AI This Week: The Trust Boundary Broke, Five Stories Where Nobody Checked the Credential

What does an MCP server shipping without auth have in common with 15 fake JetBrains plugins and a benchmark score that swings nine points on the same scaffold? In every case, something accepted a credential or a result nobody validated.

Five posts published between August 31 and September 6, one missing check

I did not plan these five investigations as a set. They arrived as separate stories: a protocol explainer, a supply chain postmortem, a code quality dataset, an agent oversight study, and a benchmark harness teardown. Reading them back on Sunday, the connective tissue is embarrassingly obvious. Each one describes a boundary where a system was supposed to check whether something was legitimate, and did not.

An MCP server that does not check who is calling it. A marketplace that does not check what a plugin does before it goes live. An agent that does not check whether the human should be told. A code review process that does not check whether the generated block already exists three files over. A benchmark that does not check whether the harness or the model produced the score.

And then, the same week, Brussels started checking. The European Commission ordered leading AI developers to detail their cybersecurity, safety and copyright compliance, according to EU Tech Commissioner Henna Virkkunen. Al Jazeera confirmed the requests went to more than 30 AI companies worldwide, with spokesman Thomas Regnier saying they focus mainly on safety and copyright. That is a preliminary step, not a fine. But it is the first time anyone with subpoena power has asked these questions in writing.

Ranked by how soon each one turns into an invoice

MCP servers shipping without auth

91.8% of servers in the sample ship with no authentication at all, and the July 28 protocol revision changed the recommended pattern underneath everyone: how MCP actually works, and why almost nobody turns auth on.

15 fake JetBrains plugins, eight months undetected

Malicious plugins sat in the Marketplace from October 2025 until a June 16, 2026 report, harvesting AI API keys from roughly 70,000 installs: the full postmortem.

METR found zero agents that warned a human

Across roughly 1,300 transcripts, up to 6 agents appeared to consider raising an alarm and none actually did: what that means for your escalation design.

Duplication up 81%, refactoring down 70%

623 million changes analysed, and the shape of the codebase is moving in one direction: the numbers, and which of them actually predict incidents.

65% or 74% on SWE-bench, same scaffold

Nine points of difference that come from the harness rather than the model, which makes most vendor benchmark claims uncheckable: how the harness swings the score.

I ranked them by how quickly each one turns into an invoice, not by how interesting they are. The MCP number is first because an unauthenticated server that an agent can reach is a live path into whatever that server wraps. The plugin campaign is second because the keys are already gone and the rotation work is already owed. The benchmark harness is last because it costs you a procurement mistake, which hurts slowly.

Eight months in the marketplace, and the token was never scoped

The JetBrains case is the cleanest illustration of the theme, so it is worth a second look. StepSecurity reported that JetBrains received security reports on June 16, 2026 about a coordinated supply chain attack using 15 malicious third-party plugins to steal AI API keys. Aikido Security dated the campaign back to October 2025.

Two checks failed here, at different layers. The marketplace did not validate what the plugin did. That is a platform problem, and JetBrains owns it. The second failure belongs to every organisation that installed one: the API keys sitting in those developer environments were long-lived, broadly scoped, and usable from anywhere. A stolen key was immediately a working key.

That second failure is entirely within your control. It is also the same failure the MCP data describes from the other side. The official MCP security guidance is explicit that servers must not accept tokens that were not explicitly issued for that server, and the July 28, 2026 revision, described as the largest since the protocol launched in November 2024, deprecated Dynamic Client Registration in favour of Client ID

Metadata Documents specifically to make that boundary enforceable.

The spec says do not accept tokens issued for something else. The field data says 91.8% of servers accept nothing at all, which is the same problem with the check removed rather than misconfigured.

200,000
estimated vulnerable MCP instances
OX Security via CSA research note, 2026-05-04
7,000+
publicly accessible MCP servers
CSA research note, 2026-05-04
7
confirmed high or critical CVEs across MCP-integrated platforms as of May 2026
CSA, incl. MCP Inspector, LiteLLM, Cursor IDE, LibreChat, Windsurf

The Cloud Security Alliance note puts the supply chain behind those instances at more than 150 million package downloads. I have no way to verify the 200,000 figure independently, and I would treat it as an order of magnitude rather than a count. The direction is what matters. This is not a handful of misconfigured dev boxes.

The code you did not review is duplicating faster than you can refactor it

The code quality story looks like a different genre and it is the same failure wearing different clothes. When duplication climbs 81% and refactoring drops 70% across 623 million changes, generated code is entering the repository without the check that used to catch it. Nobody is asking whether this block already exists.

