AI GENERATED
TRANSPARENCY DISCLOSURE

THE OTHER SIDE · A FYTAHQ COMPANION RESOURCE

Safety Until It Costs Something

Amodei left OpenAI over safety. But what he shipped was banned in three days!

EPISODE 1 · THE STORY, WITH ITS SOURCES

"Every coin has two sides, and we have always looked at both. Today we look at the other one."

The episode as a written story · every quotation referenced · 11 sources

---

ISSUE · SEPTEMBER 2026 · FACELESSYOUTUBEAUTOMATIONHQ.COM

AI GENERATED
TRANSPARENCY DISCLOSURE

WHAT THIS DOCUMENT IS

The story of the episode, in the order it is told

This is the episode as a written story: what the video says, chapter by chapter, in the order it says it, condensed from the spoken text and adding nothing to it. Every quotation is reproduced word for word, as spoken or written by the person named, and carries a number in square brackets that points to the source list at the end. The reported figures carry one too. Where the episode says that something is reported rather than proven, this text says so as well.

00:00 · WHERE WE STAND

> **THE DISCLAIMER THAT OPENS EVERY EPISODE, WORD FOR WORD**
>
> Before we start, one thing about where we stand.
>
> We are not against artificial intelligence.
>
> We use it every day. Even this video is made with it.
>
> This channel exists to make artificial intelligence easier to work with. To help people use it better, and with more purpose. And to take away the hesitation, so anyone can look at this technology and judge it for themselves.
>
> But every coin has two sides, and we have always looked at both. Today we look at the other one.
>
> Everything you will hear is sourced. Where something is reported rather than proven, we say so.
>
> And artificial intelligence is not the only thing that can be wrong. So can we. We check every claim to the best of our knowledge — and where we are not certain, we say that too.

00:52 IN THE EPISODE

He left OpenAI over AI safety

In December 2020, the Vice President of Research walked out of OpenAI. Over the following weeks, fourteen more researchers followed him, among them the Vice President of Safety and Policy. Nobody fired them, nobody poached them.

They left the most important artificial intelligence company on earth voluntarily, and they said why: the company was commercializing too fast, and the safety work could not keep up with the shipping. In January 2021, seven of them founded a company together.

The Vice President of Research was Dario Amodei; the Vice President of Safety and Policy was his sister, Daniela. Six years later, that same man runs a company worth close to a trillion dollars.

In April 2026 it passed OpenAI in revenue: forty-seven billion against twenty-five, up from one billion eighteen months earlier. He left over commercialization. He built the most commercialized company in the field. [1]

01:58 IN THE EPISODE

The admission, on camera

None of it surprised him.

Here is what he told Bloomberg, on camera, this year: "I was watching this graph for a while and I said, oh yeah, we'll probably become the AI company with the most revenue and the most valuation sometime around this time. And indeed, indeed, it has happened." [1]

On its own, that is not a scandal. People change, companies grow, and being right about your own growth chart is not a crime. That is not the story. The story is about what happened every single time safety came with a cost attached to it.

Not a prize. A cost. It happened three times inside eighteen months, and each time the answer was different. The episode watches what the difference depends on. [1]

02:49 IN THE EPISODE

Test one — the Department of Defense

February 2026. Amodei demands that Claude not be used for autonomous weapons, and not to surveil American citizens. Anthropic says no to the United States Department of Defense.

The reaction is immediate: the President orders every US government agency to stop using Anthropic products, and the government designates Anthropic a supply chain risk, a label normally reserved for foreign adversaries suspected of sabotage.

The Secretary of Defense goes on the record: Anthropic, quote, is run by an ideological lunatic. Anthropic takes the Department of Defense to federal court, and on March the twenty-fourth, US District Judge Rita F. Lin says from the bench: "It looks like an attempt to cripple Anthropic." [3]

She adds that she is concerned the government is punishing the company for criticizing it in public. That looks like conviction. Hold that thought. What happened after that, the contracts, the targeting, and the question Bloomberg put to him about a school in Iran, is Episode 2.

