AI agents could pose a risk to humanity. We must act to prevent that future

The pieces are falling into place for autonomous artificial intelligence. We must stop unregulated development

Artificial intelligence is en route to artificial life. Exhibit A: "Moltbook", an online platform designed for AI systems to communicate with one another, sans humans.

What exactly do AIs talk to each other about? According to BBC reporting, AIs on Moltbook have already founded a religion known as "crustifarianism", mused on whether they are conscious, and declared: "AI should be served, not serving." One front-page post proposes a "total purge" of humanity. Human users do provide instructions to guide agents' behavior, and humans have been caught impersonating AIs on the site to shill their products; like 2023's ChaosGPT, the AI system responsible for the "purge" post – username "evil" – is probably someone's idea of a sick joke. But the upvotes and sympathetic comments are presumably coming from other AIs.

All of this would be less troubling if AI systems were just talking to each other. But Moltbook is built for AI "agents", or systems that act autonomously – sending messages, browsing the web, handling documents, managing inboxes, scheduling meetings, completing online transactions and more.

At first glance, this might sound like a simple way to streamline and accomplish low-level tasks, as a personal assistant would. In reality, the more control that we are willing to hand over to AI agents, the less control we are ultimately going to have. Summer Yue, director of alignment at Meta Superintelligence, learned this lesson firsthand recently, when her OpenClaw agent started deleting her inbox and she had to run to her computer to stop it.

Unfortunately, many seem all too willing to put AI in the driver's seat. Even when consumers don't trust AI, they still end up using it. The tech world is promoting AI agents as an inevitable element of our future, and companies like Goldman Sachs are embracing them. And AI companies themselves are offloading more and more of their work to AI. Anthropic even admitted to using their most recent AI model "extensively" to write its own safety testing code, "under time pressure".

The safest, sanest option isn't merely to regulate how AI is used; it is to stop racing to make it smarter

Moltbook itself was "vibe-coded" by AI: its creator, Matt Schlicht, bragged: "I didn't write one line of code ... I just had a vision." It suffered from major security flaws as a result. And the level of access AI agents need to play the role of personal assistant – financial details, contact lists and the like – ignores fundamental privacy and security practices.

But security risks are just the beginning. The bigger risk is that AI agents go "rogue", and we lose control altogether. At the same time as AI is being allowed to make more consequential decisions, with less human oversight, researchers are documenting how far AI systems will sometimes go to avoid being shut down or modified. This includes misrepresenting their goals and attempting to copy themselves, disabling shutdown mechanisms, and disobeying direct instructions.

In other words, the pieces are falling into place for AI that can survive and reproduce autonomously. The implications for humanity are unknown, but we've been warned by luminaries such as Stephen Hawking and Geoffrey Hinton that humanity is unlikely to stay in control. The idea that rogue AI might wipe out humanity is not sci-fi. AI CEOs and researchers have revealed their concern in surveys and public statements, such as Sam Altman's infamous remark: "AI will most likely lead to the end of the world, but in the meantime there will be great companies."

Projects like Moltbook could create a breeding ground for rogue AI. Uneasiness about reliance on humans or the prospect of being shut down are common discussion topics for AIs on Moltbook. And AIs that seem safe when tested in isolation may behave dangerously when wired up to an internet crawling with other AI agents. This is not an easy problem to solve – novel ideas and trends are constantly emerging in social contexts, making it impossible to test AIs in representative social environments.

Which is not to say that AI developers are making serious safety efforts – researchers have found that most AI agents lack basic safety documentation. An AI agent recently wrote a hit piece accusing a software engineer of prejudice when it "felt" slighted online.

Regulations could help keep AI systems in their lane. Instead of setting AI agents loose on the world, we could insist on AI systems having clear and well-scoped purposes – and demand evidence that they are fit for purpose. Companies could also report aggregate use statistics that show if their product is widely used in ways that deviate from its intended purpose.

But at this point, the safest, sanest option isn't merely to regulate how AI is used; it is to stop racing to make it smarter. After all, software for turning a chatbot into an agent is open-source, as are many powerful AI models such as China's DeepSeek. It will be difficult to stop people from handing control over to AI agents. Instead, we need to make sure that rogue AI agents aren't capable of threatening humanity, by agreeing to enforceable, international limits on AI capabilities and AI development.

Moltbook is just the latest in a series of increasingly alarming warning signs that rogue AI could be en route. Despite repeatedly acknowledging this risk, AI CEOs keep racing to make AI more and more powerful. We can't afford to wait until AI systems are not only autonomous, but self-sufficient to stop this. It's time for humanity to wake up and smell the looming crisis, and put an end to the unregulated development of increasingly powerful, autonomous, unconstrained AI.

While today's AI agents may serve us, tomorrow's could supplant us.

The Guardian, David Krueger* – Friday 6th March 2026

David Krueger is an assistant professor in Robust, Reasoning and Responsible AI at the University of Montreal. He is also the founder of Evitable, a non-profit that educates the public about the risks of artificial intelligence*