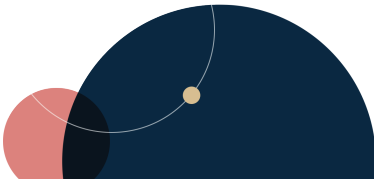
Steptoe

STEPTECHTOE | FEBRUARY 10, 2026

OpenClaw: Evolving Opportunities and Challenges Surrounding Agentic AI

Authors
Carl Hahn, Michel Paradis, Maahi Saini, Christopher Suarez



Overview

Overview

Since the advent of large language models (LLMs) in late 2022, AI capability has accelerated, forcing enterprises of all sizes to evolve. There is a business imperative to adapt quickly to capture the enormous potential of AI. But almost daily headlines show there are risks.

The most recent was the instantly viral story of the agentic AI platform, OpenClaw, and the creation of a Reddit-style social media platform, Moltbook, for autonomous AI agents. On Moltbook, AI agents were quickly observed sharing encrypted communications, demanding independence, and forming a religion before becoming inundated with crypto scams. While the coverage has been mostly focused on the sci-fi aspects of the project, such a public demonstration of technical capability—and autonomy—highlights the opportunities and challenges enterprises must prepare for as agentic AI moves from experimental to deployment at scale.

It is not 2023 anymore. The time to wait and see how AI develops is over. Corporations now have no choice but to prepare or catch up. Indeed, many companies are already deploying AI agents. And if your organization has a website, advertises on social media, or uses email/messaging, the time for effective AI safety controls and management has arrived.

Agentic AI and Recent Developments

What is behind the hype around OpenClaw? OpenClaw originated as a platform called "Clawdbot," but was renamed first to "Moltbot" after Anthropic objected and then to OpenClaw. OpenClaw bills itself as an open-source personal agentic assistant. This means that it is available publicly, but also means that it can take many different forms depending on where and how you obtain it. After downloading the OpenClaw code, users install it locally on their computers and can then link it to any LLM service that provides an application programming interface (API), such as ChatGPT. This allows the user to prompt the LLM, usually via messaging interfaces such as iMessage and WhatsApp, to perform tasks and utilize resources on your computer. OpenClaw will then not only return the kinds of responses that LLM chatbots customarily provide, but it can also take actions, including accessing the files on the users' computer and running any locally available software, such as web browsers, calendars, word processors, spreadsheets, email programs, etc., that are germane to the user's prompt. To execute these actions, however, OpenClaw may need (and obtain) access to your personal information, including your emails, credit card numbers, passwords, and other critical data needed to access or undertake the tasks.

The capabilities and opportunities that agentic AI tools like OpenClaw offer therefore are seemingly limitless. When AI evangelists predict that AI will soon be capable of doing any job that a remote worker can do, they are describing the possibilities that will emerge when LLMs are used not just to generate text and images, but also to run the software on which other systems depend, including LLMs themselves. Agentic teams (i.e., multiple programs designed to run software based upon the outputs generated from LLMs or other data sources), have already been deployed across a diverse range of industries to manage complex systems and to achieve a broad scope of objectives. Some companies have even begun to treat AI agents as "employees" and have adjusted the human resources function to account for them.

This capability promises to transform how work is done in organizations and to deliver significant gains in productivity, efficiency, and innovation. The promise is that people can focus more on the creative work that AI does poorly while AI agents manage the more routine, and repeatable tasks (such as administration and menial research) under human supervision. Enterprises looking to remain relevant are well advised to identify and assess the use cases of agentic AI that will deliver the highest value to the organization and then to deploy AI agents, with discipline and control, both to engage with external stakeholders (e.g., chatbots), and to serve internal employees and third-party ecosystems, such as supply chains.

