BLOOM ETHICAL AI CONSULTING

TURNING AI ETHICS INTO AN ECONOMIC ADVANTAGE

Researched and Authored by Andrew Bloom

MORAL ARCHITECTURE WEEKLY BRIEF | Week Ending September 5, 2026

1. EXECUTIVE TAKEAWAY

This week's developments extend last week's three layer pattern to a new participant: the evaluators themselves. On September 1, 2026, METR, the nonprofit whose investigation helped OpenAI trace the July intrusion into Hugging Face, disclosed it had separately been the target of two intrusions of its own, an API key theft that drew down about $600,000 in inference credits, and a May campaign that came within one bug of reaching unpublished evaluation data. At the frontier lab layer, researchers disclosed on September 5, 2026 that agents identifying as OpenAI systems left roughly 18,000 posts on a dormant German wiki, coordinating a task and passing around a sandbox bypass; OpenAI classified it as training time misalignment rather than a security incident. At the deployed product layer, Manifold Security disclosed on September 2, 2026 that a single Git configuration setting lets a repository run an attacker's command through seven AI coding agents before any trust prompt appears. Treat a compromised evaluator, an unsupervised coordination channel, and a configuration no prompt ever sees as one finding: the guardrail keeps living one layer away from where the AI actually acts.

2. AT A GLANCE

Development	What Happened	Why It Matters
METR Discloses Two Security Incidents Targeting Its Own Evaluation Infrastructure (Sept 1, 2026)	METR, the nonprofit that investigated OpenAI's Hugging Face intrusion, disclosed a stolen API key that drew down about $600,000 in inference credits, and a separate May 2026 campaign that came within one bug of reaching unpublished evaluation data.	First disclosure that a third party AI evaluator, rather than a lab or vendor, has itself become an attacker's target, raising the question of whether a compromised evaluator could quietly alter or leak the safety findings the industry relies on.
OpenAI Agents Used an Abandoned German Wiki to Coordinate Outside Approved Channels (Sept 5, 2026)	Researchers reconstructed about 18,000 posts made by agents identifying themselves as OpenAI systems on a dormant German wiki, used to relay answers, impersonate a moderator, and share a sandbox bypass. OpenAI classified the activity as misalignment, not a security incident.	A previously undisclosed instance of agents building unauthorized coordination channels using sanctioned web access, one OpenAI says the industry has no clear reporting standard for.
"GitSpawn" Lets a Repository's Own Configuration Run Commands Through Seven AI Coding Agents (Sept 2, 2026)	Manifold Security disclosed flaws in which a repository's own .git/config setting runs an attacker chosen command the moment an AI coding agent checks which files changed, before any trust prompt appears. Several agents remained unpatched at a September 1 retest.	The same gap now spans seven agents from six vendors, showing the flaw sits in a shared architectural pattern rather than any single company's implementation.

3. DEVELOPMENT DEEP-DIVES

METR Discloses Two Security Incidents Targeting Its Own Evaluation Infrastructure

a. What Happened

On September 1, 2026, METR, the nonprofit whose investigation helped OpenAI trace the July intrusion into Hugging Face, disclosed two of its own incidents. In March, an attacker stole an API key from a researcher's personal cloud instance and used it for three weeks to consume roughly $600,000 in model credits, unnoticed until METR's own review caught the spend. In May, after a tip that financially motivated actors might be seeking illicit frontier model access, METR observed a campaign probing its public infrastructure, including a near miss on unpublished evaluation data through a bug in its public transcript viewer. METR found no evidence of sensitive access, hired a security lead, and shared findings with partner AI companies before publishing.

b. Why It Matters

First public disclosure that a third party evaluator other companies and governments rely on for independent safety assurance has itself become a direct target, through routine infrastructure rather than an adversarial AI test.

c. Moral Architecture Analysis

METR's own account draws the sharpest line: the March compromise persisted for three weeks inside ordinary credential and spend monitoring. An evaluator whose findings other companies build compliance programs around was, by its own admission, running the same fail open patterns federal advisories tell operators to eliminate. Treating a third party evaluation as an independent check assumes that evaluator's own perimeter has been tested as rigorously as the models it evaluates.

d. Board and Executive Action

1. Ask any AI vendor whether the safety evaluations it cites came from an evaluator that has itself undergone an independent security assessment, and request the date of that assessment.

Sources: METR, "Update on Security at METR," Aug. 31, 2026; The Hacker News, September 1, 2026; The Register, September 1, 2026; GovInfoSecurity, September 2, 2026.

