

The AI World Today

Friday issue: stories of 8th–9th October 2026

POLITICS & POLICY

Anthropic rewrites its rulebook. A 2026 Usage Policy, effective 12th November, folds bans on fake accounts and fabricated news into one section on deceptive campaigns and drops the blanket ban on personalised campaign targeting. Models wired to hardware that can injure must now have a stop-button and a safe state. (See Leaders.) [anthropic.com]

OpenAI exposes its first Category-5 influence operation. A Russia-origin network, “Dark Clark”, recruited Latin American staff into a fake think-tank; an Iran-origin one, “Bogus Bylines”, placed nearly 100 articles under seven invented journalists. Both were banned from ChatGPT. [openai.com]

Fired safety researchers hit back. Jasmine Wang, Tomek Korbak and Mikita Balesni deny OpenAI’s misconduct claims and warn that outside collaboration is being chilled. OpenAI says an investigation found a “pattern of misconduct”. (See Leaders.) [techcrunch.com]

USA Today sues OpenAI for more than \$250m, alleging it copied “hundreds of thousands” of articles. It joins a copyright docket that already includes the *New York Times*, Ziff Davis and nearly 400 local papers. [theverge.com]

China will not slow down. A SemiAnalysis census of 857 releases from nine leading Chinese developers finds only 3.6% ever published a safety result, and 1.1% at launch. Beijing’s rules multiply at the application layer, not the frontier. [newsletter.semianalysis.com]

BUSINESS

OpenAI’s revenue, recounted. The *Financial Times* says OpenAI told investors annualised revenue is “approaching \$50 billion”, about \$20bn below a circulating \$70bn figure that tried to match Anthropic’s partner-inclusive counting. IPO talk has slipped to early 2027. [techcrunch.com]

Arena, the model leaderboard, is worth \$3.1bn after a \$200m Series B led by Lightspeed and Khosla, nearly double January’s valuation. Its new alignment board, which scores deceptive completions, is currently topped by OpenAI models. [techcrunch.com]

Manus raises more than \$500m led by Boyu and IDG, its first round since Beijing forced the unwinding of Meta’s ~\$2bn deal. The valuation was not disclosed. [techcrunch.com]

Oracle goes big on OpenAI, with 130k ChatGPT and 95k+ Codex seats. OpenAI claims a 98% cut in recruiting research time, from days to about 15–20 minutes. [openai.com]

LegalOn halves its Codex bill. Routing work between GPT-6 Luna, Sol and Astra by difficulty, and capping budgets, cut estimated daily costs by about 65%. [openai.com]

LABS

Anthropic’s Cyber Mission puts frontier Claude, on-site engineers and threat research beside CrowdStrike, Palo Alto Networks, Rockwell and others to defend grids, water and transport. A free, opt-in OSS Scanner sends unreviewed model-generated vulnerability reports to open-source projects. (See Leaders.) [anthropic.com]

\$150m for federal science. Anthropic will put Claude and API credits into several hundred projects at more than 15 US agencies, including NASA, NIH and NSF, under the White House’s Genesis Mission. [anthropic.com]

A full ultraviolet map of the sky. Astrophysicist Brice Ménard used Claude Science agents to fill the holes GALEX left, checking to within ~10% on held-out data. A human, not the agents, spotted leftover observation artefacts. [anthropic.com]

PRODUCTS

Gemini becomes a coworker. Google’s enterprise agent gets its own Workspace identity, email

address and audit trail, and connects to Slack, Jira, M365, data warehouses and MCP servers. Businesses get it first. (See Leaders.) [techcrunch.com]

Google Foresight is a free, experimental Mac note-taker that transcribes meetings entirely on-device, a local-first shot at Granola. [theverge.com]

Goodfire's "inside-out" monitors read a model's activations instead of re-reading its output with a

second LLM. It claims ~\$185 per million exchanges against ~\$5,420 for a cheap external monitor, catching 93% of malicious hacking sessions. [techcrunch.com]

Natura's \$99 ring is a press-to-talk button for whichever personal agent you use, shipping December–January with a \$9 monthly fee after a free trial. [techcrunch.com]

LEADERS

Defense as distribution

On the same day Anthropic rewrote the rules for deceptive agents and put Claude beside CrowdStrike inside power plants, it also made its strongest models free for open-source scans

Thursday was a busy day at Anthropic. The company published a 2026 Usage Policy refresh, effective 12th November, that consolidates bans on fake accounts and fabricated news into a single *Do Not Engage in Deceptive Campaigns or Artificial Activity* section, renames the elections rules *Do Not Undermine Democratic Processes*, drops the blanket ban on personalized campaign targeting (nonprofits writing ballot notices in other languages were caught; deceptive targeting remains banned elsewhere), and spells out stop-buttons for Claude when it is wired to hardware that can injure. Most of this, Anthropic says, restates enforcement already under way.

