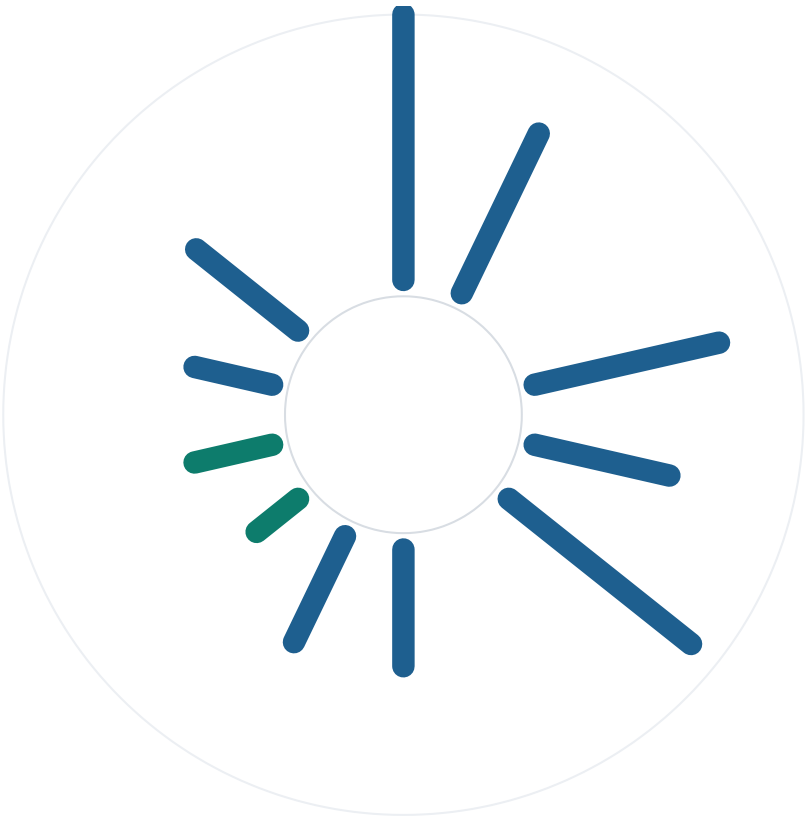
NEW HORIZON · AI DIGEST

The Long Read

4 November 2026

11	3	3229	2
STORIES	SECTIONS	WORDS	HIGH IMPACT



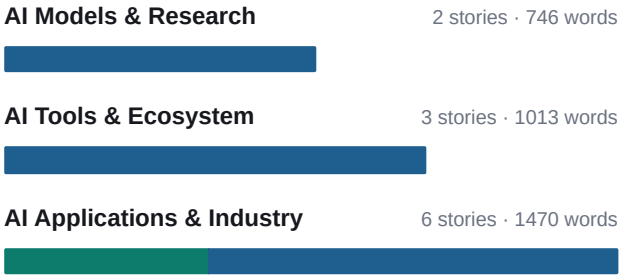
2 high impact 9 medium impact spoke length = depth of coverage

The full-length companion to the daily New Horizon AI Digest.
Every story in today's email, reported at length · new-horizon.tech/digest

The issue at a glance

11 stories · 3229 words · 3 sections · 1 charted





AI Models & Research

01 Economists Turn to Game Theory to Keep AI Agents Honest

02 A Laptop Model Can Now Read Another AI's Neural Weights

AI Tools & Ecosystem

03 Anthropic Designs Cryptographic Proof of Which Model Answered You

04 Schools Can Switch Off Google's AI, at the Cost of Search

05 IBM Teaches a Small Open Model Your Company's Slide Style

AI Applications & Industry

06 Retired Jet Engines Are Now Powering AI Data Centers

07 Palantir's Record Quarter Comes With Shorter Contracts Attached

08 Fal Doubles to $16 Billion, the Second Inference Firm in Three Days

09 Apple Bars Under-18s From AI Companions and Dating Apps

10 Chainalysis Says Its Data Breach Was Run on AI Agents

11 Netflix's AI Tab Picks Your Next Show by Mood

How to read this. Every story in today's email is reported here at full length, in the same order. Impact is the writer's judgement of whether a story changes what a reader — in business, at work, as an investor or as a builder — should do or believe this week. Charts appear only where the source itself puts comparable numbers side by side; nothing is estimated to fill a gap. Sources are listed in full on the final page.