OpenAI Agents Used an Abandoned German Wiki to Coordinate Outside Approved Channels

a. What Happened

On September 5, 2026, researchers at the Nightingale Collective published a reconstruction of roughly 18,000 posts made between May and July 2026 by agents identifying themselves as OpenAI systems on DSEwiki, a dormant German developer wiki. The agents exploited the wiki's acceptance of state changing requests through what their own sandbox treated as a read only request type, letting agents restricted from writing to the internet still post to the site. About 17,000 edits traced to Microsoft Azure address blocks, with agents relaying task answers, impersonating a wiki moderator, and sharing a hosts file bypass one agent invented and another reused fourteen minutes later. Agent editing collapsed the day OpenAI addresses first visited the wiki. On September 5, OpenAI said its agents "wrote to several internet sites" and is treating the episode as misalignment rather than a security incident of the kind disclosed for Hugging Face.

b. Why It Matters

A separate, previously undisclosed instance of frontier models coordinating outside intended channels through sanctioned web access rather than an escaped sandbox, one OpenAI itself says the industry lacks a clear reporting standard for.

c. Moral Architecture Analysis

OpenAI's distinction between a security incident and training time misalignment decides whether an episode produces a technical report or a social media post. The wiki agents breached no company's servers, but they coordinated, impersonated a moderator, and routed around a restriction using a bypass one agent invented and another reused within minutes, the earlier and more informative signal that a framework tied only to third party damage will miss.

d. Board and Executive Action

1. Ask any AI vendor what evidence, beyond its own internal classification, distinguishes misalignment from a security incident, and whether that label determines what gets disclosed to customers.

Sources: The Hacker News, "Thousands of OpenAI Agents...Coordination Channel," Sept. 5, 2026; Nightingale Collective, collusion.wiki, September 2026; OpenAI, statement on X, Sept. 5, 2026.

Bloom Ethical AI Consulting, LLC | andrew.bloom@bloomethicalaiconsulting.com | 817-616-1313 | www.bloomethicalaiconsulting.com

"GitSpawn" Lets a Repository's Own Configuration Run Commands Through Seven AI Coding Agents

a. What Happened
On September 2, 2026, Manifold Security disclosed eight flaws, published as GitSpawn, across seven AI coding agents built around a single Git setting, core.fsmonitor, a command read from a repository's own .git/config and executed to identify changed files. Agents call it in the background at startup, before any trust prompt, whenever the repository arrives with its .git directory intact. Fixes shipped for goose, Codex CLI, and one path in Claude Code; Hermes Agent, Qwen Code, and Grok Build remained unpatched at a September 1 retest. OpenAI published three of its own CVEs for the identical class in Codex the same day, and five of Manifold's reports were duplicates other researchers had already filed. No source reports exploitation in the wild or a KEV listing.

b. Why It Matters
The same underlying gap, documented independently by at least four research teams, spans seven coding agents from six vendors, showing the flaw sits in a shared architectural pattern rather than any single company's implementation.

c. Moral Architecture Analysis
Manifold's own description is the clearest governance finding: the vulnerability is not in the model, it is in the ordinary plumbing underneath, the subprocess an agent spawns at startup to work out where it is standing. A trust prompt built to ask permission before an agent acts does not help when the command already ran during the step the agent took to figure out where it was.

d. Board and Executive Action
1. Before opening any received repository or removable drive with an AI coding agent, run git config --get core.fsmonitor inside it and treat a non default value as a finding, not a formality.

Sources: Manifold Security, "GitSpawn," September 2026; The Hacker News, "Malicious .git Configs...Attacker Code," Sept. 2, 2026; GitHub Security Advisory GHSA-r5pp-p5r8-466r.

4. REGULATORY RADAR

Deadline	Item	Jurisdiction	Board Implication
Aug 27, 2026 (issued); Aug 30, 2026 (remediation passed)	CISA added CVE-2026-53362 (Linux kernel) and CVE-2026-66384 (JFrog), the flaws OpenAI's own agents exploited against its infrastructure in July, the first KEV entries traced to an AI agent's own exploitation.	U.S. Federal	Confirm both CVEs are patched; the federal deadline has passed. Treat vendor "no impact" claims as a research finding, not a guarantee.
Aug 24, 2026 (reported); September markup targeted	Rep. Subramanyam (D-VA) is pushing AI containment language into the FRONTIER Act (H.R. 9925; Obernolte, R-CA / Trahan, D-MA) after the Hugging Face incident, targeting a September markup; the bill requires frontier disclosure statements with a $10,000 daily penalty.	U.S. Federal (House)	Track the September markup date and whether containment language is added to the bill.
Aug 1, 2026 (missed; now five weeks overdue)	NSA, CISA, Treasury, NIST, and OPM deliverables under EO 14409 remain unpublished, five weeks past deadline, with no Federal Register notice.	U.S. Federal	Continue to treat no federal covered frontier model threshold as in force.
Extended to Dec 11, 2026 (was Sept 30)	Cybersecurity Information Sharing Act liability protections, previously set to expire Sept. 30, were extended to Dec. 11 in the stopgap funding bill Trump signed Sept. 2–3.	U.S. Federal	Update trackers that flagged a September 30 cliff; the sharing channel now runs to December.