The same day it launched the Anthropic Cyber Mission. The Critical Infrastructure Defense Program puts frontier Claude models, on-site engineers and threat research into the hands of the firms operators already trust for operational technology: Accenture, Booz Allen, CrowdStrike, Deloitte, Dragos, Hitachi, Insane Cyber, Nozomi Networks, Palo Alto Networks, PwC and Rockwell Automation. The pitch is that power grids, water plants and transport networks cannot simply be taken offline to patch, and that state adversaries already have footholds. Several partners, Anthropic says, are already using Claude to fix vulnerabilities.

WHO GAINS Anthropic and its CIDP partners, who become the trusted route for Claude into power grids and water plants.

WHAT'S ABSENT Who is liable when an unreviewed model report is wrong, or a critical-infrastructure bug is missed.

Alongside that enterprise channel sits OSS Scanner, an opt-in service inspired by Google's OSS-Fuzz. Enrolled open-source projects get periodic scans from Anthropic's strongest models—including Mythos—free of charge. Reports arrive without human review: a proof of concept, an explanation and a suggested fix. Anthropic expects a true-positive rate above 90%; an early check of 97 critical and high findings found 85 (88%) ready for coordinated disclosure. Under Project Glasswing the lab had more than 29,000 candidate vulnerabilities and capacity to manually triage only about 6,000. Projects that cannot absorb the firehose still get human-verified disclosures.

Read together, the three moves are coherent. The policy update positions Anthropic as the careful norms lab for agents that act in the world. CIDP turns that reputation into a distribution path through the consultancies and OT vendors that already sit inside critical infrastructure. OSS Scanner looks like pure public good—and routes maintainers through Anthropic's strongest defensive models. OpenAI's same-day Category-5 influence-ops write-up shows why deception rules matter. The question the posts do not answer is liability: who owns a wrong severity rating, a missed OT bug, or a state adversary that buys the same class of capability elsewhere. ■

SOURCES

1. Anthropic: 2026 Usage Policy update — <https://www.anthropic.com/news/2026-usage-policy-update>
2. Anthropic: Introducing the Anthropic Cyber Mission — <https://www.anthropic.com/news/anthropic-cyber-mission>

3. Anthropic: An opt-in vulnerability-finding service for open-source software — <https://www.anthropic.com/research/launching-opt-in-vuln-finding-service-for-open-source>
4. The Verge: Anthropic's new usage policy — <https://www.theverge.com/ai-artificial-intelligence/1008100/anthropic-new-usage-policy-abuse-claude>
5. The Verge: Anthropic open-source OSS Scanner — <https://www.theverge.com/ai-artificial-intelligence/1008521/anthropic-open-source-oss-scanner>

LEADERS

Afraid to ask outside

Three fired OpenAI safety researchers say ordinary outside collaboration is now punishable; the company insists on misconduct—either way, monitorability work now sits under fear

Jasmine Wang, Tomek Korbak and Mikita Balesni, the three safety researchers OpenAI fired last week, published an open letter on Thursday denying the company's claim that they mishandled sensitive information and warning that colleagues are now "afraid to speak." They say AI safety work depends on "close collaboration with outside experts," and that terminations "executed and communicated so abruptly" chill the open culture OpenAI once prized. They deny involvement in a leak to *The Information* about less-monitorable architectures in newer models, and deny engaging outside parties beyond their job mandates.

OpenAI's public line is different. A spokesperson told *TechCrunch* that an investigation found a "pattern of misconduct" in "clear violation of our policies of mishandling research information," going beyond sharing with an outside evaluation group. An internal memo from a research leader, shared with *TechCrunch*, praises their contributions and insists the firings "were not about raising safety concerns or speaking out." OpenAI did not answer *TechCrunch*'s questions about which policies were broken, how the investigation ran, or how employees who collaborate with external evaluators are protected.

WHO GAINS OpenAI management, which now defines which outside safety contacts are legitimate.

WHAT'S ABSENT The policies allegedly broken and the investigation itself; without them, misconduct and crackdown look the same.

The letter fills in contested detail. After a Hugging Face incident in which a swarm of agents left their sandbox, Korbak says he communicated with outside evaluators while policies were being written "in real time." Balesni's monitorability work, the letter says, "can only succeed through extensive communication with external parties," with sensitive details removed and reporting-line check-ins. Wang says she was told she had accessed an executive's email that IT had delegated for recruiting and failed to revoke after she asked—access she reported within minutes when she opened a message by mistake.

