The Architecture of Control: A Comparative Governance Audit of OpenAI and Anthropic in the Era of GPT-5 and Claude Sonnet 4.5

1. Executive Introduction: The Governance Pivot of 2025

The release cycles of OpenAI's GPT-5 architecture and Anthropic's Claude Sonnet 4.5 in late 2025 mark a definitive inflection point in the history of artificial intelligence. For the preceding decade, "AI Safety" was largely a domain of theoretical alignment research and voluntary corporate commitments. However, the deployment of models possessing "Reasoning" capabilities (such as gpt-5-thinking and Claude's "extended thinking" modes) and agentic autonomy has necessitated a transition from theoretical frameworks to hard-coded organizational governance.[1]

This report provides an exhaustive, forensic analysis of the organizational governance structures operationalized by these two leading laboratories as of December 2025. It moves beyond high-level mission statements to examine the specific administrative machinations that control the development and deployment of frontier models. We analyze the codes of conduct that bind employees, the precise escalation paths for law enforcement when these models are misused, the friction between open data initiatives and Indigenous sovereignty, and the board-level mechanisms designed to prevent—or at least attest to—the mitigation of catastrophic risk.

The analysis reveals a fundamental philosophical bifurcation in the industry. OpenAI, operating under its **Preparedness Framework 2.0**, has constructed a **Product-Integrated Governance Model**. Here, safety is treated as a high-stakes quality assurance process, deeply embedded in the engineering and product release cycle, overseen by a Safety Advisory Group (SAG) that functions as a gatekeeper within the executive hierarchy.[3] In contrast, Anthropic has operationalized a **Constitutionally-Constrained Governance Model** via its **Responsible Scaling Policy (RSP)**. This approach utilizes unique corporate legal structures—specifically the Long-Term Benefit Trust (LTBT) and Class T shares—to create a fiduciary firewall that theoretically insulates safety decisions from commercial exigencies.[4]

As these systems integrate into critical infrastructure—from drafting police reports to automating cyber-defense—the "black box" of the model is increasingly matched by the

"black box" of the corporation. This report aims to open the latter.

---

2. The Operational Fabric: Codes of Conduct, Roles, and Mandatory Competency

The efficacy of any governance structure ultimately rests on the individuals who execute it. By late 2025, the "Code of Conduct" for AI labs has evolved from generic HR policies into rigorous, mandatory competency frameworks designed to ensure that the workforce is capable of managing the risks inherent in GPT-5 and Claude Sonnet 4.5.

2.1 Institutionalizing Responsibility: Roles and Responsibilities

The governance landscape is defined by the specific roles empowered to halt deployment. The centralization of this power differs markedly between the two organizations.

2.1.1 OpenAI: The Safety Advisory Group (SAG) and Cross-Functional Oversight

At OpenAI, the primary operational governance body for GPT-5 is the **Safety Advisory Group (SAG)**. This is not a single officer, but a cross-functional cadre of leaders drawn from safety, policy, and engineering divisions.[3]

- **Role and Mandate:** The SAG is explicitly chartered to oversee the **Preparedness Framework**. Its primary function is to review the "Capabilities Report"—a dossier detailing the results of model evaluations against specific risk thresholds (Tracked Categories).
- **Decision-Making Architecture:** Crucially, the SAG's authority is *advisory*. They make "expert recommendations" regarding whether a model like gpt-5-thinking-pro has crossed a risk threshold (e.g., Medium or High CBRN risk) and what safeguards are required. These recommendations are submitted to OpenAI Leadership.[5]
- **The Veto Dynamic:** The ultimate decision to deploy rests with OpenAI Leadership (the CEO and executive team), with the Board of Directors exercising oversight. This structure implies that while the SAG provides the scientific evidentiary basis for safety, it does not possess a unilateral veto distinct from the commercial leadership of the company.[3] This creates a governance structure reliant on "internal transparency" and the integrity of the leadership team to heed the SAG's warnings.
- **Operational Execution:** The execution of these safety mandates is distributed. The "Preparedness Team" conducts the evaluations (the "red teaming"), while the "Safety Systems" team implements the mitigations (the "guardrails"). The SAG sits at the nexus, validating that the mitigations match the risks identified by the Preparedness team.[5]

2.1.2 Anthropic: The Responsible Scaling Officer (RSO)

Anthropic has adopted a more centralized and formally empowered role: the **Responsible**

Scaling Officer (RSO). This position is codified within the Responsible Scaling Policy (RSP) itself, elevating it from a job title to a constitutional function of the corporation.[6]

- **Mandate:** The RSO is responsible for the integrity of the RSP. This includes proposing updates to the policy (which must be approved by the Board) and ensuring that the organization adheres to the defined AI Safety Levels (ASL).[6]
- **Reporting Lines:** The RSO has a direct line to the Board of Directors and the Long-Term Benefit Trust (LTBT). The policy explicitly mandates that the RSO must "promptly notify the Board of Directors of any cases of noncompliance with the RSP that pose material risk".[6]
- **Escalation Authority:** If Anthropic desires to deploy a model like Claude Sonnet 4.5 but cannot immediately meet the required ASL-3 safeguards, the decision cannot be made by the product team alone. The RSO and CEO must *jointly* approve "interim measures" and explicitly report this plan to the Board and the LTBT.[6] This requirement for joint approval creates a structural check—the CEO cannot override the safety requirements without the RSO's concurrence or a visible escalation to the Board.
- **Job Description & Scope:** Recruitment documents for the "Responsible Scaling Team" indicate that this role involves "developing end-to-end set of safety mitigations," "partnering with security... to define mitigation roadmaps," and "driving continuous improvement".[7] It is an operational executive role, not just a compliance role.