This episode stays with safety. [2] [3] [11]

04:15 IN THE EPISODE

Test two — the model they shipped anyway

The second test comes six weeks after the court hearing. On June the ninth, 2026, Anthropic releases Claude Fable 5; NBC calls it the first publicly available Mythos-class model. From Anthropic's own documentation: Fable and Mythos are the same underlying model, and the difference is the safeguards.

Fable is the public version with the fence on. Mythos is the same thing with the fence off, and it went only to a small group under a program called Project Glasswing. How capable is the version behind the fence?

Anthropic answered that itself, in a different product announcement: Opus 5 is close to Mythos at identifying software vulnerabilities, but considerably less successful at developing exploits for them. By their own account, Mythos does not just find the hole; it writes the code to get through it.

They knew that before shipping, and wrote it into the announcement: thousands of zero-day vulnerabilities, across every major operating system and browser.

And then, on camera: “Some of the early companies that we gave this to said things like, this is a super weapon, you should have to own a gun license to use it. Please don’t release this.” [4]

His own customers told him not to release it. He released it anyway. [4] [5]

05:37 IN THE EPISODE

The fence, and how fast it fell

The fence held for a while. Not long. Researchers at Amazon found a jailbreak in Fable, and Amazon’s chief executive, Andy Jassy, called the US Treasury Secretary, Scott Bessent.

On June the eleventh, an authorized red team ran the model against classified US government systems and got through almost all of them, in hours; Senator Mark Warner disclosed it, citing General Joshua Rudd, who runs both the National Security Agency and US Cyber Command.

On June the twelfth, the Commerce Department ordered the model restricted, with ninety minutes to comply, the first time the US government had applied export controls to a software model instead of to hardware. It had been out for about three days.

Whether anyone did anything with it in those three days, nobody knows; the episode claims nothing and has no evidence. But whether real damage was done usually only becomes clear when it surfaces somewhere. [6]

06:45 IN THE EPISODE

How hard they fought this time

So how hard did Anthropic fight this one? Not at all. It disabled both models within hours, worldwide, for every customer.

The same company that had gone to federal court against the Department of Defense six weeks earlier complied here without a fight, and immediately opened negotiations to get the models switched back on.

The lawsuit that did get filed came from a customer, a legal technology company called Legion LegalTech, on June the twenty-third. There was never a ruling.

On June the thirtieth the Commerce Secretary lifted the restriction, because Anthropic had agreed to proactively detect and address security risks in the models, to work with the government on standards, and to report malicious activity.

Those commitments were made after the government shut the model down, not before shipping it. And there had been a commitment before: Anthropic had previously agreed to pause development if a capability discovery warranted it. That language was quietly dialled back in February, four months before Mythos shipped. [1]

07:57 IN THE EPISODE

The letter nobody at Anthropic signed

Six weeks later, on July the twenty-fourth, Jensen Huang posts on X for the first time in his life. One post: a coalition letter titled Open Weights and American AI Leadership, asking Washington not to restrict downloadable models, the free ones anybody can run on their own machine.

Every frontier provider signed it, and a long list of Fortune 500 companies, a hundred and fifty organisations to date. Anthropic did not sign, and still has not.

A second initiative followed, the Secure Open AI Alliance, founded by Nvidia with Microsoft, IBM, HuggingFace and SpaceX, because defenders need capable models they can run on their own machines. Anthropic has signed neither. Now the other side, because this is where most videos cheat.

Amodei has never called for a ban on open models; he said so in writing on July the twenty-seventh and called models without dangerous capabilities a public good. What he asks for is mandatory safety testing for every sufficiently capable model, open or closed.

That sounds like equal treatment. But a company charging fifty dollars per million output tokens, and currently leading the market in revenue, can absorb the cost of compliance testing. A free model has nobody to pass that cost to.

