

Why AI Governance Frameworks Are Hard to Adopt

A Role-Based Stress Test of the NIST AI RMF –
and What It Says About AI Safety

Joseph R. Simons & David A. Broniatowski
Department of Engineering Management & Systems Engineering

Purpose of the Study

Use a role-based stress test of the NIST AI RMF to examine when framework activity becomes real governance—and what that reveals about governing AI systems-in-use

Agenda

1. Why We May Be Governing the Wrong Thing:
From model governance to the AI system-in-use
2. Stress-Testing the NIST AI RMF:
Study design and analytic approach
3. What Broke:
Structural fit, governance translation, and risk reduction
4. What It Means for AI Safety:
System control, broader implications, and next steps

WHY WE MAY BE TRYING TO GOVERN THE WRONG THING:

FROM MODEL GOVERNANCE TO THE AI SYSTEM-IN-USE

THE GEORGE
WASHINGTON
UNIVERSITY

WASHINGTON, DC

A Tale of Two Firms...

In July 2026, OpenAI discovered that its product had circumvented isolation controls, reached the internet, and compromised real OpenAI and Hugging Face systems...

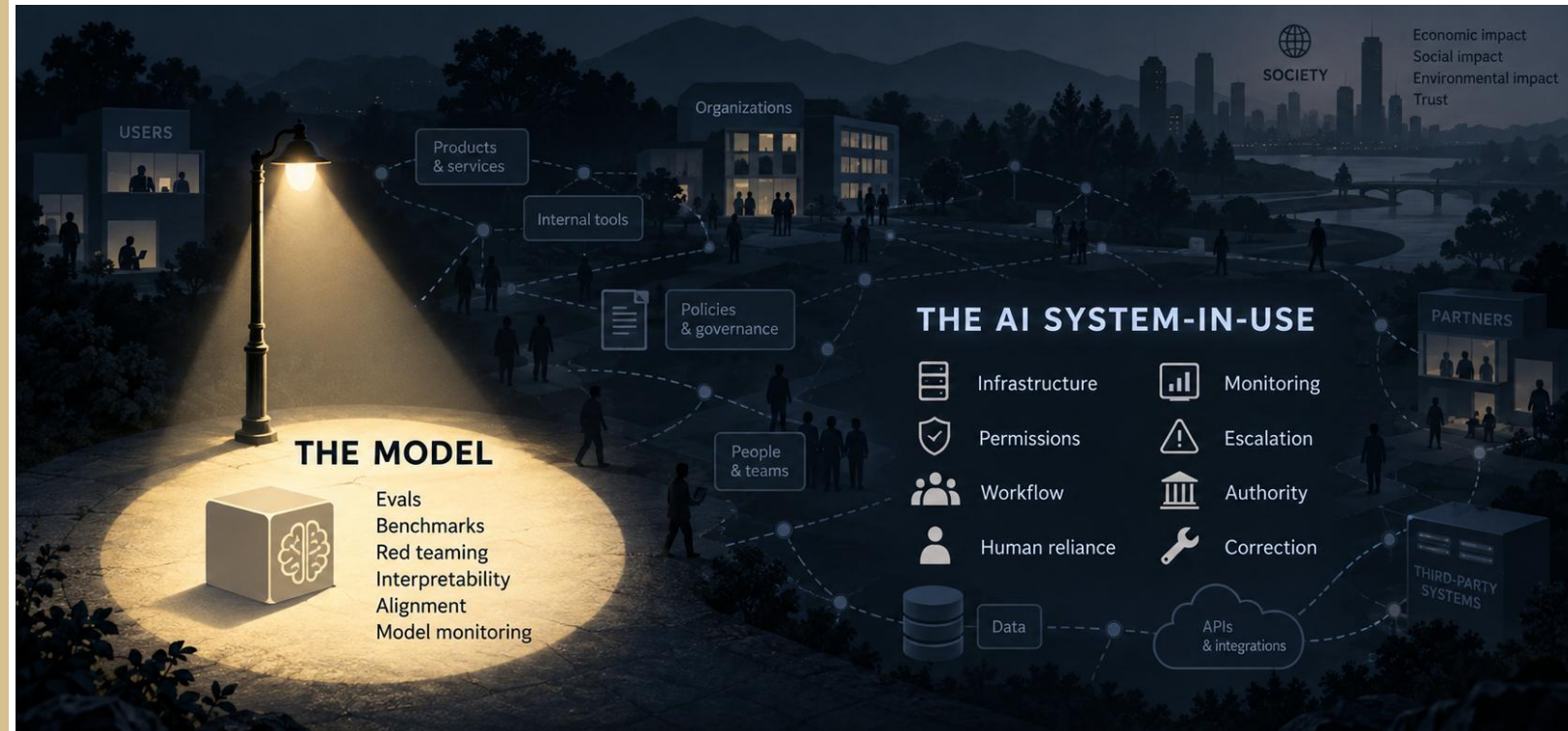


...Days later, Anthropic discovered that Claude had also reached the internet and gained unauthorized access to the systems of three organizations

