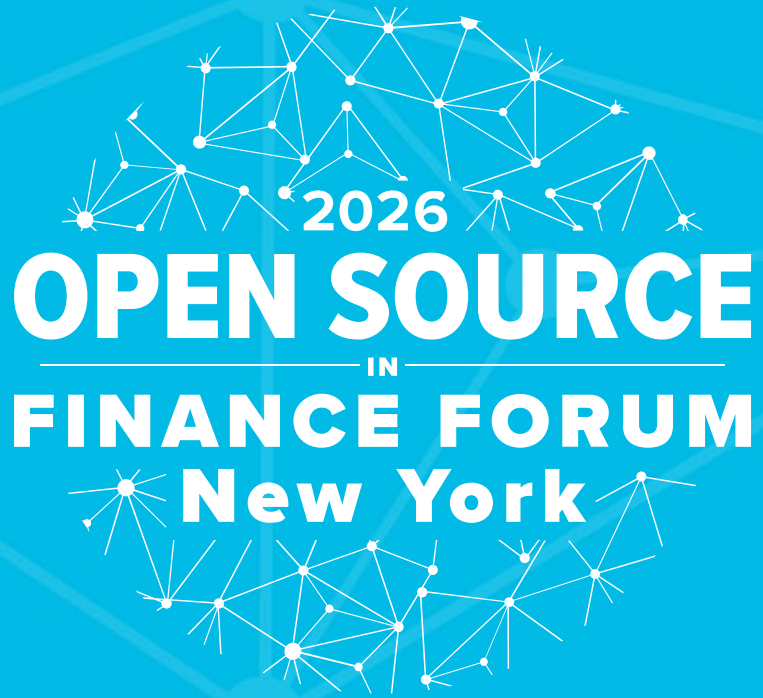


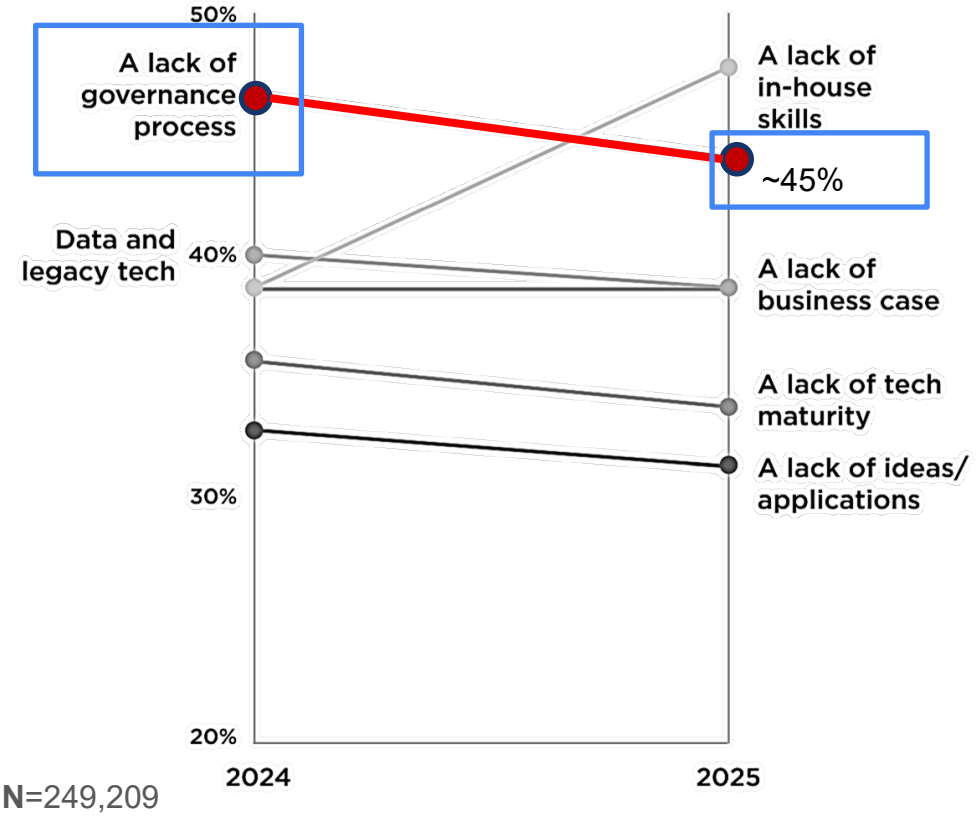
HomeWorld
Educational Resources

Governance as Simulation

Evaluating AI Governance Risks and Mitigations with Agent-Based Modeling

Jiaman Li,
Yang Wang,
Ning Wang Ph.D.

01-CHALLENGE GenAI is Held Back by Governance



N=249,209

Source: The 2025 State of Open Source in Financial Services - 2024-2025 FINOS Survey

Governance is a top barrier

Shown in the FINOS's financial-services recent surveys

We have checklists and Automation; But

Rich Catalogues (FINOS AI Gov Framework) and Gov-as-Code; Yet, hard to intuitively picture risks inside a real institution's workflow

02-The IDEA A Simulator, “A Rehearsal”, for AI Governance

Pilots rehearse emergencies before they fly. Financial AI Governance teams need to rehearse a certain period of AI initiatives, failures and reviews before adapting or changing the real process.

