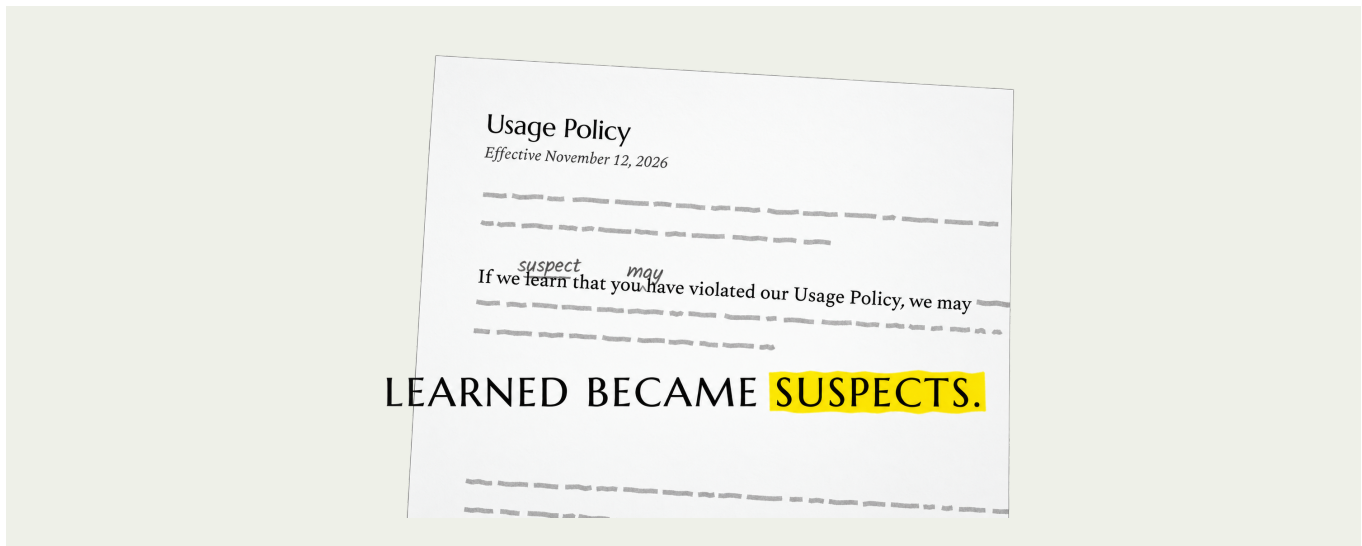




Trust in AI Is Checked Now, Not Assumed

Users, labs, and lawmakers each have to check whether the others can be trusted. Anthropic's new usage policy shows all three doing it at once.

Anthropic's new usage policy takes effect on November 12. Read beside the threat report the company published in September, it shows trust in AI being inferred at three levels at once: by the company about its users, by the public about the company, and by governments about both. This essay asks the same question at each level: what can a person on the outside check, and how far should the checking go?

The sentence that changed. Anthropic's usage policy, effective November 12, 2026.

The last essay by this author ended on one word: inferred. [21] Americans no longer assume that the technology companies can be trusted. They infer it, from what the companies do, and the inference is fraying. *Governance, Bureaucracy, and Organization* draws the same line between competence that is assumed and competence that has to be inferred. Trust works the same way. Anthropic's new usage policy, which takes effect on November 12, is a good place to watch that inference being made, because it is being made at three levels at once. [1] At the first level, a company decides whether to trust the people who use its tool. At the second, those people, and everyone else, decide whether to trust the company. At the third, governments decide how far to trust either one and how soon to write rules. Each level has a different worry and needs a different answer. The policy, read beside the threat report Anthropic published in September, shows all three.

The individual

At the first level, the company is the one inferring. It decides whether to trust the people who use its tool, and the evidence is what they do with it, watched by software. The policy says so in a sentence that changed. The version in force since September 2025 said that if the company learned you had violated it, it could throttle, suspend, or terminate your access. [2] The new one says that if the company suspects you may have violated it, it may warn, throttle, limit, suspend, or terminate you. [1] Learned became suspects. Had became may have.