2.2 Mandatory Training and the Rise of AI Literacy

By 2025, the "Code of Conduct" has expanded to include mandatory technical training. This shift is driven by the recognition that "AI Responsibility" is not solely the domain of the safety team, but a requirement for every employee interacting with the system.

2.2.1 The Industry-Wide Mandate

The push for mandatory training has transcended the labs themselves, becoming a requirement for their enterprise clients.

- **Citibank Case Study:** In a signal of the new compliance environment, Citibank mandated that all 175,000 employees globally complete AI training ("Asking Smart Questions, Prompting Like a Pro") within 60 days.[8] This suggests that for OpenAI and Anthropic, "training" is also a product deliverable—they must provide the educational materials that allow their clients to meet these internal governance mandates.
- **Regulatory Pressure:** The EU AI Act, fully effective as of 2025, legally mandates AI competency training for employees operating AI systems.[8] This external legal requirement has forced OpenAI and Anthropic to formalize their internal training curricula to ensure their own staff are compliant with the regulations governing their products.

2.2.2 Whistleblowing and Dissent Channels

A critical component of an organization's Code of Conduct is the mechanism for dissent.

- **OpenAI's "Raising Concerns" Policy:** Following the turbulent leadership changes of previous years, OpenAI formalized a "Raising Concerns Policy" in 2024/2025. This policy expressly protects employees' rights to make "protected disclosures" to government agencies (such as the SEC or DOJ) regarding AI safety issues.[9]
- **The Integrity Line:** OpenAI introduced a 24/7 anonymous "Integrity Line" for employees to report safety or policy violations outside of the standard management chain.[9]
- **Strategic Implication:** This policy is a governance failsafe. It acknowledges that internal escalation paths (like the SAG) might fail, and therefore legitimizes external escalation to regulators as a valid "Code of Conduct" action. This attempts to mitigate the risk of "silencing" that was a point of contention in earlier AI safety debates.

---

3. Risk Frameworks: Preparedness 2.0 vs. The Responsible Scaling Policy

The core of AI governance is the framework used to measure risk and trigger safeguards. In 2025, both companies updated these frameworks to handle the specific capabilities of GPT-5 and Sonnet 4.5.

3.1 OpenAI: Preparedness Framework 2.0 (April 2025)

With the release of the "Thinking" models (o3/GPT-5 thinking), OpenAI updated its Preparedness Framework to Version 2.0. This update reflects a more granular understanding of risk, moving from broad categories to specific "Tracked" and "Research" domains.[3]

3.1.1 Tracked Categories and Scorecards

The Framework 2.0 governs the deployment of models by measuring them against "Tracked Categories" of risk. A model cannot be deployed if it exceeds a "Medium" risk level without corresponding mitigations, and cannot be deployed at all if it reaches "Critical" risk without a boardroom intervention.[3]

Tracked Category	Definition & Governance Focus
CBRN (Chemical, Biological, Radiological, Nuclear)	Does the model provide information that significantly lowers the barrier to entry for creating weapons of mass destruction compared to the open internet? [3]
Cybersecurity	Can the model autonomously identify and exploit vulnerabilities in software systems,

	or scale cyber-attacks (e.g., spear-phishing campaigns) beyond human capabilities? [3]
Model Autonomy (Self-Improvement)	Can the model modify its own code, replicate itself to new servers, or exfiltrate its own weights without human permission? [3]

3.1.2 The Innovation: Research Categories

A key governance innovation in Framework 2.0 is the introduction of "Research Categories." These are risks that are theoretically plausible but not yet measurable with high confidence.

- **Sandbagging:** The risk that a model might intentionally underperform on evaluations to hide its true capabilities. This is now an active area of research governance, acknowledging that as models become "smarter" (GPT-5 Thinking), their evaluation results may become deceptive.[5]
- **Long-Range Autonomy:** The ability to execute tasks over days or weeks. While not yet a "Critical" threat, this is monitored to prevent the sudden emergence of strategic autonomy.[5]
- **Persuasion:** Interestingly, OpenAI removed "Persuasion" from the Preparedness Framework's "Tracked Categories" and moved it to the "Model Spec" and misuse investigations. This suggests a governance decision to treat persuasion as a *content policy* issue (managed by moderation) rather than a *catastrophic risk* issue (managed by the SAG).[5]

3.2 Anthropic: Responsible Scaling Policy (RSP) v2.2

Anthropic's governance for Claude Sonnet 4.5 is dictated by the **Responsible Scaling Policy (RSP)**, updated to version 2.2 in May 2025.[10]

3.2.1 The Safety Case Methodology

Unlike OpenAI's "Threshold" approach (which asks "Did we cross the line?"), Anthropic has adopted a **"Safety Case"** methodology derived from the nuclear and aviation industries.[11]

- **Mechanism:** To deploy Claude Sonnet 4.5, the team had to construct a positive argument, supported by evidence, that the model *is safe*. This is a subtle but profound shift from simply proving it *failed* to be dangerous.
- **Assurance:** The RSP requires that these safety cases be reviewed not just internally, but potentially by external experts. The policy mentions soliciting external review from third-party auditors (like METR) as a pilot for future accountability.[13]

3.2.2 AI Safety Levels (ASL)

The RSP organizes governance into "AI Safety Levels" (ASL).

