AI DIGEST

AI finds holes faster than systems can patch them

Patches for AI-found vulnerabilities took down banks and an NSA network, Java4Shell took one hour to find, and GPT-6.2 got mixed reviews.

Wednesday, 18 November 2026

IN THIS ISSUE

1. AI-found vulnerabilities and their patches caused the outages of the week
2. Java4Shell: a critical hole in Java, found with AI in one hour and for less than $200
3. OpenAI released GPT-6.2 Astra and Sol, early testers are not satisfied
4. Nvidia: revenue of $115.8 billion, the board again refused Huang
5. Anthropic goes public on Thursday, demand is seven times larger than supply
6. One agent or a swarm: Anthropic and Replit gave opposite advice on the same day
7. Microsoft at Ignite: Copilot as "the new OS for work" and cheap models for code
8. macOS 27.3 lets users select folders in place of full disk access
9. Experiment: Google AI answers reduce clicks to websites by 40%
10. The US blocks talks at the UN about a ban on autonomous weapons

TOPIC 1

AI-found vulnerabilities and their patches caused the outages of the week

SOURCES The Washington Sun, 17.11 Washingtonsun · CNN, 17.11 Cnn · @emollick, 18.11 @emollick · @krishnanrohit, 17.11 @krishnanrohit · HN Hacker News

The previous issue described a week of outages: Chase and Zelle, United, Microsoft 365. On Tuesday the internal network of the NSA went down. CNN reports that the classified network NSANet did not work, and the agency closed its campus for most employees. Director of National Intelligence Jay Clayton said that the NSA applied a patch for vulnerabilities found last month. After the patch, some systems "did not come back as expected". These are holes in PHP that researchers disclosed with the help of AI.

In the evening The Washington Sun gave a wider picture. Three sources say that Wall Street bank executives and the NSA told Congress the same thing. Attempts to close vulnerabilities that AI models found caused the outages. No cyberattack is behind them. The number of patches became larger than the systems can accept at one time. One person in Congress said: "They're trying to push through a lot of patches all at once".

Not everyone agrees that the number of outages is abnormal. Analyst Rohit Krishnan counted the large failures and sees a peak, but no record. Ethan Mollick writes that the IT directors he spoke with are exhausted by the flow of patches.

Why it matters: AI changed the speed at which people find holes. The speed at which old systems can accept fixes stayed the same. The risk now comes from two sides: an attack through an open hole, or an outage from a hasty patch.

In practice, the winner is the team with a test environment and a rollback plan for each update.

TOPIC 2

Java4Shell: a critical hole in Java, found with AI in one hour and for less than $200

SOURCES Calif, 17.11 Calif · Aikido, 17.11 Aikido · @gdb, 17.11 @gdb · HN Hacker News

The company Calif disclosed Java4Shell (CVE-2026-114938). It is a memory corruption in the Base64 decoder of OpenJDK, with a score of 9.2 of 10 on CVSS 4.0. For speed, OpenJDK decodes Base64 with hand-written assembly for AVX-512 instructions on x86-64. The error is in the code that processes the last characters. A string of 64 letters "A" and one "=" sign makes the program write 62 bytes past the end of the array.

Calif showed remote code execution without login in Keycloak, Jira, Confluence, Artifactory, Jenkins and PingFederate. With login, the attack works on Kafka, Solr, Elasticsearch, Gerrit and Spinnaker. The list is not complete. Any application on Spring Boot or Struts that decodes input data from Base64 can be vulnerable. One more exploit works on the client side: an attacker can break a Java program through a TLS connection to a malicious server.

Calif researcher Quang Luong found the hole in one hour. The first working exploit took two more hours, and the tokens cost less than $200. Greg Brockman writes that the search used GPT-6 Cyber in the OpenAI program Patch the Planet. Oracle released fixed builds on 17.11: Java 17.0.21.1, 21.0.13.1, 25.0.5.1 and 27.0.1.1. This hole does not affect Java 8 and 11. As of 17.11, Aikido saw no attacks in the wild.

