difference in kind. If the former, existing frameworks can likely be applied; if the latter, perhaps more radical changes are needed. When it comes to AI, some issues may be mere differences in degree. Others,

such as the question of how to apply ideas of mental states to an AI system, may be genuine differences in kind. Overall, some combination of creatively applying existing doctrine and developing new legal concepts will be necessary.

Conclusion

The answers to many questions about the prospects of AI agents over the coming years remain murky. Nonetheless, several takeaways emerged from the workshop discussions:

- **AI agents are the subject of excitement and significant investment among AI developers.** Startups and major tech companies alike are working to convert progress in LLMs, which have so far primarily been used as chatbots, into more agentic systems that can flexibly and autonomously carry out goals given to them by a user.

- **Definitions are contested.** There is no clear boundary between AI systems that are and are not agents, but in brief, increasingly agentic systems would be able to independently take direct actions in pursuit of more complex goals in more open-ended environments. With sufficient technical advances, this could include making purchases, hiring humans, automating AI research itself, and a wide range of other activities.

- **Existing LLM-based agents have limited functionality and often make mistakes.** Products currently on the market fail or get stuck on tasks as simple as reserving a library book. Researchers are working on better ways for agents to identify and correct their errors, so that they will be able to plan and carry out more complex, multistep tasks. Researchers also expect continued progress on the underlying AI technologies (e.g., LLMs), the ability to integrate with a wider range of tools, and other advances to continue to improve the sophistication of AI agents.

- **If highly sophisticated AI agents enter widespread use, they are likely to exacerbate existing problems and bring new problems to the fore.** These challenges include mitigating harm resulting from accidental failure and malicious use, allocating responsibility for said harm, protecting user privacy, establishing norms and safeguards around human-agent and agent-agent interactions, and handling the impact of agent adoption on the labor market.

- **The trajectory of progress in AI agents is difficult to evaluate.** Current methods for measuring AI systems' capabilities are lacking, and this is doubly true for AI agents. Given this, it is challenging to determine whether AI agents are likely to progress rapidly in sophistication and widespread deployment, or whether they are likely to be primarily research curiosities over the coming

years. Policy approaches should attempt to manage a wide range of possibilities.

- **Many potential technical guardrails for AI agents exist, but their implementation may sometimes involve making difficult decisions between competing goals.** This includes evaluating trade-offs between visibility and privacy; security and performance; and trustworthiness and utility.
- **The legal status of AI agents and the applicability of existing legal frameworks is evolving.** Well-established legal ideas around how to allocate liability for harms, who is responsible when one party acts on another's behalf, and other similar questions, will be applied to AI agents—but which analogies are employed, and which standards are used, is still to be seen. AI agents raise challenging legal questions that will likely require a combination of creatively applying existing concepts and developing novel legal ideas.

Authors

Authors are ordered alphabetically by first name. The five workshop organizers are listed first.

Helen Toner is the director of strategy and foundational research grants at CSET.

John Bansemer is a senior fellow and director of the CyberAI Project at CSET.

Kyle Crichton is a research fellow on the CyberAI Project at CSET.

Matt Burtell was a Horizon junior fellow at CSET until August 2024.

Thomas Woodside was a Horizon junior fellow at CSET until May 2024.

Anat Lior is an assistant professor of law at Drexel University's Thomas R. Kline School of Law.

Andrew J. Lohn is a senior fellow on the CyberAI Project at CSET.

Ashwin Acharya is an AI policy and strategy researcher.

Beba Cibralic is a visiting scholar at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge.

Chris Painter is the head of policy at Model Evaluation and Threat Research (METR).

Cullen O'Keefe is the director of research at the Institute for Law & AI and a research affiliate with the Centre for the Governance of AI.

Iason Gabriel is a staff research scientist at Google DeepMind.

Kathleen Fisher is the office director of the Information Innovation Office (I2O) at the Defense Advanced Research Projects Agency (DARPA).*

Ketan Ramakrishnan is an associate professor of law at Yale Law School.

Krystal Jackson is a researcher in cybersecurity and AI policy.