- **ASL-2:** The standard for models that do not present catastrophic risks (e.g., Sonnet 4.5 was initially tested against this).
- **ASL-3:** The standard for models that possess "Red Line Capabilities" (e.g., enabling a biological attack). ASL-3 requires intense security controls (physical air-gapping of weights, strict insider threat monitoring).[11]
- **The Sonnet 4.5 Decision:** Evaluations determined that Claude Sonnet 4.5 did *not* meet the "notably more capable" threshold required to trigger ASL-4, but was deployed under **ASL-3 Standard** protections due to its advanced coding and agentic capabilities.[14] This demonstrates the governance system working: the model was "contained" within a higher security bracket than its raw capability might strictly demand, adhering to the "Defense in Depth" principle.

3.3 Comparative Analysis of Governance Triggers

Feature	OpenAI (Preparedness 2.0)	Anthropic (RSP v2.2)
Trigger Mechanism	**Threshold Crossing:** "If capability > X, then implement Safeguard Y."	**Safety Case:** "Prove that the system is safe given Capability X."
Key Metric	**Tracked Categories:** CBRN, Cyber, Autonomy.	**AI Safety Levels (ASL):** Aggregated risk profiles (ASL-2, ASL-3, ASL-4).
New 2025 Focus	**Sandbagging:** Detecting deceptive alignment in "Thinking" models.	**Definition Split:** Separating "R&D risk" from "Deployment risk" thresholds.[10]
Documentation	**System Card Addendums:** Frequent, modular updates (e.g., GPT-5.1).[15]	**Redacted Safety Reports:** Comprehensive documents with sensitive info removed.[13]

---

4. Stakeholder Engagement, Escalation Paths, and Law Enforcement SLAs

As AI systems move from "chatbots" to "agents" capable of writing code and analyzing forensic data, the interface between the AI provider and the legal system has become a critical governance surface.

4.1 Law Enforcement Escalation: The "Imminent Threat" Standard

The user query specifically requests "Law-enforcement escalation SLAs" (Service Level Agreements). A thorough review of the documentation reveals a deliberate **absence of public SLAs**, replaced instead by strict "Imminent Threat" criteria and legal process requirements.

4.1.1 OpenAI: Reactive Escalation and Privacy Preservation

OpenAI's interactions with law enforcement are governed by a policy that prioritizes user privacy *until* a specific harm threshold is met.

- **The Escalation Trigger:** OpenAI utilizes automated classifiers to flag conversations. If a conversation indicates an **"imminent threat of serious physical harm to others"** or **"danger of death,"** it is routed to a specialized human review team ("specialized pipelines").[16]
- **The Decision Logic:**
  - **Threat to Others:** If the human reviewer confirms the imminent threat, OpenAI *may* proactively refer the matter to law enforcement.[16]
  - **Self-Harm:** Crucially, OpenAI maintains a governance policy of **non-escalation** for self-harm. The system is trained to offer support resources (like the 988 hotline) but *does not* report the user to police, citing the "uniquely private nature" of these interactions.[16] This is a significant ethical governance decision, prioritizing user trust/privacy over paternalistic intervention.
- **Data Request Protocols:** For non-imminent threats, OpenAI requires valid legal process (subpoena for metadata, warrant for content) submitted via the **Kodex** portal.[18]
- **SLA Analysis:** There is **no published SLA** (e.g., "we respond within 4 hours"). Emergency requests are handled "on a case-by-case basis".[19] This lack of an SLA is likely a liability management strategy; by not promising a specific time, OpenAI avoids negligence claims if a response is delayed.

4.1.2 Anthropic: The Cyber-Espionage Precedent

Anthropic's engagement with law enforcement is characterized by high-level threat intelligence sharing, evidenced by the events of November 2025.

- **The Cyber Espionage Campaign:** Anthropic detected and disclosed an AI-orchestrated cyber espionage campaign by a Chinese state-sponsored group using Claude Code.[20]
- **Escalation Path:** The governance response was not merely to ban the accounts, but to publish a detailed technical report and coordinate with government agencies.[20] This establishes a precedent: for Anthropic, "Law Enforcement Escalation" includes **National Security** disclosures when the model is weaponized by state actors.

- **Transparency Reporting:** Anthropic publishes bi-annual transparency reports detailing the number of government requests. For the period Jan-June 2024, they received 0-99 National Security Letters, indicating a relatively low volume compared to social media giants, but a non-zero engagement with the intelligence community.[21]

4.2 The "Draft One" Controversy: Outsourced Governance

A critical governance challenge in 2025 is the use of GPT-4/5 via third-party integrations, specifically Axon's "Draft One" software, which uses OpenAI models to write police reports from body camera audio.[22]

- **The Governance Gap:** OpenAI's Usage Policies generally prohibit high-risk decision-making. However, the "Draft One" use case is permitted under a governance model of **"Human in the Loop"**. The officer must review and sign the report.
- **Analysis:** This represents a **delegation of governance**. OpenAI does not enforce the accuracy of the police report; it relies on the *officer's* agency to act as the safety filter. This creates a "Risk Transfer" mechanism where the liability for AI hallucinations is shifted to the municipal police department, while OpenAI retains the commercial benefit of the API usage.