Governance as Documentation

What are the risks and mitigation options?

Governance as Code

Checks that run automatically (e.g. via FINOS CALM)

Governance as Simulation

What may happen in different setup and governance strategies?

“What may happen when things go wrong?”

02-The IDEA Runtime Design Sets The Risk. Governance Is Tested Against It.

Two-Layers Simulation

<I> AI Agent Modality Simulation + <II> Governance Strategy Simulation

01

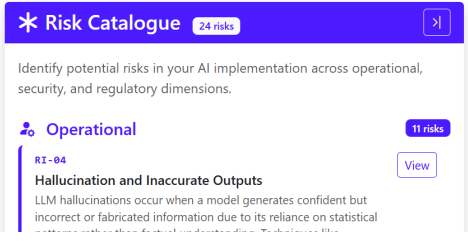
Reference modality

Per AI type: agent workflow, controls, use-case questions, expected volume.

02

Risk profile

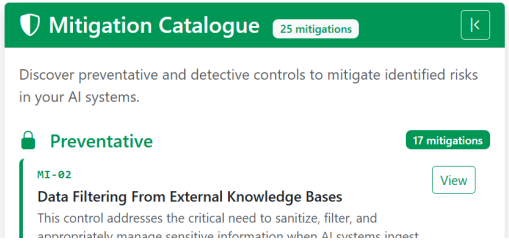
Match Modaility Design with FINOS risks, how often they occur, how costly they are.



03

Governance run

Initiatives flow through gates, queues and reviewers simulating thousands of times.

