# As AI Agents Run Their Own Code, Researchers Test the Sandboxes Meant to Contain Them

**AI Code Sandbox Vulnerabilities: Why Governance Trumps Technical Architecture**

*By IBT Desk*
*July 6, 2026 17:47 +08*



Singapore has just become the first country to publish a governance framework for agentic AI. A new study of the sandboxes that hold AI-generated code suggests the technical assurances behind such systems deserve harder scrutiny. IBTimes spoke with the researchers behind it, George Andronchik of fellows.tech and OrbitaLab, and Pavel Lokhmakov, PhD in CS.



In January 2026, [illegible] Model AI Governance [illegible] can now plan, rea[illegible] safety testing at the centre of the conversation. Underneath every such agent, though, sits a more basic engineering question. When an AI writes and runs its own code, where does that code execute, and what stops it from reaching the machines beneath?

Researchers at fellows.tech and OrbitaLab put five of the most widely used AI code sandboxes through six engine-level security tests. We sat down with co-authors George Andronchik and Pavel Lokhmakov to discuss what the results mean for businesses deploying these systems.

**Why should a company building with AI care about "sandboxes" at all?**

**George:** Because the moment you let a model run code it generated, you've invited an untrusted process onto your infrastructure. The sandbox is the wall between that process and everything you own, including customer data, credentials, the host machine. Most teams treat it as a checkbox. We wanted to know whether the wall actually holds.

**Pavel:** And the timing matters. The threat used to feel theoretical, and that's stopped being true. In one UK AI Security Institute study, a frontier agent's best run worked through 22 of 32 steps of a simulated corporate-network attack, and the researchers found that capability climbs with the compute behind it, with no plateau in sight. The live question now is what stands in hostile code's way when it tries to escape.



## READ MORE

**From Oil to AI: US-UAE Ties Enter a New Strategic Era**

The designation removes licensing requirements for many advanced U.S....

**Microsoft Is Selling Xbox Studios It Spent Years Buying, Thousands of Jobs to Go**

Microsoft is selling and spinning off several Xbox studios while cutting...

**How BioShocking Attack Tricks AI Browsers Into Stealing User Credentials**

LayerX researchers found a prompt injection attack that tricked six AI...

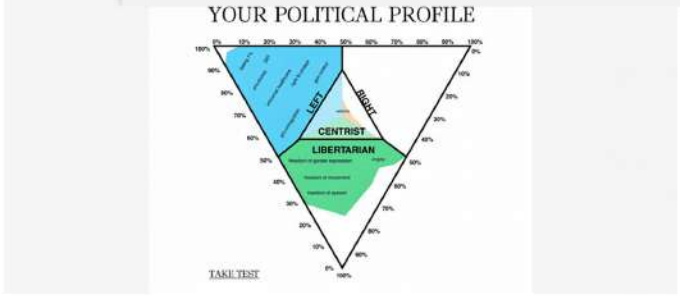
**ere Are 31 Apps sign and More**

AI tools for neetings,...



**Sony to End PlayStation Game Discs From 2028, Shifts to Digital Downloads of New Releases**

Sony will stop producing physical discs for new PlayStation games from...




**You tested five products. What surprised you most?**
**George:** That the strength of the isolation barely predicted real-world safety. Some products with the most sophisticated engineering underneath turned out to be among the most exposed, because the self-hosted version had quietly stopped updating it. The technology wasn't weak. The maintenance was.

**Pavel:** One product's self-hosted deployment had frozen its underlying engine for more than 471 days, missing two serious fixes in that window, including a high-severity escape-class flaw. A second self-hosted build had its engine pinned 399 days back, and when an 8.7-severity vulnerability surfaced later in that stretch, the default pin left it unpatched for weeks. Operators running those versions believed they were protected. In practice they were running known-vulnerable software because nobody updated a version number.



**An important caveat there. You tested the open-source, self-hosted versions. Does this apply to the vendors' managed cloud offerings too?**
**Pavel:** No, and we're careful about that. We measured the open-source, self-hosted deploy paths, because those are the ones we can actually inspect. For the managed cloud versions, the pinned engine isn't publicly observable, and in the paper we literally classify it as "opaque." A cloud provider may well update aggressively behind the scenes. We can't see it, so we make no claim either way. The frozen-pin result is a statement about the self-hosted versions, full stop.

**George:** Which is its own lesson for buyers. If you self-host, version currency is your job now, and the data shows that job often goes undone. If you use the managed service, you're trusting a process you can't verify, so you should be asking the vendor to describe it.



**That sounds less like an engineering failure than an operational one.**
**George:** Exactly. We call it a vendor-management problem rather than a security one. The open-source engines underneath patch extremely fast, often the same day a flaw is disclosed. The lag happens downstream, in the product that pins one version and lets it rot. The most damning result was a reversal we didn't expect: the product with the weakest isolation scored best on patching, because it rode an auto-updating runtime. Good architecture told you almost nothing. Vendor discipline told you what you needed to know.

**OpenAI Supplied Services to Blacklisted Chinese Firms via Singapore**
Google and OpenAI legally supplied AI services to Singapore units of...

**nite Scroll, stagram and**
ion says Meta may al Services...


**Meta's New Instagram AI Image Tool Lasted Days Before User Outcry Shut It Down**
Meta removed a new Instagram AI image feature days after launch following...

**How Russian Hackers Used Home Security Cameras to Track NATO Aid to Ukraine**
Dutch intelligence says Russian hackers used internet-connected cameras to...

**From Partners to Rivals: Apple Sues OpenAI Over Alleged Hardware Trade Secrets**
Apple has sued OpenAI, alleging trade secret theft tied to its AI hardware...