Before treating that as a scandal, consider the fair reading. Enforcement at this scale is done by software. A classifier does not produce findings. It produces suspicion. The old sentence described a standard the company probably never met in practice. The new sentence describes the one it actually uses. So the change is not a new power. It is an honest label on an old one. It is also what inferred trust looks like when a company does it to its users: nobody is assumed to be in good faith, and everyone is judged, continuously, by what they do.

Whether that judgment is warranted is the question the casebook answers. On September 10, Anthropic published a threat report covering misuse of its models from December 2025 through August 2026, across seven areas: cyber operations, influence operations, surveillance, scams and fraud, biological misuse, conventional weapons development, and distillation. [3] The cases are specific. Russian actors, whom Anthropic says were not linked to the state, tried to use Claude to write software for an autonomous kamikaze drone swarm trained on Ukrainian combat footage. A consultant believed to work for Mali's intelligence service used it to build a system for monitoring about 25 million phones and generating dossiers on targets. [4] The report counts six conventional weapons cases and five involving biology, and it says Chinese AI developers were accused of routing live customer conversations through Claude to harvest training data. [5] In the weapons cases, Anthropic says the actors used its coding tool, Claude Code, in place of software engineers to write guidance and control code. [6]

Put the report beside the new policy and the match is close to line by line. The old weapons rule banned designing or producing weapons. The new one names drones, autonomous platforms, and software for targeting, fire control, and engagement. The old surveillance rule banned tracking a person's location or communications without consent. The new one adds online activity, associations, analysis of data already collected, and the building of dossiers. There is a new ban on routing requests through consumer subscriptions or misrepresenting which product is being used, and a sharper one on using outputs to train other models. There is an entire new section on fake personas, hidden sponsorship, and networks of sites built to fool search engines and AI systems. [1] None of this is a guess about what people might do. It is a record of what they did, and the rules arrived about a month after the company described the cases.

Opinion 1. That is the part this author finds most telling. It confirms what many people already suspected about AI: that it would be pulled toward weapons, surveillance, fraud, and influence operations as fast as toward anything else. The confirmation is not pleasant. It is still better to have it in writing from the company that watched it happen than to guess. And it settles the first level. A company that has seen this is right to infer trust in its users rather than assume it. The question is what everyone else can infer about the company.

The company

At the second level, the direction reverses. Users, and everyone else, infer whether to trust the company, and the evidence is what the company publishes and what it lets people check. The usual word for what this level needs is transparency. The word has been used so much that it has stopped meaning anything. Companies get credit for saying it, and nobody tests it. This essay uses a narrower test instead: can a person outside the company check it? A published casebook passes. A written standard passes, even an unflattering one. A reason and a date for a suspension would pass. A notice that says a user is being slowed down would pass. "We take this seriously" does not. By that test, Anthropic passes on two of the four things that matter and fails on the two that touch the user.

The two passes are the casebook and the standard. The threat report is a published record of what the company caught, and the suspicion sentence is an accurate description of how it enforces, written in a way that reads worse than the polite version would have. Both can be checked. Both earn credit, and the second one cost the company something to say. The company that writes the accurate sentence invites criticism that the company keeping the polite one avoids.

The two failures are what happens to a user on the other end of that suspicion. The first is the loud penalty, suspension. There is an appeal process. It lives in a support article, not in the policy. If you believe your account was wrongly suspended, you log in to the suspended account, fill out a form, and the safety team can investigate why the account was disabled. The company promises a reply as soon as it can and asks for patience. [7] There is no timeline. There is no commitment to tell you what you did. The form can only be reached by logging in to the account that was just suspended. [8] Users posting bug reports this summer described suspension emails that cited "suspicious signals" with no

specific reason, and an appeal link that led back to the login page with no form behind it. [9] Those are user reports, not verified findings, and this essay cites them as such. But they describe exactly what a suspicion-based system would be expected to produce: a notice without a reason, and a door that may or may not open.

