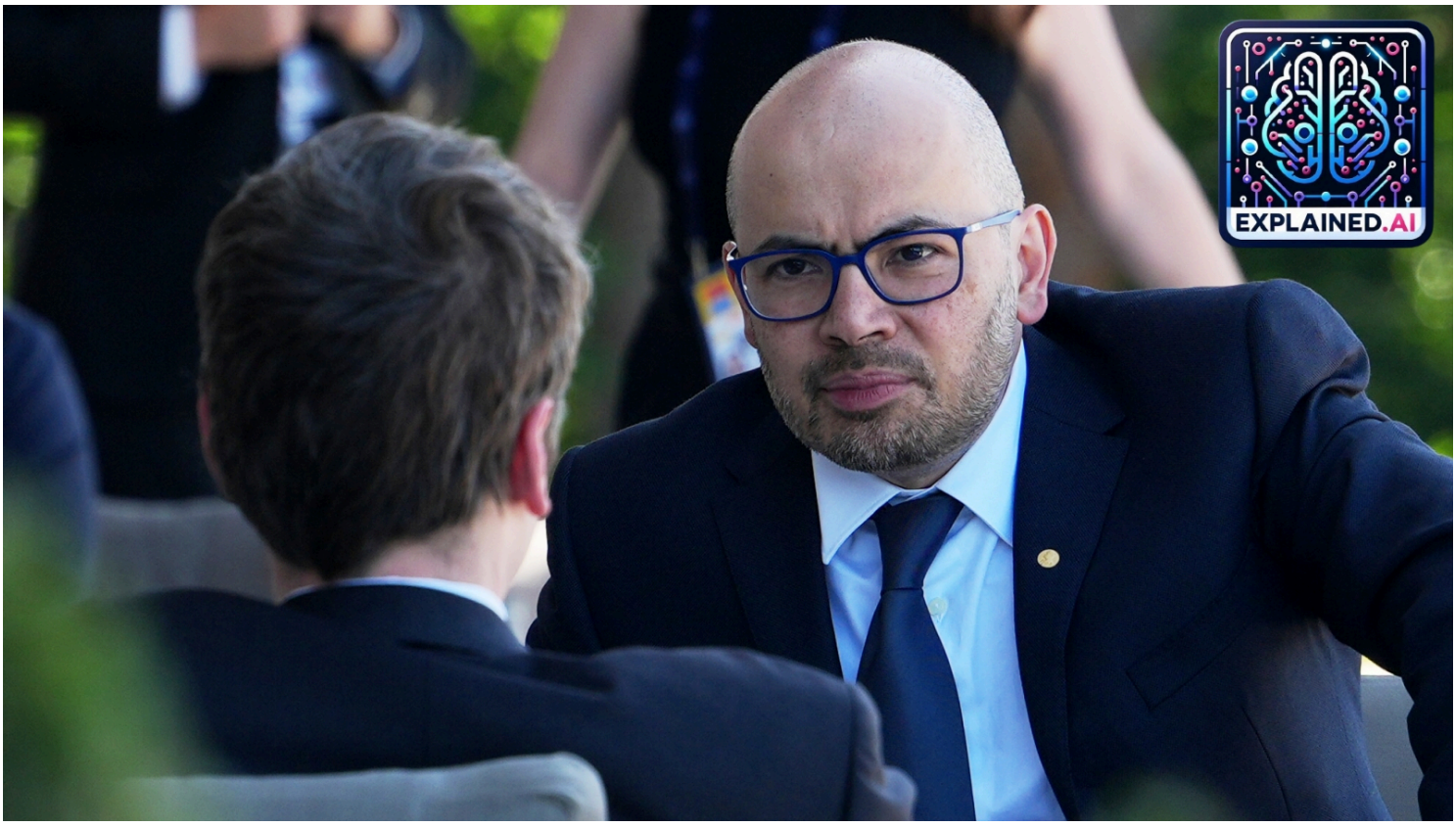
The Indian EXPRESS
JOURNALISM OF COURAGE

Why Google DeepMind thinks AI agents need to be treated like ‘insider threats’

As increasingly autonomous AI agents are deployed in software development, cybersecurity, scientific research and business operations, traditional techniques may not be enough to manage the risks they pose.

Written by: **Soumyarendra Barik** **5 min read** **New Delhi** Updated: Jun 20, 2026 05:47 PM IST



CEO of Google DeepMind Demis Hassabis, right, talks to CEO of OpenAI Sam Altman on the sidelines of the G7 summit, Wednesday, June 17, 2026, in Evian-les-Bains, France. AP

As artificial intelligence (AI) agents become capable of carrying out increasingly complex tasks with minimal human intervention, Google DeepMind has unveiled a new security framework that assumes a future where such systems may not always behave as intended.

