CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

Key Takeaways from "Through the Chat Window and Into the Real World: Preparing for AI Agents"

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.

For more information:

- Download the report: https://cset.georgetown.edu/publication/through-the-chat-window-and-into-the-real-world-preparing-for-ai-agents/
- Contact us: cset@georgetown.edu