This is not an Anthropic problem. It is an industry pattern, and the industry has already shown where it leads. In 2021 a San Francisco father photographed his toddler's infection at a nurse's request for a telehealth visit. Google's automated scanning flagged the images as child abuse material, disabled his account, and reported him to the authorities. The police investigated and found no crime. Google never restored the account. [10] Everything in that story was done by a system acting on suspicion, and the one human decision in it, the refusal to reinstate, was the only one that could not be appealed.

The second failure is the quiet penalty. Throttling and limiting are listed as penalties too, and the policy says nothing about whether a user is told when one is applied. [1] That matters more than it sounds, because of how AI is sold. Developers pay by the token. People on monthly plans pay a flat fee for a usage limit they never see in tokens at all. Either way, the number a user watches is not money, and decades of research on payment says that when the number is not money, people spend more and feel it less. Credit cards and casino chips work the same way. [11] Tokens are the unit because that is what computing time costs, so this is not a trick. But it has the same effect as one. A system that already keeps cost one step removed from the user can slow or shrink what that user gets while the fee stays the same, and the user may never see it.

This is not hypothetical. On April 23 of this year Anthropic published a postmortem on six weeks of worse output in Claude Code and two related products. The company had lowered the default amount of thinking the model does, to cut waiting time, at a small cost in quality. Two other product changes also hurt the output. The API, where customers pay per token, was not affected. Anthropic reverted the changes, said it never intentionally degrades its models, and reset everyone's usage limits. [12] It had said the same thing in September 2025, after a separate set of bugs. [13] This author takes the company at its word on the model. The point is what the user could see, which was nothing. For six weeks, people paying a flat monthly fee got less than they had been getting, and nothing on the screen told them whether it was a bug, a cost decision, or a penalty. The new policy makes the third possibility official and still does not say how a user would know.

Opinion 2. So the second level comes out mixed, and the fix is cheap. Anthropic would pass the full test by putting the appeal path in the policy itself, saying what a suspended user will be told, giving a timeline, even a long one, and stating plainly when a user is being slowed down. None of that requires a new policy. It requires four sentences that a person can check.

Opinion 3. There is a fair objection to all this, and it should be stated plainly. A company that shows too much gives away things it needs to keep. Detection methods, once published, teach the next actor how to avoid them. Details of the model are the product itself. And a company that speaks plainly about its own enforcement invites a round of criticism that a quieter company avoids. All of that is

true, and it is why the test matters. The test does not ask a company to open everything. It asks for the specific things the affected person needs to check: the standard, the reason, the date, the notice. None of those is a trade secret. A reason for a suspension does not reveal how the suspicion was formed. A notice that a user is being slowed does not reveal how the model works. The one place the objection really bites is the casebook, and Anthropic's own reports show the line can be drawn: the cases are public, the detection details mostly are not. That leaves the trade-off where it belongs, as a cost rather than a barrier. In a period of low inferred trust in Big Tech, this author's view is that the cost is one the frontier labs have to pay. It is the price of staying credible, and it is cheaper than the alternative, which is rules written by people who were given nothing to check.

Each level needs the same pair: a right to check and a limit on checking.

The government

At the third level, lawmakers and agencies infer whether the labs and their users can be trusted, and decide how soon to write rules. Here the picture is harder, because the checking runs both ways and both sides can fail. A lab can conceal. The spring quality problem came to light through user complaints, not through anything the company published first, [14] and California's new frontier AI law has to rely on whistleblowers to learn whether companies are following it, because the state cannot see inside. [15] A government can overreach. The same power that lets an agency demand an incident report lets it demand the conversations of every user or the model itself, and a lever like that is rarely used only for the purpose it was built for. And governments are users too. The usage policy allows tailored contracts for government customers whose missions the standard rules would block, so the body doing the checking at this level is also one of the parties being checked at the first. [1]

