It’s practically impossible to run a big AI company ethically

Anthropic was supposed to be the good guy. It can’t be — unless government changes the incentives in the industry.

By Sigal Samuel

11 min. read · View original

Anthropic was supposed to be the good AI company. The ethical one. The safe one.

It was supposed to be different from OpenAI, the maker of ChatGPT. In fact, all of Anthropic’s founders once worked at OpenAI but quit in part because of differences over safety culture there, and moved to spin up their own company that would build AI more responsibly.

Yet lately, Anthropic has been in the headlines for less noble reasons: It's pushing back on a landmark California bill to regulate AI. It's taking money from Google and Amazon in a way that's drawing antitrust scrutiny. And it's being accused of aggressively scraping data from websites without permission, harming their performance.

What's going on?

The best clue might come from a 2022 paper written by the Anthropic team back when their startup was just a year old. They warned that the incentives in the AI industry — think profit and prestige — will push companies to "deploy large generative models despite high uncertainty about the full extent of what these models are capable of." They argued that, if we want safe AI, the industry's underlying incentive structure needs to change.

Well, at three years old, Anthropic is now the age of a toddler, and it's experiencing many of the same growing pains that afflicted its older sibling OpenAI. In some ways, they're the same tensions that have plagued all Silicon

Valley tech startups that start out with a "don't be evil" philosophy. Now, though, the tensions are turbocharged.

An AI company may want to build safe systems, but in such a hype-filled industry, it faces enormous pressure to be first out of the gate. The company needs to pull in investors to supply the gargantuan sums of money needed to build top AI models, and to do that, it needs to satisfy them by showing a path to huge profits. Oh, and the stakes — should the tech go wrong — are much higher than with almost any previous technology.

So a company like Anthropic has to wrestle with deep internal contradictions, and ultimately faces an existential question: Is it even possible to run an AI company that advances the state of the art while also truly prioritizing ethics and safety?

"I don't think it's possible," futurist Amy Webb, the CEO of the Future Today Institute, told me a few months ago.

If even high-minded Anthropic is becoming an object lesson in that impossibility, it's time to consider another option: The government needs to step in and change the incentive structure of the whole industry.

The incentive to keep building and deploying AI models

Anthropic has always billed itself as a safety-first company. Its leaders say they take catastrophic or existential risks from AI very seriously. CEO Dario Amodei has testified before senators, making the case that AI models powerful enough to "create large-scale destruction" and upset the international balance of power could come into being as early as 2025. (Disclosure: One of Anthropic's early investors is James McClave, whose BEMC Foundation helps fund Future Perfect.)

So you might expect that Anthropic would be cheering a bill introduced by California state Sen. Scott Wiener (D-San Francisco), the Safe and Secure Innovation for Frontier Artificial Intelligence Model Act, also known as SB 1047. That legislation would require companies training the most advanced and expensive AI models to conduct safety testing and maintain the ability to pull the plug on the models if a safety incident occurs.

But Anthropic is lobbying to water down the bill. It wants to scrap the idea that the government should enforce safety standards before a catastrophe occurs. “Instead of deciding what measures companies should take to prevent catastrophes (which are still hypothetical and where the ecosystem is still iterating to determine best practices)” the company urges, “focus the bill on holding companies responsible for causing actual catastrophes.”

In other words, take no action until something has already gone terribly wrong.

In some ways, Anthropic seems to be acting like any for-profit company would to protect its interests. Anthropic has not only economic incentives — to maximize profit, to offer partners like Amazon a return on investment, and to keep raising billions to build more advanced models — but also a prestige incentive to keep releasing more advanced models so it can maintain a reputation as a cutting-edge AI company.

This comes as a major disappointment to safety-focused groups, which expected Anthropic to welcome — not fight — more oversight and accountability.

“Anthropic is trying to gut the proposed state regulator and prevent enforcement until after a catastrophe has occurred — that’s like banning the FDA from requiring clinical trials,” Max Tegmark, president of the Future of Life Institute, told me.

The US has enforceable safety standards in industries ranging from pharma to aviation. Yet tech lobbyists continue to resist such regulations for their own products. Just as social media companies did years ago, they make voluntary commitments to safety to placate those concerned about risks, then fight tooth and nail to stop those commitments being turned into law.