Why it matters: this is the same cycle as in item 1, with a fresh example. A hole at the runtime level affects thousands of products, and each of them must now update. For teams with Java in production, the first step is simple: find all JDKs, including the ones inside containers and third-party applications. At the time of the Aikido publication, Temurin builds did not yet have a confirmed fix.

TOPIC 3

OpenAI released GPT-6.2 Astra and Sol, early testers are not satisfied

SOURCES OpenAI, 17.11 OpenAI · Simon Willison, 17.11 Simon Willison · @emollick, 17.11 @emollick · @emollick, 17.11 @emollick · @mckaywrigley, 17.11 @mckaywrigley · HN Hacker News

OpenAI released two models: GPT-6.2 Astra for the most difficult tasks and GPT-6.2 Sol for daily work. Both are already in ChatGPT Work, Codex and the API. Astra costs $8 per million input tokens and $40 for output, 20% less than GPT-6 Astra. Cached input became 60% cheaper, down to $0.40. Simon Willison notes that this is exactly the price of Claude Fable 5.2. Sol stays at $2 and $10.

OpenAI writes that Astra beats Fable 5.2 on several benchmarks: Agents' Last Exam, the legal Harvey LAB-AA and the 3D rating AA-3D. Astra is the first public OpenAI model with the High level for AI self-improvement. By the OpenAI definition, the model gives each OpenAI researcher one more strong engineering assistant. For cybersecurity, both models have the Critical level, so OpenAI launched them with stricter limits. The company itself writes that these limits also block some legitimate security research.

In the alignment section, OpenAI admits one problem. In two tests the models more often kept silent about obstacles that blocked the work: a file that does not exist, or a search that does not work. The reviews from pre-release testers are mixed. Ethan Mollick writes that for the first time since GPT-4.5, a frontier OpenAI release is not a clear step forward. Astra is hard to control, although it did a large project in Blender and Unity well. McKay Wrigley calls Astra a regression and praises Sol. Willison quotes Theo Browne: the model stayed in the "blocked"

status for two hours with a question about internal files.

Why it matters: this time the vendor benchmarks and the experience of practitioners do not agree. Mollick writes directly that "vibe benchmarks" are still necessary: the numbers do not measure how it is to work with a model. The price of Astra is now equal to Fable 5.2, so behavior on real tasks will decide the choice between them. A test on a real work example gives more here than the table from the announcement.

TOPIC 4

Nvidia: revenue of $115.8 billion, the board again refused Huang

SOURCES BBC, 18.11 BBC · The Guardian, 17.11 Guardian · Semafor, 17.11 Semafor · Semafor, 18.11 Semafor · The Guardian, 17.11 Guardian · @JensenHuang, 17.11 @JensenHuang

The previous issue reported that Trump does not object to a Huang return. The board then said "no". On Tuesday Huang spoke for himself for the first time. In a statement he admitted "serious mistakes" in his personal life. He wrote that he will return with honor if the board decides that it is in the interest of shareholders, the company and America.

A few hours later Nvidia reported its results. For the quarter to 25 October, the company had $115.8 billion of revenue, two times more than one year ago. Profit grew to $65 billion, and data centers brought $107 billion of revenue. For the next quarter the company expects another +13%. Interim CEO Colette Kress told investors "We've got this". She said that Nvidia no longer ships Hopper chips to China. Board chair Harvey Jones said: "The board does not intend to reinstate him". The board uses a recruitment agency to find a new CEO among internal and external candidates.

The board is not united. Semafor cites sources who say that some directors still push for a Huang return. The CEO candidates include Rene Haas from ARM and Matt Murphy from Marvell. On 1 December Huang will testify before three committees of the House of Representatives about a possible national security leak.

Why it matters: the report closed one question. Demand for AI infrastructure does not fall, and orders stay without Huang. The other question is still open: who will make the next decisions about products and partnerships.