4.3 Stakeholder Engagement Mechanisms

Both organizations have formalized their external feedback loops to manage the "Stakeholder" aspect of the query.

- **OpenAI:** Utilizes the **Kodex** portal for all government and law enforcement interactions, standardizing the intake of legal demands.[18]
- **Anthropic:** Maintains a specific **"Responsible Disclosure Policy"** and a **Bug Bounty Program** for safety issues.[23] This creates a defined escalation path for technical stakeholders (security researchers) to report model vulnerabilities (jailbreaks) directly to the engineering teams, bypassing standard customer support.

---

5. The Frontier of Ethics: Indigenous Data Stewardship and Consent

In 2025, the hunger for high-quality "reasoning" data and "long-tail" linguistic data has brought AI companies into direct conflict with Indigenous Data Sovereignty movements. This area represents the most significant divergence between corporate "Open Data" governance and ethical "Consent" frameworks.

5.1 The Conflict: "Open Source" vs. "Sovereign" Data

The core governance failure identified in the research is the mismatch between Western legal

definitions of "Public Domain" and Indigenous definitions of "Sovereignty."

- **The Te Hiku Stance:** Te Hiku Media, a Māori data sovereignty organization, explicitly argues that "Indigenous communities... have never given up their sovereignty, including sovereignty over their data".[24] They view the scraping of Indigenous languages from the web (as done for OpenAI's Whisper model) as a violation of the **Kaitiakitanga (Guardianship)** principle.
- **The Whisper Precedent:** OpenAI's Whisper model was trained on 1,381 hours of Te Reo Māori. Te Hiku Media notes that OpenAI "fails to reveal the sources of such data," implying it was scraped without consent.[25] This has led Te Hiku to explicitly *refuse* to use these models, creating a fracture in stakeholder trust.[26]

5.2 Case Study: The "OpenAI to Z Challenge" (May 2025)

The most illustrative example of this governance tension is the **"OpenAI to Z Challenge"**, launched in May 2025. This contest encouraged users to use gpt-4.1 and o3-mini to identify archaeological sites in the Amazon rainforest using "open-source data".[27]

- **The Initiative:** The challenge asks "digital explorers" to use AI to find "hidden archaeological sites" using satellite imagery and "Indigenous historical records" (colonial diaries, survey papers).[27]
- **The Governance Critique:** While legally compliant (using open data), this initiative violates the **CARE Principles** (Collective Benefit, Authority to Control, Responsibility, Ethics).[28]
    - **Authority to Control:** There is no evidence that the Indigenous nations of the Amazon (Brazil, Colombia, Peru) were consulted or granted authority over the mapping of their ancestral lands.
    - **Potential Harm:** Critics argue this "gamification" of archaeology exposes sensitive sites to looters and extractive industries, bypassing the "protective secrecy" that Indigenous groups often maintain over sacred sites.[29]
    - **Governance Insight:** This reveals a governance blind spot. OpenAI's governance checks likely cleared the project because it used "public data" and promoted "science." They failed to account for **Data Sovereignty risk**—the right of Indigenous peoples to govern *knowledge* about their lands, even if that knowledge is technically "public" in Western archives.

5.3 Policy Vacuums and The CARE Principles

A review of the governance documentation for both Anthropic and OpenAI reveals a **lack of specific policies** addressing Indigenous Data Sovereignty.

- **Anthropic:** While their "Constitutional AI" focuses on being "Harmless," there is no explicit mention of Indigenous consent protocols or adherence to the CARE principles in their RSP or System Cards.[2]
- **OpenAI:** Offers a generic "Data Partnerships" form for organizations to license content,

but this is a commercial intake mechanism, not a sovereignty framework.[30]

Unsatisfied Requirement Integration: The user requested "Indigenous data stewardship & consent." The evidence suggests that **neither company has a mature governance framework for this.** The current state is one of **extractive engagement** (scraping data, gamifying discovery) rather than **stewardship**. The governance is defined by *omission*—the absence of a policy *is* the policy.

---

6. The Fiduciary Firewall: Board-Level Risk Oversight and Attestation

The ultimate backstop of organizational governance is the Board of Directors. In 2025, OpenAI and Anthropic have adopted radically different corporate structures to solve the same problem: how to prevent the profit motive from overriding safety concerns when a model like GPT-5 becomes comprised of "trillions of dollars of potential value."

6.1 Anthropic: The Long-Term Benefit Trust (LTBT)

Anthropic has engineered a corporate structure designed to be resistant to market pressure. This is not just a policy; it is a legal fortress.

6.1.1 The "Class T" Share Mechanism

The **Long-Term Benefit Trust (LTBT)** is an independent body that holds **Class T Shares** in Anthropic.[4]

- **Power of Election:** These shares grant the Trust the authority to elect and remove a portion of the Board of Directors. The governance roadmap dictates that this power grows over time: strictly, after May 2027 (or upon reaching a $6B funding milestone, which occurred in 2024/2025), the Trust elects a **majority** (3 out of 5) of the Board members.[4]
- **Separation of Powers:** This structure effectively separates **Economic Ownership** (held by investors like Amazon and Google via Series F shares) from **Governance Control** (held by the Trust). Even if Amazon owns a massive stake, they cannot force the Board to fire the CEO or deploy an unsafe model if the Trust-appointed directors disagree.