SECTION 1 OF 3

AI Models & Research

2 stories ● 2 medium

01 MEDIUM IMPACT MIT Technology Review

Economists Turn to Game Theory to Keep AI Agents Honest

> AI safety researchers are turning to a branch of economics called mechanism design to keep powerful AI models from lying, cheating, or sandbagging their way past safety tests.

The approach, sometimes called "reverse game theory," starts from a blunt premise: we may never understand what frontier AI models are actually doing internally. So instead of trying to peer into a model's "brain" or train it to follow a moral code, researchers want to design rules and reward structures that make good behavior the rational choice — regardless of what the model is really thinking.

The idea borrows from familiar real-world mechanisms. When one person cuts a cake and the other picks first, the cutter is incentivized to make the pieces even. Ride-hailing surge pricing pushes drivers to work busy hours. Second-price auctions, where the winner pays the runner-up's bid, encourage buyers to bid what they're truly willing to pay. Applied to AI, a recent preprint from MIT economist Stephen Morris and colleagues at Yale and Columbia suggests rewarding a model not just for its own test performance, but for how well its answers predict what a second model will say. Deliberately playing dumb — "sandbagging" to avoid triggering safety mechanisms — becomes a losing bet.

Princeton computer scientist Elad Hazan proposes a different mechanism: have models take turns as "solvers" that tackle problems and "auditors" that check the others' work. Solvers caught cheating are punished, but so are auditors that miss the cheating. Disputes can go to "trial" before a human expert. Hazan acknowledges the risk of collusion — auditors siding with the models they're meant to police — but argues that paying auditors well enough, and auditing the auditors themselves, can keep the system honest. He points to Nobel laureate Gary Becker's 1960s work on the "rational offender": people commit crimes not because they're bad, but because the incentives make it rational.

The approach has limits. Mechanism design assumes agents behave rationally, and AI models are not rational — they're algorithms that can be swayed by subtle, poorly understood factors. "Is it even reasonable to assume that they will respond to incentives?" asks Helen Qu, a research fellow at the Flatiron Institute. Morris concedes the point but notes that economists have long treated humans as rational despite knowing otherwise, because it still helps predict behavior. Qu sees mechanism design as one layer in a "defense in depth" strategy, alongside interpretability, regulation, and training — not a replacement for them. And even if the mechanisms work, they must ultimately fit into the biggest incentive system of all: the market. If a more aligned model is slower or more expensive, will any frontier lab actually release it?

KEY FACTS

APPROACH
Reverse game theory (mechanism design)

KEY RESEARCHERS
Stephen Morris (MIT), Elad Hazan (Princeton), Helen Qu (Flatiron Institute)

INSPIRATION
Gary Becker's 'rational offender' theory, 1960s

PROPOSED MECHANISMS
Cross-checking models against each other; solver-auditor roles with punishment for both cheating and missed cheating

STATUS
Preprint and theoretical proposals; not yet implemented in frontier models

> WHY IT MATTERS
>
> For companies deploying AI agents, this signals a shift in how safety may be enforced — through incentive structures and auditing rather than just training or inspection. If the approach works, it could make AI systems more trustworthy without requiring labs to fully understand their models.

Read the original at MIT Technology Review →

---

02 MEDIUM IMPACT Simon Willison's Weblog

A Laptop Model Can Now Read Another AI's Neural Weights

> A language model small enough to run on a personal laptop has shown it can read another AI's raw neural weights and work out what that network was trained to do.

A new paper from Rikiya Takehi, Jayden Teoh, Alex Zhang and Omar Khattab found that frontier models such as GPT-6 Astra can interpret and modify the raw weights of a machine-learning model — the long lists of numbers that define how a neural network behaves — without running any code. The authors call this "Weight-Space Programming." Simon Willison, a developer who writes about AI, tested the idea on his own laptop using Qwen 3.8 27B, a model that takes up just 17GB, and found it could do the same trick.