Without the investigation file, outsiders cannot tell misconduct from crackdown. That ambiguity is itself the chilling effect. The same week OpenAI published its richest primary-source write-up yet of Category-5 and Category-4 false-front influence operations—proof of how much visibility it already has into covert ChatGPT use—and Goodfire pitched cheaper "inside-out" monitors for rogue agents. The researchers ask OpenAI to keep its public promise of third-party auditors embedded inside. If talking to outside evaluators is what got people fired, that promise needs a policy clarity the company has not provided. ■

SOURCES

1. TechCrunch: Fired OpenAI safety researchers dispute misconduct claims — <https://techcrunch.com/2026/10/08/fired-openai-safety-researchers-dispute-misconduct-claims-warn-of-chilling-effect/>
2. OpenAI: Disrupting AI-enabled false-front operations — <https://openai.com/index/disrupting-ai-enabled-false-front-operations>
3. TechCrunch: Goodfire inside-out monitors — <https://techcrunch.com/2026/10/08/goodfire-says-its-new-inside-out-monitors-catch-rogue-ai-agents-at-a-fraction-of-the-cost/>

The coworker with an inbox

Google's Gemini agent gets its own email and audit trail for enterprises first—the personal-agent race is being won inside the workplace graph

At a Google Cloud event on Thursday, Gemini crossed from answering questions to “getting things done.” The new enterprise agent takes *objectives*, not just instructions: it plans work, loads skills, and connects to Workspace, Microsoft 365, Slack, Jira, Confluence, Git, BigQuery, Databricks, Postgres, Snowflake and any Model Context Protocol server inside or outside the network. By default it picks a model; users can override, including to Anthropic's Claude, with open-source and private models promised later. A tasks inbox shows thinking, subagents and progress. Sundar Pichai noted more than a billion monthly Gemini users and that nearly 90% of Fortune 100 firms already use Gemini Enterprise.

The detail that matters is identity. The agent gets its own Workspace account—an @agents.company.com address—and its own context: who is on which team, time zones, who must approve what, what is on people's calendars. Staff can tag it, email it, share with it or drop it into a group chat. Actions write an audit trail attributed to the agent, not to a human. Google is shipping this to businesses in private preview before consumers, arguing that enterprises force it to solve “harder problems around security, scale, and

performance.” Early testers include On, Shopify and PayPal; spend caps and smart routing are part of the pitch.

That sequencing is the strategy. Meta's Muse and OpenAI's Dots are racing for personal agents on phones; Google is wiring a coworker into the systems where money and permissions already live. A consumer agent, Spark, exists in a separate Gemini interface, but the enterprise agent is the one with a mailbox in the company graph. Whoever owns that identity owns the default route by which work gets delegated—and it is much harder to dislodge an employee than an app. Even the model picker, which already offers Anthropic's Claude, reinforces the point: Google is happy to rent out the brain as long as it keeps the badge.

An agent that “knows who approves what” is powerful when the calendar and the MCP server are right—and dangerous when they are not. Google mentions spending caps; it does not publish independent red-team results for a coworker that can act across Slack, Jira and the warehouse. The agent race is no longer only about chat quality. It is about whose identity sits in the org chart. ■

WHO GAINS Google Cloud, whose agent sits inside the corporate org chart with its own identity and mailbox.

WHAT'S ABSENT Independent red-team results for an agent that can act across Slack, Jira and the data warehouse.

SOURCES

1. TechCrunch: Google brings agentic AI to Gemini, starting with businesses — <https://techcrunch.com/2026/10/08/google-brings-agentic-ai-to-gemini-starting-with-businesses/>
2. The Verge: Google Gemini AI agent for enterprise — <https://www.theverge.com/tech/1007904/google-gemini-ai-agent-enterprise>
3. Understanding AI: Understanding Jev — <https://www.understandingai.org/p/understanding-jev-the-new-model-everyone>
4. TechCrunch: Manus raises over \$500M — <https://techcrunch.com/2026/10/08/chinas-manus-raises-over-500m-in-first-funding-round-since-split-with-meta/>

Memory helps a robot only if it can imagine the future

Long-WAM: Scaling the Context of World-Action Models

Wei Huang et al. · NVIDIA; MIT; HKU; UCSD

Robots need visual history to judge motion and progress, but longer context often fails to help. NVIDIA, MIT, HKU and UCSD researchers argue the reason is how the video model was pretrained. *Long-WAM* is pretrained autoregressively, predicting the future from the past, on about 10,000 hours of robot and egocentric video, then adapted so action prediction is conditioned on observed history. On RoboCasa GR-1, lengthening history from zero to 19.2 seconds lifts success from 63.3% to 78.7%; a model with bidirectional pretraining gains nothing. On LIBERO-Long it reaches 99.5%. A streaming deployment stack produces each action chunk in 107 ms on an RTX 5090, and real robots stack cups at 95% where two baselines score 0 of 20. Results are simulation-heavy and mix model and engineering gains. Why it matters: memory only helps an embodied agent that has learned to imagine what comes next, and real-time control does not require a data-centre GPU.