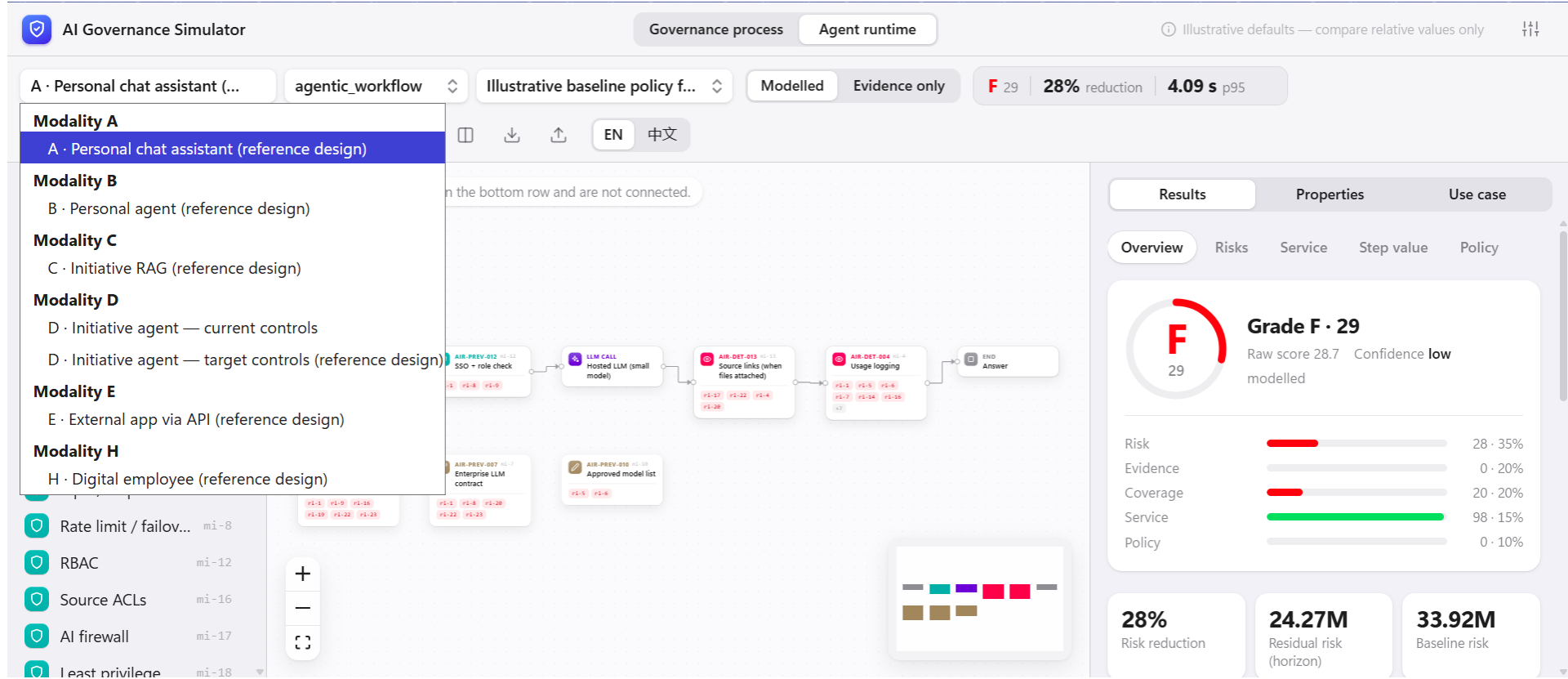


04

Decision Comparison

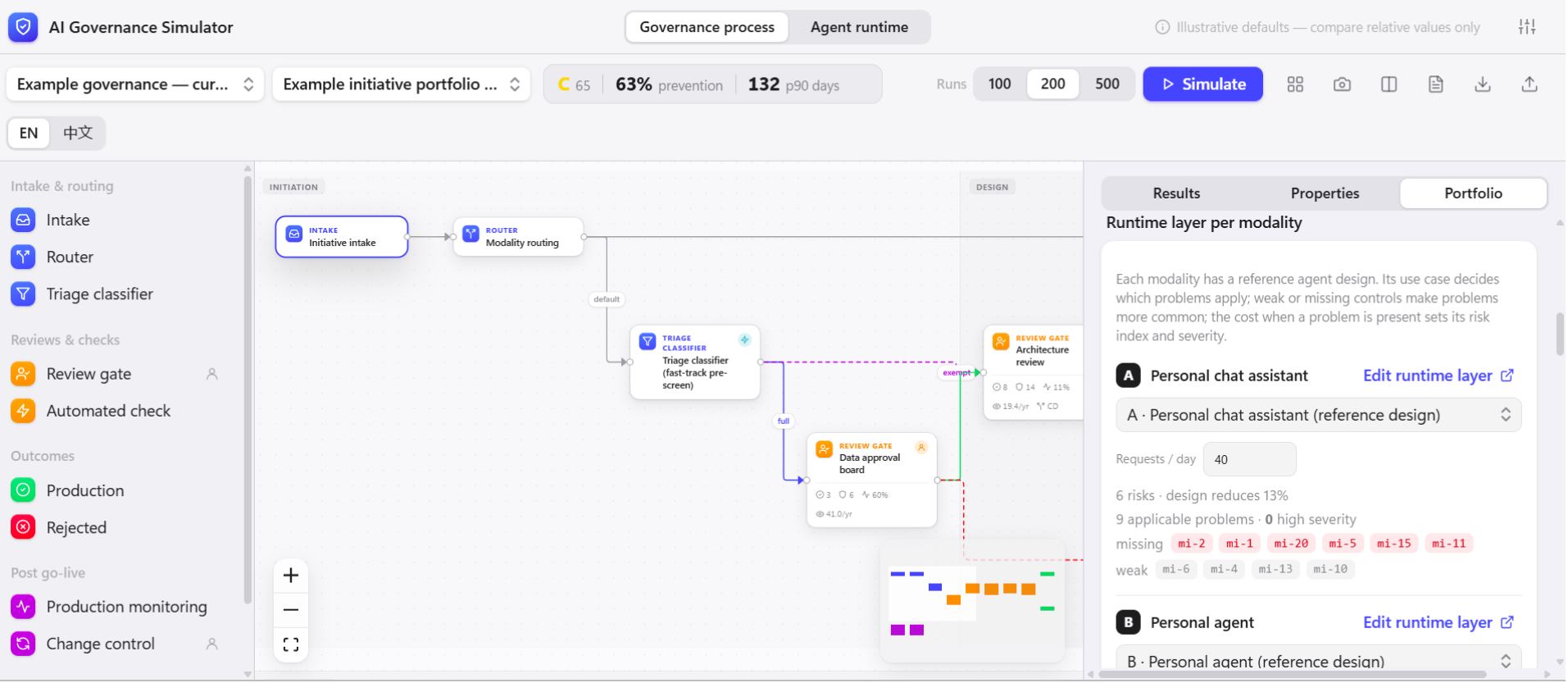
e.g., Grade, escapes by AI type, effort vs. risk, and the checks that help most.

03-The PROTOTYPE <I> AI Agent Modality Simulation



Click to see the video demo of AI Agent Modality Simulator: Link

03-The PROTOTYPE <II> Governance Strategy Simulation



HomeWorld Educational Resources

Click to see the video demo of AI Proposal Governance Simulator: Link

02-The PROTYPE Two Layers, Joined By the AI Modality Design

AI Modality Decides the Risk Baseline

Prototype Types:
(A) chat assistant · (B) RAG · (C) tool-using agent · (D-H) customer-facing app

AI Agent Runtime layer (Modality Design)