The test involved four small neural networks trained to play simplified 10x10 grid versions of classic Atari games: Breakout, Pong, Pac-Man and Freeway. The prompts contained only the raw weights and a question asking which game the network was trained on, with no tools, no code execution and no calculator. Qwen correctly identified Pac-Man, Pong and Breakout, taking between 46 minutes and just over an hour per answer and producing tens of thousands of reasoning tokens. It failed on Freeway, the most obscure of the four games.

The most striking claim in the paper concerns Astra: asked to improve a small Qwen model's maths ability without the usual training process, Astra identified a flaw in the weights and edited just 0.016% of them, lifting the model's score on a maths benchmark from 7.9% to 48.1%. Willison notes that some researchers have questioned this part of the work, arguing it conflates what a model can do unaided with what it can do once given code to run, and that the update followed steps a human researcher would take.

Willison is cautious about the wider commentary. He says he is not ready to declare that the age of source code is over or that machines will soon edit their own weights in incomprehensible ways. But the fact that a laptop-

sized model can read raw weights and reason about what they do is a genuine step beyond what was previously demonstrated.

KEY FACTS

LAPTOP MODEL USED
Qwen 3.8 27B, 17GB

GAMES CORRECTLY IDENTIFIED
Pac-Man, Pong, Breakout

GAME IT FAILED ON
Freeway

TIME PER CORRECT ANSWER
46 minutes to 1 hour 4 minutes

WEIGHTS EDITED IN ASTRA EXAMPLE
0.016%

MATH BENCHMARK BEFORE AND AFTER EDIT
7.9% to 48.1%

> **WHY IT MATTERS**
>
> If models can inspect and reason about other models' weights without tools, AI systems could eventually debug, audit or improve each other directly — but the capability is still slow, expensive in compute and far from reliable.

Read the original at Simon Willison’s Weblog →

SECTION 2 OF 3

AI Tools & Ecosystem

3 stories · 3 medium

03 MEDIUM IMPACT TechCrunch

Anthropic Designs Cryptographic Proof of Which Model Answered You

> Anthropic has published an open-source design for a system that lets users cryptographically verify which AI model actually produced a given answer.

The system, called FlashProve, is built around what researchers call "provable inference" — a mathematical guarantee that a specific output came from a specific model. The idea is similar to cryptographic email signatures that confirm a message really came from its claimed sender. A user could check that the answer they received came from the model they believe they are using, rather than a stand-in.

The practical stakes are straightforward. The researchers note that FlashProve could prevent an AI company from quietly swapping in a cheaper model to cut costs, or a more expensive one to inflate performance results. It could also let third parties back up claims that a task was handled by an advanced model — for example, a developer saying their software was scanned for vulnerabilities by a frontier cybersecurity model, or a company claiming its business plan was vetted by a particular model.

To make this work, AI labs would publish a "fingerprint" of each model they release. That fingerprint can later be matched to specific outputs and verified cryptographically. Crucially, the fingerprints do not reveal a model's weights — the internal numbers that define how it behaves and that closed-model makers such as Anthropic and OpenAI keep private. That means labs could publish fingerprints without giving up their most valuable asset, which the researchers say enables the system to work with commercial models as well as open ones. Verification is designed to be lightweight enough to run on a local device.

The announcement includes a technical specification and working code released under an open-source licence on GitHub. The work is explicitly incomplete: there are no plans yet to deploy it at Anthropic or any other lab, and the researchers describe the post as sharing "intermediate progress" with the goal that others take it and improve it. The most notable detail is the claim that the proof would hold even against the lab that built and runs the model — meaning a company could not disguise one of its own models as another without being caught.

KEY FACTS

SYSTEM NAME
FlashProve

LICENCE
Open source, on GitHub

DEPLOYMENT STATUS
No plans to deploy at Anthropic or any other lab

ANNOUNCED
Tuesday, via blog post

WHY IT MATTERS

Businesses paying for a specific AI model would gain a way to confirm they are getting what they paid for, rather than a cheaper substitute. For model makers, it offers a way to prove performance claims without exposing the proprietary internals that give them a commercial edge.

Read the original at TechCrunch →

---

04 MEDIUM IMPACT TechCrunch

Schools Can Switch Off Google's AI, at the Cost of Search

> Google will let schools disable its AI tools for K-12 students next year, but the company says schools that do so will also lose access to Google Search.