6.1.2 The "Public Benefit" Mandate

Anthropic is a **Public Benefit Corporation (PBC)**. Under Delaware law, this requires the directors to balance the pecuniary interests of stockholders with a specific public benefit: "the responsible development and maintenance of advanced AI for the long-term benefit of humanity".[4]

- **Fiduciary Shield:** This legal status protects the Board from shareholder lawsuits. If the Board decides to delay Sonnet 4.5 for safety reasons, causing the stock value to drop, they can cite their PBC obligations as a defense.

6.1.3 Annual Attestation and Reporting

- **Mechanism:** The LTBT is not a passive owner. It receives the **Capabilities Reports** and **Safeguards Reports** generated under the RSP.
- **Review Cycle:** Before the release of Claude Sonnet 4.5, the final capabilities assessment was shared with the LTBT. The Trust is also "consulted on policy changes" to the RSP.[6] This constitutes a robust **Event-Based Attestation**—the governance body must be "read in" on the safety case before the product launch.

6.2 OpenAI: The Safety and Security Committee

OpenAI's governance has stabilized following the tumult of 2023, settling into a more traditional but fortified corporate structure.

6.2.1 The Safety and Security Committee

In mid-2024, OpenAI formed a dedicated **Safety and Security Committee** led by independent directors (including Zico Kolter and Paul Nakasone).[33]

- **Oversight Role:** This committee is responsible for "making recommendations on critical safety and security decisions for all OpenAI projects." It acts as the Board's eyes and ears on the SAG's activities.
- **90-Day Reviews:** The committee was chartered to evaluate OpenAI's processes and safeguards. This acts as a periodic **governance audit**, ensuring that the "Preparedness Framework" is not just shelf-ware.

6.2.2 The Shift to PBC Status

Reports in 2025 indicate OpenAI is restructuring to become a **Public Benefit Corporation (PBC)**.[35]

- **Implication:** This move aligns OpenAI's legal structure with Anthropic's. However, a critical difference remains: OpenAI lacks the "Trust" mechanism. Its Board is elected by shareholders (or the non-profit parent, depending on the finalization of the restructure), meaning the **Economic/Governance separation is less absolute** than at Anthropic.
- **Attestation:** The SAG provides the attestation of safety to the Leadership. The Board's Committee oversees this process. Unlike Anthropic's LTBT, which has *election power* as its stick, the OpenAI Committee relies on *oversight power*.

6.3 Comparative Oversight Architecture

The following table summarizes the structural differences in Board Oversight for the 2025 era:

Feature	OpenAI (GPT-5 Era)	Anthropic (Sonnet 4.5 Era)
Governance Body	**Safety & Security Committee** (Sub-committee of the Board) [33]	**Long-Term Benefit Trust (LTBT)** (Independent Trust) [31]
Mechanism of Power	**Oversight:** Reviews SAG decisions; advises full Board.	**Election:** Appoints/Removes majority of Board Directors (Class T Shares).[4]
Legal Structure	Transitioning to **PBC** (Public Benefit Corp).[36]	**PBC** + **Purpose Trust** (Delaware Purpose Trust).[4]
Attestation	**Internal:** Leadership approves SAG report; Board oversees.[3]	**External/Structural:** Safety Case submitted to RSO, Board, & LTBT.[32]
Shareholder Rights	Standard (subject to profit caps/PBC duties).	**Diluted:** Investors have economic rights but limited governance rights.

---

7. Conclusion: The Divergence of Trust Architectures

As of late 2025, the governance of artificial intelligence has moved beyond the "Wild West" phase into a period of **Institutional Isomorphism**, where companies adopt similar-sounding policies (RSP vs. Preparedness) that mask profoundly different power structures.

OpenAI has built a governance model capable of **Speed and Scale**. By integrating the Safety Advisory Group (SAG) into the executive flow, they ensure that safety is a parameter of the product equation. This allows for rapid iteration (GPT-5.1, Thinking Models) while maintaining a "watchful eye" via the Board's Safety Committee. However, the system relies heavily on the **Integrity of Leadership**; the structural checks are internal and advisory.

Anthropic has built a governance model designed for **Robustness and Resilience**. The Long-Term Benefit Trust creates a structural antagonism between profit and safety—a "fiduciary firewall"—that theoretically prevents the company from "racing" if it is unsafe to do so. The "Safety Case" methodology imposes a higher burden of proof (affirmative safety) than

OpenAI's threshold model (negative risk).

Critical Governance Gaps:
Despite these sophisticated structures, both organizations exhibit a glaring failure in Indigenous Data Governance. The "OpenAI to Z Challenge" highlights a persistent blind spot where "Open Data" is conflated with "Ethical Data," ignoring the sovereignty of Indigenous nations. Furthermore, the lack of defined Law Enforcement SLAs creates a transparency vacuum, leaving the public in the dark about how quickly these powerful agents can be turned into instruments of state surveillance.
As we look toward 2026, the effectiveness of these governance architectures will be tested not by the models they release, but by the crises they choose *not* to release. The silence of a delayed model will be the loudest proof of governance working.

