Google DeepMind Admits Large-Scale AI Agent Deployment Is Unsafe Today via @sejournal, @martinibuster

In a recent interview, Nenad Tomašev, Senior Staff Research Scientist at Google DeepMind, described the sorts of traps that malicious actors are setting in order to take control of systems, take money, and jailbreak models without any of it being visible to the average user. Tomašev said this is already happening.

Agentic AI Agents At Scale Tips Them Toward Failure

Host Hannah Fry asked about traps that malicious actors are setting for AI agents and Tomašev responded that it's true, people are setting traps for AI agents in order to take advantage of them for criminal purposes. He remarked that complete reliability of every interaction is necessary but that the scale of what's happening tips it statistically toward failure.

Fry asked:

> *"Just looking at the other side of this, I also wanna think about the sort of cybersecurity element of this, because as more and more agents are out there interacting in the world on the internet and so on, there are inevitably gonna be people who are trying to exploit the vulnerabilities of agents.*

> *Tell me a little bit about agentic traps that people are laying.”*

Nenad Tomašev answered that the topic is both scary and fascinating:

> *“This is a scary and a fascinating topic at the same time, I would say. And I think it’s one of the main reasons why these kinds of deployments at scale cannot work, right?*
>
> *Because as we said, if there is not complete reliability of individual interactions, any system at scale that has many interactions is naturally going to statistically fail.*
>
> *And because these systems take a lot of compute and therefore energy and money to run, if they’re not reliable, it’s just a non-starter.*
>
> *And agentic traps are something that we have been thinking about for quite a while now. They can manifest in different ways.*
>
> *There are many types of traps, but it boils down to agents operate within an environment. And in this context, the environment is the web.*
>
> *If the environment itself is poisoned, if the traps are laid, agents may stumble upon them when interacting with the web.*
>
> *And then yes, malicious people or malicious agents deployed by malicious people can place those traps and then compromise systems really.”*

Kinds Of Agentic Traps To Beware Of

Host Hannah Fry then asked Tomašev how these traps are set and Tomašev provided examples, remarking that the traps aren’t going to be visible on a website but are nonetheless available

to AI agents. Some of what he described will sound familiar to old-school SEOs who engaged in things like cloaking in the early days of search engines.

Tomašev said that hidden tokens could be hidden for AI agents to consume. Tokens in this context is a reference to how AI breaks words into representations of words. When an AI reads words on a page what it does is to break it down into tokens. Hidden tokens could be completely invisible to humans.

He mentioned three ways that traps could be set for AI agents:

1. Hidden tokens
2. Dynamic cloaking
3. Content that induces jailbreaking

Fry asked:

> *"So I don't know, the sort of the wine buying agent for the wedding goes on to a particular wine merchant where there is some, essentially a prompt injector in the website that changes the agent's goals? Is that the sort of thing that we're talking about here?"*

Tomašev answered:

> *"That is one way this could happen, yes. And the reason why that may potentially go unnoticed is, you know, in terms of how web pages are encoded, there are elements there that are just not rendered visually.*
>
> *So if we're talking about an agent that isn't a visual computer use agent that sees the webpage, I mean, the pixels the same way a human does, rather consumes the actual format of the page in its raw format, then it could inadvertently consume those hidden tokens that can make it do different things than what the intention was, right?*
>
> *But this is not the only way it may happen because what*

malicious websites could potentially do, they could do what we refer to as dynamic cloaking as well, where they display pages differently for humans and agents.

Because you can, based on the behavior on a page, make a very good guess as to whether it is a human or it is an agent interacting with the page. And then only if an agent is interacting with the page with a specific intent, do tweak the content in such a way so as to induce some kind of jailbreaking."

Exploiting AI Agents To Steal Money From Humans

Tomašev confirmed that not only can criminals steal money from humans who are deploying AI agents, he confirms that it has already happened. He said that this kind of criminal activity isn't always something that is anticipated when testing a system out in a trusted environment but it becomes apparent out on the web, which is not a trusted environment.

The host asked:

"But just kind of going a little bit further on this, you could have agentic traps out there that, I don't know, are designed sort of… take money from you to do all kinds of things.""

Tomašev answered:

"Yes, and this has happened to people who have experimented with agents and have given them access to wallets, right, to do things.

As you say, in the early days of this all, when we are especially experimenting internally or anyone else's, this is done in a trusted environment. So you don't necessarily, in

> *your early prototyping, have to deal with any of this.*
>
> *…but once you deploy on the web, especially now with, AI really being used in all sorts of places, the more agents there are, the more incentives there are for malicious people to do malicious things because they have a higher surface area to target."*

The More AI Agents The Higher The Incentive

That last part about higher incentive to target AI agents makes sense. Systems that are used on a large scale quickly become targets for scammers and hackers, which is why systems like WordPress and Windows are frequently targeted. What Tomašev indicates is once AI agents become more prevalent at scale we will probably begin to see more criminal activities focusing on exploiting AI agents on the web.

Watch the interview at the 23 minute mark:

[embedded content]

Featured Image/Screenshot

https://www.searchenginejournal.com/google-deepmind-admits-large-scale-ai-agent-deployment-is-unsafe-today/580321/