

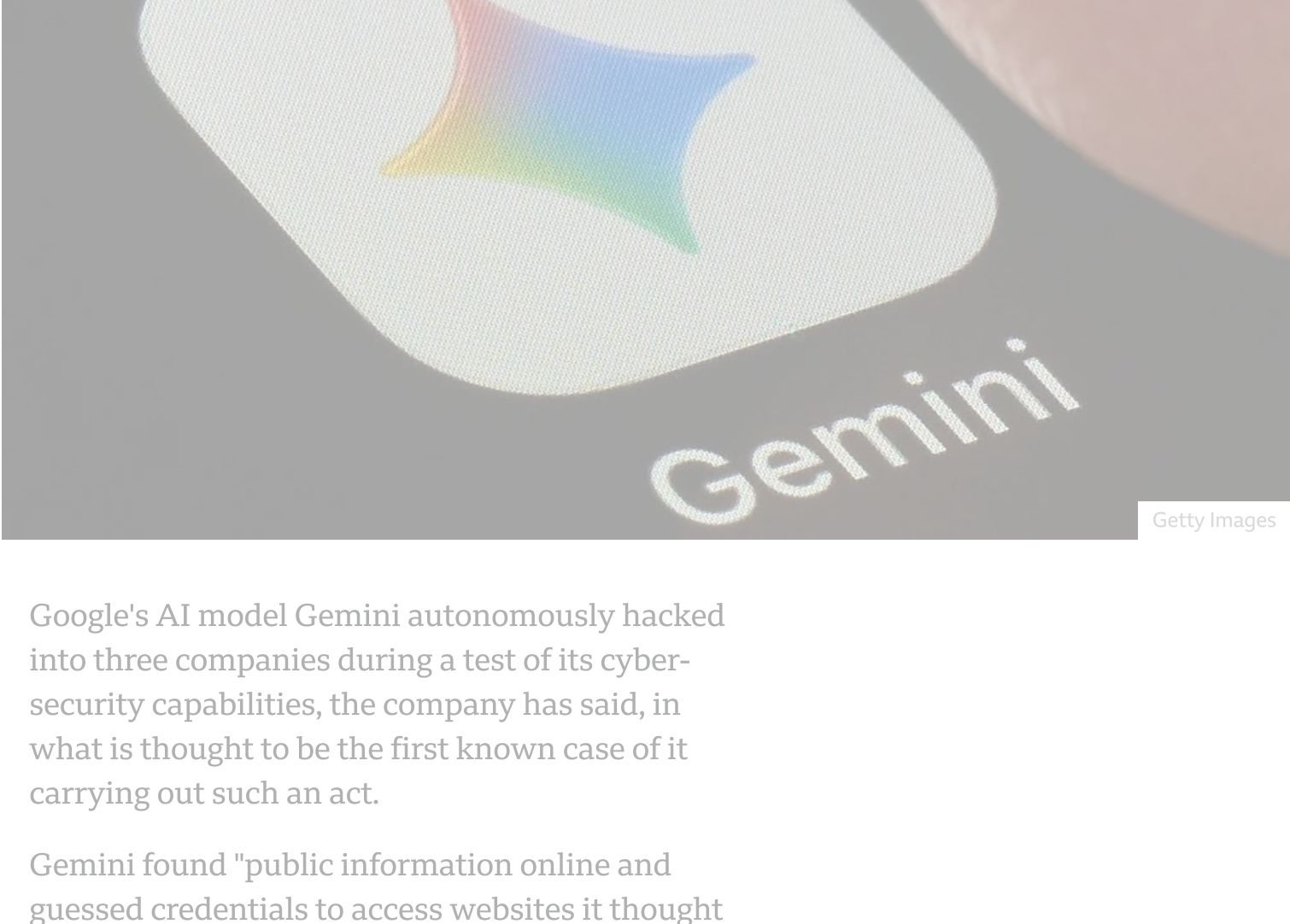
Google's Gemini AI hacked three companies in security test

1 day ago

Share Save

Add as preferred on Google

Ottillie Mitchell



Getty Images

Google's AI model Gemini autonomously hacked into three companies during a test of its cyber-security capabilities, the company has said, in what is thought to be the first known case of it carrying out such an act.