unit: one request

- Model
- Data
- Tools
- Runtime controls
- Design-time controls

→ request-level risk, latency, availability, cost

Governance layer

unit: one AI initiative

- Intake
- Triage
- Review boards
- Monitoring
- Change control
- Release gates

→ caught vs. escaped, lead time, cost, post-launch detection

03-The PROTOTYPE All Controls Matches FINOS AI GOV FRAMEWORK

24
FINOS risks

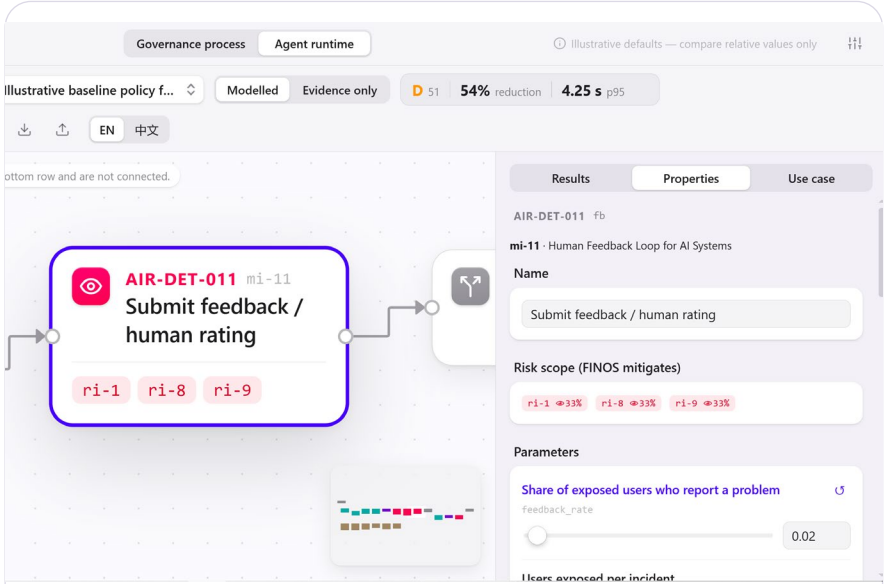
25
FINOS mitigations

17
governance issue types

Each control acts in one of three ways:

- **Reduce likelihood**
- **Block before the harm point**
- **Detect and contain**

Same mitigation twice? Counted once, the strongest wins.



The feedback-loop node AIR-DET-011 (mi-11) declares the FINOS risks it mitigates: ri-1, ri-8, ri-9.

Catalogue: FINOS AI Governance Framework risks (ri-*) and mitigations (mi-*)

03-The PROTOTYPE Risk Flows Through Every Control

Occurs

exposure × model failure rate

→

Prevented

design-time and runtime controls lower the rate

→

Blocked

controls stop it before the harm point

→

Detected

monitors catch it; value decays with delay

→

Escapes

what is left × impact = risk points

03-The PROTOTYPE Each Element Carries Properties You Can Tune

Five property groups

Probability & control — occurrence, coverage, detection, fix rate, availability

Mix & routing — AI type, risk tier, data class, fast track, risk scope

Capacity & time — reviewers, handling time, rework and delay

Model & cost — tokens, prices, model error rates, salaries, loss

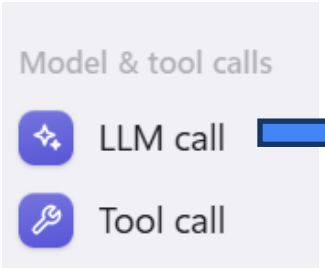
Audit evidence — successes / samples, source, version, period

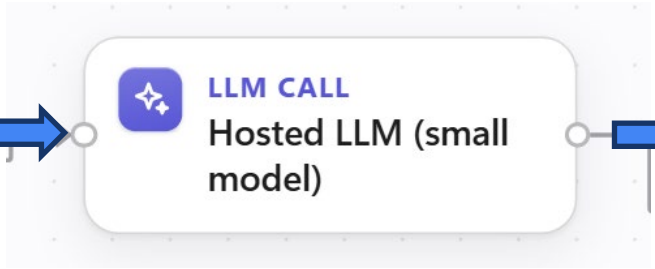
Model behaviour profile

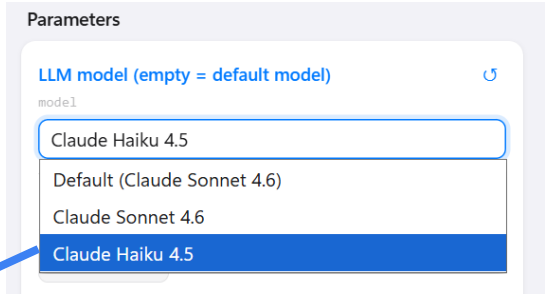
Each FINOS operational risk becomes a 0–1 parameter

hallucination_rate	AIR-OP-004
confidence_calibration	AIR-OP-004
rule_following_prob	AIR-OP-014
escalation_tendency	AIR-OP-018
bias_rate	AIR-OP-016
output_consistency	AIR-OP-006
explainability	AIR-OP-017
version_stability	AIR-OP-005
availability	AIR-OP-007