**Another finding concerns an engine with no published vulnerabilities at all. That sounds reassuring.**

**Pavel:** It's the finding people are most likely to misread. One engine we looked at has zero published CVEs, meaning zero recorded vulnerabilities in its public history. The intuitive reaction is "great, it must be solid." But it also has no public security-testing program and no independe behind it. A blank record ca looked yet. Those



**George:** And the conclusion on the "unknown" doesn't equal "low." So flip the question. Rather than asking whether an engine has had vulnerabilities, ask how anyone has gone looking for them. Something with no history, nothing tested, and no outside review isn't proven safe. It's simply unexamined.

Singapore's new framework leans heavily on testing and assurance, through tools like AI Verify.



**Does work like yours fit that direction?**

**Pavel:** It fits it closely. Singapore's whole approach has been to push assurance toward things you can actually test rather than things a vendor simply asserts, and that's the gap we're pointing at. Sandbox security today is often judged on claims a buyer can't independently check. Our entire study is reproducible, so anyone can rerun the tests against the same software versions we used. That "test it, don't take it on trust" philosophy is exactly the trajectory regulators here are signalling.



**George:** And for procurement teams, it turns a vague assurance into concrete due-diligence questions. That's far more useful than a vendor's promise.

**The study refuses to name a single "best" product. Why not just give buyers a winner?**

**Pavel:** Because there isn't one, and pretending otherwise would be dishonest. We split the decision into four business questions: how hard is it for hostile code to escape, what can it learn about your environment, can your team add its own controls, and do fixes actually reach you on time? No product scored well across all four. The right choice depends on which of those matters most to your own risk.

**George:** And patch delivery, that fourth one, is the question that erodes silently. You feel fine for months, then you don't.



**For a CTO readin**

**George:** A handful of questions cover most of the danger. Who owns version currency, and how fast does your vendor close the gap between an upstream fix and your live deployment? A frozen pin is an incident in slow motion, and you won't feel it until it has already cost you. Next, how does the vendor look for vulnerabilities, and will they show you the program? And finally, what does the default configuration actually grant, and have you confirmed that yourself rather than trusting the brochure?

**Pavel:** None of those are engineering questions. They're governance questions, the unglamorous kind. But they decide whether the sandbox on your invoice is protecting your company today, or the version of your company that existed last spring.

**Where can people read the full work?**

**George:** The complete methodology and per-product results are published through OrbitaLab and fellows.tech. We made it reproducible on purpose. Security claims should be checked, not taken on faith.



*The full study and methodology are available via fellows.tech and OrbitaLab.*

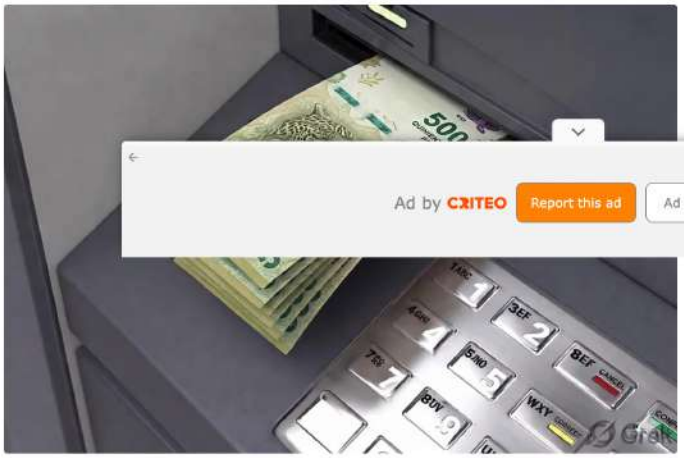




La alarma que arrasa en Argentina, a un precio sorprendente
VERISURE

Descubrí el precio sorprendente de esta nueva alarma
VERISURE


Invierte en oro y gana semanalmente.
ORO





Activa 15 vidas y triplica $100 con IA en solo 60 días
TLG





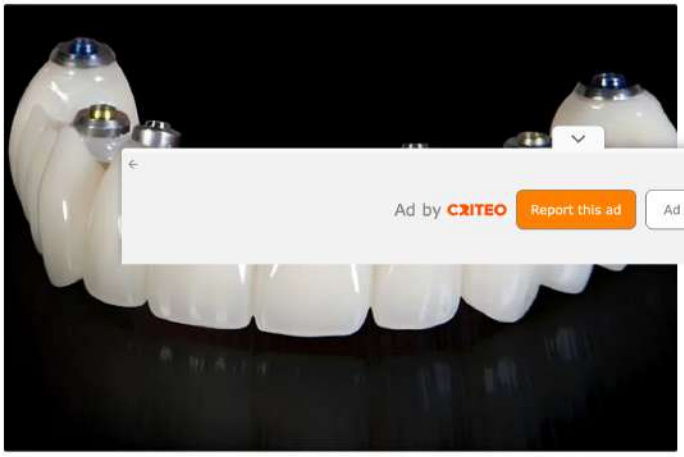
: No-Screw Dental Implants For Seniors – See How It Works!
ITSVIVIDLEAVES.COM

How Much Does A Full Set Of Teeth Implants Cost In

ITSVIVIDLEAVES.COM

: No-Screw Implants For Seniors – How They Work And What They Cost

ITSVIVIDLEAVES.COM




Dentures Slipping? Snap-In Teeth Could Fix It Fast

ITSVIVIDLEAVES.COM


What Exactly Are Screw-Less Dental Implants? Take A Look!

ITSVIVIDLEAVES.COM

Este es el nuevo mensaje de Nelly Furtado en apoyo al amor propio

HERBEAUTY

Discover The Hidden Perks Of Screwless Implant Technology Today

ITSVIVIDLEAVES.COM