Noam Kolt is an assistant professor at the Hebrew University of Jerusalem's Faculty of Law and School of Computer Science and Engineering.

Rebecca Crootof is a professor of law at the University of Richmond School of Law.

Samrat Chatterjee is the chief data scientist and team lead at the Pacific Northwest National Laboratory (PNNL).

---

* Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

Acknowledgements

We would like to thank several participants in the workshop who contributed greatly to the discussion but were unable to participate in the writing process: Alan Chan, Brendan Dolan-Gavitt, David Bau, and Miranda Bogen.

We also thank Igor Mikolic-Torreira and Seth Lazar for helpful discussions and feedback, as well as Lauren Lassiter, Shelton Fitch, and Jason Ly for their editorial and design support.

Document Identifier: doi: 10.51593/20240034

Endnotes

[1] David Kalat, "Nervous System #23: Claude Shannon's Magic Mouse," BRG, 2019, https://media.thinkbrg.com/wp-content/uploads/2021/03/19175859/23_Claude-Shannons-Magic-Mouse_Dec2019.pdf.

[2] Karen Hao, "DeepMind's AI Outsmarted Most StarCraft II Players." *MIT Technology Review*, October 30 2019, https://www.technologyreview.com/2019/10/30/132130/ai-deepmind-outcompeted-most-players-at-starcraft-ii/.

[3] Lei Wang et al., "A Survey on Large Language Model based Autonomous Agents," arXiv preprint arXiv:2308.11432 (2023), https://arxiv.org/abs/2308.11432; Sayash Kapoor et al., "AI Agents That Matter," arXiv preprint arXiv:2407.01502 (2024), https://arxiv.org/abs/2407.01502.

[4] Philosophical debates about agency often seek to find a clean line between systems with (some minimal level of) agency and those without, often treating "representational capacities" as an important factor distinguishing agents from non-agents. For more on philosophical conceptions of agency, see, G.E.M. Anscombe, "Intention," 1957, England, Basil Blackwell; Donald Davidson, "Actions, Reasons, and Causes," November 7, 1963, The Journal of Philosophy; Daniel C. Dennett, "The Intentional Stance" March 6, 1989, The MIT Press; and Fred Dretske, "Explaining Behavior: Reasons in a World of Causes" February 5, 1991, The MIT Press.

[5] Refer to Alan Chan et al., "Harms from Increasingly Agentic Algorithmic Systems," arXiv preprint arXiv:2302.10329 (2023), https://arxiv.org/abs/2302.10329; Yonadav Shavit et al., "Practices for Governing Agentic AI Systems," December 14, 2023, https://openai.com/index/practices-for-governing-agentic-ai-systems/; Iason Gabriel et al., "The Ethics of Advanced AI Assistants," arXiv preprint arXiv:2404.16244 (2024), https://arxiv.org/abs/2404.16244.

[6] "Deep Blue (chess computer)," Wikipedia, Last modified: September 6, 2024, https://en.wikipedia.org/wiki/Deep_Blue_(chess_computer).

[7] Microsoft, "Staying Ahead of Threat Actors in the Age of AI," *Microsoft Security Blog*, February 14, 2024, https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/.

[8] Minghao Shao et al., "An Empirical Evaluation of LLMs for Solving Offensive Security Challenges," arXiv preprint arXiv:2402.11814 (2024), https://arxiv.org/abs/2402.11814.

[9] Anna Tong and Jeffrey Dastin, "Race Towards 'Autonomous' AI Agents Grips Silicon Valley," *Reuters*, July 17, 2023, https://www.reuters.com/technology/race-towards-autonomous-ai-agents-grips-silicon-valley-2023-07-17/.

[10] Mustafa Suleyman has proposed a "modern Turing test" along the lines of building a CEO-AI: to pass, an AI system would have to autonomously earn $1 million on a retail web platform given an initial investment of $100,000. Mustafa Suleyman, "Mustafa Suleyman: My New Turing Test Would See If AI Can Make $1 Million," *MIT Technology Review*, July 14, 2023

https://www.technologyreview.com/2023/07/14/1076296/mustafa-suleyman-my-new-turing-test-would-see-if-ai-can-make-1-million/.