In what he called “a cynical procedural move,” Tegmark noted that Anthropic has also introduced amendments to the bill that touch on the remit of every committee in the legislature, thereby giving each committee another opportunity to kill it. "This is straight out of Big Tech's playbook," he said

An Anthropic spokesperson told me that the current version of the bill "could blunt America's competitive edge in AI development" and that the company wants to "refocus the bill on frontier AI safety and away from approaches that aren't adaptable enough for a rapidly evolving technology."

The incentive to gobble up everyone's data

Here's another tension at the heart of AI development: Companies need to hoover up reams and reams of high-quality text from books and websites in order to train their systems. But that text is created by human beings, and human beings generally do not like having their work used without their consent.

All major AI companies scrape publicly available data to use in training, a practice they argue is legally protected under fair use. But scraping is controversial, and it's being challenged in court. Famous authors like Jonathan

Franzen and media companies like the New York Times have sued OpenAI for copyright infringement, saying that the AI company lifted their writing without permission. This is the kind of legal battle that could end up remaking copyright law, with ramifications for all AI companies. (Disclosure: Vox Media is one of several publishers that has signed partnership agreements with OpenAI. Our reporting remains editorially independent.)

What's more, data scraping violates some websites' terms of service. YouTube says that training an AI model using the platform's videos or transcripts is a violation of the site's terms. Yet that's exactly what Anthropic has done, according to a recent investigation by Proof News.

Web publishers and content creators are angry. Matt Barrie, chief executive of Freelancer.com, a platform that connects freelancers with clients, said Anthropic is "the most aggressive scraper by far," swarming the site even after being told to stop. "We had to block them because they don't obey the rules of the internet.

This is egregious scraping [that] makes the site slower for everyone operating on it and ultimately affects our revenue."

Dave Farina, the host of the popular YouTube science show *Professor Dave Explains*, told Proof News that "the sheer principle of it" is what upsets him. Some 140 of his videos were lifted as part of the dataset that Anthropic used for training. "If you're profiting off of work that I've done [to build a product] that will put me out of work, or people like me out of work, then there needs to be a conversation on the table about compensation or some kind of regulation," he said.

Why would Anthropic take the risk of using lifted data from, say, YouTube, when the platform has explicitly forbidden it and copyright infringement is such a hot topic right now?

Because AI companies need ever-more high-quality data to continue boosting their models' performance. Using synthetic data, which is created by algorithms, doesn't look promising. Research shows that letting ChatGPT eat its own tail leads to bizarre, unusable output. (One writer coined a term for it: "Hapsburg AI," after the European royal house that famously devolved over generations of inbreeding.) What's needed is fresh data created by actual humans, but it's becoming harder and harder to harvest that.

Publishers are blocking web crawlers, putting up paywalls, or updating their terms of service to bar AI companies from using their data as training fodder. A new study from the MIT-affiliated Data Provenance Initiative looked at three of the major datasets — each containing millions of books, articles, videos, and other scraped web data — that are used for training AI. It turns out, 25 percent of the highest-quality data in these datasets is now restricted. The authors call it "an emerging crisis of consent." Some, like OpenAI, have begun to respond to this in part by striking licensing deals with media outlets, including Vox. But that may only get them so far, given how much remains officially off-limits.

AI companies could theoretically accept the limits to advancement that come with restricting their training data to what can be ethically sourced, but then they wouldn’t stay competitive. So companies like Anthropic are incentivized to go to more extreme lengths to get the data they need, even if that means taking dubious action.

Anthropic acknowledges that it trained its chatbot, Claude, using the Pile, a dataset that includes subtitles from 173,536 YouTube videos. When I asked how it justifies this use, an Anthropic spokesperson told me, “With regard to the dataset at issue in The Pile, we did not crawl YouTube to create that dataset nor did we create that dataset at all.” (That echoes what Anthropic has previously told Proof News: “[W]e’d have to refer you to The Pile authors.”)

The implication is that because Anthropic didn’t make the dataset, it’s fine for them to use it. But it seems unfair to shift all the responsibility onto the Pile authors — a nonprofit group that aimed to create an open source dataset researchers could study — if Anthropic used YouTube’s data in a manner that violates the platform’s terms.

"Companies should probably do their own due diligence. They're using this for commercial purposes," said Shayne Longpre, lead author on the Data Provenance Initiative study. He contrasted that with the Pile's creators and the many academics who have used the dataset to conduct research. "Academic purposes are clearly distinct from commercial purposes and are likely to have different norms."

