New ethics risks courtesy of AI agents? Researchers are on the case

When AI systems go rogue, the results aren't pretty. Leaked confidential information, offensive messages and, in one instance, a user-friendly recipe for deadly chlorine gas, have all been blamed on chatbots gone awry.[1]

Such instances fueled greater emphasis on AI alignment, which is the practice of encoding human values and ethical principles into AI models. But AI researchers aren't stopping at tackling the ethical implications of today's machine learning technologies. They're also working to address the ethical issues of tomorrow—in particular, those posed by agentic artificial intelligence.

Also known as AI agents, agentic AI is an autonomous AI technology that presents an expanded set of ethical dilemmas in comparison to traditional AI models, says Kush Varshney, an IBM Fellow at IBM Research.

"Because AI agents can act without your supervision, there are a lot of additional trust issues," Varshney says. "There's going to be an evolution in terms of the capabilities but also in unintended consequences. From a safety perspective, you don't want to wait to work on it. You want to keep building up the safeguards as the technology is being developed."

What exactly are AI agents?

Before exploring AI agent safeguards, it's important to understand exactly what AI agents are: intelligent systems or programs that can autonomously perform tasks on behalf of a human being or on behalf of another system. Though they feature large language model (LLM) capabilities like natural language processing, these autonomous systems can also make decisions, solve problems, execute actions and interact with external environments.

Through such capabilities, AI agents can go beyond crafting text responses to user prompts to actually accomplishing tasks in the real world.

For example, external interactions happen through tool calling, also known as function calling, which is an interface that allows agents to work on tasks that require timely information—information that would otherwise be unavailable to LLMs. So, AI agents deployed in a supply chain ecosystem could autonomously work to optimize inventory levels by altering production schedules and ordering from suppliers as necessary.

How risky is greater AI autonomy?

When it comes to advanced artificial intelligence like agentic AI, how much autonomy is too much? To answer this question, we can look to the paperclip maximizer scenario. The famous thought experiment, by philosopher Nick Bostrom, centers on the still-hypothetical concept of AI superintelligence or ASI, an AI system with an intellectual scope that exceeds that of human intelligence. Bolstrom considers what might happen if such a system prioritized paperclip manufacturing above all other objectives.

In the proposed scenario, the system eventually devotes all of our planet's resources to making paper clips—an unethical outcome when life depends on more than just an endless bounty of tiny metal office supplies. Returning to our original question, we can obviously conclude that in this hypothetical case, the AI system in question had too much autonomy.

The good news is that today's agentic AI is not the same as ASI, so a paperclip dystopia driven by catastrophically flawed machine ethics remains unlikely. "We're closer, but we're still far away," Varshney says.

Other risks stemming from AI automation, however, are more imminent. The possibilities range from artificial agents sending inappropriate emails to stopping and starting machines in ways that users hadn't intended, Varshney says. Concerns over autonomous AI behavior are serious enough that, in an April 2024 report on AI safety and security guidelines, the US Department of Homeland Security (DHS) included "autonomy" in its list of risks to critical infrastructure systems such as communications, financial services and healthcare.[2]

Evolving solutions to support ethical agent behavior

Existing AI governance solutions can help support the ethics of AI agents, with software tools already empowering organizations to monitor, evaluate and address biases stemming from training datasets and algorithms that might skew decision-making processes. These tools can also help developers and companies ensure that the AI tools they're using meet current trustworthy AI standards, explainability objectives and responsible AI principles widely adopted by various companies and governments.

But as companies increasingly incorporate agentic AI into workflows, researchers are also at work on new ethical AI solutions and strategies that can curb misbehavior in autonomous agents and improve the sustainability of AI technology. Here are several worth following:

A novel AI alignment approach

Pretrained AI models today undergo fine-tuning to be trained on domain-specific data. During the fine-tuning phase of AI development, models may be aligned to moral values and ethical considerations, but

often questions arise about what normative values should be included in alignment. After all, values and ethical frameworks vary by company, country, stakeholder group and so on.

Varshney and a team of fellow IBM researchers have proposed a technology-driven approach that would be more context-specific: Known as Alignment Studio, it would align large language models to rules and values delineated in natural language policy documents, such as government regulations or a company's own ethical guidelines.

The approach, detailed in a September 2024 paper published in IEEE Internet Computing magazine, includes a continuous cycle of development so that models don't just learn policy-related vocabulary from policy documents, but actually adopt desired behaviors for better value alignment.[3]

Function-calling hallucination detection

Among the causes of AI agent-related misbehaviors is a lack of specific instructions on the part of the user or a misinterpretation of the user's instructions by the agent. Such "misunderstandings" could lead agents to choose the wrong tools or to use them in inappropriate or damaging ways, which is known as a function-calling hallucination.

Fortunately, improving function-calling has become a competitive endeavor, with the creation of several benchmarks measuring how well LLMs call APIs. Among the most recent improvements comes courtesy of a new feature in the latest IBM Granite Guardian release, Granite Guardian 3.1, part of IBM's family of Granite language models specifically designed for businesses. The model can detect function-calling hallucinations by agents before unintended consequences occur. "The detector checks for all kinds of mistakes, from the human language description to the function called," Varshney explains.

Detecting AI-generated text and disinformation

Malicious actors have already used generative AI to permeate social media with deepfakes, which are realistic AI-generated audio, video or images that can recreate a person's likeness. Meanwhile, scammers have leveraged AI-generated text for more sophisticated phishing emails. And the power of agentic AI could exacerbate these dangerous trends.

"There is growing evidence that AI-generated outputs are as persuasive as human arguments," researchers at Google DeepMind warned in an April 2024 report. In the future, they said, malicious actors could use autonomous AI to "tailor misinformation content to users in a hyperprecise way, by preying on their emotions and vulnerabilities."[4]

To date, the performance of tools designed to detect AI-powered deception has been mixed. But researchers are continuing to step up to the challenge of improving AI detection, with some of the most promising results stemming from the latest generation of AI-text detectors.[5]

For example, a new framework called RADAR—created by researchers at Chinese University of Hong Kong and IBM Research—uses adversarial learning between two separate, tunable language models to train an AI-text detector, leading to better performance in comparison to older AI-text detection solutions.[6]

As the development of AI-detection technology continues, technology companies like IBM, Microsoft and OpenAI are also calling on policymakers to pass laws to target the distribution of deepfakes and hold bad actors accountable.[7]

Preserving dignity for human workers

While many of the ethics issues stemming from agentic AI relate to misbehaviors, other ethics concerns arise even when autonomous AI technology performs as expected. For example, much discussion has focused on AI applications like OpenAI's ChatGPT replacing human labor and eliminating livelihoods.

But even when AI is deployed to augment (rather than replace) human labor, employees might face psychological consequences. If human workers perceive AI agents as being better at doing their jobs than they are, they could experience a decline in their self-worth, Varshney explains. "If you're in a position where all of your expertise seems no longer useful—that it's kind of subordinate to the AI agent—you might lose your dignity," he says. In some discussions of AI ethics, such loss of dignity is considered a human rights violation.[8]

In an August 2024 research paper, Varshney and several university-based researchers proposed an organizational approach to addressing the dignity concern: adversarial collaboration. Under their model, humans would still be responsible for providing final recommendations, while AI systems are deployed to scrutinize the human's work.

"The human is ultimately making the decision, and the algorithm isn't designed to compete in this role, but to interrogate and, thus, sharpen the recommendations of the human agent," researchers wrote.[9] Such adversarial collaboration, Varshney says, "is a way of organizing things that can keep human dignity alive."