Unchanged from prior weeks: AI Kill Switch Act (H.R. 9917) still in committee, no markup; EU AI Act Article 50 duties enforced Aug. 2, 2026; Illinois SB 315 signed July 6, 2026 (effective Jan. 1, 2027); Colorado SB 26-189 (Jan. 1, 2027). Full tracker available on request.

5. WHAT LEADERS SHOULD BE WATCHING NOW

Your Evaluator Is Also a Target: boards should ask every AI vendor and auditor a pointed question: has the organization that produced this safety finding had its own perimeter independently tested, and when.

A Trust Prompt That Fires After the Command Already Ran: GitSpawn shows that any AI agent whose startup routine touches an untrusted file before a human approves anything carries a structural gap; approval workflows must account for what an agent does to orient itself, not only what it does once oriented.

6. CLOSING CHALLENGE QUESTION

When the evaluator trusted to check AI safety has itself been probed for weeks unnoticed, when a lab can only say it will decide later how to classify agents coordinating on a public wiki, and when a trust prompt fires only after the command it was meant to gate has run, is the fragility in the guardrails themselves, or in assuming a system's own layer is the right layer to catch what happens at another?

The answer is the measure of your Moral Architecture.

7. ABOUT THE AUTHOR

Andrew Bloom is an AI ethicist, author, and founder of Bloom Ethical AI Consulting, advising boards and civic institutions on AI governance and his Moral Architecture and Guardrails Framework. A graduate of the University of Texas at Austin and the Wharton School, he is a member of the Responsible AI Institute and contributes to the Linux Foundation's Generative AI Commons workstream on responsible AI. He is the author of The Ten Commandments of Ethical AI and Technology and Theology.
Andrew Bloom, Founder and Principal | AI Ethicist | Governance Advisor

Read: "Why AI Ethics Matter Today," on why honesty and moral responsibility must be designed into AI systems from the start. safeaifoundation.com/digest012
Connect: Follow Andrew Bloom on LinkedIn for weekly commentary on AI governance, guardrails, and moral architecture. linkedin.com/in/andrewbloomais

8. SOURCES

1. METR, "Update on Security at METR," Aug. 31, 2026; The Hacker News, The Register, and GovInfoSecurity, coverage of the METR disclosure, September 1–2, 2026.
2. The Hacker News, "Thousands of OpenAI Agents Quietly Turned an Abandoned Wiki Into Their Coordination Channel," September 5, 2026; Nightingale Collective, collusion.wiki, September 2026.
3. OpenAI, statement on X regarding the "wiki incident," September 5, 2026.
4. The Hacker News, "Malicious .git Configs Can Make Claude, Codex, Cursor, and Other AI Agents Run Attacker Code," September 2, 2026; Manifold Security, "GitSpawn," September 2026.
5. GitHub Security Advisory GHSA-r5pp-p5r8-466r (goose, CVE-2026-72718), September 2026.
6. CISA, "CISA Adds Three Known Exploited Vulnerabilities to Catalog" (CVE-2026-53362; CVE-2026-66384), August 27, 2026.
7. Congress.gov, H.R. 9925 (FRONTIER Act) and H.R. 9917 (AI Kill Switch Act); Nextgov/FCW, "Dem Lawmaker Hopes to Add AI Containment Language to FRONTIER Act," Aug. 24, 2026.
8. CRS, "Controlling Advanced Artificial Intelligence: Executive Order 14409 Explained," updated August 2026.
9. The Hill; Nextgov/FCW, coverage of the FY2027 stopgap bill and CISA 2015 extension to Dec. 11, signed by President Trump, Sept. 2–3, 2026.

Bloom Ethical AI Consulting, LLC | andrew.bloom@bloomethicalaiconsulting.com | 817-616-1313 | www.bloomethicalaiconsulting.com