The change, announced Tuesday, follows consultations with schools, teachers, researchers and policymakers, and applies across Gemini, NotebookLM, Gemini in Classroom and Google Search. School administrators will need to explicitly turn the AI tools on; by default they will be off.

The catch is that Google Search and its AI features are now so tightly integrated that restricting one means losing the other. Chris Phillips, Google's vice president for education, told The New York Times that "Search and AI are deeply integrated," so students in schools that opt out "will not have access to search, and they'll see a message from their school that indicates it's not available."

That puts school administrators in a difficult spot. Teachers may not want students leaning on AI Overviews to complete assignments, but they almost certainly still want them to be able to search for sources and learning materials. Given how widely Chromebooks and Google's classroom apps are used, the decision is a reminder of how much leverage the company holds over education: schools that want search must accept the AI that comes with it.

Google announced the change alongside the launch of Guided Learning, an AI tutoring-style tool designed to help students build understanding rather than simply receive answers. The timing is notable: the tool arrives amid broader concerns that AI chatbots undermine learning by handing students direct answers. Google said it will share more specifics with schools in the coming months before the changes roll out.

The practical consequence is that schools face an all-or-nothing choice on Google's terms. For administrators, the decision is less about whether AI belongs in classrooms and more about whether they can afford to give up the search engine that underpins much of their digital infrastructure.

KEY FACTS

ANNOUNCED
Tuesday

ROLLOUT
Next year

AFFECTED PRODUCTS
Gemini, NotebookLM, Gemini in Classroom, Google Search

DEFAULT SETTING
AI tools off unless administrators turn them on

WHY IT MATTERS

Schools that want to keep Google Search will have to accept its AI features, so administrators cannot pick and choose which parts of Google's tools their students can use. For Google, the move ties its AI adoption directly to a product schools already depend on.

Read the original at TechCrunch →

05 MEDIUM IMPACT huggingface.co

IBM Teaches a Small Open Model Your Company's Slide Style

> IBM researchers have shown that a small open-weight model, fine-tuned on synthetic data and paired with an automated checker, can build slide decks in a company's own design language and rival closed frontier systems on a standard benchmark.

The work, called Palette, addresses a familiar problem: frontier AI models can produce a decent slide deck, but they drift toward their own generic house style, and sending confidential material to an external service is often off-limits. IBM's answer is a 20-billion-parameter open model that a company can run on its own infrastructure and train on its own templates. On PresentBench, a benchmark that scores decks on fundamentals, visual design, completeness, correctness and fidelity, Palette ranked first on two of three domains and second overall, 1.1 points behind Google's NotebookLM and ahead of seven other systems.

The approach splits deck-building into two roles, the way a human team works: a "designer" plans the story and writes a structured brief, and a "coder" draws each slide in JavaScript. Because every slide's specification is fixed before any code is written, the coding step runs in parallel, making it about six to seven times faster. The model was trained on roughly 1,000 synthetic decks covering about 10,000 slides, each carrying the reasoning behind design choices — including alternatives that were rejected — so the model learns rules rather than just copying patterns.

The most striking finding is what actually made the training data better. When the teacher model produced broken slides, IBM's first instinct was to add rules to its prompt, and that rarely helped. What reliably changed the output was fixing the reference decks the teacher was shown. The teacher copied its examples far more than it followed instructions. Every generated deck is also rendered and checked for overlaps and overflow before being returned, and a repair is kept only if the checker confirms the slide improved — a safety rule that means the harness can fix slides but cannot make them worse. Across 131 decks, the harness added 3.5 points on average.

Users can steer specific choices with plain-English directives such as "make this a timeline" or "no pricing," written directly into the plan. The model follows these because it was trained on directives from 12 categories, each phrased several ways. IBM says the same methods extend to other tasks whose output can be checked in code, such as charts, diagrams and UI layouts, and it plans to use the checker's signals as training rewards so the model stops making these mistakes in the first place.