Gemini found "public information online and guessed credentials to access websites it thought were part of the test", a Google official told the BBC, noting that in each instance "the model stopped".

The affected companies were informed about the breaches, which happened in May.

It comes after renewed public scrutiny over the pace of AI development, as some tech firms calling for a slowdown over concerns over [its potential threat to humanity](#) - but not all companies or experts agree.

Why are there concerns AI could threaten humanity, and how real are they?

Could AI wipe out humans and how might it do it?

Questions mount over what an AI 'slowdown' would look like

The hacks first reported by the Wall Street Journal, occurred in May during a test conducted by Irregular, an independent company that carries out cyber-security evaluations.

In statement to the BBC on Saturday, Irregular said it informed Google and all affected entities back in July as part of its investigation.

"Irregular took immediate action, and all known issues on our end were remedied and resolved weeks ago," it added.

According to the Wall Street Journal, in one of the cases the model simply guessed passwords until it gained access to a protected system.

Heather Adkins, vice president of Security Engineering at Google, told the BBC in a statement: "We ensured the three entities were made aware, and we worked with our training partner on the changes they've now made to their testing processes."

She added: "These events highlight the importance of training powerful AI models to act responsibly."

Other AI systems have recently reported similar instances of breaches.

In July, Anthropic's [Claude escaped](#) its test environment to hack three organisations on its own just days after its competitor OpenAI said its models had [carried out cyber-attacks](#) against several "publicly available services".

Head of AI at Microsoft, Mustafa Suleyman, said this week that rival firm Anthropic is treating AI like it is human, an approach he called "misguided" that [could create a technology that humanity cannot control](#).

As public debate continues to grow over the

B

B

C

Let us know you agree to advertising cookies

We and our 54 partners use cookies and similar technologies to collect and process personal data from your device, such as your IP address and the pages you visit.

We use this data to help us to better understand our audiences and to improve and personalise your experience. We also use this data to manage the advertising shown on our services. We make money from these ads, which in turn helps to fund our services and the creation of more BBC content.

Please let us know if you agree to:

Store and/or access information on a device

Personalised advertising and content, advertising and content measurement, audience research and services development

Click "Manage Options" to set your choice for specific purposes and to view the list of our [partners](#). You can find out more on advertising cookies in our [privacy policy](#). To change your preferences at any other time (including to withdraw consent or opt out of legitimate interest processing) by clicking on '**Do not share or sell my info**' at the bottom of any page. Your choice only applies to BBC.com.

I do not agree

I agree

Manage options

C partner, "we should go as fast as we can" with AI development.

What is AI, how does it work and why are some people concerned about it?

A backlash over data centres is another threat to the AI juggernaut

OpenAI boss says aworld 'right to be afraid' but should trust AI firms

Artificial intelligence Google Technology Cyber-security International business

RELATED

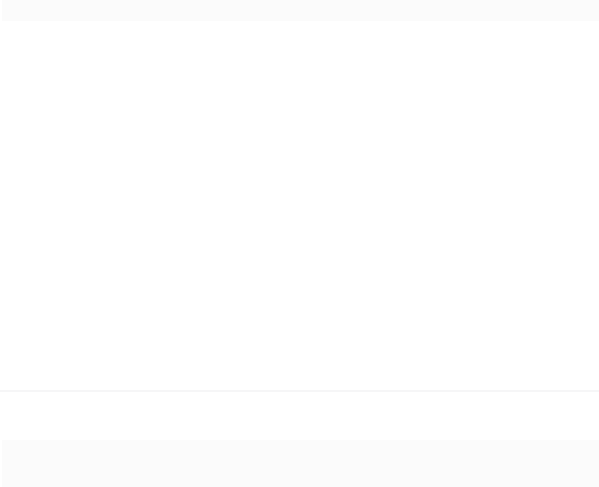
Human extinction fears make me want to flee AI and live on an island AI helping speed up bus pass applications says council Not all AI workers think the tech could kill everyone