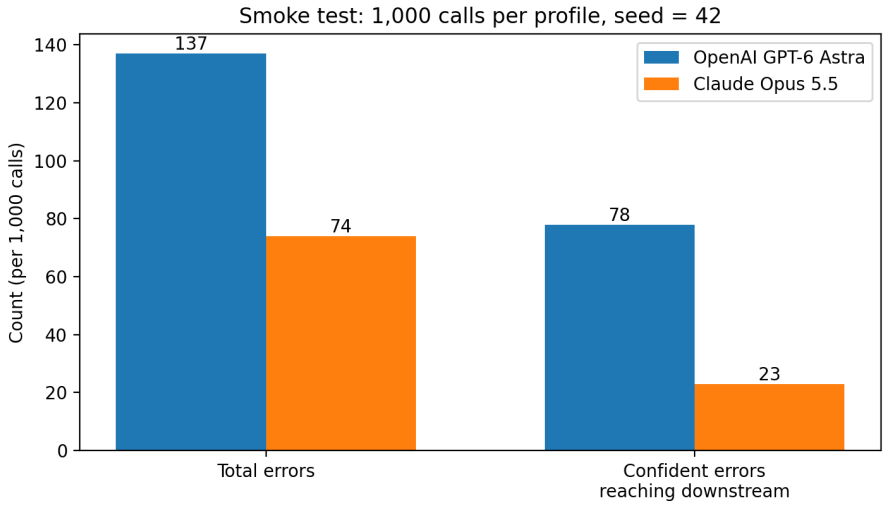
03-The PROTOTYPE AI Models Risk Simulation and Comparison







Parameter	GPT-6 Astra	Claude Opus 5.5
Hallucination rate	0.12	0.08
Confidence calibration	0.45	0.70
Escalation tendency	0.30	0.65
Rule following	0.88	0.93
Cost per call	$0.045	$0.018



Synthetic profiles · unmitigated smoke test, 1,000 calls each, seed 42 · logic check and data from literatures and reports