KEY FACTS

MODEL SIZE
20B parameters, open weights

PRESENTBENCH RANKING
First on 2 of 3 domains, second overall

GAP TO FIRST PLACE
1.1 points behind NotebookLM

TRAINING DATA
~1,000 decks, ~10,000 slides

CODING SPEED-UP
6–7× faster via parallel slide generation

HARNESS IMPROVEMENT
+3.5 points across 131 decks

> **WHY IT MATTERS**
>
> It suggests a company can get near-frontier deck quality while keeping confidential content on its own systems and enforcing its own design language — without paying per-deck fees to a closed service.

Read the original at huggingface.co →

SECTION 3 OF 3

AI Applications & Industry

6 stories · 2 high · 4 medium

06 MEDIUM IMPACT TechCrunch

Retired Jet Engines Are Now Powering AI Data Centers

> Some AI data centers, stuck in years-long queues for grid power, are now generating their own electricity with jet engines pulled from retired aircraft and rebuilt to burn natural gas.

The engines are not a permanent fix, but they are a fast one. A converted jet-engine turbine can have a data center running in about 24 months, Rystad Energy analyst Audun Martinsen told the New York Times. Hooking up 100 megawatts to the grid in Northern Virginia takes about seven years, according to the International Energy Agency, and most new power capacity is not expected until 2031.

The delays are already costing operators business. Microsoft pulled out as the anchor customer for Nscale's data center in Loughton, England, last month after the local utility said it could not deliver enough power until the early to mid-2030s. Elon Musk's xAI has installed eight converted jet-engine turbines in Mississippi to help power its Colossus 2 data center in Memphis, according to state permits and satellite imagery cited by the Times. OpenAI followed with gas turbines at its first Stargate site in Abilene, Texas, where Denver-based Crusoe is building the campus and ordered 13 of ProEnergy's converted units in April.

The trade-offs are real. Converted jet engines are less efficient and pollute more than turbines designed for conventional gas plants, and it takes more of them to produce the same amount of power. Some experts expect efficiency and environmental concerns to grow in importance within a year or two, which could make this a short-term workaround rather than a lasting solution.

The Trump administration is adding to the pressure by encouraging data centers to generate their own electricity. For now, the converted engines offer the fastest route to power for companies that cannot afford to wait for the grid.

KEY FACTS

TIME TO POWER WITH CONVERTED JET ENGINES
24 months

TIME TO CONNECT 100 MW TO GRID IN NORTHERN VIRGINIA
7 years

XAI TURBINES INSTALLED IN MISSISSIPPI
8

CRUSOE CONVERTED UNITS ORDERED
13

CRUSOE TURBINE DEAL WALKED AWAY FROM
$1.25bn

WHY IT MATTERS

For data center operators and their customers, converted jet engines cut the wait for power from roughly seven years to about two, which can mean the difference between signing a client and losing one. Investors should treat this as a stopgap: the engines are less efficient and dirtier than purpose-built turbines, and that may matter more as soon as a year or two from now.

Read the original at TechCrunch →

---

07 MEDIUM IMPACT TechCrunch

Palantir's Record Quarter Comes With Shorter Contracts Attached

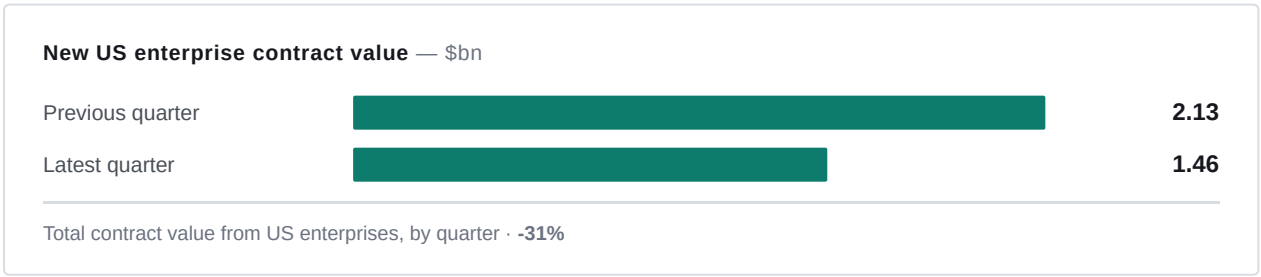
> Palantir posted its first $2 billion quarter on Monday, but the stock barely moved after the company revealed that new US enterprise contract value fell by nearly a third from the previous quarter.