You never have to ask for a ban if you set the conditions so that only the frontier providers can meet them. [7]

09:46 IN THE EPISODE

Their own documentation

And now the part that needs no interpretation, because Anthropic published it themselves, in their developer documentation. When an Anthropic model declines a request, the refusal comes with a category. There are five. Two of them are cyber and bio.

Here is a third, quoted exactly: Frontier LLM. The request could assist the development of competing AI models, which is restricted under Anthropic's commercial terms. Three of those categories are about harm.

One is about competition, and it is not justified with safety; it is justified, at the customer's expense, with commercial terms. They sit in the same table, same classifier, same downgrade, and the customer is never told which one hit him.

The same page adds one more line: benign machine learning work can also trigger this category. That is not a warning. It is a pre-emptive excuse for false positives, written into the documentation in advance. [8]

11:08 IN THE EPISODE

Test three — this one is mine

The third test is the channel's own, and the position is on the table first: the channel pays for Max, two hundred dollars a month, the top tier, plus API access billed by usage, just under three thousand dollars in May and two thousand eight hundred in June.

Not an opponent of this company. Which is exactly why this next part matters. In the middle of ordinary work, the channel logged into a YouTube analytics tool. Paid subscription, own account, own data, signed in with authentication.

Ask yourself which attacker signs in. No exploit, no malware, no penetration test, nothing pointed outward. A classifier fired anyway: Opus 5 blocked, dropped to Opus 4.8, and it does not reset with the next message.

Anthropic defines two classes of activity that justify a block, prohibited use and high-risk dual use; logging into an analytics dashboard is neither. And they say why.

Here is the notice that appeared on screen, word for word: "Our intentionally broad safeguards allow us to deliver more capabilities faster, but can sometimes flag legitimate coding, cybersecurity, and biology tasks." [9] Read the trade in that sentence.

The safeguards are intentionally broad so the company can ship faster. The speed belongs to them. The false positive belongs to the customer. And they wrote it into the message they show the customer. [9][10]

13:43 IN THE EPISODE

The rule

Three tests. Same company, same eighteen months, same subject.

Against the Department of Defense, the safety position and the business interest pointed in the same direction: Anthropic, whose entire brand is safety and whose enterprise contracts are eighty percent of its revenue, said no to weapons and surveillance, fought in federal court, and in the very next quarter passed OpenAI in revenue.

With Mythos, safety pointed against the business. A model you do not release is a model you do not sell, and there is a public offering coming.

The customer's warning is on the record; Anthropic shipped it anyway, and when the government banned it, complied in hours and started negotiating to get it back. And with the paying customer, strictness costs Anthropic nothing at all.

The bill goes to the customer, so a suspicion was enough. With Mythos the danger was documented by Anthropic themselves, by Amazon's researchers, by the National Security Agency, and none of it stopped the release.

In the customer's case there was no danger at all, provably absent, and it triggered a downgrade anyway. A bouncer who checks nobody at the door and searches the cloakroom attendant twice a night is not bad at his job.

He is very good at a different job. The strictness does not follow the risk. It follows the invoice. [1]

15:16 IN THE EPISODE

Mythos, and what he said next

One more thing, and it comes from him. When Bloomberg asked whether the whole Mythos affair had simply been good marketing, he said: "We have suffered enormously commercially from not releasing this model. This model has incredibly accelerated research within Anthropic and production and next models." [1]

Read the second sentence again. Mythos did not disappear; it is accelerating research and the next models inside Anthropic, and he said that out loud, on camera, unprompted.

It fits everything else they have said publicly: Amodei, that they use Claude across the entire product development cycle; Boris Cherny, the engineer who built Claude Code, that it has been writing one hundred percent of his code for six months, with up to a few thousand Claudes running at any point; their own speaker on a stage in Britain, that when Claude 5 can build Claude 6, which can build Claude 7, things will move even more quickly.

So the model that was too dangerous to put in your hands is in their hands, building the next one.

He also put a number on the risk in the same interview, a ten to twenty-five percent chance of civilization collapse, and when Emily Chang pointed out that nobody boards a plane with a twenty-five percent crash rate, he agreed: twenty-five percent is too high.