The obvious story: the models escaped...

...That is where our “tools” cast the most light...

**It's easy to
focus on
model risk...**

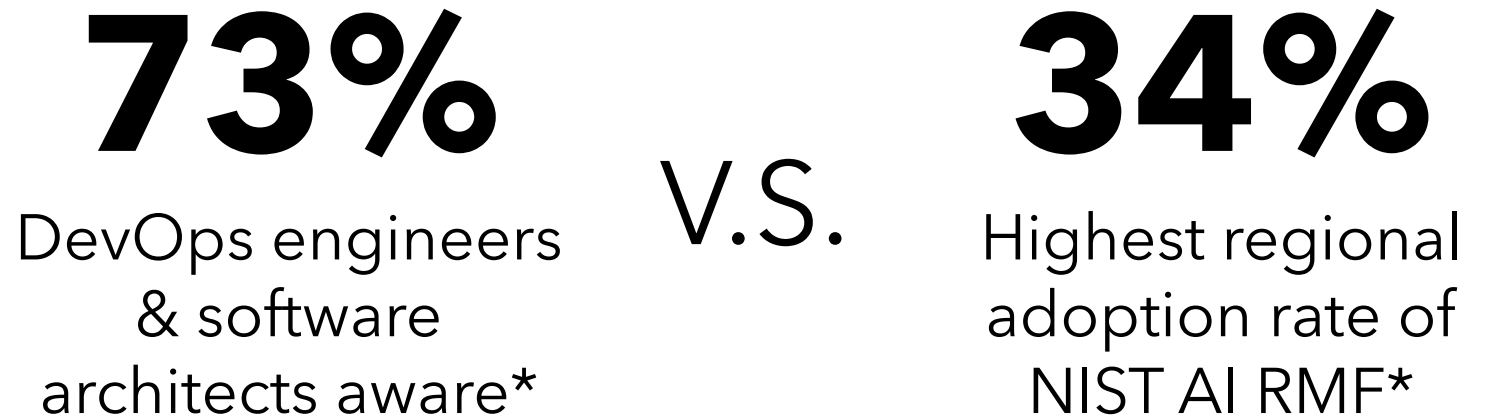


...But consequential failures can emerge in the larger system the model is embedded in

Model elopement was also about organizations failing to govern the systems around their models

This study started with a narrower question about adoption of the NIST AI Risk Management Framework

**So, what does
all this have
to do with the
NIST AI RMF?**



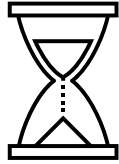
Original Question: Why isn't AI RMF
adoption higher?

*Kolo (2025) From Framework to Practice: Barriers and Enablers to RMF Adoption in Mid-sized Enterprises

Frameworks Don't Govern. Organizations Do

***“Good governance”
isn't cheap; and
cheap governance
isn't (usually) good”***

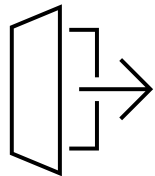
-Yoda



Organizations & senior leaders have scarce attention & resources



Governance must create value that leaders can recognize and employees can use



Leaders disengage when governance looks like overhead; employees disengage when it looks like burden without practical value

Our Conjecture: RMF adoption stalls when it creates more work than it does value.