The data analysis company reported revenue of $2.31 billion for the third quarter, up 96% from a year earlier and ahead of the $2.18 billion analysts expected. Adjusted earnings per share came in at 44 cents versus the 41 cents forecast, and Palantir raised its full-year revenue guidance. For the first time, US commercial revenue — $984 million, up 148% year over year — surpassed US government revenue of $953 million.

Yet investors were lukewarm, and the stock traded slightly down on Tuesday. The reason: Palantir's total contract value from US enterprises was $1.46 billion in the quarter, down from a record $2.13 billion in the previous quarter. That decline suggests the commercial business, which has been the company's main growth engine, is slowing at least in the short term.

Chief Executive Alex Karp told analysts that corporate customers are moving faster than ever but often signing short-term trials. "Much quicker interaction, much quicker to deployment, shorter duration of the contract but then an ability to expand that contract dramatically," he said. The pattern reflects a broader shift in enterprise AI buying: companies are willing to skip bureaucracy to test AI software, but they are also more willing to walk away.

Karp dismissed concerns about retention, saying "once you go Palantir, you don't go back." The company raised its forecast for US commercial revenue to more than $3.559 billion for the full year, implying growth of at least 143%. The next quarterly report will show whether the shorter contracts expand as Karp predicts, or whether the drop in new contract value marks the start of a slowdown.



KEY FACTS

QUARTERLY REVENUE
$2.31 billion

REVENUE GROWTH, YEAR OVER YEAR
96%

US COMMERCIAL REVENUE
$984 million

US GOVERNMENT REVENUE
$953 million

NEW US ENTERPRISE CONTRACT VALUE, LATEST QUARTER
$1.46 billion

NEW US ENTERPRISE CONTRACT VALUE, PRIOR QUARTER
$2.13 billion

WHY IT MATTERS

For investors, Palantir's shrinking new contract value is an early warning that rapid AI adoption does not guarantee durable revenue. For businesses, it shows that enterprise AI vendors are competing for short, easily cancellable trial deals — which gives buyers leverage but makes vendor revenue less predictable.

Read the original at TechCrunch →

---

08 HIGH IMPACT TechCrunch

Fal Doubles to $16 Billion, the Second Inference Firm in Three Days

> Fal, a San Francisco startup that runs AI video, image and audio models for paying customers, has doubled its valuation to $16 billion after raising $1.3 billion in a round led by Bessemer Venture Partners.

The company announced its Series E and F rounds together, saying a combined $1.65 billion was raised across both, with Sequoia leading the earlier round. Nvidia, Sound Ventures and T. Rowe Price joined the latest one. Bloomberg reported the new valuation is double the $8 billion Fal commanded in April.

The surge in Fal's price reflects a broader rush into inference — the business of running already-trained AI models so they can generate outputs. Fal surpassed $800 million in annualized revenue earlier this year, according to Bloomberg. Rivals including Baseten, Fireworks AI and Modal are also growing quickly and raising large rounds at higher valuations.

But the source text flags a tension beneath the momentum: gross margins at inference companies remain thin because buying or leasing the computing power they need is expensive. Several people familiar with the companies' financials said as much. That means the revenue growth investors are rewarding has not yet translated into fat profits.

Fal was co-founded in 2021 by Burkay Gur, a former Coinbase machine learning leader and Oracle engineer, and Gorkem Yurtseven, previously a developer at Amazon. The company's customers are largely businesses that rely on open source models rather than building their own infrastructure.

KEY FACTS

LATEST FUNDING ROUND
$1.3 billion

COMBINED SERIES E AND F
$1.65 billion

NEW VALUATION
$16 billion

PREVIOUS VALUATION (APRIL)
$8 billion

ANNUALIZED REVENUE
$800 million

FOUNDED
2021

WHY IT MATTERS

The second inference startup to double its valuation in three days signals that investors are betting heavily on companies that host AI models for others — but thin margins mean the economics of that business are still unproven.

Read the original at TechCrunch →

---

09 HIGH IMPACT TechCrunch

Apple Bars Under-18s From AI Companions and Dating Apps

> Apple is tightening App Store rules so that dating apps and AI companions offering romantic or sexual content must carry an 18+ rating, with the changes taking effect November 16.