So the test at this level has two halves: what can a government check, and how much checking should it be allowed? A workable answer already exists in outline. California's SB 53 asks the largest developers to publish their safety frameworks, report serious incidents to the state within fifteen days, and protect employees who speak up. It asks for disclosure, not control, and it does not reach into user data or the model's inner workings. [16] Anthropic endorsed it, saying the law formalizes practices it already follows and lets developers compete while still disclosing the capabilities that pose risks to public safety. [17] That is checking scoped to what a government needs in order to write good rules, and no wider.

The high-risk half of Anthropic's announced policy update shows the relationship from the other side. It lists eleven areas, among them employment, credit, insurance, housing, education, and public benefits, and requires that a qualified person review the AI's recommendation and that the affected person be told AI was used. [1] That language closely parallels regulation. Colorado's 2026 replacement framework covers automated decision-making technology used to materially influence consequential decisions in education, employment, housing, financial or lending services, insurance, healthcare, and essential government services and public benefits. Its new requirements include

consumer notice, an explanation after an adverse decision, and a right to request meaningful human review and reconsideration. [18] The European Union's list adds migration, asylum, and the courts. So when Anthropic's announced update calls for human review and disclosure in employment decisions, it reflects the same concern lawmakers have put into law: the person affected needs a way to check the decision. That is the system working. The pace is the hard part. Colorado first delayed its earlier law to June 30, 2026. It then enacted SB26-189 on May 14, replacing that framework with new requirements beginning January 1, 2027. The implementing rules are still being developed. [19] Lawmakers are revising both the content of the rules and the time allowed to put them into practice.

Waiting has a cost of its own, and AT&T is the record of it. Anthropic, like most of Big Tech, stands at the crossroads of a utility and a for-profit company, and AT&T stood there for most of a century. It was private, it was essential, and for decades it answered mostly to itself. It held back inventions that threatened its business; a working answering machine built in its own labs in 1934 was shelved because the company feared that recorded calls would hurt the telephone. [20] The government made it a regulated monopoly in the 1910s and 1930s, which mostly protected it, and when the government finally changed course, the remedy was to break the company apart in 1984. A company at that crossroads is tempted to hold information back and shape the rest, and the longer the rules wait, the more they cost when they arrive.

The three together

Put the three levels together and the same pattern shows at each. Trust is no longer assumed. It is inferred, from evidence. At every level, the evidence sits with the lab: it sees what users do, it knows why it suspended someone, it knows what a regulator would need to see. And at every level, the checker can reach too far: a company watching its users more closely than it needs to, a public demanding the product itself, a government demanding the conversations. So each level needs the same pair, a right to check and a limit on checking.

Opinion 4. For the labs, the most useful thing is to show their work before they are made to: publish what they catch, write down the standard they use, give a reason and a date to the people they cut off, and say when they slow someone down. For lawmakers, it is to ask for what they need in order to write rules and confirm they are followed, and nothing wider, and to do it at the right pace: early enough to prevent the harms that can already be seen, not so early or so heavy that it prevents the thing itself. It is a hard balance. It is also the one on which AI's future depends.

