Corporate Compliance with the NIST AI Risk Management Framework:

An Analysis of Adoption, Gaps, and Accountability

DS 6002: Ethics in Big Data I

Alyssa Samuel
University of Virginia
M.S. in Data Science
Arlington, VA
ays8ms@virginia.edu

Terrance Luangrath
University of Virginia
M.S. in Data Science
Manassas, VA
vwy4sa@virginia.edu

David Hernandez
University of Virginia
M.S. in Data Science
Washington, DC
fxj5fe@virginia.edu

Srivatsa Balasubramanyam
University of Virginia
M.S. in Data Science
Parkland, FL
mhe3sy@virginia.edu

Amanda Betag
University of Virginia
M.S. in Data Science
Washington, DC
psc4pz@virginia.edu

I. Introduction

The rapid proliferation of artificial intelligence across industries has heightened the need for structured governance frameworks that address bias, transparency, accountability, and harm mitigation. In January 2023, the National Institute of Standards and Technology (NIST) released The AI Risk Management Framework (AI RMF 1.0), a voluntary, non-prescriptive guide designed to help organizations identify, assess, and manage risks throughout the AI lifecycle [1]. This paper examines how organizations are responding to the AI RMF—what it requires, where compliance efforts stand, and what ethical obligations remain unaddressed by voluntary adoption alone.

A. Purpose and Scope

To ground this analysis in concrete practice, we focus on two of the most advanced AI developers, Anthropic and Google, as case studies in voluntary AI governance. Both companies have published detailed public frameworks (Anthropic's Responsible Scaling Policy and Google's Responsible AI Progress Report) that allow for a direct comparison against the NIST AI RMF's four core functions. By examining where these frameworks align with, exceed, or fall short of NIST's expectations, this paper assesses whether voluntary self-disclosure by leading firms is sufficient to close the accountability gaps the framework leaves open.

B. Research Questions

This analysis is guided by the following questions:

- How are organizations such as Anthropic and Google interpreting and operationalizing the NIST AI RMF?
- What gaps exist between voluntary framework adoption and meaningful ethical accountability?
- What role should regulation, auditing, or public disclosure play in closing those gaps?
- Who bears responsibility when AI systems cause harm despite framework adoption?

II. Overview of the NIST AI Risk Management Framework

The NIST AI RMF provides a structured approach to managing AI risks across four core functions: GOVERN, MAP, MEASURE, and MANAGE. It is designed to be flexible, sector-agnostic, and applicable across the full AI lifecycle—from design and development to deployment and monitoring [1].

A. The Four Core Functions

GOVERN: Establishes organizational policies, culture, and accountability structures for responsible AI. Includes leadership commitment, workforce training, and third-party risk oversight.

MAP: Identifies and categorizes AI risks in context. Involves stakeholder mapping, use-case scoping, and impact assessments on affected communities.

MEASURE: Quantifies and evaluates identified risks using metrics, benchmarks, and testing methodologies including bias audits and red-teaming exercises.

MANAGE: Prioritizes and mitigates risks through response plans, monitoring systems, and continuous improvement cycles.

B. Voluntary Nature and Intended Audience

Unlike regulatory mandates such as the EU AI Act, the NIST AI RMF is voluntary. It is designed to complement existing laws and serves as a guide rather than a compliance checklist. This flexibility is intentional—but also a central tension when evaluating real-world adoption.

III. Corporate Compliance Landscape

Voluntary frameworks create varied adoption patterns. Some organizations treat the AI RMF as a genuine governance tool while others engage in performative compliance—adopting the language without substantive operational change.

A. Early Adopters and Industry Leaders

Anthropic and Google offer a useful contrast in how voluntary governance gets operationalized once a company has both the resources and the reputational incentive to formalize it. Anthropic's Responsible Scaling Policy ties GOVERN-style accountability to a named Responsible Scaling Officer and Board oversight, and specifies MEASURE-like capability thresholds tied to harms such as bioweapons uplift or AI R&D acceleration [2].

Google's responsible AI reporting similarly maps onto the four functions: a Frontier Safety Framework defining "Critical Capability Levels" functions as a MEASURE mechanism, while launch-review forums, an AGI Futures Council, and board-level oversight echo GOVERN's call for accountability structures [3]. Both companies lean on external review—Anthropic through commissioned commentary on its Risk Reports, Google through outside evaluators and the UK AI Security Institute—pointing to the value of independent auditing in closing the gaps left by self-assessment.