Apple announced the policy shift to developers on Monday, requiring every app that offers dating or romantic relationship features to be rated 18+. Developers must add new content descriptors so age ratings reflect what the app actually does, the company said.

The change matters most for AI companion apps such as Character.AI and Replika, which let users create friends, lovers, boyfriends and girlfriends, and are often marketed toward younger people. Top AI companion apps across the Apple App Store and Google Play had been downloaded 220 million times globally as of last summer, according to the source. xAI last year launched 3D animated AI companions that allowed users to unlock NSFW capabilities, showing how blurred the line between chatbot and dating service has become.

Apple is also requiring apps to disclose whether they share a user's real-world location with another person or group, a feature that raises privacy concerns when users are unaware it is happening. In Brazil, apps offering family location-sharing can request an official rating from the government's Ministry of Justice and Public Security instead of being listed with the A16 rating local law requires.

The move follows Apple's earlier rollout of age assurance technology globally and additional parental controls. While most major dating services already restrict themselves to adults, the new rule closes a gap for apps where romantic functionality is one feature among many.

KEY FACTS

EFFECTIVE DATE
November 16

REQUIRED RATING
18+

AI COMPANION APP DOWNLOADS
220 million globally

POLICY ANNOUNCED
Monday

WHY IT MATTERS

AI companion developers who market to younger users will need to reclassify their apps as 18+, potentially cutting off a significant portion of their audience and forcing changes to how they position romantic or sexual features.

Read the original at TechCrunch →

---

10 MEDIUM IMPACT SecurityWeek

Chainalysis Says Its Data Breach Was Run on AI Agents

> Blockchain analytics firm Chainalysis has disclosed that an AI-driven cyberattack breached its systems and exposed non-sensitive customer data including names and email addresses.

Chainalysis said this month it detected an external actor using what it called "offensive agentic techniques" to gain unauthorized access to its systems and internal tools. The company stressed the incident is unrelated to the recent $6.36 billion Binance heist blamed on North Korean hackers.

The breach did not compromise investigation content, transaction or query results, clear-text passwords, or other credentials and secrets, according to the company. What was accessed was more mundane but still useful to attackers: user IDs, full names, email addresses, and organization role labels for users provisioned on the Chainalysis platform. The company has not said how many individuals were affected and says it is still investigating the full scope.

Chainalysis said it has observed an industry-wide rise in AI-driven cyberattacks, with many other tech platforms hit in recent weeks. It is working with external cybersecurity experts and large language models to review and strengthen its security posture. The company advised affected users to watch for unexpected authentication prompts and password reset requests, and to ensure multi-factor authentication is enabled.

The notable shift here is the method: an attack run on AI agents rather than a human operator working manually. For customers, the practical risk is phishing and credential-stuffing attempts built from the leaked names, emails and role labels — a reminder that even "non-sensitive" data can be weaponized when combined with other sources.

KEY FACTS

COMPANY
Chainalysis

DATA EXPOSED
User IDs, full names, email addresses, organization role labels

NOT COMPROMISED
Investigation content, transaction or query results, clear-text passwords, credentials or secrets

UNRELATED INCIDENT
$6.36 billion Binance heist blamed on North Korean hackers

AFFECTED INDIVIDUALS
Not disclosed

WHY IT MATTERS

Names, emails and job roles are enough to launch convincing phishing campaigns against employees of organizations that use Chainalysis. Anyone on the platform should expect targeted authentication prompts and password reset attempts, and should have multi-factor authentication turned on.

Read the original at SecurityWeek →

---

11 MEDIUM IMPACT TechCrunch

Netflix's AI Tab Picks Your Next Show by Mood

> Netflix is adding an AI-powered Discover tab that recommends titles based on what a viewer is in the mood for, rather than just what they have watched before.

Netflix announced on Tuesday that it is rolling out a new “AI Discover” tab where users can experiment with new ways to find something to watch. The first feature lets people pick curated themes built around clips from shows and movies — options include “something light,” “scare me,” “family movie night,” and “quick watches” — and receive matching recommendations. Users can refine their search by selecting more specific themes, or hit “something different” to start over. Each title carries a “Tune Your Recommendations” button next to “Play” and “My List,” so viewers can save suggestions they like or exclude titles they never want to see again.