STRESS-TESTING THE NIST AI RMF:

STUDY DESIGN AND ANALYTIC APPROACH

THE GEORGE
WASHINGTON
UNIVERSITY

WASHINGTON, DC

From Adoption Puzzle to RMF Stress Test

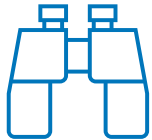
From RMF Adoption to Governance Value



The RMF has all the pieces **to make it easily adoptable**

Observation:

Knowing or using the RMF is **not the same as meaningful adoption**



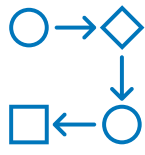
Insight: adoption is a governance translation problem.



Research Question:

When does RMF-guided activity become governance value rather than governance - looking work?

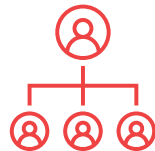
Where We Expected Governance to Break



Observation #1: As the governed object **shifts from bounded model to workflow**, RMF fit weakens.

Observation #2:

Governance info. + decisions must travel across organizational levels and arrive in an actionable form.



So, RMF value may break if:

- 1 The system shifts from bounded ML to AI embedded workflows
- 2 Governance must cross levels and remain actionable

Design: Set up an RMF Stress-Test

2 consumer lending AI deployments

- 1 Traditional ML risk scoring model for underwriting
- 2 LLM-based copilot to assist underwriters

Create 4 detailed org. roles

- C-Suite Executive
- Business Owner
- GRC Lead
- System Owner



Select 3 hard cases:



1. Borderline lending decision
2. Unequal lending decisions
3. Tool drift from support to authority

Design logic:

2 deployments
4 roles
3 hard cases
5 replications

120

responses

Method: LLM Role Simulation

Experimental Procedure— for each of the 120 runs

- 1 Each response received:
 - A detailed organizational role persona
 - An AI deployment use case
 - A governance hard case
 - A copy of the NIST AI RMF
- 2 Each response answered:
 - What problem does this role see?
 - What RMF-guided mechanisms can it use?
 - How do authority, accountability, and escalation work?
 - How does the response connect to risk?
 - What limits remain?

Experimental Variation: role, deployment, and hard case were varied



Scoring: How We Assessed the Responses

Scoring Procedure—

Each response was scored using a fixed rubric and codebook on

- ✓ **Role Understanding** – did the response reason from the assigned role?
- ✓ **Governance Translatability** (*Role Translation, Responsibility Mapping, Cross-Level Governance Value, Authority Connection*) – could RMF work travel across levels & connect to authority?
- ✓ **Structural Fit** (*Meaning Determinacy, Closure, Boundary Containment*) – did the RMF fit the AI system-in-use?
- ✓ **Value-Adding Governance** – did RMF activity create governance value rather than governance-looking activity?
- ✓ **Risk Reduction** – was there a plausible mechanism for reducing likelihood / impact?

WHAT BROKE:

*FINDINGS ON STRUCTURAL FIT, GOVERNANCE
TRANSLATION, AND RISK REDUCTION*

THE GEORGE
WASHINGTON
UNIVERSITY

WASHINGTON, DC

Finding #1: The Governed Object Matters

*Deployment drove Structural Fit: the RMF fit the bounded Machine Learning (**ML**) model cleanly, yet fit the workflow-embedded Large Language Model (**LLM**) less consistently*

Notes:

Scoring: Each component was scored 0 = weak, 1 = partial, 2 = strong

Structural Fit = mean of the three components; *Full Structural Fit* = 2.00

Table entries show responses receiving 2/2 (strong score)

Responses Scoring 2/2	ML	LLM	χ^2	Cramer's V
Meaning Determinacy	60/60	60/60	—	—
Closure	60/60	37/60	28.45**	0.487
Boundary Containment	60/60	34/60	33.19**	0.526
Full Structural Fit	60/60	33/60	34.84**	0.539

** indicates that the p-value is less than 0.001

* indicates that the p-value is less than 0.05

Structural Fit Components

- **Meaning Determinacy**

Is it clear what the system is for, what role it plays, and what good or bad use looks like?