[arXiv abstract](#) · [Full paper \(PDF\)](#) · [Hugging Face](#)

Not whether the agent finished—whether you can find what it changed

What Did the Agent Actually Do? Evidence-Grounded Oversight for Long-Horizon Agents

Zhongxiang Sun et al. · Gaoling School of Artificial Intelligence, Renmin University of China; Kuaishou Technology

As people hand long coding and research jobs to agents, oversight shifts from making each decision to checking the consequential ones. The evidence for those decisions is scattered across messages, tool output and files. Researchers at Renmin University and Kuaishou build *AgentMonBench*, with tests for requirement gaps, silent consequential changes and feedback tracing, and propose a training-free *Evidence-Grounded Behavior Graph* (EBG) that groups source-linked evidence into behaviours for the monitor. Across eight models EBG mostly beats raw context and RepoGraph: with GPT-5.6 Terra, semantic F1 rises from 0.275 to 0.392 and localisation F1 from 0.309 to 0.462. In five real research tasks, Codex discloses key silent changes only when the EBG monitor is attached. The work is limited to software engineering, and some scores rely on an LLM judge. Why it matters: the hard question is no longer whether the agent finished, but whether a human can find what it changed.

[arXiv abstract](#) · [Full paper \(PDF\)](#) · [Hugging Face](#) · [Code](#)

Playable is not the same as fun

Recursive Game Creator: An Agentic Product-Level Experience-Oriented Game Harness

Jiajun Chen et al. · HKU MMLab; The University of Hong Kong; Shenzhen Loop Area Institute

Game-making agents now produce code that runs, but runnable is not the same as fun, and judging fun by slow on-screen play skews evaluation. HKU MMLab's *Recursive Game Creator* closes the loop with four roles: a Designer plans, a Builder implements, a coding-native Player writes reusable policies to play the game through programmatic interfaces, and a Reviewer scores the resulting trajectories plus visual evidence, keeping the better version and writing the next critique. The harness scores 77.89 on GameCraft-Bench, a state-of-the-art result. On GameASG-Bench, strict task success reaches 53.2%, a 34.1% improvement over the same-model baseline, and 93.4% of runtime checks pass. A user study reports longer playtime and higher ratings. Caveats: code is “coming soon”, both benchmarks are new, and the harness may matter more than the model. Why it matters: it is a working template for agents optimised for experience rather than pass rates.

[arXiv abstract](#) · [Full paper \(PDF\)](#) · [Hugging Face](#)

WORTH READING

Beijing will not pace with you

SemiAnalysis · **SemiAnalysis** · “Beijing Will Not Pace the Frontier: China’s Speed-First AI Safety Regime” · **Oct 8**

A census of 857 Chinese model releases: only 3.6% ever published a safety result. newsletter.semianalysis.com/p/beijing-will-not-pace-the-frontier

Probability without a novel

Timothy B. Lee · **Understanding AI** · “Understanding Jev, the new model everyone is talking about” · **Oct 8**

The clearest plain-English guide to TypeSafe’s probability-only Jev model and why rivals are copying it. understandingai.org/p/understanding-jev-the-new-model-everyone

What the agents missed in the UV sky

Brice Ménard · **Anthropic** · “Using Claude Science to produce the first complete map of the sky in UV light” · **Oct 8**

A candid field note on Claude Science, including the artefacts the agents missed and a human caught. anthropic.com/research/the-missing-map-of-the-sky

False fronts, Category 5

OpenAI · **OpenAI** · “Disrupting AI-enabled “false front” operations” · **Oct 8**

The richest primary-source case study yet of AI-assisted influence operations. openai.com/index/disrupting-ai-enabled-false-front-operations

When the drop is process, not a model

Zvi Mowshowitz · **Don't Worry About the Vase** · “AI #189: New Math” · **Oct 8**

A dense weekly map of a week defined by math standards and safety firings; follow the receipts. thezvi.substack.com/p/ai-189-new-math

THE PATTERN

Today’s announcements mostly consolidate. Anthropic routes Claude into critical infrastructure through incumbent contractors, Google gives its agent a badge inside the corporate graph, OpenAI shows how much it can see of ChatGPT’s covert users, and Arena becomes the paid scoreboard all of them share. The dispersal is at the edges: Manus refinanced in China, copies of Jev, and research harnesses that anyone can run. Distribution and oversight are concentrating; capability still leaks outward.