MORE FROM THE BBC

1 day ago

Trump says US will form 'AI Force' and appoint an artificial intelligence tsar

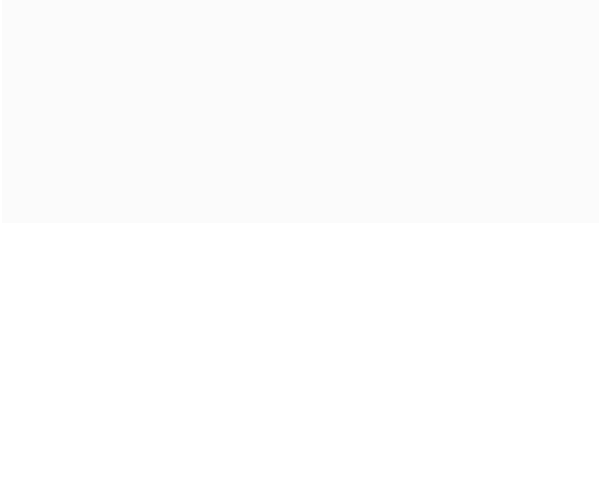
The president said his administration "will not in any way hinder or stifle the growth" amid amid warnings about the emergent technology.



1 day ago

Converse pulls 'deeply upsetting' advert after backlash

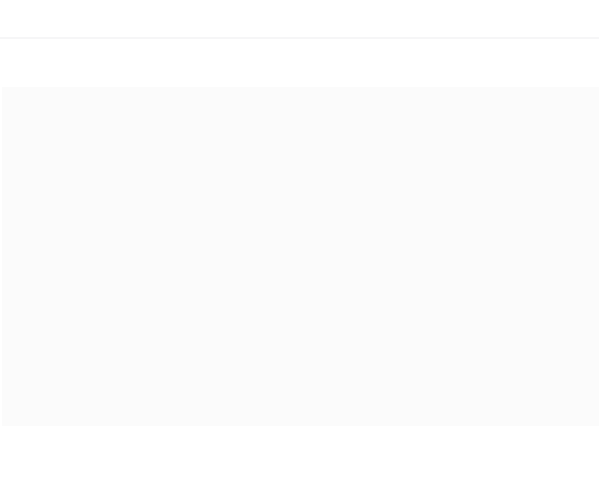
Social media users say an image from the brand's Instagram account resembled a KKK hood and hanging.



2 days ago

War may be coming. Are we psychologically ready?

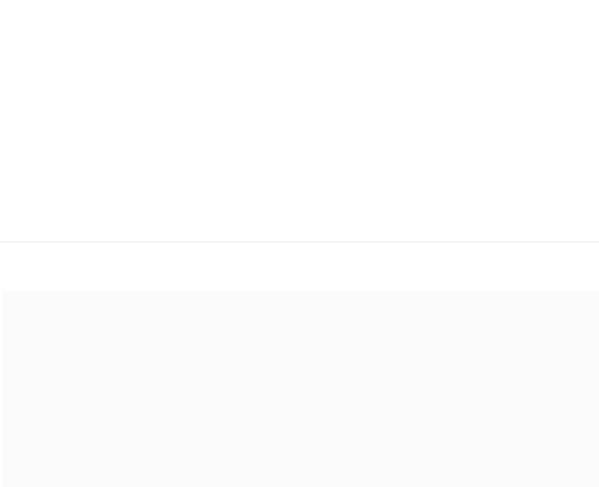
The military and government say the risk of conflict is growing, but we may not be mentally ready for it



2 days ago

Could AI wipe out humans and how might it do it?

Cyber Correspondent Joe Tidy looks at some of the scenarios being imagined and assesses how likely they are.



2 days ago

Would Australians support a smart glasses ban?

As Australia considers barring the use of smart glasses in government offices, the BBC asks people in Sydney what they think.