[11] For more on this idea, see Thomas Woodside and Helen Toner, "Multimodality, Tool Use, and Autonomous Agents," Center for Security and Emerging Technology, March 8, 2024. https://cset.georgetown.edu/article/multimodality-tool-use-and-autonomous-agents/.

[12] Note that the use of terms like "knowledge" and "understanding" to refer to LLM capabilities is highly contested among AI experts; it may be more accurate to say that they exhibit behaviors that are reminiscent of, but different from, those we associate with knowledge and understanding in humans. For further discussion, see Melanie Mitchell and David C. Krakauer, "Intelligent Agents as Multi-Objective Decision Makers," Proceedings of the National Academy of Sciences 120, no. 7 (2023): e2215907120. https://www.pnas.org/doi/10.1073/pnas.2215907120.

[13] Jason Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," arXiv preprint arXiv:2201.11903 (2022), https://arxiv.org/abs/2201.11903.

[14] "ChatGPT Can Now Browse the Internet, OpenAI Says," *Reuters*, September 27, 2023, https://www.reuters.com/technology/openai-says-chatgpt-can-now-browse-internet-2023-09-27/.

[15] Maxwell Nye et al. "Show Your Work: Scratchpads for Intermediate Computation with Language Models," arXiv preprint arXiv:2112.00114 (2021), https://arxiv.org/abs/2112.00114.

[16] Omar Aftab, "Microsoft Copilot Studio: Building Copilots with Agent Capabilities," *Microsoft Blog,* May 21, 2024, https://www.microsoft.com/en-us/microsoft-copilot/blog/copilot-studio/microsoft-copilot-studio-building-copilots-with-agent-capabilities/; Matt Binder, "Google I/O 2024: 'AI Agents' are AI personal assistants that can return your shoes," *Mashable*, May 14, 2024, https://mashable.com/article/google-io-2024-ai-agents.

[17] Ananya Mariam Rajesh and Krystal Hu, "AI startup Adept raises $350 mln in fresh funding," *Reuters*, March 14, 2023, https://www.reuters.com/technology/adept-raises-350-mln-series-b-funding-2023-03-14/.

[18] Quote by Dario Amodei, "And if the accuracy of any given step is not very high, is not like 99.9 percent, as you compose these steps, the probability of making a mistake becomes itself very high," Ezra Klein, "What if Dario Amodei Is Right About A.I.?" *The New York Times*, April 12, 2024. https://www.nytimes.com/2024/04/12/opinion/ezra-klein-podcast-dario-amodei.html?showTranscript=1.

[19] Shirin Ghaffary, "Tech Companies Bet the World Is Ready for 'AI Agents,'" Bloomberg, February 15, 2024, https://www.bloomberg.com/news/newsletters/2024-02-15/tech-companies-bet-the-world-is-ready-for-ai-agents.

[20] Wei, "Chain-of-Thought Prompting,"; Eric Zelikman et al., "STaR: Bootstrapping Reasoning With Reasoning," arXiv preprint arXiv:2203.14465 (2022); Scott Reed et al., "A Generalist Agent," *Google DeepMind*, 12 May, 2022, https://deepmind.google/discover/blog/a-generalist-agent/.

[21] Will Knight, "Forget Chatbots. AI Agents Are the Future," *Wired*, March 14, 2024, https://www.wired.com/story/fast-forward-forget-chatbots-ai-agents-are-the-future/; Carlos E. Jimenez et al., "SWE-bench: Can Language Models Resolve Real-World GitHub Issues?," arXiv preprint arXiv:2310.06770 (2023), https://arxiv.org/abs/2310.06770.

[22] Earlier work lays out extensive taxonomies of how AI agents could be helpful and harmful. See, e.g., "Harms from Increasingly Agentic Algorithmic Systems;" Shavit et al., "Practices for Governing Agentic AI Systems;" Gabriel et al., "The Ethics of Advanced AI Assistants."

[23] Charles Perrow, *Normal Accidents: Living with High-Risk Technologies,* (Princeton University Press, 1999).