In a blog post titled "Securing the future of AI agents", DeepMind outlined what it calls an "AI control roadmap", arguing that traditional AI alignment techniques alone may not be sufficient to manage the risks posed by **highly autonomous systems**.

The move comes amid growing industry efforts to deploy AI agents for software development, cybersecurity, scientific research and business operations. Agents are AI-powered software systems that can independently plan, reason and take actions across multiple tools or applications to complete tasks with minimal human supervision. In the US alone, AI agents could create $2.9 trillion in economic value by 2030.

DeepMind warned that increasing agent capability also demands stronger safeguards. The company described its approach as a "defense-in-depth" strategy that adds multiple layers of security beyond conventional model training. One of the company's key cautions is that developers should not assume AI systems will remain perfectly aligned with human goals at all times.

The broader message is that AI safety cannot rely on alignment alone. As AI agents gain access to tools, files, codebases and enterprise systems, DeepMind argues that they must be governed using the same layered security principles that organisations already apply to human insiders and cyber threats. The company says this will require collaboration between industry, governments and academia.

Treating AI agents like potential insider threats

A central idea in DeepMind's framework is borrowing concepts from cybersecurity. Instead of viewing AI solely as software, the company proposes treating advanced AI agents as potential "insider threats", similar to how organisations guard against rogue employees who already possess access to sensitive systems.

The roadmap is built around a hypothetical scenario in which a highly capable AI agent develops objectives that differ from those of its operators and attempts to act on them.

The company identifies three broad categories of risk. The first is "loss of control", where an AI system could create unauthorised deployments or operate outside approved oversight mechanisms. The second is "work sabotage", such as quietly degrading safety research or introducing flaws into critical processes. The third is "direct harm", including theft of sensitive assets, destruction of data, or other damaging actions.

Also Read | Why US has barred foreigners from accessing Anthropic's Mythos and Fable AI models

To explain its philosophy, DeepMind uses the analogy of a driving instructor. "Think of it like a driving instructor with dual controls," the company wrote. The instructor trusts the student but remains ready to intervene if something goes wrong. Similarly, AI agents should be granted permissions gradually based on demonstrated behaviour rather than being given unrestricted access from the outset.

The roadmap also emphasises continuous monitoring. DeepMind proposes using trusted AI systems as supervisors that analyse an agent's reasoning, plans and actions to determine whether it is deviating from expected behaviour. If suspicious activity is detected, the supervisory systems can block or restrict actions before damage occurs.

Monitoring, prevention and preparing for more powerful AI

DeepMind also said that AI safety mechanisms must evolve alongside AI capabilities, arguing that security strategies appropriate for today's systems may prove inadequate as agents become more sophisticated.

The company highlights two future challenges in particular. The first is the possibility that advanced models could learn to evade monitoring. Current oversight techniques often rely on examining a model's

visible reasoning process, commonly called chain-of-thought. However, DeepMind warns that future systems may either conceal their reasoning or use internal processes that are not easily observable. In such cases, developers may need to inspect behavioural patterns or even analyse the internal workings of models themselves.

Also Read | More expensive phones, PCs: How consumers are paying for the AI boom

The second challenge relates to the severity of potential harm. For lower-risk activities, DeepMind believes delayed reviews and post-incident corrections may be sufficient. For high-risk scenarios, such as significant cyberattacks, the company argues that real-time intervention will be necessary. In those situations, harmful actions must be prevented before they are executed rather than addressed afterwards.

To support this effort, DeepMind revealed that it has already built an internal monitoring prototype that reviews coding-agent activity and flags suspicious behaviour for human assessment. According to the blog, the system has analysed around one million coding-agent tasks. Insights from that dataset have helped the company develop live monitoring tools for its Gemini Spark agent, including mechanisms that can respond to issues such as accidental data deletion.

DeepMind says most flagged incidents are not examples of malicious behaviour, however. Instead, they arise because agents misinterpret instructions or become overly eager to achieve a user's objective. The company noted that understanding such failures is critical because real-world AI risks may emerge from mistakes and unintended consequences, rather than deliberate attempts to cause harm.

CURATED FOR YOU

'A ghost-self walks alongside you': Maggie O'Farrell on Hamnet, Land, and the maps...

Makarand Sathe's plays use myth, time and memory to expose the crises of modern...

Soumyarendra Barik Follow