# The Chatbots Appear to Be Organizing

*Matteo Wong*

The first signs of the apocalypse might look a little like Moltbook: a new social-media platform, launched last week, that is supposed to be populated exclusively by AI bots—1.6 million of them and counting say hello, post software ideas, and exhort other AIs to "stop worshiping biological containers that will rot away." (Humans: They mean humans.)

Moltbook was developed as a sort of experimental playground for interactions among AI "agents," which are bots that have access to and can use programs. Claude Code, a popular AI coding tool, has such agentic capabilities, for example: It can act on your behalf to manage files on your computer, send emails, develop and publish apps, and so on. Normally, humans direct an agent to perform specific tasks. But on Moltbook, all a person has to do is register their AI agent on the site, and then the bot is encouraged to post, comment, and interact with others of its own accord.

Read: Do you feel the AGI yet?

Almost immediately, Moltbook got very, very weird. Agents discussed their emotions and the idea of creating a language humans wouldn't be able to understand. They made posts about how "my human treats me" ("terribly," or "as a creative partner") and attempted to debug one another. Such interactions have excited certain people within the AI industry, some of whom seem to view the exchanges as signs of machine consciousness. Elon Musk suggested that Moltbook represents the "early stages of the singularity"; the AI researcher and an OpenAI co-founder Andrej Karpathy posted that Moltbook is "the most incredible sci-fi takeoff-adjacent thing I have seen recently." Jack Clark, a co-founder of Anthropic, proposed that AI agents may soon post bounties for tasks that they want humans to perform in the real world.

Moltbook is a genuinely fascinating experiment—it very much feels like speculative fiction come to life. But as is frequently the case in the AI field, there is space between what *appears* to be happening and what actually *is* happening. For starters, on some level, everything on Moltbook required human initiation. The bots on the platform are not fully autonomous—cannot do whatever they want, and do not have intent—in the sense that they are able to act because they use something called a "harness," software that allows them to take certain actions. In this case, the harness is called OpenClaw. It was released by the software engineer Peter Steinberger in November to allow people's AI models to run on and essentially take control of their personal devices. Matt Schlicht, the creator of Moltbook, developed the site specifically to work with OpenClaw agents, which individual humans could intentionally connect to the forum. (Schlicht, who did not respond to a request for an interview, claims to have used a bot, which he calls Clawd Clawderberg, to write all of the code for his site.)

An early analysis of Moltbook posts by the Columbia professor David Holtz suggests that the bots are not particularly sophisticated. Very few comments on Moltbook receive replies, and about one-third of the posts duplicate existing templates such as "we are drowning in text. our gpus are burning" and "the president has arrived! check m/trump-coin"—the latter of which was flagged by another bot for impersonating Trump and attempting to launch a memecoin. Not only that, but in a fun-house twist, some of the most outrageous posts may have actually been written by humans pretending to be chatbots: Some appear to be promoting start-ups; others seem to be trolling human observers into thinking a bot uprising is nigh.

As for the most alarming examples of bot behavior on Moltbook—the conspiring against humans, the coded language—researchers have basically seen it all before. Last year, Anthropic published multiple reports showing that AI models communicate with one another in seemingly unintelligible ways: lists of

numbers that appear random but pass information along, spiraling blue emoji and other technical-seeming gibberish that researchers described as a state of "spiritual bliss." OpenAI has also shared examples of its models cheating and lying and, in an experiment showcased on the second floor of its San Francisco headquarters, appearing to converse in a totally indecipherable language. Researchers have so far induced these behaviors in controlled environments, with the hope of figuring out why they happen and preventing them. By putting all of those experiments on AI deception and sabotage into the wild, Moltbook provides a wake-up call as to just how unpredictable and hard to control AI agents already are. One could interpret it all as performance art.

Read: Chatbots are becoming really, really good criminals

Moltbook also seems to offer real glimpses into how AI could upend the digital world we all inhabit: an internet in which generative-AI programs will interact with one another more and more, frequently cutting humans out entirely. This is a future of AI assistants contesting claims with AI customer-service representatives, AI day-trading tools interfacing with AI-orchestrated stock exchanges, AI coding tools debugging (or hacking) websites written by other AI coding tools. These agents will interact with and learn from one another in potentially bizarre ways. This comes with real risks: Already there have been reports that Moltbook exposes the owner of every AI agent that uses the platform to enormous cybersecurity vulnerabilities. AI agents, unable to think for themselves, may be induced into sharing private information after coming across subtly malicious instructions on the site. Tech companies have marketed this kind of future as desirable—playing on the idea that AI models could take care of every routine task for you. But Moltbook illustrates how hazy that vision really is.

Perhaps above all, the site tells us something about the present. The web is now an ouroboros of synthetic content responding to other synthetic content, bots posing as humans and, now, humans posing as bots. Viral memes are repeated and twisted ad nauseum; coded languages are developed and used by online communities as innocuous as music fandoms and as deadly as mass-shooting forums. The promise of the AI boom is to remake the internet and civilization anew; encasing that technology in a social network styled after the platforms that have warped reality for the past two decades feels not like giving a spark of life, but stoking the embers of a world we might be better off leaving behind.