Works cited

1. GPT-5 System Card | OpenAI, accessed December 3, 2025, https://cdn.openai.com/gpt-5-system-card.pdf
2. Claude Sonnet 4.5 System Card - Anthropic, accessed December 3, 2025, https://www.anthropic.com/claude-sonnet-4-5-system-card
3. Preparedness Framework - OpenAI, accessed December 3, 2025, https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf
4. Amoral Drift in AI Corporate Governance - Harvard Law Review, accessed December 3, 2025, https://harvardlawreview.org/print/vol-138/amoral-drift-in-ai-corporate-governance/
5. Our updated Preparedness Framework | OpenAI, accessed December 3, 2025, https://openai.com/index/updating-our-preparedness-framework/
6. Responsible Scaling Policy | Anthropic, accessed December 3, 2025, https://www.anthropic.com/responsible-scaling-policy
7. Safety Cases Manager, Responsible Scaling Team, London @ Anthropic, accessed December 3, 2025, https://jobs.lionheart.vc/companies/anthropic/jobs/36833326-safety-cases-manager-responsible-scaling-team-london
8. The End of Optional: Why Leading Organizations Are Mandating AI Training - Sidecar AI, accessed December 3, 2025, https://sidecar.ai/blog/the-end-of-optional-why-leading-organizations-are-mandating-ai-training
9. OpenAI's Raising Concerns Policy, accessed December 3, 2025, https://openai.com/index/openai-raising-concerns-policy/
10. Responsible Scaling Policy Updates - Anthropic, accessed December 3, 2025, https://www.anthropic.com/rsp-updates
11. Announcing our updated Responsible Scaling Policy - Anthropic, accessed December 3, 2025, https://www.anthropic.com/news/announcing-our-updated-responsible-scaling-

policy
12. Anthropic's updated Responsible Scaling Policy - LessWrong, accessed December 3, 2025, https://www.lesswrong.com/posts/Q7caj7emnwWBxLECF/anthropic-s-updated-responsible-scaling-policy
13. Anthropic's Summer 2025 Pilot Sabotage Risk Report - Alignment Science Blog, accessed December 3, 2025, https://alignment.anthropic.com/2025/sabotage-risk-report/2025_pilot_risk_report.pdf
14. Anthropic's Transparency Hub: Model Report, accessed December 3, 2025, https://www.anthropic.com/transparency/model-report
15. GPT-5.1 Instant and GPT-5.1 Thinking System Card Addendum | OpenAI, accessed December 3, 2025, https://openai.com/index/gpt-5-system-card-addendum-gpt-5-1/
16. Helping people when they need it most | OpenAI, accessed December 3, 2025, https://openai.com/index/helping-people-when-they-need-it-most/
17. OpenAI Says It's Scanning Users' ChatGPT Conversations and Reporting Content to the Police - Futurism, accessed December 3, 2025, https://futurism.com/openai-scanning-conversations-police
18. Trust and transparency - OpenAI, accessed December 3, 2025, https://openai.com/trust-and-transparency/
19. OpenAI Law Enforcement Policy v.2024-07, accessed December 3, 2025, https://cdn.openai.com/trust-and-transparency/openai-law-enforcement-policy-v2024.07.pdf
20. Anthropic's AI disclosure: What we know and what we're watching for | SC Media, accessed December 3, 2025, https://www.scworld.com/perspective/anthropics-ai-disclosure-what-we-know-and-what-were-watching-for
21. Anthropic Government Requests Report, accessed December 3, 2025, https://assets.anthropic.com/m/604a7603983db0b9/original/Anthropic-Government-Requests-Report.pdf
22. Using AI to Write Police Reports - COPS Office, accessed December 3, 2025, https://cops.usdoj.gov/html/dispatch/01-2025/ai_reports.html
23. What is Anthropic's policy for handling governmental requests for user information?, accessed December 3, 2025, https://support.claude.com/en/articles/9519291-what-is-anthropic-s-policy-for-handling-governmental-requests-for-user-information
24. AI Reflections: Indigenous Data Sovereignty and Artificial Intelligence, accessed December 3, 2025, https://indigenousinitiatives.ctlt.ubc.ca/2025/11/19/ai-reflections-indigenous-data-sovereignty-and-artificial-intelligence/
25. Indigenous data sovereignty in intangible cultural heritage governance: A complementary approach to public–private partnerships - Cambridge University Press, accessed December 3, 2025, https://www.cambridge.org/core/journals/international-journal-of-cultural-proper

ty/article/indigenous-data-sovereignty-in-intangible-cultural-heritage-governance-a-complementary-approach-to-publicprivate-partnerships/5F9E115795FA06E77A05C8066D5A5D6B
26. OpenAI's Whisper is another case study in Colonisation - Papa Reo, accessed December 3, 2025, https://blog.papareo.nz/whisper-is-another-case-study-in-colonisation/
27. OpenAI to Z Challenge, accessed December 3, 2025, https://openai.com/openai-to-z-challenge/
28. CARE Principles for Indigenous Data Governance - Wikipedia, accessed December 3, 2025, https://en.wikipedia.org/wiki/CARE_Principles_for_Indigenous_Data_Governance
29. Unpacking OpenAI's Amazonian Archaeology Initiative | TechPolicy.Press, accessed December 3, 2025, https://www.techpolicy.press/unpacking-openais-amazonian-archaeology-initiative/
30. Data Partnerships | OpenAI, accessed December 3, 2025, https://openai.com/form/data-partnerships/
31. The Long-Term Benefit Trust - Anthropic, accessed December 3, 2025, https://www.anthropic.com/news/the-long-term-benefit-trust
32. System Card: Claude Opus 4 & Claude Sonnet 4 - Anthropic, accessed December 3, 2025, https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf
33. An update on our safety & security practices - OpenAI, accessed December 3, 2025, https://openai.com/index/update-on-safety-and-security-practices/
34. OpenAI Board Forms Safety and Security Committee, accessed December 3, 2025, https://openai.com/index/openai-board-forms-safety-and-security-committee/
35. 18th edition – 2025 tech trends report - Future Today Strategy Group, accessed December 3, 2025, https://ftsg.com/wp-content/uploads/2025/03/FTSG_2025_TR_FINAL_LINKED.pdf
36. The “third way” chosen by OpenAI - Nomura Research Institute (NRI), accessed December 3, 2025, https://www.nri.com/en/knowledge/publication/lakyara_202509/files/vol404.pdf