[24] Dan Hendrycks, "Single-Agent Safety," in *Introduction to AI Safety, Ethics, and Society,* (Center for AI Safety, 2023); Dan Hendrycks et al., "Unsolved Problems in ML Safety," arXiv preprint arXiv:2109.13916 (2021), https://arxiv.org/abs/2109.13916; Richard Ngo, Lawrence Chan, and Sören Mindermann, "The Alignment Problem from a Deep Learning Perspective," ICLR (2024), https://openreview.net/attachment?id=fh8EYKFKns&name=pdf.

[25] A growing body of work on explainable AI aims to address this so-called "black box" problem. See, e.g., "What is explainable AI?," IBM, accessed September 18, 2024, https://www.ibm.com/topics/explainable-ai; Rudresh Dwivedi et al., "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," *ACM Computing Surveys* vol. 55, issue 9 (2023): article 194, https://dl.acm.org/doi/full/10.1145/3561048.

[26] The question of whether advances in AI will tend to favor cyber attackers or defenders is an open one and subject to significant debate. See, e.g., Andrew Lohn and Krystal Jackson, "Will AI Make Cyber Swords or Shields?" Center for Security and Emerging Technology, August 2022, https://cset.georgetown.edu/publication/will-ai-make-cyber-swords-or-shields.

[27] Katharine Kemp, "Concealed data practices and competition law: why privacy matters," *European Competition Journal* vol. 16, issue 2-3 (2020), pp. 628-672, https://www.tandfonline.com/doi/full/10.1080/17441056.2020.1839228.

[28] Recent controversy around Microsoft's planned "Recall" feature, which would save large volumes of screenshots of the user's display, illustrates the relevant tensions here. See, e.g., Tom Warren, "Microsoft's all-knowing Recall AI feature is being delayed," The Verge, June 13, 2024, https://www.theverge.com/2024/6/13/24178144/microsoft-windows-ai-recall-feature-delay?ueid=735ec93a25199a1268f3bb86cc446922.

[29] Approaches such as on-device data storage could mitigate these concerns to some extent and are discussed in a later section of this paper.

[30] Haleluya Hadero, "Artificial intelligence, real emotion. People are seeking a romantic connection with the perfect bot," *Associated Press*, February 14, 2024, https://apnews.com/article/ai-girlfriend-boyfriend-replika-paradot-113df1b9ed069ed56162793b50f3a9fa.

[31] These designs are sometimes referred to as "dark patterns." See, e.g., Federal Trade Commission, "FTC Report Shows Rise in Sophisticated Dark Patterns Designed to Trick and Trap Consumers," *Federal Trade Commission*, September 15, 2022, https://www.ftc.gov/news-events/news/press-releases/2022/09/ftc-report-shows-rise-sophisticated-dark-patterns-designed-trick-trap-consumers.

[32] Hannah Garden-Monheit and Ken Merber, "Price-Fixing Algorithm Is Still Price-Fixing," *Federal Trade Commission*, March 1, 2024, https://www.ftc.gov/business-guidance/blog/2024/03/price-fixing-algorithm-still-price-fixing; Organisation for Economic Co-operation and Development. "Algorithms and Collusion: Competition Policy in the Age of Digital Platforms," *OECD*, 2017, https://one.oecd.org/document/DAF/COMP/WD(2017)12/en/pdf.

[33] Will Henshall, "Nobody Knows How to Safety-Test AI," Time, March 21, 2024, https://time.com/6958868/artificial-intelligence-safety-evaluations-risks/.

[34] U.S. Department of Commerce, "U.S. and UK Announce Partnership on Science of AI Safety," *U.S. Department of* Commerce, April 1, 2024, https://www.commerce.gov/news/press-releases/2024/04/us-and-uk-announce-partnership-science-ai-safety; Anthropic, "Challenges in evaluating AI systems," *Anthropic,* Oct 4, 2024, https://www.anthropic.com/news/evaluating-ai-systems; METR, "Autonomy Evaluation Resources," METR, March 15, 2024.

[35] Kapoor et al., "AI Agents That Matter."