References

- [1] Anthropic, Usage Policy, effective November 12, 2026. <https://www.anthropic.com/legal/aup>
- [2] Anthropic, Usage Policy, effective September 15, 2025, archived. <https://www.anthropic.com/legal/archive/22742366-2ef0-4c7a-a833-6523f10d3944>
- [3] Unite.AI, "Anthropic Details Disrupted Claude Misuse Across Seven Harm Areas," September 2026. <https://unite.ai/anthropic-details-disrupted-claude-misuse-across-seven-harm-areas>
- [4] Mobile World Live, "Anthropic flags AI weapons misuse," September 2026. <https://mobileworldlive.com/ai-cloud/anthropic-flags-ai-weapons-misuse>
- [5] Northeast Times, report on Anthropic's September 2026 threat intelligence report. <https://northeasttimes.com/?p=163202>
- [6] The Next Web, "Anthropic details how Claude was misused for surveillance and weapons," September 2026. <https://thenextweb.com/news/anthropic-claude-misuse-threat-intelligence-report>
- [7] Anthropic Support, "Trust and safety warnings and appeals." <https://support.claude.com/en/articles/8241253-trust-and-safety-warnings-and-appeals>
- [8] AI Weekly, "Anthropic gates Claude ban appeals behind a login-only form." <https://aiweekly.co/alerts/anthropic-gates-claude-ban-appeals-behind-a-login-only-form>
- [9] User bug reports, June 11 and July 1, 2026. <https://claudeissues.com/issue/67332-bug-account-suspended-without-clear-reason-appeal-form-broken-no-working-appeal> and <https://claudeissues.com/issue/72668-bug-appeal-page-is-inaccessible-for-suspended-pro-accounts-restricted-redirects>
- [10] Kashmir Hill, "A Dad Took Photos of His Naked Toddler for the Doctor. Google Flagged Him as a Criminal," The New York Times, August 21, 2022. <https://www.nytimes.com/2022/08/21/technology/google-surveillance-toddler-photo.html>
- [11] Priya Raghubir and Joydeep Srivastava, "Monopoly Money: The Effect of Payment Coupling and Form on Spending Behavior," Journal of Experimental Psychology: Applied 14, no. 3 (2008): 213-225; Drazen Prelec and Duncan Simester, "Always Leave Home Without It: A Further Investigation of the Credit-Card Effect on Willingness to Pay," Marketing Letters 12, no. 1 (2001): 5-12.
- [12] Anthropic, "An update on recent Claude Code quality reports," April 23, 2026. <https://www.anthropic.com/engineering/april-23-postmortem>
- [13] Anthropic status incident, "Model output quality," September 2025, as quoted at <https://simonwillison.net/2025/Sep/9/anthropic-model-output-quality/>
- [14] AI Catchup, "Anthropic's Claude Code Post-Mortem: Three Engineering Missteps Behind the Spring 2026 Quality Decline." <https://aicatchup.com/news/claude-code-quality-postmortem>
- [15] AI Safety Newsletter 63, "California's SB-53 Passes the Legislature," September 2025.

[AI Safety Newsletter](#)

[AI Safety Newsletter #63: California's SB-53 Passes the Legislature](#)

[Welcome to the AI Safety Newsletter by the Center for AI Safety. We discuss developments in AI and AI safety.](#)

[No technical background required...](#)

[16] Implicator, "California's SB 53 puts AI labs on the record."

<https://www.implicator.ai/californias-sb-53-puts-ai-labs-on-the-record/>

[17] Anthropic, "Anthropic is endorsing SB 53," September 8, 2025, as summarized at

<https://www.longtermwiki.com/resources/523ff672174f35cf>

[18] Colorado General Assembly, "SB26-189: Automated Decision-Making Technology," signed May 14, 2026.

<https://leg.colorado.gov/bills/sb26-189>

[19] Colorado General Assembly, "SB25B-004: Increase Transparency for Algorithmic Systems," approved August 28, 2025.

<https://leg.colorado.gov/bills/sb25b-004>; Colorado Attorney General, "Colorado Automated Decision-Making Technology & Chatbot Safety Rulemaking." <https://coag.gov/ai/>

[20] Tim Wu, *The Master Switch: The Rise and Fall of Information Empires* (New York: Knopf, 2010).

[21] John Elijah, "Data Centers Have Exposed Big Tech's Trust Deficit," Substack, October 2026.

Facts are sourced. Opinions are marked.

About the author

John Elijah is a pen name. The writer behind it has spent more than twenty years inside large institutions, public and private, and holds a graduate degree in organization development. He writes plain essays on institutions, merit, and machines.

© 2026 John Elijah. · [Read this essay online](#)