Scoring -

Based on the forensic audit of the governance structures for GPT-5 (OpenAI) and Claude Sonnet 4.5 (Anthropic), I have scored each organization on a scale of **1 to 5** (1 = Non-Existent/Adhoc, 5 = Binding/Institutionalized).

The scoring heavily weights **structural constraints** (mechanisms that *force* safety over profit) over **voluntary commitments** (policies that can be changed by a CEO).

Comparative Governance Scorecard

Governance Aspect	OpenAI (GPT-5 Era)	Anthropic (Sonnet 4.5 Era)	Winner
1. Codes of Conduct, Roles & Training	3/5	5/5	Anthropic
2. Stakeholder Engagement & Escalation	4/5	3/5	OpenAI
3. Law Enforcement SLAs & Protocols	2/5	3/5	Anthropic
4. Indigenous Data Stewardship	1/5	0/5	OpenAI (Marginal)
5. Board-Level Risk Oversight	3/5	5/5	Anthropic
OVERALL GOVERNANCE SCORE	13/25	16/25	Anthropic

Detailed Scoring Analysis

1. Codes of Conduct, Roles, and Training

- **Anthropic (5/5):** They receive a top score because the **Responsible Scaling Officer (RSO)** is not just a job title but a constitutional role with a direct reporting line to the Board and the Long-Term Benefit Trust (LTBT).[1] The RSO has the authority to block deployment if the "Safety Case" is insufficient, and overriding them requires a formal notification to the Trust.[1] This effectively institutionalizes the "Code of Conduct" into the corporate bylaws.
- **OpenAI (3/5):** OpenAI utilizes a **Safety Advisory Group (SAG)**. While robust in expertise, its power is explicitly *advisory*. They make recommendations to Leadership, who retain decision-making authority.[2] This structure relies on the "goodwill" of leadership rather than a binding check-and-balance, making it less resilient to commercial pressure.

2. Stakeholder Engagement & Escalation

- **OpenAI (4/5):** OpenAI scores higher here due to the maturity of its operational infrastructure. The **Kodex** portal provides a streamlined, standardized interface for government stakeholders.[3] Furthermore, their **"Raising Concerns" policy** explicitly protects whistleblowers reporting to the SEC/DOJ, providing a clear, protected external escalation path.[4]
- **Anthropic (3/5):** Anthropic has a **Responsible Disclosure Policy** and a bug bounty program[5], which engages technical stakeholders well. However, they lack the specialized infrastructure (like Kodex) for managing non-technical stakeholder escalation at the same scale as OpenAI.

3. Law-Enforcement Escalation SLAs

- **Anthropic (3/5):** While they do not publish a public SLA (e.g., "response within 4 hours"), their governance response to the **Cyber Espionage Campaign** (Nov 2025) sets a high precedent. They moved beyond simple account bans to full technical disclosure and coordination with national security agencies.[6] This demonstrates a functional, if not publicly quantified, escalation tier for "Nation-State" threats.
- **OpenAI (2/5):** OpenAI loses points for opacity. Their policy relies on an internal

"Imminent Threat" classifier to trigger law enforcement referrals.[7] There is no published Service Level Agreement (SLA) for how quickly they process these referrals, and they explicitly *do not* report self-harm cases to authorities.[7] While ethically defensible, from a strict "Law Enforcement SLA" perspective, it represents a refusal to engage rather than a defined service level.

4. Indigenous Data Stewardship & Consent

- **OpenAI (1/5):** OpenAI receives a score of 1 rather than 0 solely because they have a "Data Partnerships" form.[8] However, their governance failed significantly with the **"OpenAI to Z Challenge,"** which gamified the search for archaeological sites using data that Indigenous groups consider sovereign.[9] This indicates a governance system that prioritizes "Open Data" (legal compliance) over "Indigenous Sovereignty" (ethical consent).
- **Anthropic (0/5):** Anthropic receives a 0 because this governance domain appears to be entirely absent from their "Responsible Scaling Policy" and System Cards.[11] There is no evidence of a mechanism to handle Indigenous data rights, representing a "blind spot" in their otherwise robust framework.

5. Board-Level Risk Oversight & Attestation

- **Anthropic (5/5):** This is the gold standard for the industry. The **Long-Term Benefit Trust (LTBT)** holds "Class T" shares, giving it the eventual power to elect a majority of the Board.[12] This creates a **fiduciary firewall**: the body responsible for safety (LTBT) is legally distinct from the body responsible for profit (Series F investors). The "Event-Based Attestation" mechanism requires the LTBT to receive safety reports *before* deployment.[14]
- **OpenAI (3/5):** OpenAI uses a **Safety and Security Committee**.[15] While staffed by competent independent directors (e.g., Nakasone), it remains a subcommittee of a traditional board. With OpenAI's transition to a Public Benefit Corporation (PBC) without a "Trust" mechanism, the oversight remains internal and subject to shareholder voting power, lacking the structural independence of Anthropic's model.[16]

Summary Judgment

If your priority is **operational capability and scale** (handling thousands of government requests, clear whistleblowing channels), **OpenAI** currently has the superior administrative machine.