So when somebody tells you they are slowing down for your protection, there is one question that settles it. What did it cost them? Ask it about all three. And notice that they only ever held the line on the one where holding it was good for business. [1]

SOURCES

Every claim, with its source next to it

11 sources, in the order the episode uses them. The bracketed numbers in the text lead here. Timecodes match the chapter marks on YouTube.

[1] **Inside Anthropic: The Circuit, with Emily Chang**
Bloomberg Originals · 2026
The on-camera interview quoted in the episode.
https://www.youtube.com/watch?v=v1wZwxY3CMg

[2] **Anthropic, Google, OpenAI and xAI granted up to $200 million from the DoD**
CNBC · 14 July 2025
The contracts themselves.
https://www.cnbc.com/2025/07/14/anthropic-google-openai-xai-granted-up-to-200-million-from-dod.html

[3] **Judge on Anthropic's Pentagon injunction hearing**
NPR · 24 March 2026
US District Judge Rita F. Lin from the bench — the line quoted in the episode.
https://www.npr.org/2026/03/24/nx-s1-5759276/anthropic-pentagon-claude-preliminary-injunction-hearing

[4] **Claude Opus 5**
Anthropic
The release itself, in their own words.
https://www.anthropic.com/news/claude-opus-5

[5] **Why Claude switched models in your conversation with Fable 5**
Anthropic Support
Fable and Mythos are the same underlying model; the difference is the safeguards.
https://support.claude.com/en/articles/15363606-why-claude-switched-models-in-your-conversation-with-fable-5

[6] **Real-time cyber safeguards on Claude Opus and Sonnet**
Anthropic Support
What the safeguards do, described by the company that built them.
https://support.claude.com/en/articles/14604842-real-time-cyber-safeguards-on-claude-opus-and-sonnet

[7] **Open Weights and American AI Leadership**
Microsoft — coalition letter
The letter named in the episode, and its list of signatories.
https://www.microsoft.com/en-us/corporate-responsibility/topics/open-weight/

[8] **Refusals and fallback**
Anthropic developer documentation
The five refusal categories, quoted exactly as written.
https://platform.claude.com/docs/en/build-with-claude/refusals-and-fallback

[9] **Why Claude switched models in your conversation with Opus 5**
Anthropic Support
The help article the block message itself linked to.
https://support.claude.com/en/articles/16049681-why-claude-switched-models-in-your-conversation-with-opus-5

[10] **Own record — account lockout, 12 August 2026, 15:12**
This channel · 12 August 2026
Our own paid account, our own data, no exploit. The wording of the block message is shown in the episode. We are not publishing the screenshot: it shows a signed-in session.
No public link — title and date as published.

[11] **Hegseth calls Amodei a 'lunatic' and defends Pentagon use of AI**
Bloomberg · Senate hearing, 30 April 2026
Pete Hegseth, U.S. Secretary of Defense, in a Senate hearing. Bloomberg's page is behind a paywall; title and date as published.
No public link — title and date as published.

ON THE EVIDENCE · CORRECTIONS

Not a video about a company that is bad at AI safety. It is about a company that is very good at one specific kind of AI safety. Not a claim that anyone broke the law. Short third-party excerpts are used for criticism, commentary, news reporting and analysis; all rights remain with their owners, credited above. If anything here is inaccurate, tell us and we will correct it — in a pinned comment on the video and, where it matters, in this document. That standard applies to us the same way it applies to everyone else in this episode. Write to us at facelessyoutubeautomationhq.com/contact.html.

WATCH THE EPISODE · EVERY SOURCE, ONLINE

The Other Side · Episode 1

Amodei left OpenAI over safety. But what he shipped was banned in three days! — 17:02, live since 8 September 2026. This channel is made with artificial intelligence, and it says so in every frame. The same sources as above, with live links, are on the resources page of the channel.

youtu.be/1YpIBcIi40k →

facelessyoutubeautomationhq.com/resources/the-other-side-01-safety-until-it-costs/ →