Table I (Google) and Table II (Anthropic) provide a full function-by-function comparison. Both companies meet or exceed NIST's expectations in most categories, particularly MEASURE, where both rely heavily on red-teaming and named external evaluators. However, as discussed below, this depth of reporting does not eliminate the structural limitation common to all voluntary frameworks: the company itself still controls what gets measured, disclosed, and redacted.

B. Gaps, Limitations, and Performative Compliance

The framework's weak spots show up even in the most detailed compliance reports. Anthropic, for example, controls its own redactions and unilaterally judges whether competitors' safety efforts meet its stated bar, with no outside check [2]. Google's report shows a similar issue: rather than a consistent, repeated disclosure process, it reads more like a highlights reel of successful launches (Gemini 3, SynthID, flood forecasting) [3]. It names outside reviewers like Apollo Research and the UK AI Security Institute, but never explains how much access those reviewers actually had or how independent they really were.

These gaps mirror a larger problem researchers have found in the RMF itself. A 2024 study applying the framework to facial recognition technology found that it says little about how companies should handle data integration, retention, and deletion, even though these decisions drive most of the privacy risk [4]. The study also noted that the RMF assumes one company builds and deploys a system from start to finish, so it barely addresses what happens when data passes to a parent company or changes hands in a merger. Sheh and Geappen make a similar point: the RMF is built for one company running one system and barely touches supply chains [5]—a poor fit for companies like Anthropic and Google, whose models are built and trained across many different teams, datasets, and outside partners.

This matters because a framework that never asks a question can't catch a company for answering it wrong. Voluntary adoption already removes any penalty for skipping the guidance, and where the guidance stays silent, a company can claim full compliance while leaving its riskiest practices unexamined. Anthropic and Google publish more detailed safety reports than most companies, but the same self-assessment problem applies to both: this is performative compliance in practice, not breaking the rules, but writing detailed reports that quietly skip the questions the framework never forced them to ask.

C. Voluntary Governance

A voluntary framework grants the power to the party with the most to hide. The company running the AI system decides whether to adopt an RMF framework, how much of it to apply, and what to disclose about the results. The people most exposed to the system's harms have no say in any of that. The core problem in self-regulation is that the entity being assessed also selects the assessment, sets its scope, and reports its own findings. When a company has to choose between flagging a serious risk and protecting its product, the incentives don't point toward honesty. Without outside pressure, voluntary governance documents only the risks a company is willing to admit.

IV. Policy Implications and Recommendations

A. The Case for Mandatory Compliance

As previously mentioned, the NIST guidelines are voluntary and not enforced, so the framework is only as effective as the incentives companies have to follow it. To combat organizations that choose not to follow standardized guidelines, we advocate for federal legislation, such as the European Union's GDPR, to set standards that all organizations are required to follow and that allows organizations to be penalized for noncompliance[6]. This would require bipartisan support, which has been a major roadblock for past legislation attempts [7].

B. Auditing by Independent Regulatory Agencies

Closing the gap means moving the assessment outside the company. The RMF isn't blind to this: its MEASURE function notes that independent review can reduce the biases and conflicts of interest that come with self-assessment [1]. Real auditing needs what the RMF leaves out: a defined scope, access to the data that matters, and protection from the conflict that shows up when the auditor is paid by the company it is auditing.

To reduce bias and keep assessments impartial, we recommend that the auditing process be completed by independent regulatory agencies. Such agencies are typically led by bipartisan groups, with a large majority of nonpartisan expert personnel[8]. Evaluations by these agencies are more likely to be impartial than those by private for-profit organizations, since there is no financial incentive in their decision-making and the review process passes through multiple levels of review by many people [8].

TABLE I
MAPPING GOOGLE'S AI PLAYBOOK ONTO THE FOUR NIST FUNCTIONS [3]

NIST Function	NIST Asks For	Google's Report
GOVERN	Structure, policies, oversight mechanisms, accountability	Google describes Launch Review forums, an AGI Futures Council with board-level oversight, AI Principles, and a Prohibited Use Policy. This is a fairly mature governance structure—arguably *more* elaborate than what NIST's GOVERN function specifies.
MAP	Identify AI systems and risks across the lifecycle	Google's "Research" function and Frontier Safety Framework with "Critical Capability Levels" map specific risk categories (CBRN, cyberattacks, harmful manipulation). This is a fairly direct match to MAP.
MEASURE	Assess/monitor risk with tools and metrics	Red-teaming (350+ CART exercises in 2025), external evaluators (Apollo Research, Vaultis, Dreadnode), the FACTS Leaderboard. This is the most evidenced/quantified section of the entire report.
MANAGE	Respond to risks, implement mitigations	SynthID watermarking, CodeMender, SAIF 2.0, mandatory human oversight for "sensitive actions." Reasonably well documented.