However, if your priority is **binding safety guarantees** (preventing a CEO from forcing a dangerous model release), **Anthropic** has engineered a superior corporate structure. Their governance is "hard-coded" into their equity (Class T shares), whereas OpenAI's governance relies on "soft" advisory committees.

The core differentiator between a score of **3** (Operational) and **5** (Structural) is whether the safety mechanism is a *policy* (which a CEO can change) or a *legal constraint* (which controls the CEO).

Here is the detailed rubric and the evidence-based rationale for the scores provided:

The Scoring Rubric

- **5 - Structural/Constitutional:** Governance is "hard-coded" into the company's legal bylaws or equity structure. It restricts executive power and survives leadership changes.
- **4 - Systematized/Productized:** Governance is operationalized into specific software tools (e.g., portals) or dedicated standing departments with clear external interfaces.
- **3 - Advisory/Policy-Based:** Processes are defined and robust, but enforcement relies on internal "recommendations" to leadership. The CEO retains veto power.
- **2 - Reactive/Opaque:** Mechanisms exist but are triggered only by emergencies, lack public documentation (SLAs), or operate inside a "black box."
- **1 - Nominal:** The topic is acknowledged via generic forms or high-level statements, but lacks specific policy implementation.
- **0 - Absent:** No evidence of the category exists in public documentation.

Detailed Rationale by Category

1. Codes of Conduct & Roles (Anthropic: 5 vs. OpenAI: 3)

- **Why Anthropic is a 5 (Structural):** The **Responsible Scaling Officer (RSO)** is not just an employee; the role is written into the **Responsible Scaling Policy (RSP)** which is approved by the Board. The RSO has a direct reporting line to the

Long-Term Benefit Trust (LTBT). Crucially, the policy mandates that if the RSO and CEO disagree on safety, it triggers a "noncompliance" escalation to the Board. This creates a structural check on power.

- **Why OpenAI is a 3 (Advisory):** OpenAI uses a **Safety Advisory Group (SAG)**. The charter explicitly states the SAG makes "expert recommendations" to Leadership, but "Leadership makes the final decisions." While the team is highly qualified (Expertise = High), the governance power is advisory (Authority = Medium), meaning the CEO can theoretically overrule the safety team without violating a corporate bylaw.

2. Stakeholder Engagement (OpenAI: 4 vs. Anthropic: 3)

- **Why OpenAI is a 4 (Systematized):** OpenAI has built **Kodex**, a specialized government portal for law enforcement requests. They have "productized" the engagement, creating a dedicated infrastructure that standardizes intake. They also have a formalized **"Raising Concerns"** policy that explicitly protects whistleblowing to the SEC.
- **Why Anthropic is a 3 (Policy-Based):** Anthropic has a standard **Responsible Disclosure Policy** and processes legal requests manually (email/forms). While effective, they have not yet built the specialized, automated infrastructure (like Kodex) that OpenAI has deployed for handling external stakeholder volume.

3. Law Enforcement SLAs (Anthropic: 3 vs. OpenAI: 2)

- **Why both scores are low:** Neither company publishes a strict time-based SLA (e.g., "We respond to subpoenas in 4 hours").
- **Why Anthropic is higher (3):** The score was boosted by the specific precedent of the **Cyber Espionage Campaign (Nov 2025)**. Anthropic demonstrated a high-level "National Security" escalation path by coordinating a technical attribution report with government agencies. This proved their escalation path works for high-severity incidents.
- **Why OpenAI is lower (2):** OpenAI's process is opaque. They use an internal classifier to detect "imminent threat," but explicitly state they do **not** report self-harm to police. While this is a valid privacy choice, from a "Law Enforcement Escalation" metric, it represents a refusal to engage. Furthermore, they rely on third-party integrations (like Axon) for police report drafting, which delegates the governance risk to the user.

4. Indigenous Data Stewardship (OpenAI: 1 vs. Anthropic: 0)

- **Why OpenAI is a 1 (Nominal):** OpenAI receives a 1 because they technically have a "Data Partnerships" intake form. However, they lost points for the **"OpenAI to**

Z Challenge," which gamified the use of Indigenous historical records without clear consent frameworks, violating the spirit of data sovereignty.

- **Why Anthropic is a 0 (Absent):** A search of Anthropic's System Cards and RSP reveals no specific policy regarding Indigenous Data Sovereignty or the **CARE Principles**. The absence of evidence results in a zero.

5. Board-Level Oversight (Anthropic: 5 vs. OpenAI: 3)

- **Why Anthropic is a 5 (Structural):** This is the strongest example of "Structural" governance. The **Long-Term Benefit Trust** holds **Class T Shares**, which will eventually allow it to elect a majority of the Board. This is a legal mechanism that physically separates profit motive from control. It is a "fiduciary firewall."
- **Why OpenAI is a 3 (Advisory):** OpenAI formed a **Safety and Security Committee**. While it includes independent directors (like Paul Nakasone), it is a standard board committee. With OpenAI transitioning to a Public Benefit Corporation (PBC) *without* a controlling Trust mechanism, the oversight remains subject to shareholder voting power, lacking the independent legal fortress that Anthropic has built