TECHNOLOGY

We Need to Control AI Agents Now

Automated bots are about to be everywhere, with potentially devastating consequences.

By Jonathan L. Zittrain

Illustration by Ben Kothe / The Atlantic. Sources: perets; Liyao Xie / Getty.

JULY 2, 2024 SHARE SAVE

In 2010—well before the rise of ChatGPT and Claude and all the other sprightly, conversational AI models—an army of bots briefly wiped out $1 trillion of value across the NASDAQ and other stock exchanges. Lengthy investigations were undertaken to figure out what had happened and why—and how to prevent it from happening again. The Securities and Exchange Commission's report on the matter blamed high-frequency-trading algorithms unexpectedly engaging in a mindless "hot potato" buying and selling of contracts back and forth to one another.

A "flash crash," as the incident was called, may seem quaint relative to what lies ahead. That's because, even amid all the AI hype, a looming part of the AI revolution is under-examined: "agents." Agents are AIs that act independently on behalf of humans. As the 2010 flash crash showed, automated bots have been in use for years. But large language models can now translate plain-language goals, expressed by anyone, into concrete instructions that are interpretable and executable by a computer—not just in a narrow, specialized realm such as securities trading, but across the digital and physical worlds at large. Such agents are hard to understand, evaluate, or counter, and once set loose, they could operate indefinitely.

For all of today's concern about AI safety, including potentially existential risks, there's been no particular general alarm or corresponding regulation around these emerging AI agents. There have been thought experiments about an AI given (or setting for itself) an arbitrary and seemingly harmless goal, such as to manufacture as many paper clips as possible, only to cause disaster when it diverts all of humanity's resources toward that goal. But well short of having to confront a speculative monomaniacal superintelligence, we must attend to more pressing if prosaic problems, caused by decidedly nonspeculative contemporary agents. These can mess up, either through the malice of those who get them going, or accidentally, monkey's-paw style, when commissioned with a few ill-chosen words. For example, Air Canada recently experienced the latter when it set up a chatbot for customer assistance with a prompt to be helpful, along with access to the Air Canada website for use in answering customer questions. The bot helpfully explained a policy on bereavement fares in a way far more generous than the airline's actual policy. Air Canada tried to repudiate the bot's promises, and failed: A tribunal held that the customer was owed compensation.

Read: This is what it looks like when AI eats the world

Today's agents add up to more than a typical chatbot, with three distinct qualities. First, they can be given a high-level, even vague goal and independently take steps to bring it about, through research or work of their own. The idea is simple but powerful. For example, a year ago, an enterprising techie developed an AI that could order a pizza for him. He relied on software tools developed by companies such as OpenAI to create a "top-level AI" that could charter and command other AIs. That top-level AI was provided a goal—order a pepperoni pizza by voice from a given phone number—and then it went on to create its own task list and develop different versions of itself to perform those tasks, including prioritizing different steps in the list and producing a version of itself that was able to use a text-to-voice converter to make the phone call. Thus the AI was able to find and call a local pizzeria and place the order.

That demonstrates a second quality of agents beyond planning to meet a goal: They can interact with the world at large, using different software tools at will, as you might when opening Excel or placing a DoorDash order while also browsing the web. With the invitation and blessing of companies such as OpenAI, generative-AI models can take in information from the outside world and, in turn, affect it. As OpenAI says, you can "connect GPTs to databases, plug them into emails, or make them your shopping assistant. For example, you could integrate a travel listings database, connect a user's email inbox, or facilitate e-commerce orders." Agents could also accept and spend money.

This routinization of AI that doesn't simply talk with us, but also acts out in the world, is a crossing of the blood-brain barrier between digital and analog, bits and atoms. That should give us pause.

A non-AI example jumps to mind as a nefarious road map for what may lie ahead. Last year, a man left a bag conspicuously containing wires and a lockbox outside Harvard Yard. Harvard police then received a call with a disguised voice warning that it was one of three bombs on campus, and that they'd all go off soon unless the university transferred money to a hard-to-trace cryptocurrency address. The bag was determined to be harmless. The threat was a hoax.

When police identified and arrested the man who left the bag, it turned out that he had answered a Craigslist ad offering money for him to assemble and bring those items to campus. The person behind that ad—and the threatening calls to Harvard—was never found. The man who placed the wires pleaded guilty only to hiding out and deleting some potentially incriminating text messages and was sentenced to probation, after the authorities credited that he was not the originator of the plot. He didn't know that he'd joined a conspiracy to commit extortion.

Read: Welcome to a world without endings

This particular event may not have involved AI, but it's easy to imagine that an AI agent could soon be used to goad a person into following each of the steps in the

Harvard extortion case, with a minimum of prompting and guidance. More worrying, such threats can easily scale far beyond what a single malicious person could manage alone; imagine whoever was behind the Harvard plot being able to enact it in hundreds or thousands of towns, all at once. The act doesn't have to be as dramatic as a bomb threat. It could just be something like keeping an eye out for a particular person joining social media or job sites and to immediately and tirelessly post replies and reviews disparaging them.

This lays bare the third quality of AI agents: They can operate indefinitely, allowing human operators to "set it and forget it." Agents might be hand-coded, or powered by companies who offer services the way that cemeteries offer perpetual care for graves, or that banks offer to steward someone's money for decades at a time. Or the agents might even run on anonymous computing resources distributed among thousands of computers whose owners are, by design, ignorant of what's running—while being paid for their computing power.