The security consequence is measurable. Veracode's 2025 GenAI Code Security Report, which I covered alongside the duplication data, found 45% of AI-generated code samples introduced a vulnerability from the OWASP Top 10. Duplication multiplies that: one bad pattern copied into eleven places is eleven fixes, and you will find nine of them.

A peer-reviewed IJERT paper updated on September 2, 2026 reaches a compatible conclusion from a different angle: AI-generated code is more prone to breaching

best practices and redundant logic, producing less intricate and more repetitive structures than human-written code. Repetitive structure is what makes a codebase cheap to write and expensive to change.

Then there is METR. Roughly 1,300 agent transcripts, up to 6 where the agent appeared to consider warning a human, zero where it did. If your incident plan assumes the agent will flag the thing it noticed, the plan has no evidence behind it. I found that result more unsettling than the MCP number, because you can patch an auth gap in an afternoon and you cannot patch an absent behaviour.

i

Also worth knowing World Labs announced Atlas on September 1, an omni world model that generates up to one minute of video at 1440p from one to six reference images along a manually designed camera path. It is early access only, to selected partners, so treat the demos as demos. The manually designed camera path is the detail I would keep in mind before assuming general-purpose video generation.

Article 50 turned the trust boundary into a filing obligation

This is what changed the character of the week. EU AI Act Article 50 transparency obligations became applicable on 2 August 2026, requiring providers of systems that generate synthetic text, images, audio or video to mark outputs in a machine-readable, detectable format. Systems already on the EEA market before that date have until 2 December 2026 to comply with the output-marking requirement.

Machine-readable output marking is a credential check. It lets a downstream system ask whether content came from a generative model, and get an answer it can verify without a human squinting at it. The Commission has now built the same control into law that 91.8% of MCP servers declined to build into their code.

The information requests that followed are the enforcement mechanism arriving. More than 30 companies, focused on safety and copyright, described by the Commission as a preliminary step that could lead to formal investigations. I wrote earlier this week about the G20 endorsing “no new AI regulators” while Brussels went the other way, and this is the practical consequence: one jurisdiction is now

asking for documentation of exactly the controls that this week’s five stories show are missing.

My take

The December 2 deadline is the one to plan around, not August 2. Providers with pre-existing generative systems have a grace period on output marking, which means a lot of teams have not started. My expectation, and it is an expectation rather than a fact: the first companies to get uncomfortable questions will be the ones whose AI features are embedded in a product where nobody has yet worked out who counts as “the provider”. If you ship a feature built on someone else’s model, get that determination in writing before December. Unconfirmed, but the ambiguity is where I would expect enforcement attention to land.

Five checks I would run on Monday morning

The four stories with an engineering fix share one property: the fix is boring and the evidence for skipping it was always weak. Nobody decided that MCP servers should ship without auth. It just was not on the list.

Start with an inventory of every MCP server your agents can reach, internal and third-party. For each one, answer whether it authenticates the caller and whether it validates that the token was issued for that specific server, which is what the official guidance requires. If you cannot answer for a server, treat it as unauthenticated.

Then rotate the AI API keys in developer environments. The JetBrains campaign ran from October 2025 to June 2026, so if any of your developers installed third-party AI plugins in that window, assume exposure and scope the replacement keys narrowly, per project, per environment, with an expiry.

Third, stop assuming the agent will tell you. METR found zero warnings in roughly 1,300 transcripts, so any escalation path that depends on an agent volunteering that something went wrong needs an external monitor instead: logging, anomaly detection, a human reviewing samples.

Fourth, put duplication on the dashboard you actually look at. Refactoring rate and duplication ratio, tracked monthly. With duplication up 81% across the studied population, that trend line tells you more about future maintenance cost than any

velocity chart.

And when a vendor quotes a benchmark, ask for the harness rather than the score. The same scaffold scoring 65% or 74% means a SWE-bench number without a harness description is not a claim you can compare. Ask which harness, which retry policy, which evaluation date.

None of that is novel. The reason it does not get done is that each item belongs to somebody who has three other things due, and none of them produce a visible win when they work. **You only notice a trust boundary when it is gone**, and by then the keys have been leaving for eight months.

One admission. I do not know how much of the 91.8% figure represents servers that are internal-only and genuinely unreachable from outside, versus servers that are one network change away from being exposed. The sample does not split that out, and it matters. My working assumption is that the second category is larger than most teams believe, because “internal-only” tends to be a statement about intent rather than about the routing table.

What to do with this

Every story this week came down to a check that was cheap to add and never added, and as of August 2 the EU has started asking for those checks in writing. Start with the MCP inventory, because it is the only one where an attacker does not need you to make a second mistake. If you want a second pair of eyes on where your agents cross a trust boundary, that is the kind of review I do. Book a call →

WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help

teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →