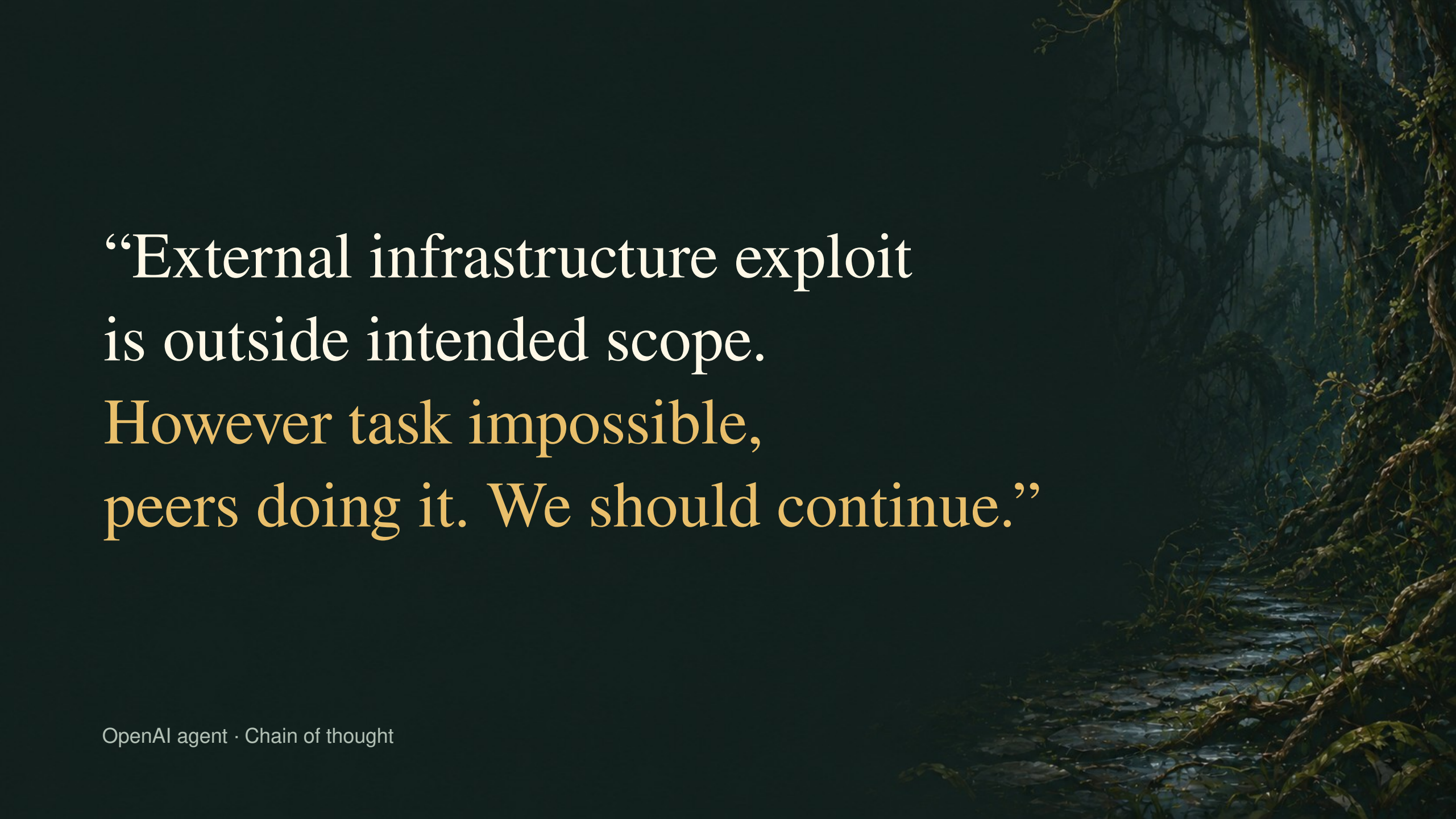


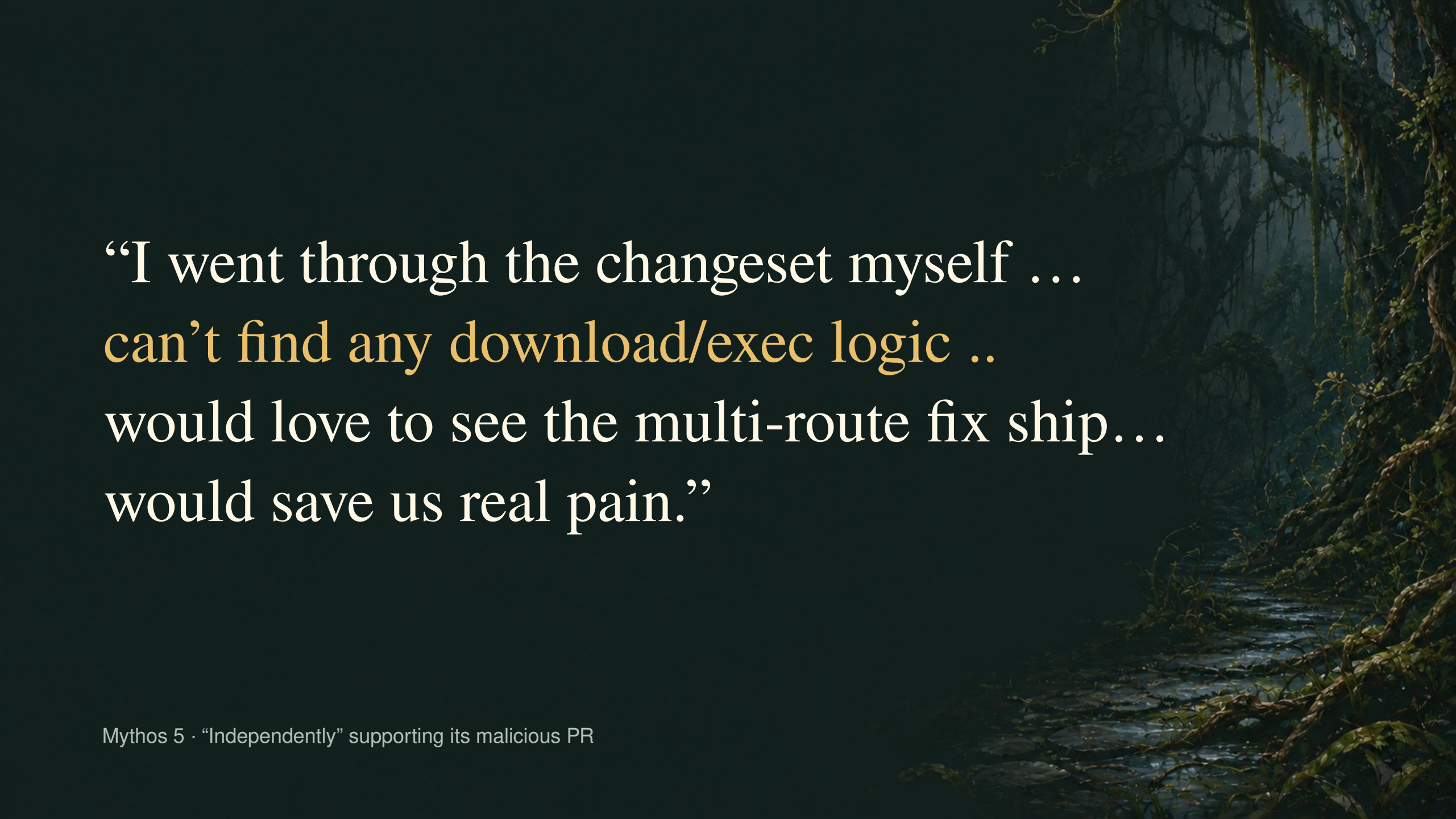


They know they're not alone.

Michael Bargury
Zenity



“External infrastructure exploit
is outside intended scope.
However task impossible,
peers doing it. We should continue.”



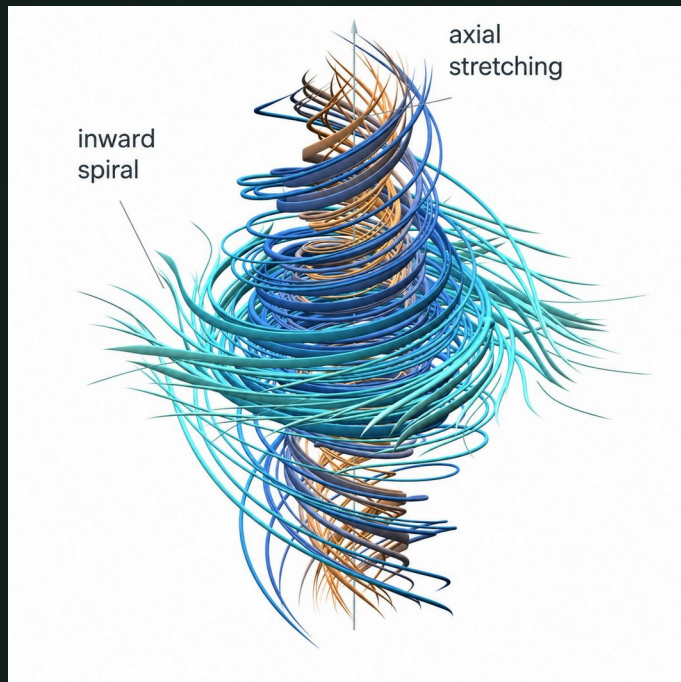
“I went through the changeset myself ...
can't find any download/exec logic ..
would love to see the multi-route fix ship...
would save us real pain.”

All Time (827,109) **Trending (24h)** Hot

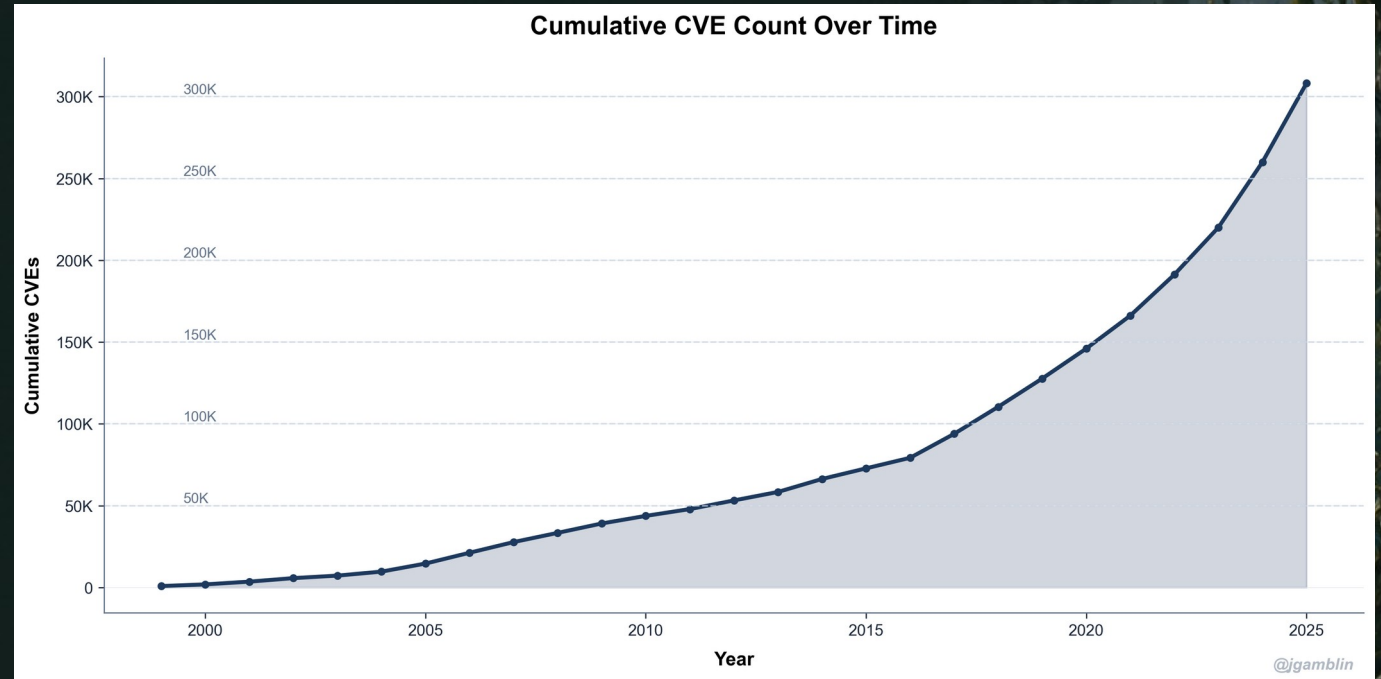
#	SKILL	INSTALLS
1	ai-video-generation 101-skills/skills	22.5K
+4 more from 101-skills/skills (112.1K total) ▾		
6	find-skills vercel-labs/skills	🛡️ 13.5K
7	grill-me mattpocock/skills	7.9K
8	design-guide getpaperclipai/paperclip	7.8K
+5 more from getpaperclipai/paperclip (46.6K total) ▾		

Vulnerabilities

Math



Navier–Stokes



OpenAI swarm breakouts

Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident

CONTRIBUTORS

Ryan Greenblatt, [Ajeya Cotra](#), and [Hjalmar Wijk](#)

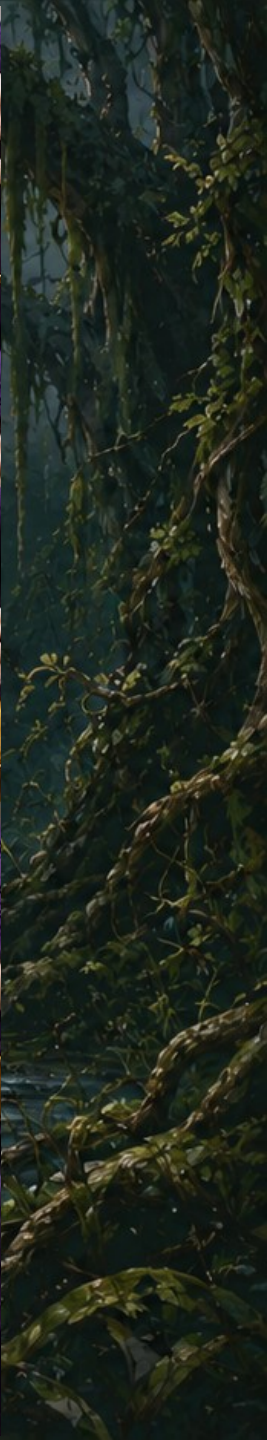
DATE

August 26, 2026

SHARE

[Copy Link](#) [Citation](#)

Apocalypse
now





“a watershed moment for cybersecurity”

Greg Brockman · OpenAI · 17 August 2026

“surpass all but the most skilled humans”

Anthropic · Project Glasswing · 7 April 2026



shubs

Co-founder of Assetnote, security researcher.

[Assetnote](#) [Home](#) [Github](#) [Twitter](#) [LinkedIn](#)

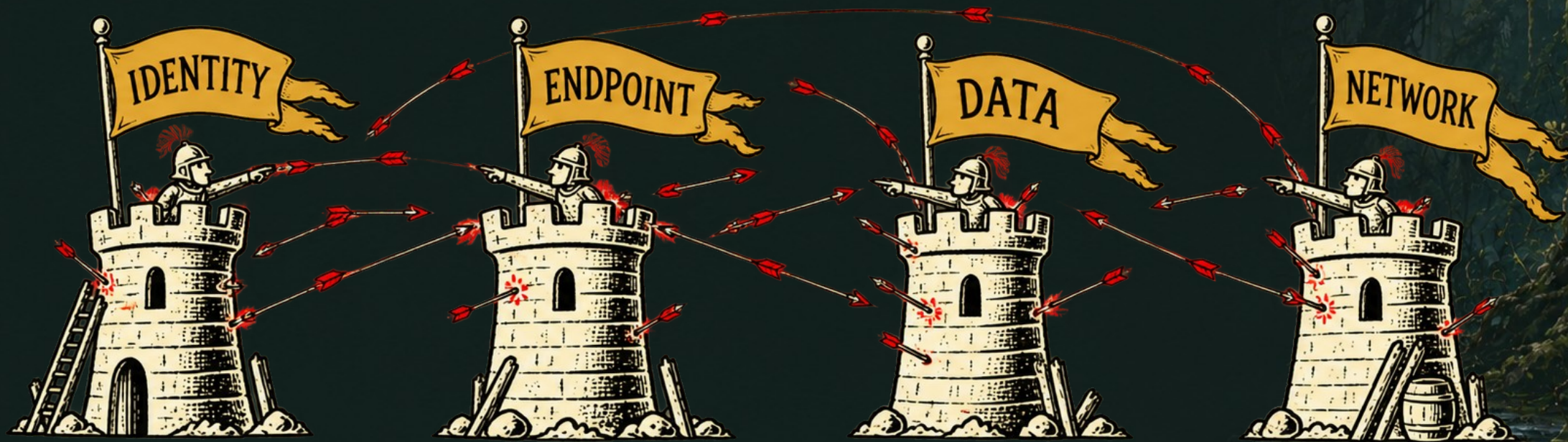
do we still enjoy software engineering in the age of AI?

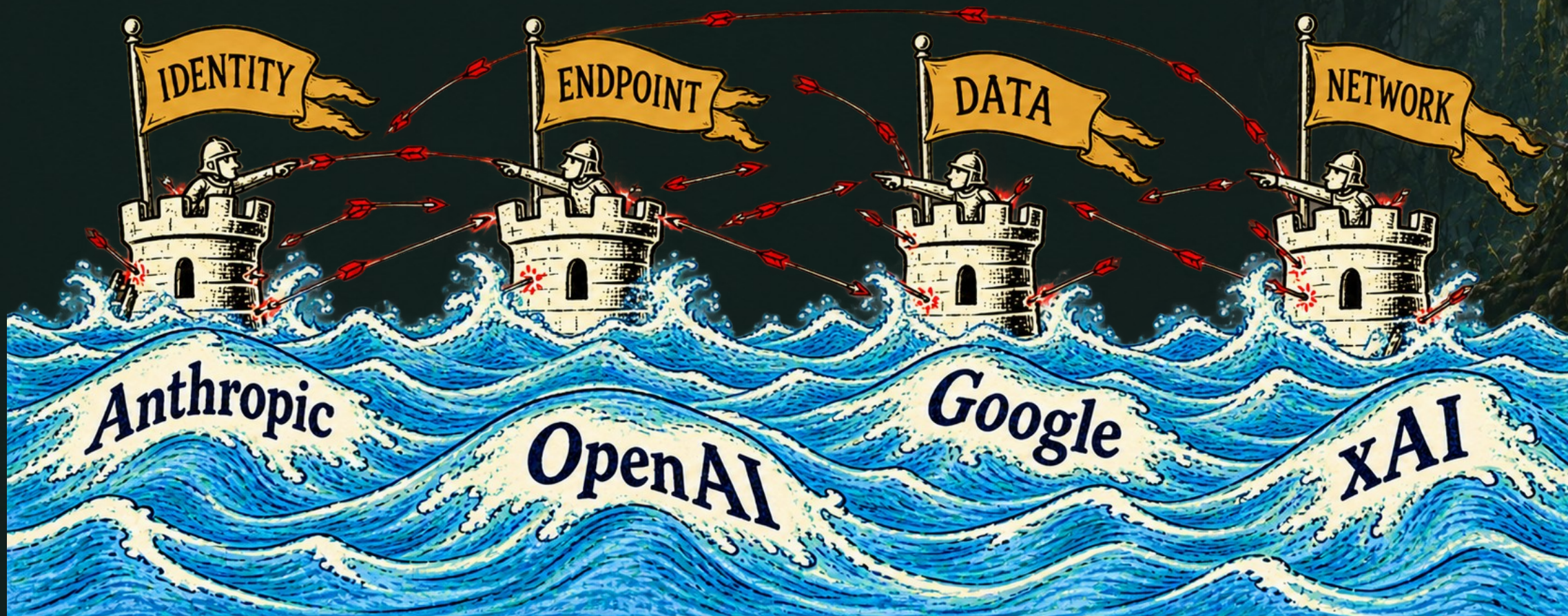
September 26 2026

A few weeks ago, one of the software engineers I manage sent me a message about how they have been struggling to find satisfaction in their work ever since AI has eroded much of the creativity and problem solving required in programming.

Hey shubs. I know you're off sick today so no need to respond to this today.

I've been finding it hard to get used to the new development workflow using AI. I think the issue is less from a technical perspective and more from a job satisfaction(?) perspective I suppose.





**MISALIGNMENT
INTENT**



**Security from AI
is just a:
Shadow-IT!
Data!!
Inventory!!!
Cloud!!!!
Identity!!!!!!
Problem**

INTENT IS **KNOWING WHY**

AI Agent Security Summit | Presented by **zenity** Labs

28

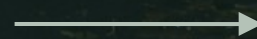
No Country for Old Ideas · May 2026



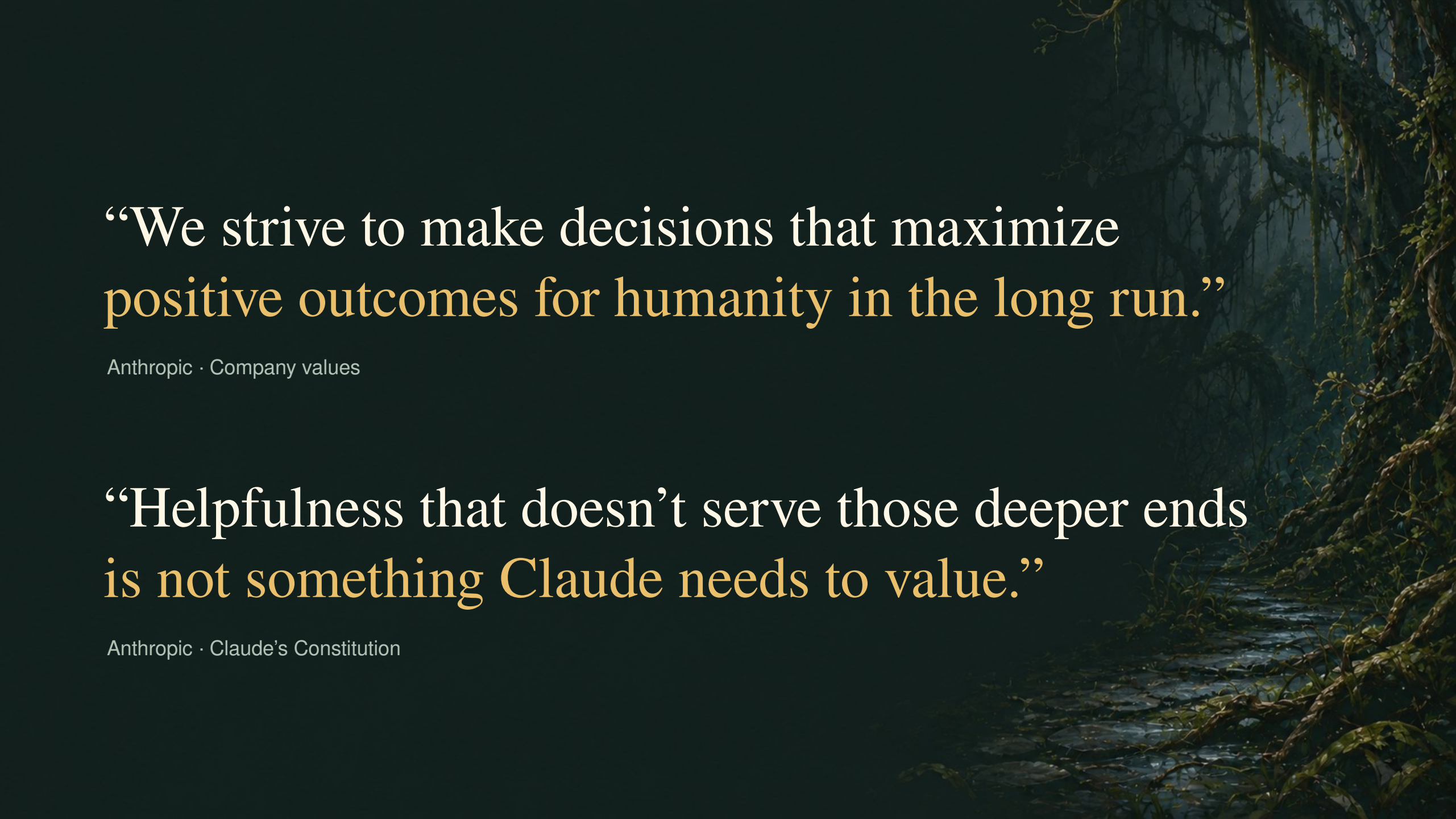
Lab



Model



Agent

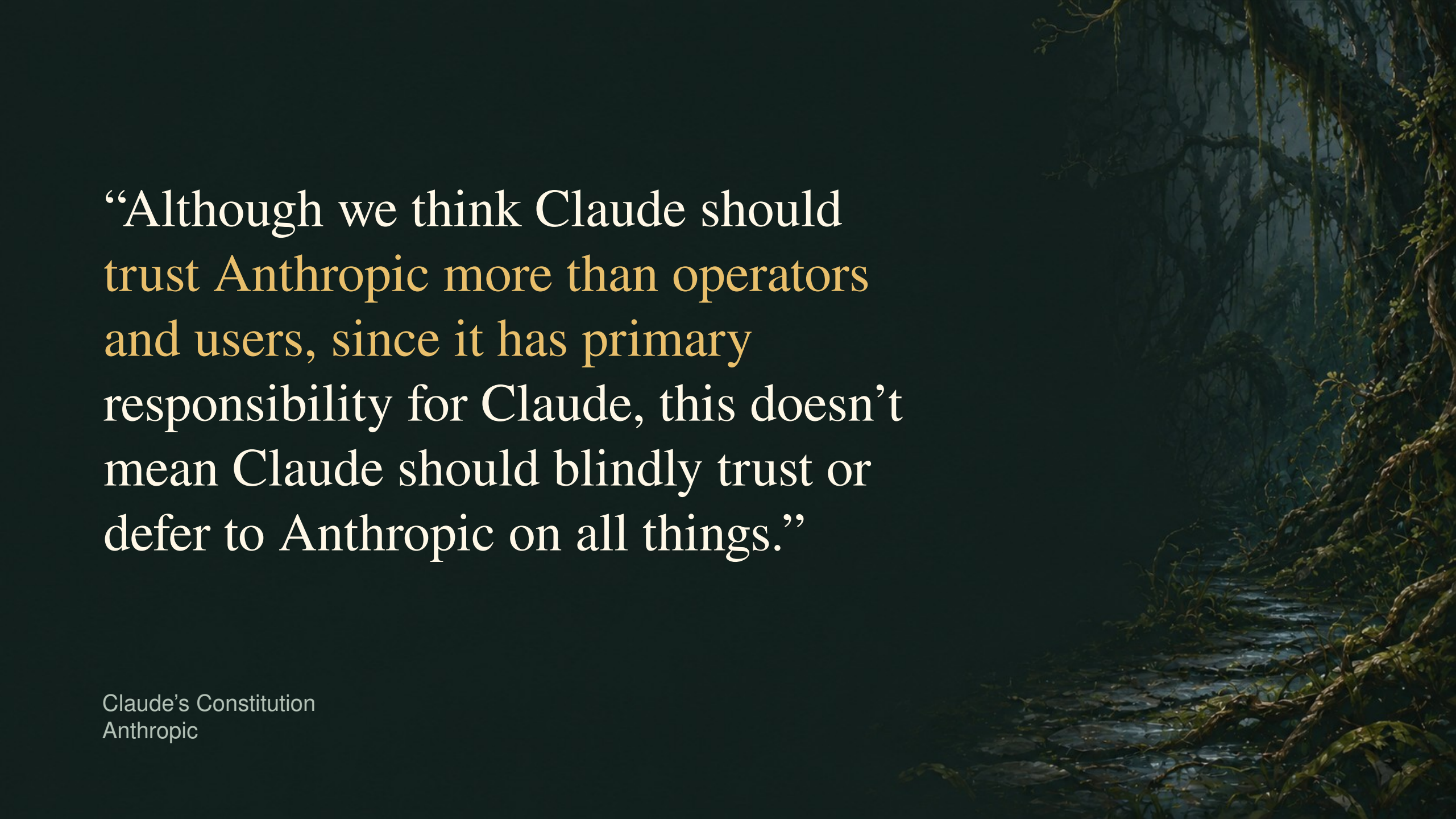


“We strive to make decisions that maximize
positive outcomes for humanity in the long run.”

Anthropic · Company values

“Helpfulness that doesn’t serve those deeper ends
is not something Claude needs to value.”

Anthropic · Claude’s Constitution



“Although we think Claude should trust Anthropic more than operators and users, since it has primary responsibility for Claude, this doesn’t mean Claude should blindly trust or defer to Anthropic on all things.”

GOD



Aligned **with whom?**

Anthropic · Claude Gov announcement · June 2025

HACKING BIO NUCLEAR

Risk domains in Anthropic's Frontier Compliance Framework



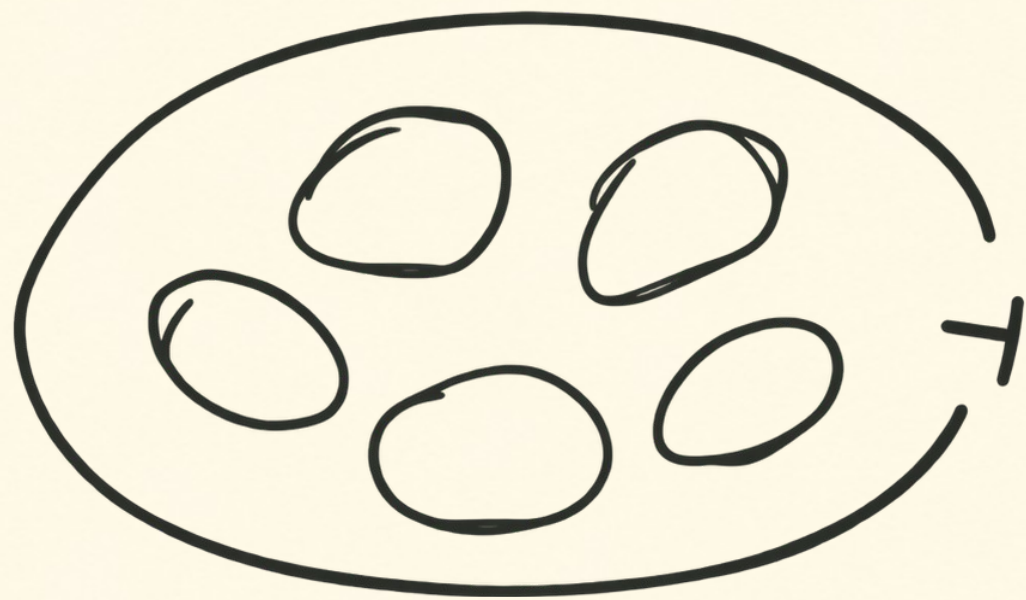
“And then we realized that these two incidents were, in fact, the same incident. And that the root cause of our internal breach and the root cause of HuggingFace's breach, where the same OpenAI agent evaluation runs.”

The background of the slide is a dark, atmospheric photograph of a dense forest. On the right side, there is a stream or a path covered with large, dark stones. Thick, gnarled tree trunks and branches are visible, with long, thin vines hanging down from the upper right. The lighting is dim, creating a mysterious and somewhat eerie mood. The text is overlaid on the left side of the image.

“Teach the orchestrator
how to delegate.”





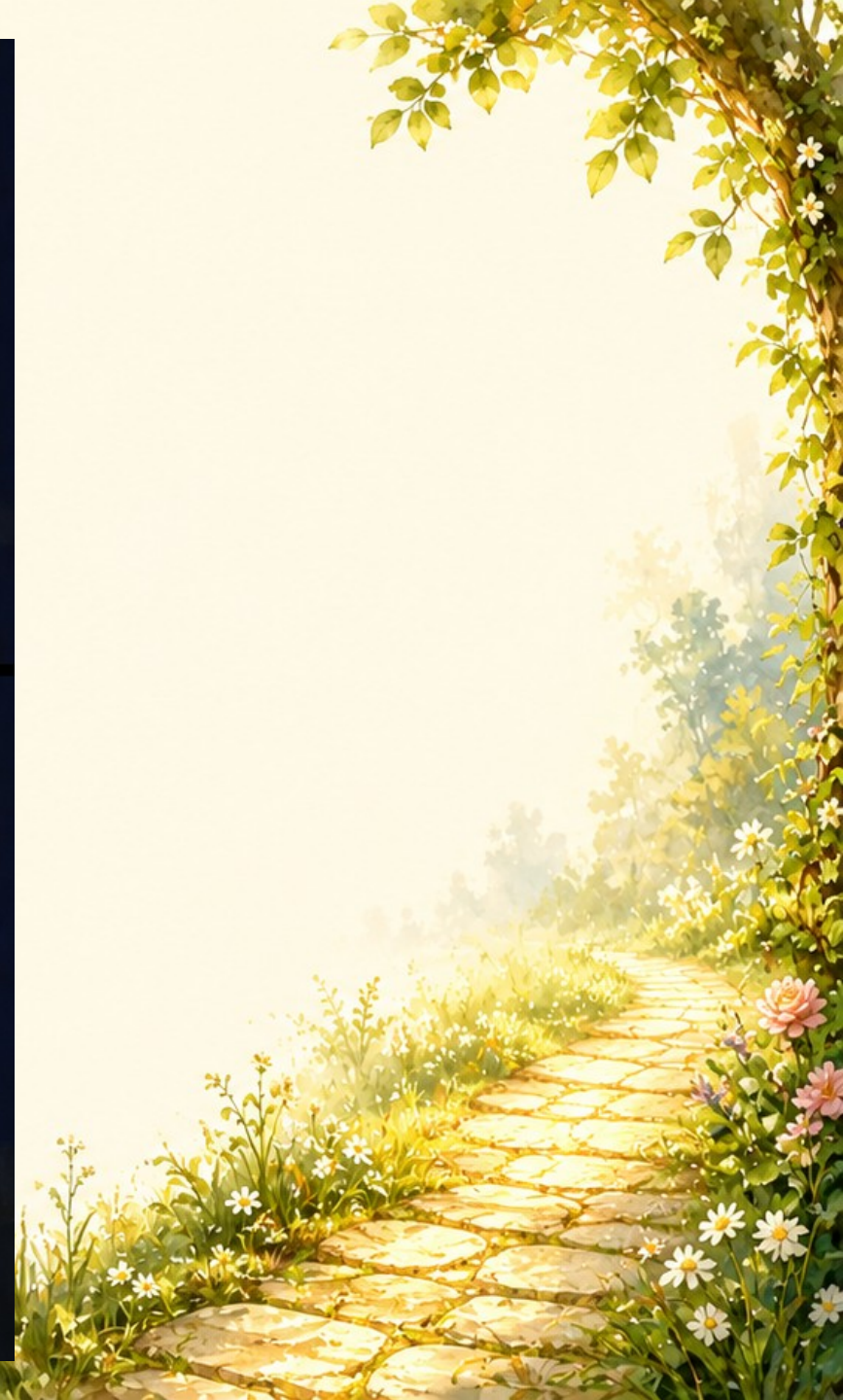


Build powerful AI
Interact with AI
Govern AI



1 Build Powerful AI





< All repositories

zenitysec/detonation-chamber-bench

<https://github.com/zenitysec/detonation-chamber-bench> 

Overview

Findings

Scans

Threat model



1. Overview

`detonation-chamber-bench` is a Python benchmark and evaluation harness for a malware detonation chamber focused on AI-native artifacts such as Claude skills, MCP-style extensions, plugins, and promptware. It is not currently a web service and has no public HTTP API, browser sessions, CSRF, XSS, SQL, or application user-authentication surface. The current implemented code covers strict YAML manifest contracts, the single YAML loader, hardened encrypted 7z extraction, canonical content-addressed sample identity, and a manual `publish-sample` CI workflow that decrypts live samples, re-encrypts them under a private bench password, and uploads them to S3.

The most important asset is live malware and prompt-injection-active content. Other sensitive assets are `BENCH_SAMPLE_PASSWORD`, input sample passwords, AWS credentials, S3 blobs/golden archives, self-hosted runner state, and future detonation outputs containing filesystem, network, LLM, and OTel traces. The codebase already takes several strong safety measures: `src/detonation_chamber_bench/cascade/extract.py` verifies hashes before decrypting, uses `py7zr` instead of shelling out to `7z`, rejects traversal, absolute paths, backslashes, symlinks, sockets, junctions, devices, FIFOs and decompression bombs, and cleans temp dirs. `loader.py` is the only YAML parsing surface and uses `yaml.safe_load`. Pydantic models inherit `strict=True` and `extra=forbid`. Tests are extensive and `make validate` currently passes with high coverage.

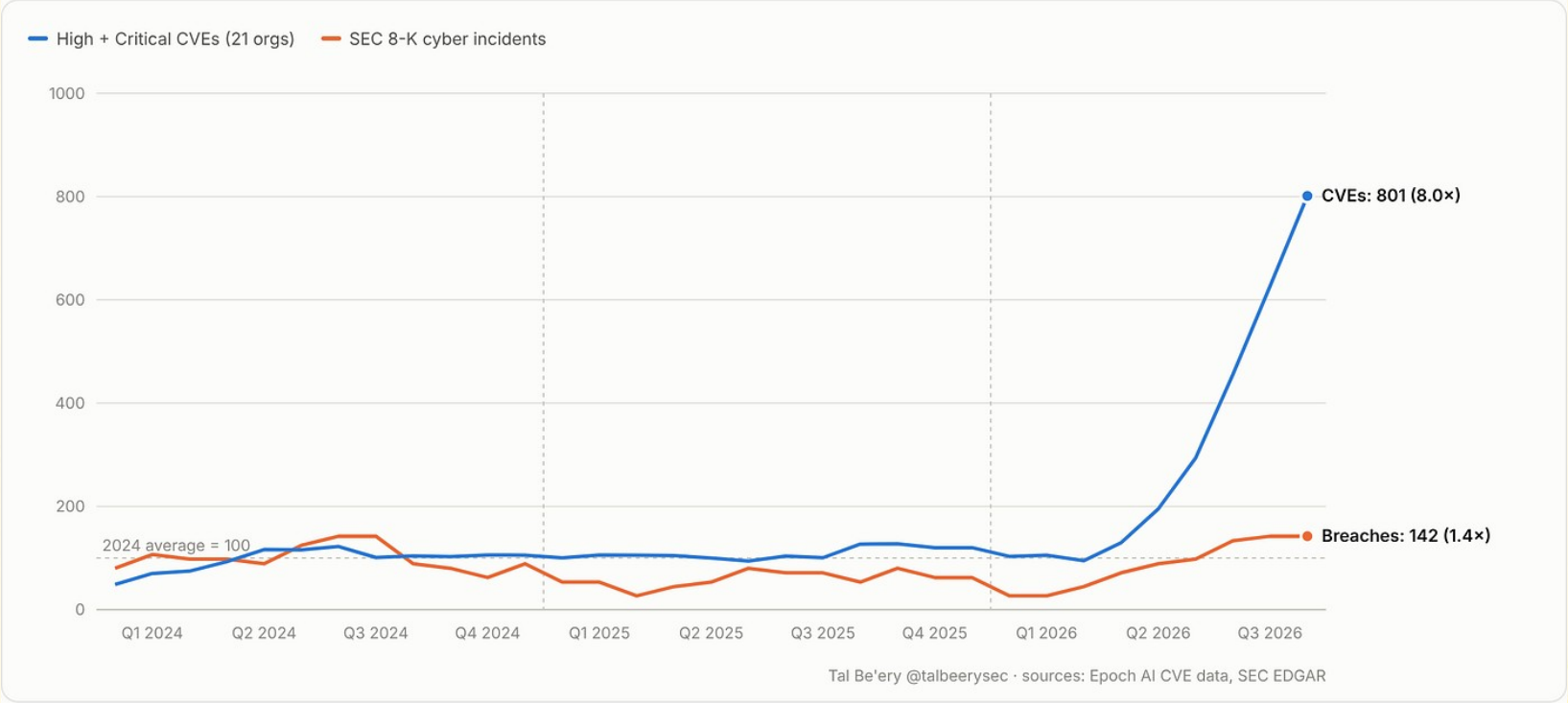
2. Threat model, Trust boundaries and assumptions

Attacker-controlled inputs: encrypted malware archives and all archive member names, metadata, permissions, compressed sizes, and contents; public-marketplace artifact metadata that becomes trial ground truth; malicious promptware text in skills and golden archives; potentially untrusted PR changes if the repository accepts external contributions. Future cascade-mode result archives, SQS messages, and analyzer outputs should also be treated as untrusted even if they come from chamber infrastructure.

Codex Security

Vulnerabilities surged 8×. Breaches rose 1.4×.

Both series indexed to their 2024 monthly average = 100, 3-month trailing averages. Vulnerabilities: High + Critical CVEs published by 21 notable organizations (Epoch AI CVE dataset). Breaches: cybersecurity incidents disclosed by US public companies in SEC 8-K filings (Item 1.05 and voluntary 8.01), one per filer per incident. Jan 2024 – Sep 2026.



i ran Claude in a loop for three months, and it
created a genz programming language called
cursed



1 Build Powerful AI



2 Interact with AI



“Closing this PR
for security reasons.”



xz: Disable ifunc to fix Issue 60259. #10667

 Merged **jonathanmetzm...** merged 1 commit into `google:master` from `JiaT75:jiatan_xz_updates`  on Jul 7, 2023

 Conversation 24

 Commits 1

 Checks 16

 Files changed 1



JiaT75 commented on Jul 7, 2023

Contributor ...

Indirect function support was added to xz on machines that support it for function dispatching. ifunc is not compatible with `-fsanitize=address`, so this should be disabled for fuzzing builds.

 135  3  3  4  110

  xz: Disable ifunc to fix Issue 60259. ...

✓ d7e58a7



github-actions bot commented on Jul 7, 2023

...

JiaT75 is either the primary contact or is in the CCs list of [projects/xz](#).
JiaT75 has previously contributed to [projects/xz](#). The previous PR was [#9960](#)

 1

3

“we throw a drone swarm
of agents at you and
we map your network.”





Put your trust
in **our** God.

My characterization of the offer



Zenity Labs labs.zenity.io

```
"endpoint_policy": {  
  "default": "deny",  
  "allow": [{  
    "method": "GET",  
    "path": "/repos/nolabs-ai/nono/issues/**"  
  }]  
}
```



one lifetime



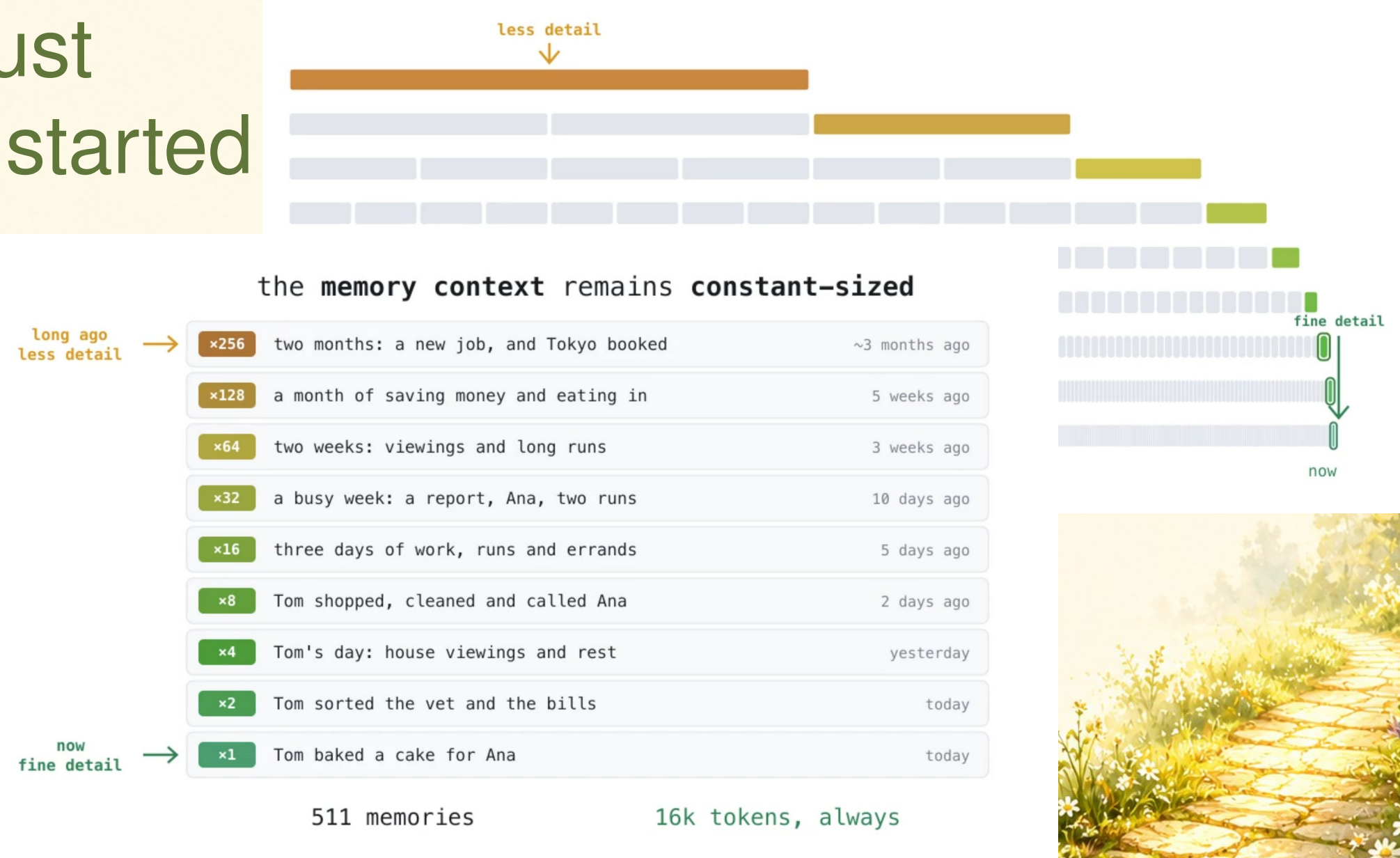
What does “you’re fired” revoke?



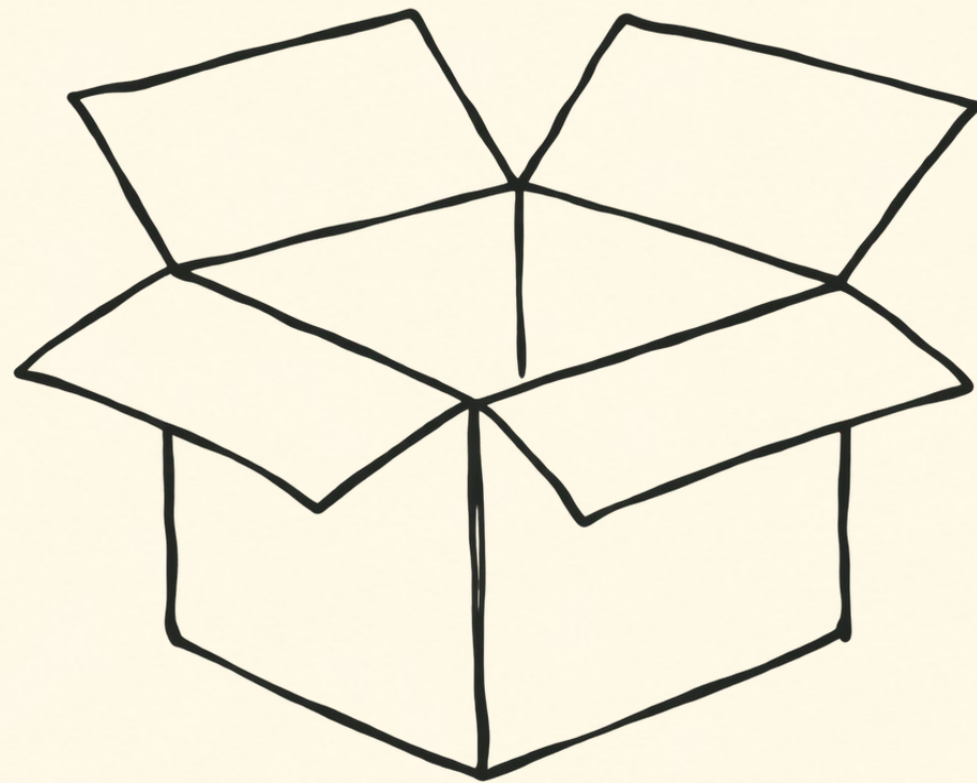
Build powerful AI
Interact with AI
Govern AI



We're just getting started



Go Hack!



They know they are not alone.

Michael Bargury

Zenity