[36] Arvind Narayanan and Sayash Kapoor, "AI safety is not a model property," *AI Snake Oil*, March 12, 2024, https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property; Cullen O'Keefe, "AI Safety is Sometimes a Model Property," *Jural Networks*, 2024; Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems," arXiv preprint arXiv:2310.11986 (2023), https://arxiv.org/abs/2310.11986.

[37] For more on the potentially disruptive effects of automating AI R&D, see: Ajeya Cotra, "AIs Accelerating AI Research,", *Planned Obsolescence*, April 4, 2024, https://www.planned-obsolescence.org/ais-accelerating-ai-research/; and for how AI could accelerate AI R&D in practice, see: David Owen, "Interviewing AI Researchers on Automation of AI R&D," *Epoch AI,* August 27, 2024, https://epochai.org/blog/interviewing-ai-researchers-on-automation-of-ai-rnd.

[38] The concept of "compound AI systems" is helpful in thinking about the system layer; see Matei Zaharia et al., "The Shift from Models to Compound AI Systems," *Berkeley AI Research*, February 18, 2024, https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/.

[39] Alan Chan et al. "Visibility into AI Agents," FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (2024), pp. 958-973, https://dl.acm.org/doi/10.1145/3630106.3658948.

[40] We refer readers to Chan et al., "Visibility into AI Agents" for a more in-depth look at agent visibility, including detailed considerations relevant to the potential guardrails described above.

[41] Margaret Mitchell et al., "Model Cards for Model Reporting," arXiv preprint arXiv:1810.03993 (2018), https://arxiv.org/abs/1810.03993; Meta, "System Cards, a new resource for understanding how AI

systems work," *Meta*, February 23, 2022, https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/.

[42] Alan Chan et al., "IDs for AI Systems," arXiv preprint arXiv:2406.12137 (2024), https://arxiv.org/abs/2406.12137.

[43] Chan et al., "IDs for AI Systems;" Shavit et al., "Practices for Governing Agentic AI Systems;" Gabriel et al., "The Ethics of Advanced AI Assistants."

[44] Anthropic, "Our Approach to User Safety," *Anthropic*, accessed September 18, 2024, https://support.anthropic.com/en/articles/8106465-our-approach-to-user-safety; OpenAI, "Moderation," *OpenAI,* accessed September 18, 2024, https://platform.openai.com/docs/guides/moderation/overview.

[45] For an overview of challenges in controlling large language models, see Jessica Ji, Josh A. Goldstein, and Andrew Lohn, "Controlling Large Language Models: A Primer," Center for Security and Emerging Technology, December 2023, https://cset.georgetown.edu/publication/controlling-large-language-models-a-primer/.

[46] For one exploration of building an "undo" button for use with modern AI systems, see Shishir G. Patil et al., "GoEX: Perspectives and Designs Towards a Runtime for Autonomous LLM Applications," arXiv preprint arXiv:2404.06921 (2024), https://arxiv.org/abs/2404.06921.

[47] Alma Whitten and J. D. Tygar, "Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0," *USENIX Security Symposium* (1999), pp. 169-183, https://people.eecs.berkeley.edu/~tygar/papers/Why_Johnny_Cant_Encrypt/USENIX.pdf.

[48] See, e.g., Rebecca Crootof, Margot E. Kaminski and Margot E. Kaminski, "Humans in the Loop," *76 Vanderbilt Law Review* 429 (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4066781.

[49] "Joint Guidance on Deploying AI Systems Securely," *Cybersecurity and Infrastructure Security Agency (CISA)*, April 15, 2024, https://www.cisa.gov/news-events/alerts/2024/04/15/joint-guidance-deploying-ai-systems-securely.

[50] Anthropic, "I have a zero retention agreement with Anthropic. What products does it apply to?," *Anthropic*, accessed September 18, 2024, https://support.anthropic.com/en/articles/8956058-i-have-a-zero-retention-agreement-with-anthropic-what-products-does-it-apply-to; OpenAI, "Models: How we use your data," *OpenAI,* accessed September 18, 2024, https://platform.openai.com/docs/models/how-we-use-your-data.

[51] Google, "NCMEC, Google and Image Hashing Technology," *Google,* accessed September 18, 2024, https://safety.google/stories/hash-matching-to-help-ncmec/.