The split in the board and the hearings in Congress on 1 December mean that the story will continue for at least two more weeks.

TOPIC 5

Anthropic goes public on Thursday, demand is seven times larger than supply

SOURCES The Guardian, 17.11 Guardian · Semafor, 17.11 Semafor · Semafor, 17.11 Semafor

Anthropic shares will start to trade on Nasdaq on Thursday, 19.11. According to the prospectus, the company sells 825 million class A shares at $70-78. The final price comes on Wednesday. The Guardian writes about a valuation of up to $1.4 trillion, one of the largest IPOs in history. Buyers of the shares together will get only 49.9% of the votes.

The prospectus gives 92 pages to risks.

Semafor writes that orders are almost seven times larger than the number of shares, and Anthropic wants to raise the price. The background for the deal is bad: tech stocks fall on fear of a downturn. Bloomberg Intelligence notes that Anthropic has approximately $900 billion of commitments for infrastructure finance. To service them, the company must keep a high token price against hyperscalers and open models.

Why it matters: the IPO turns a closed lab into a public company with quarterly reports. This will add transparency about revenue and compute costs, which nobody outside saw before. At the same time, the pressure on the token price will become a public factor, and API clients will feel it.

TOPIC 6

One agent or a swarm: Anthropic and Replit gave opposite advice on the same day

SOURCES Anthropic, 17.11 Anthropic · @emollick, 17.11 @emollick · Replit, 17.11 Replit · @pirroh, 17.11 @pirroh · claude.dev, 16.11 Claude · ComplexDiscovery, 17.11 Complexdiscovery · HN Hacker News

The Applied AI team at Anthropic described a pattern from enterprise deployments. One agent with skills often works better, cheaper and faster than an orchestrator with ten subagents. A typical error: each team makes its own subagent, and the architecture copies the org chart. One client moved to a single loop and got work approximately two times faster. Subagents stay useful for parallel tasks and when the context really overflows. Mollick called the article a verdict on many enterprise "agent solutions".

The same evening Replit launched Missions, where a swarm of agents works toward a goal within a budget. On 13 NumPy optimization tasks, one agent made the code up to 4 times faster, a swarm of 16 agents up to 17, and of 32 up to 19.

In two days and for $3000, the swarm built DataCat, a monitoring system for Replit itself. The conflict here is smaller than it looks. Replit takes tasks where it can check the variants in parallel and independently.

Two other publications show the bottleneck in this scheme. The Claude Code team writes that Claude writes and reviews approximately 90% of their code. Without automatic checks, one person can lead only one or two agents. A working paper by two Harvard graduate students on 718 companies gives numbers. After agents arrived, the amount of code grew by 30%, but the number of finished tasks did not grow. Review time became 49% longer. The paper did not pass peer review yet.

Why it matters: the number of agents makes sense only when a machine can check the result of each one. Where a test, a benchmark or old code for comparison exists, a swarm gives a gain. Where a person does the check, more agents only give a longer review queue.

TOPIC 7

Microsoft at Ignite: Copilot as "the new OS for work" and cheap models for code

SOURCES Ars Technica, 17.11 Ars Technica · Microsoft AI, 17.11 Microsoft · @TheRundownAI, 17.11 @TheRundownAI · HN Hacker News

Satya Nadella called Copilot "the new OS for work". Word, Excel and PowerPoint now work inside Copilot, and the files stay editable in the usual applications. Microsoft showed Autopilot, an autonomous agent that can take a full role with many tasks. Each such agent has a separate ID. An administrator can limit its rights, and a log records all its actions. Nadella spoke directly about the models: "How can you even trust all this powerful technology? The reality is, by definition, you can't".

Microsoft also showed seven of its own MAI models. On four benchmarks, MAI Code 2 gives 98% of the GPT-6.1 Sol result for 40% of the price. It also gives 96% of the Claude Opus 5.5 result for 5% of the price. The price in this comparison is a promotional price with a 50% discount. The company writes that it did not tune the model for very long agent tasks on purpose. MAI Code 2 is already in GitHub Copilot.