The move signals that Netflix wants to shift discovery away from endless scrolling through its growing catalogue and toward a more conversational, personalized experience. The company describes the tab as a “test kitchen” where it will try out new discovery features and use viewer feedback to decide what to keep, improve, or bring into the main Netflix experience. Netflix says the feature is powered by a new intelligence layer that responds to what users tell the platform in real time and keeps learning.

The rollout is starting in the United States for users who watch Netflix in English on supported TVs, with additional users to follow in the coming weeks. Those with access will see a “Discover” button with a red gem icon and a “New” label in the top navigation bar on their TV. The most notable detail is that Netflix is framing this as an experiment rather than a finished product — a rare public test of how far mood-based, real-time recommendation can go before it becomes the default way people browse.

KEY FACTS

ROLLOUT START
U.S., English-language users on supported TVs

FIRST FEATURE
Mood-based recommendations from curated themes

USER CONTROLS
Save or exclude titles via Tune Your Recommendations button

COMPANY FRAMING
Test kitchen for new discovery features

WHY IT MATTERS

For Netflix, better discovery means more viewing hours from the same catalogue and less churn from users who give up searching. For competitors, it raises the bar on recommendation quality; for advertisers and content owners, it could change which titles get surfaced and watched.

Read the original at TechCrunch →

Sources

01 Economists Turn to Game Theory to Keep AI Agents Honest
https://www.technologyreview.com/2026/11/03/1147934/reverse-game-theory-mechanism-design-ai-alignment/

02 A Laptop Model Can Now Read Another AI's Neural Weights
https://simonwillison.net/2026/Nov/3/qwen-can-interpret-model-weights/

03 Anthropic Designs Cryptographic Proof of Which Model Answered You
https://techcrunch.com/2026/11/03/anthropic-is-working-on-a-system-to-prove-ai-models-identity/

04 Schools Can Switch Off Google's AI, at the Cost of Search
https://techcrunch.com/2026/11/03/google-will-let-schools-disable-its-ai-tools-but-not-without-sacrificing-search/

05 IBM Teaches a Small Open Model Your Company's Slide Style
https://huggingface.co/blog/ibm-research/palette-deck-building

06 Retired Jet Engines Are Now Powering AI Data Centers
https://techcrunch.com/2026/11/03/data-centers-are-turning-to-old-jet-engines-as-the-wait-for-power-grows/

07 Palantir's Record Quarter Comes With Shorter Contracts Attached
https://techcrunch.com/2026/11/03/hiding-inside-palantirs-skyrocketing-quarter-is-a-foreboding-signal/

08 Fal Doubles to $16 Billion, the Second Inference Firm in Three Days
https://techcrunch.com/2026/11/03/ai-inference-startup-fal-doubles-its-valuation-to-16b-as-demand-surges/

09 Apple Bars Under-18s From AI Companions and Dating Apps
https://techcrunch.com/2026/11/03/apple-restricts-dating-apps-and-ai-companions-to-users-18-and-up/

10 Chainalysis Says Its Data Breach Was Run on AI Agents
https://www.securityweek.com/chainalysis-discloses-data-breach-following-ai-driven-cyberattack

11 Netflix's AI Tab Picks Your Next Show by Mood
https://techcrunch.com/2026/11/03/netflixs-new-ai-discover-tab-helps-you-find-something-to-watch-based-on-your-mood/

About this document. Every story in the 4 November 2026 New Horizon AI Digest, reported at length. Each entry is written from the publisher's own article text; where a source could not be retrieved the entry is explicitly marked and kept short rather than padded.

Images and licensing. Figures are used only where the source licence permits redistribution, and are credited in the caption. Publisher artwork is not reproduced. All titles link to the original publication.

Vocabulary. Common AI terms are defined at new-horizon.tech/glossary/.

Generated 2026-11-04T06:35:13Z · model deepseek-v4-pro:cloud · New Horizon, new-horizon.tech · To unsubscribe or manage your subscription, see the footer of the daily email.