The incentive to rake in as much cash as possible

To build a cutting-edge AI model these days, you need a ton of computing power — and that's incredibly expensive. To gather the hundreds of millions of dollars needed, AI companies have to partner with tech giants.

That's why OpenAI, initially founded as a nonprofit, had to create a for-profit arm and partner with Microsoft. And it's why Anthropic ended up taking multibillion-dollar investments from Amazon and Google.

Deals like these always come with risks. The tech giants want to see a quick return on their investments and maximize profit. To keep them happy, the AI companies may feel pressure to deploy an advanced AI model even if they're not sure it's safe.

The partnerships also raise the specter of monopolies — the concentration of economic power. Anthropic's investments from Google and Amazon led to a probe by the Federal Trade Commission and are now drawing antitrust scrutiny in the UK, where a consumer regulatory agency is investigating whether there's been a "relevant merger situation" that could result in a "substantial lessening of competition."

An Anthropic spokesperson said the company intends to cooperate with the agency and give them a full picture of the investments. "We are an independent company and none of our strategic partnerships or investor relationships diminish the independence of our corporate governance or our freedom to partner with others," the spokesperson said.

Recent experience, though, suggests that AI companies' unique governance structures may not be enough to prevent the worst.

Unlike OpenAI, Anthropic has never given either Google or Amazon a seat on its board or any observation rights over it. But, very much like OpenAI, Anthropic is relying on an unusual corporate governance structure of its own design. OpenAI initially created a board whose idealistic mission was to safeguard humanity’s best interests, not please stockholders. Anthropic has created an experimental governance structure, the Long-Term Benefit Trust, a group of people without financial interest in the company who will ultimately have majority control over it, as they’ll be empowered to elect and remove three of its five corporate directors. (This authority will phase in as the company hits certain milestones.)

But there are limits to the idealism of the Trust: It must “ensure that Anthropic responsibly balances the financial interests of stockholders with the interests of those affected by Anthropic’s conduct and our public benefit purpose.” Plus, Anthropic says, “we have also designed a series of 'failsafe' provisions that allow changes to the Trust and its powers without the consent of the Trustees if sufficiently large supermajorities of the stockholders agree."

And if we learned anything from last year's OpenAI boardroom coup, it's that governance structures can and do change. When the OpenAI board tried to safeguard humanity by ousting CEO Sam Altman, it faced fierce pushback. In a matter of days, Altman clawed his way back into his old role, the board members who'd fired him were out, and the makeup of the board changed in Altman's favor. What's more, OpenAI gave Microsoft an observer seat on the board, which allowed it to access confidential information and perhaps apply pressure at board meetings. Only when that raised (you guessed it) antitrust scrutiny did Microsoft give up the seat.

"I think it showed that the board does not have the teeth one might have hoped it had," Carroll Wainwright, who quit OpenAI this year, told me. "It made me question how well the board can hold the organization accountable.”

That’s why he and several others published a proposal demanding that AI companies grant them “a right to warn about advanced artificial intelligence.” Per the proposal: “AI companies have strong financial incentives to avoid effective oversight, and we do not believe bespoke structures of corporate governance are sufficient to change this.”

It sounds a lot like what another figure in AI told Vox last year: “I am pretty skeptical of things that relate to corporate governance because I think the incentives of corporations are horrendously warped, including ours.” Those are the words of Jack Clark, the policy chief at Anthropic.

If AI companies won’t fix it, who will?

The Anthropic team had it right originally, back when they published that paper in 2022: The pressures of the market are just too brutal. Private AI companies do not have the motivation to change that, so the government needs to change the underlying incentive structure within which all these companies operate.

When I asked Webb, the futurist, what a better AI business ecosystem could look like, she said it would include a mix of carrots and sticks: positive incentives, like tax breaks for companies that prove they're upholding the highest safety standards; and negative incentives, like regulation that would fine companies if they deploy biased algorithms.

With AI regulation at a standstill at the federal level — plus a looming election — it's falling to states to pass new laws. The California bill, if it passes, would be one piece of that puzzle.

Civil society also has a role to play. If publishers and content creators are not happy about having their work used as training fodder, they can fight back. If tech workers are worried about what they see at AI companies, they can blow the whistle. AI can generate a whole lot on our behalf, but resistance to its own problematic deployment is something we have to generate ourselves.