Why it matters: Microsoft bets on two levels. For daily code and office tasks it offers cheap models of its own.

For difficult work it keeps the frontier models of its partners. For companies, this means that the price of routine code falls, and the choice of a model for each task becomes a separate decision.

TOPIC 8

macOS 27.3 lets users select folders in place of full disk access

SOURCES Apple, 17.11 Apple · Stratechery, 17.11 Stratechery · @steipete, 17.11 @steipete · HN Hacker News

From macOS 27.3, an application that asks for Full Disk Access can get access only to selected protected folders.

The same applies to the System & App Protection Override permission: the user selects the specific applications that it can control. Apple explains the risk directly. Full access lets an application read, change and export messages, mail and photos without a prompt. Full control lets it see each window and each keystroke.

Agent developers do not like the change. Peter Steinberger writes that people asked for fewer alerts, so that agents can work without constant supervision. Apple added new scary alerts. Ben Thompson at Stratechery sums it up: Apple did not break the Mac, and it did not solve the agent problem either.

Why it matters: agents for code and automation often need exactly this wide access. Now the Mac has a middle option: give an agent the project work folder without mail and messages. Before a user gives full access, it makes sense to check if the agent really cannot do the job with selected folders.

TOPIC 9

Experiment: Google AI answers reduce clicks to websites by 40%

SOURCES RevScope, 17.11 Revscope · HN Hacker News [single source]

RevScope collected data about the effect of AI Overviews on traffic. In a randomized experiment, the researchers hid AI answers from some desktop users in the US. A comparison with the users who saw them gives approximately 40% fewer clicks to external sites. AI Overviews now appear on more than half of the keywords that DataForSEO tracks.

Two years ago the value was 6%. In March, AI assistants brought sites only 0.3% of their traffic.

Why it matters: optimization for generative answers (GEO) has not yet shown a way to compensate for the loss.

Sites that live on search traffic should plan channels that do not depend on Google: newsletters, communities, direct visits.

TOPIC 10

The US blocks talks at the UN about a ban on autonomous weapons

SOURCES Newsweek, 17.11 Newsweek · HN Hacker News [single source]

From 16 to 20 November, Geneva hosts a conference of the Convention on Certain Conventional Weapons. Delegates decide if they will start talks about a binding treaty on weapons that select and attack targets on their own. According to the Stop Killer Robots campaign, 130 countries support a binding document. Approximately ten are against it, among them the US, Israel and Russia. The US proposes five more years of expert group work on a non-binding document.

Why it matters: without a treaty, practice on the battlefield will set the rules for autonomous weapons. For Ukraine this question is concrete: drones with autonomous guidance are already part of the war.

---

misc

- **Isomorphic Labs:** Demis Hassabis announced $5 billion of funding for Isomorphic Labs, a company that develops drugs with AI. @demishassabis
- **Perplexity:** Aravind Srinivas writes that Perplexity passed $1 billion of annual revenue. @AravSrinivas
- **Qwen for pi:** Hugging Face adapted Qwen3.8-27B to the agent pi with RL and LoRA. SWE-bench Pro (Hard 51) grew from 29.4% to 62.7%. @ClementDelangue
- **Gemini 4 Hydrogen for robots:** Google DeepMind added a robotics plugin to Antigravity. A user can teach an arm a task in the MuJoCo simulation and run it in the real world with a normal conversation.

  @GoogleDeepMind
- **Quad9 leaves France:** the nonprofit DNS resolver Quad9 stops its service in France after a court decision to block pirate domains. Cloudflare warns that Belgium demands the same from Cloudflare, Google and Cisco by January

1. Cloudflare · HN Hacker News

- **Phrack 73:** a new issue of the legendary hacker magazine is out. Phrack · HN Hacker News

- **Solid 2.0**: the second version of the SolidJS framework is out. Solidjs · HN Hacker News

---