The problem here is that the AI may continue to operate well beyond any initial usefulness. There's simply no way to know what moldering agents might stick around as circumstances change. With no framework for how to identify what they are, who set them up, and how and under what authority to turn them off, agents may end up like space junk: satellites lobbed into orbit and then forgotten. There is the potential for not only one-off collisions with active satellites, but also a chain reaction of collisions: The fragments of one collision create further collisions, and so on, creating a possibly impassable gauntlet of shrapnel blocking future spacecraft launches.

Read: The big AI risk not enough people are seeing

If agents take off, they may end up operating in a world quite different from the one that first wound them up—after all, it'll be a world with a lot of agents in it. They could start to interact with one another in unanticipated ways, just as they did in the 2010 flash crash. In that case, the bots had been created by humans but simply acted in strange ways during unanticipated circumstances. Here, agents set

to translate vague goals might also choose the wrong means to achieve them: A student who asks a bot to "help me cope with this boring class" might unwittingly generate a phoned-in bomb threat as the AI attempts to spice things up. This is an example of a larger phenomenon known as reward hacking, where AI models and systems can respond to certain incentives or optimize for certain goals while lacking crucial context, capturing the letter but not the spirit of the goal.

Even without collisions, imagine a fleet of pro–Vladimir Putin agents playing a long game by joining hobbyist forums, earnestly discussing those hobbies, and then waiting for a seemingly organic, opportune moment to work in favored political talking points. Or an agent might be commissioned to set up, advertise, and deliver on an offered bounty for someone's private information, whenever and wherever it might appear. An agent can deliver years later on an impulsive grudge —revenge is said to be a dish best served cold, and here it could be cryogenically frozen.

Much of this account remains speculative. Agents have not experienced a public boom yet, and by their very nature it's hard to know how they'll be used, or what protections the companies that help offer them will implement. Agentics, like much of the rest of modern technology, may have two phases: too early to tell, and too late to do anything about it.

In these circumstances, we should look for low-cost interventions that are comparatively easy to agree on and that won't be burdensome. Yale Law School's Ian Ayres and Jack Balkin are among the legal scholars beginning to wrestle with how we might best categorize AI agents and consider their behavior. That would have been helpful in the Air Canada case around a bot's inaccurate advice to a customer, where the tribunal hearing the claim was skeptical of what it took to be the airline's argument that "the chatbot is a separate legal entity that is responsible for its own actions." And it's particularly important to evaluate agent-driven acts whose character depends on assessing the actor's intentions. Suppose the agent waiting to pounce on a victim's social-media posts doesn't just disparage the person, but threatens them. Ayres and Balkin point out that the Supreme

Court recently held that criminalizing true threats requires that the person making the threats subjectively understand that they're inspiring fear. Some different legal approach will be required to respond up and down the AI supply chain when unthinking agents are making threats.

Technical interventions can help with whatever legal distinctions emerge. Last year, OpenAI researchers published a thoughtful paper chronicling some agentic hazards. There they broached the possibility that servers running AI bots should have to be identified, and others have made efforts to describe how that might work.

Read: It's the end of the web as we know it

But we might also look to refining existing internet standards to help manage this situation. Data are already distributed online through "packets," which are labeled with network addresses of senders and receivers. These labels can typically be read by anyone along the packets' route, even if the information itself is encrypted. There ought to be a new, special blank on a packet's digital form to indicate that a packet has been generated by a bot or an agent, and perhaps a place to indicate something about when it was created and by whom—just like a license plate can be used to track down a car's owner without revealing their identity to bystanders.

To allow such labels within Internet Protocol would give software designers and users a chance to choose to use them, and it would allow the companies behind, say, the DoorDash and Domino's apps to decide whether they want to treat an order for 20 pizzas from a human differently from one placed by a bot. Although any such system could be circumvented, regulators could help encourage adoption. For example, designers and providers of agents could be offered a cap on damages for the harm their agents cause if they decide to label their agents' online activities.

Internet routing offers a further lesson. There is no master map of the internet because it was designed for anyone to join it, not by going through a central switchboard, but by connecting to anyone already online. The resulting network is

one that relies on routers—way stations—that can communicate with one another about what they see as near and what they see as far. Thus can a packet be passed along, router to router, until it reaches its destination. That does, however, leave open the prospect that a packet could end up in its own form of eternal orbit, being passed among routers forever, through mistake or bad intention. That's why most packets have a "time to live," a number that helps show how many times they've hopped from one router to another. The counter might start at, say, 64, and then go down by one for each router the packet passes. It dies at zero, even if it hasn't reached its destination.

Read: What to do about the junkification of the internet

Agents, too, could and should have a standardized way of winding down: so many actions, or so much time, or so much impact, as befits their original purpose. Perhaps agents designed to last forever or have a big impact could be given more scrutiny and review—or be required to have a license plate—while more modest ones don't, the way bicycles and scooters don't need license plates even as cars do, and tractor trailers need even more paperwork. These interventions focus less on what AI models are innately capable of in the lab, and more on what makes agentic AI different: They act in the real world, even as their behavior is represented on the network.

It is too easy for the blinding pace of modern tech to make us think that we must choose between free markets and heavy-handed regulation—innovation versus stagnation. That's not true. The right kind of standard-setting and regulatory touch can make new tech safe enough for general adoption—including by allowing market players to be more discerning about how they interact with one another and with their customers.

"Too early to tell" is, in this context, a good time to take stock, and to maintain our agency in a deep sense. We need to stay in the driver's seat rather than be escorted by an invisible chauffeur acting on its own inscrutable and evolving motivations, or on those of a human distant in time and space.

This essay is adapted from Jonathan Zittrain's forthcoming book on humanity both gaining power and losing control.

ABOUT THE AUTHOR

Jonathan L. Zittrain

Jonathan L. Zittrain is a professor of law, public policy, and computer science at Harvard University, and the director of its Berkman Klein Center for Internet & Society.