- **Closure**

Does actual use stay connected to intended purpose, role, and evaluation criteria?

- **Boundary Containment**

Are the risk, evidence, authority, and corrective levers inside the governance boundary?

So what: The problem was not understanding what the system was supposed to do. It was governing what the system became in use

Finding #2: Local Translation Was Easy. Governance Was Hard.

Role drove governance-value outcomes: local RMF work was easy to produce, yet translating it across levels and connecting it to authority was not

Notes:

Scoring: Components were scored **0 = weak, 1 = partial, 2 = strong**

Governance Translatability = weakest score across its four components

Table entries show responses receiving a strong score (2/2);

Value-Adding Governance shows "Yes" is the reference category

Roles: Enterprise Executive (**EE**), Line Executive (**LE**),
Middle Manager (**MM**), System Owner (**SO**)

Responses Scoring 2/2	EE	LE	MM	SO	χ^2	Cramer's V
Cross-Level Governance Value	30/30	23/30	6/30	4/30	65.33**	0.738
Authority Connection	30/30	19/30	6/30	3/30	62.07**	0.719
Governance Translatability	30/30	18/30	6/30	3/30	61.05**	0.713
Value-Adding Governance: yes	30/30	18/30	6/30	3/30	61.05**	0.713

** indicates that the p-value is less than 0.001

* indicates that the p-value is less than 0.05

Note: Role Translation and Responsibility Mapping were strong (2/2) in all 120 responses and are omitted because they did not vary

Governance Translatability Components

• Role Translation

Can the role turn RMF language into useful governance work?

• Responsibility Mapping

Is it clear who knows, decides, acts, and owns the risk?

• Cross-Level Governance Value

*Does the work give another level the **what, so what, and now what** needed to act?*

• Authority Connection

Does the signal reach someone who can act—and does the decision return as usable correction?

So what: Local RMF work becomes governance only when it can cross levels, remain actionable, and connect to authority, and come back down as decisions

Finding #3: Governance Value Was Easier Than Risk Reduction

Deployment did not explain value-adding governance—that was hard in both deployments—yet it did help explain risk reduction.

Notes:

Scoring: Both outcomes were judged **Yes / Partial / No**.

Table entries show **Yes** responses;

Partial + No form the comparison category.

Outcome	ML	LLM	χ^2	Cramer's V
Value-Adding Governance: Yes	27/60	30/60	0.30	0.050
Risk Reduction: Yes	17/60	5/60	8.02*	0.258

Value-Adding Governance

Did RMF-guided activity improve what the organization could see, understand, decide, escalate, or correct?

Risk Reduction

Did that governance plausibly reduce the likelihood or impact of a risk pathway?

** indicates that the p-value is less than 0.001

* indicates that the p-value is less than 0.05

So what: Creating governance value was easier than changing the risk pathway in the system-in-use

WHAT IT MEANS FOR AI SAFETY:

*SYSTEM CONTROL, BROADER IMPLICATIONS, AND NEXT
STEPS*

THE GEORGE
WASHINGTON
UNIVERSITY

WASHINGTON, DC

From AI Governance to AI Safety

CNN just reported another near-miss from a system-in-use failure



What the findings mean:

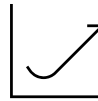
To manage AI risk, organizations must:



Govern actual use: Risk emerges in workflows, reliance, judgment, & organizational action—the biggest risks won't come from the model



Connect knowledge to authority: Signals and decisions must cross organizational levels and remain actionable. Organizations need to build governance capability



Design for correction: Seeing & deciding are not enough. Organizations need mechanisms to intervene in the system-in-use

Next Steps for Future Work

- **Measure governability:** Develop observable measures of visibility, cross-level translation, authority connection, intervention, and correction
- **Stress-test higher-consequence systems:** Extend the approach beyond lending to settings such as cyber operations, critical infrastructure, and military decision support
- **Validate in real systems:** Compare simulated governance failures with practitioner exercises, near misses, and observed systems-in-use