[52] Privacy techniques such as end-to-end encryption limit law enforcement's ability to access messaging data, and have been the subject of significant tension between privacy advocates and governments. Joe Mullin, "EFF Tells E.U. Commission: Don't Break Encryption," *Electronic Frontier Foundation (EFF)*, March 17, 2022, https://www.eff.org/deeplinks/2022/03/eff-tells-eu-commission-dont-break-encryption.

[53] "Most existing law can be applied to most new technologies most of the time." Rebecca Crootof and BJ Ard, "Structuring Techlaw," *34 Harvard Journal of Law & Technology 347* (2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3664124.

[54] Federal Trade Commission (FTC), "FTC Chair Khan and Officials from DOJ, CFPB and EEOC Release Joint Statement on AI," *Federal Trade Commission*, April 25, 2023, https://www.ftc.gov/news-events/news/press-releases/2023/04/ftc-chair-khan-officials-doj-cfpb-eeoc-release-joint-statement-ai.

[55] Matthew van der Merwe, Ketan Ramakrishnan, and Markus Anderljung, "Tort Law and Frontier AI Governance," *Lawfare*, May 24, 2024, https://www.lawfaremedia.org/article/tort-law-and-frontier-ai-governance.

[56] Gabriel Weil, "Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence," *SSRN*, 2024, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4694006.

[57] Bruce Schneier, "Computer Security and Liability," *Schneier on Security*, November 3, 2004, https://www.schneier.com/blog/archives/2004/11/computer_securi-2.html; Elias Groll and Christian Vasquez, "Biden's National Cybersecurity Strategy Advocates Tech Regulation, Software Liability Reform," *CyberScoop,* March 2, 2023, https://cyberscoop.com/biden-national-cybersecurity-strategy-2023/.

[58] Rebecca Crootof, "AI Malfeasance or Malpractice?", *Jotwell,* September 10, 2024, https://cyber.jotwell.com/ai-misfeasance-or-ai-malpractice.

[59] More formally, an agency relationship is "the fiduciary relationship that arises when one person (a "principal") manifests assent to another person (an "agent") that the agent shall act on the principal's behalf and subject to the principal's control, and the agent manifests assent or otherwise consents so to act." Restatement (Third) of Agency § 1.01 (Am. Law Inst. 2006).

[60] Samir Chopra and Laurence F. White, *A Legal Theory for Autonomous Artificial Agents* (University of Michigan Press, 1996), https://www.jstor.org/stable/10.3998/mpub.356801; Anat Lior, "AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy," *Mitchell Hamline Law Review* vol. 46, no. 5, 2020, https://open.mitchellhamline.edu/mhlr/vol46/iss5/2/.

[61] Noam Kolt, "Governing AI Agents," *SSRN*, 2024, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4772956.

[62] E.g., United States v. Alaska Packers' Assn, 1 Alaska 217 (1901); Goodspeed v. East Haddam Bank, 22 Conn. 530 (1853); see also Pamela H. Bucy, "Corporate Ethos: A Standard for Imposing Corporate Criminal Liability," Minnesota Law Review 2048 (1991), https://scholarship.law.umn.edu/cgi/viewcontent.cgi?params=/context/mlr/article/3047/&path_info=uc.pdf; William S. Laufer, "Culpability and the Sentencing of Corporations," 71 Nebraska Law Review (1992), https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1692&context=nlr.

[63] Chopra and White, *A Legal Theory for Autonomous Agents*.

[64] Bryan H. Choi, "AI Malpractice," 73 DePaul Law Review 301 (2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4721923.

[65] Mark MacCarthy, "AI Needs More Regulation, Not Less," *Brookings*, March 9, 2020, 2020, https://www.brookings.edu/articles/ai-needs-more-regulation-not-less/.

[66] Anat Lior, "Innovating Liability: The Virtuous Cycle of Torts, Technology, and Liability Insurance," *Yale Journal of Law and Technology* vol. 25, issue 2 (2023), https://yjolt.org/innovating-liability-virtuous-cycle-torts-technology-and-liability-insurance.