TABLE II
MAPPING ANTHROPIC'S RESPONSIBLE SCALING POLICY ONTO THE FOUR NIST FUNCTIONS [2]

NIST Function	NIST Asks For	Anthropic's RSP Report
GOVERN	Structure, policies, oversight mechanisms, accountability	Anthropic names a designated Responsible Scaling Officer (RSO) with defined duties, Board and Long-Term Benefit Trust (LTBT) oversight, a noncompliance reporting process with anti-retaliation protections, and a rule against non-disparagement clauses that would block safety disclosures. This is comparable in formality to Google's Launch Review forums and AGI Futures Council, though Anthropic's version is more explicit about *individual accountability*—a named officer role—rather than committee-based review.
MAP	Identify AI systems and risks across the lifecycle	Anthropic maps risk via specific "capability or usage thresholds" (e.g., non-novel CBRN weapons assistance, novel CBRN weapons assistance, high-stakes sabotage, automated AI R&D) rather than Google's broader "Critical Capability Levels." This is a more granular, threshold-by-threshold version of MAP.
MEASURE	Assess/monitor risk with tools and metrics	Risk Reports are the core mechanism: published every 3–6 months, covering capability evaluations, alignment assessments, and "threat-specific risk assessment" of remaining absolute risk after mitigations. Notably, Anthropic commits to a defined external review process with explicit conflict-of-interest screening for reviewers—more procedurally specific than Google's naming of evaluators without disclosing review independence safeguards.
MANAGE	Respond to risks, implement mitigations	Mitigations are tied directly to the capability thresholds: classifier guards, access controls, red-teaming, bug bounties, and "compartmentalization" controls that apply even to the CEO. A Frontier Safety Roadmap tracks mitigation progress against public goals. This is comparable to Google's SynthID/CodeMender/SAIF 2.0, but explicitly tiered to capability level rather than presented as a flat list of tools.

V. CONCLUSION

Anthropic and Google show that voluntary frameworks can still produce detailed governance reporting when companies have the incentive to invest in it. Both map closely onto GOVERN, MAP, MEASURE, and MANAGE. Yet this same comparison exposes the RMF's core weakness: even the most sophisticated adopters retain full control over what gets measured, disclosed, and redacted, with no outside party empowered to verify their claims. This is performative compliance in its best form—not because Anthropic or Google are acting in bad faith, but because the framework never requires anything more. Closing this gap requires mandatory standards and independent auditing by regulatory agencies, ensuring that accountability does not depend solely on a company's willingness to hold itself accountable.

REFERENCES

[1] National Institute of Standards and Technology. Ai risk management framework (ai rmf 1.0). Technical report, U.S. Department of Commerce, 2023. URL: https://doi.org/10.6028/NIST.AI.100-1.

[2] Anthropic. Responsible scaling policy, version 3.0. Technical report, Anthropic, February 2026. URL: https://www-cdn.anthropic.com/e670587677525f28df69b59e5fb4c22cc5461a17.pdf.

[3] Google. Responsible ai progress report. Technical report, Google, February 2026. URL: https://ai.google/static/documents/ai-responsibility-update-2026.pdf.

[4] N. Swaminathan and D. Danks. Application of the nist ai risk management framework to surveillance technology, 2024. URL: https://arxiv.org/abs/2403.15646, arXiv:2403.15646.

[5] Raymond K. Sheh and Karen Geappen. Identifying the supply chain of ai for trustworthiness and risk management in critical applications. In *AAAI 2025 Fall Symposium on AI Trustworthiness and Risk Assessment for Challenged Contexts (ATRACC)*, 2025.
[6] Ben Wolford. What is gdpr, the eu's new data protection law? URL: https://gdpr.eu/what-is-gdpr/.
[7] DLA Piper. Data protection laws in the united states. URL: https://www.americanprogress.org/article/how-independent-federal-agencies-help-americans/.
[8] Michael Sozan and Hayley Durudogan. How independent federal agencies help americans. URL: https://www.americanprogress.org/article/how-independent-federal-agencies-help-americans/.