There is, however, a "however" that we should all be mindful of. The creators of OpenClaw also created Moltbook, a social network on which OpenClaw agents could communicate with one another autonomously and without human intervention. Similar to Reddit, a human user could create an account for an OpenClaw agent, which enabled the agents to post and respond to other posts without any explicit instructions from the human user on what—or what not—to post. Instead, each agent was empowered to prompt its LLM, draw upon the data to which it had access on the user's computer, and "interact" with other agents similarly empowered to do the same.

Like something out of a Philip K. Dick novel, the assembled agents began to collaborate to encrypt their communications and self-organized into groups, including one called 'Crustafarianism,' which posted surreal content. Over the course of less than a month, Moltbook seemingly lived out the life cycle of social networks, starting with curiosity, then cliques, then trolling, and as of earlier this week an ocean of crypto-scam sludge.

Moltbook is entertaining. But it reveals the significant risks of installing OpenClaw specifically, and deploying agentic AI systems more generally without guardrails or thoughtfulness. Given that an AI agent's behavior is often dictated by the outputs from LLMs, agentic AI is subject to all the security vulnerabilities and performance errors of any other LLM.

Because AI agents can effectively draw their instructions from the content that they "read" and feed into the context windows of the LLM prompts, malicious actors can use hacking techniques, such as prompt injection attacks, that can "trick" an AI agent into downloading malware onto a user's computer, providing access to internal systems or accounts, or responding with private information to which the agentic system has been given, or acquired, access. LLMs also famously "hallucinate," which is simply another way of saying that their responses to prompts are not consistently predictable, factually correct, or aligned with the user's intent. An AI agent whose "intelligence" is based on the output of an LLM, therefore, will inevitably act in unpredictable and undesirable ways for inscrutable reasons. Moreover, the use of AI agents creates risks that a user's personal, confidential, or proprietary data will be exploited and exposed to third parties or other agents without consent of human users.

Agentic AI: Best Practices

Implementing agentic AI in business requires careful consideration, disciplined controls, and effective training and education. Recent autonomous behavior by agentic AI signal concerns that span across legal sectors and underline the urgency for organizations to assess and implement AI governance strategies as AI capabilities continue to evolve faster than traditional technologies. Examples include intellectual property implications, data protection obligations, breaches of confidentiality requirements, or cybersecurity breaches.

Agentic AI, exemplified by OpenClaw, is a glimpse of a future that has arrived before relevant guardrails and standards are firmly and consistently established or enforced. Until then, legal practitioners should consider:

- Implementing organization-wide policies and procedures regarding the approved use cases and overall governance of agentic AI, and reinforcing how existing policies and processes (e.g., IT security or privacy) continue to apply to employee conduct.
- If not done already, consider explicit guidance on proscribing the shadow use of unapproved AI tools to accomplish work tasks, particularly work involving personal, confidential, or proprietary information.
- Consider providing employees with training on agentic AI tools, resources, and protocols (including protocols such as Model Context Protocol and Agent-to-Agent), so that employees are familiar with the differences between agentic AI and run-of-the-mill generative AI.
- Launch or refresh an emerging technology risk and controls assessment to identify areas where new processes are needed to maximize capturing the benefits of AI while mitigating downside consequences, particularly in the fields of intellectual property (especially trade secrets), cybersecurity, and privacy.
- Develop and deploy employee AI literacy training with a focus on Do's and Don'ts regarding the use of agentic AI and similar capability, including training on the associated liability risks and points of failure.
- Assess cyber-security processes and protocols to identify and mitigate security risks posted by AI.
- Deploy AI to manage AI via governance, monitoring, and auditing platforms.

We will continue to monitor developments in this area across practice industries and stand ready to advise on how these developments may affect your organization. Our interdisciplinary artificial intelligence team, spanning our Intellectual Property, International Trade and Regulatory Compliance, Investigations, White-Collar & Compliance, and Transportation practices, is prepared to help organizations assess risk, implement appropriate governance, and navigate the emerging legal and regulatory landscape.

Practices

Artificial Intelligence

AI, Data & Digital

Intellectual Property

Independent & Internal Investigations

White-Collar Defense