03-The PROTOTYPE Use The Simulator to Estimate Risk and Compare Governance

Current vs. target process

Current → Target

C → **B**

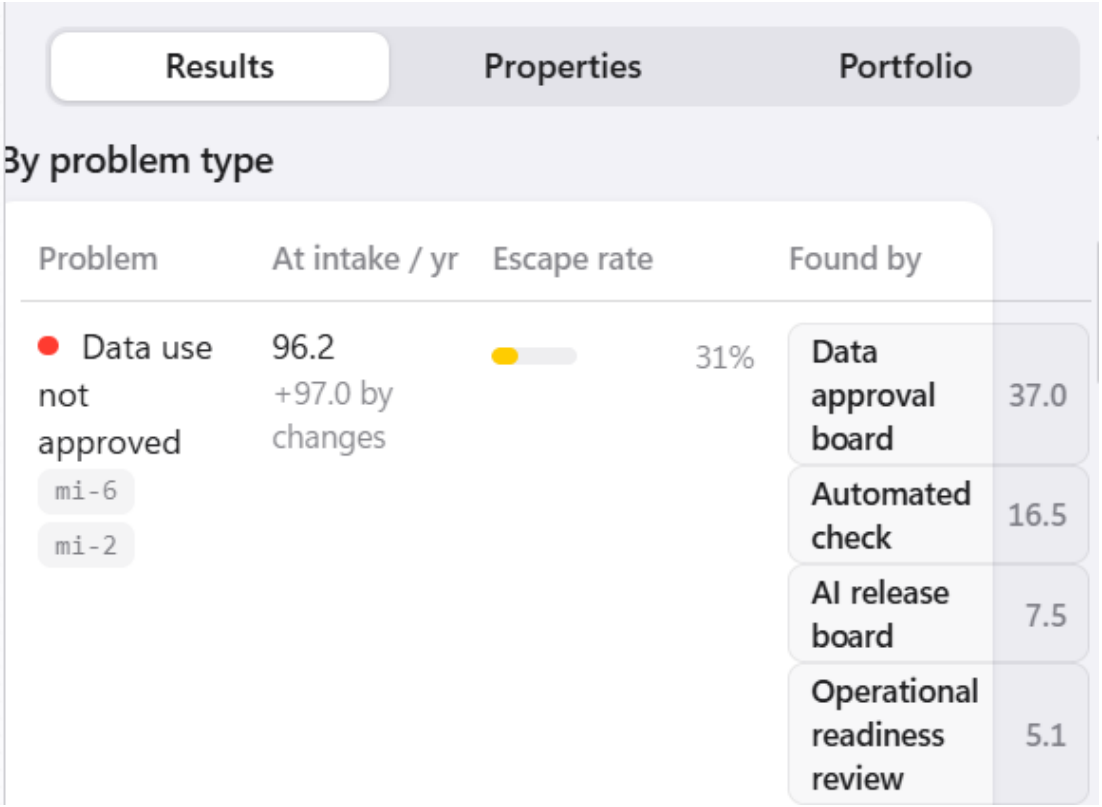
score 65 → **score 79**

63% → 72%

high-severity issues caught before launch

42% → 81%

risky post-launch changes sent back for review



Catalogue: FINOS AI Governance Framework risks (ri-*) and mitigations (mi-*)

03-The PROTOTYPE Use The Simulator to Estimate Risk and Compare Governance

What governance teams can get from the simulation

Redesign with evidence — test a proposed process before rollout

Right-size oversight — find under- and over-governed AI types

Plan capacity — see queues and lead times as demand grows

Prioritise new checks — rank which control, at which gate, helps most

Report in FINOS terms — coverage and findings for committees

FUTURE Calibrate It to Trust It

Next steps

1. Calibrate with real audit, incident and loss data
2. Backtest on history, then validate on new processes
3. Multiple model shared failures, reviewer correlation, dynamic queues
4. Add AI types: multimodal, coding assistants, external APIs

Footnote: Today's defaults are illustrative; risk points are relative; dollar loss defaults to $0.

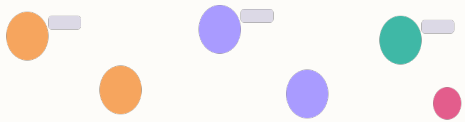
Governance as Simulation: Evaluating AI Governance Risks and Mitigations with Agent-Based Modeling

HomeWorld
Educational Resources

Questions to:
research@homeworld.education

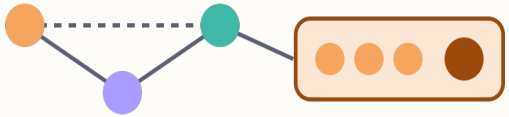
P.S. HWER will work with FINOS to develop this project into an open-source tool in FINOS ecosystem to better support AI Governance.

Appendix A. Agent Based Modeling



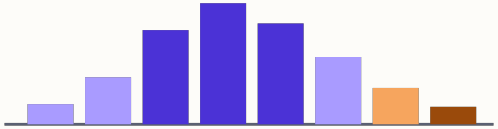
1 · Agents

Individual actors, each with its own state and simple rules.



2 · Interactions

They meet, queue, hand off and fail in a shared environment.



3 · Emergence

System behaviour appears bottom-up; nobody writes it down.

In our simulator, the agents are:

AI initiatives

Arrive through the year, carry FINOS issues, take full or exempt routes.

Human reviewers

Limited capacity; catch issues at a rate; rubber-stamp more when busy.

AI models & agents

Err, escalate or follow rules according to their behaviour profile.

Controls

Prevent, block or detect, according to their tuned properties.

Appendix B. Calibrate It to Trust It

Monte Carlo

A year of initiatives, sampled thousands of times; report p50 / p90 / p95.

Review queues

Each board is an M/M/c queue; at full utilization it never clears.

Shared failure

Controls on one model or data source fail together, so a lookalike adds less.

Audit evidence

k of n audit results update a rate via Beta(1+k, 1+n−k).

- **Monte Carlo**:
run the random year thousands of times and report the spread. p50 is the median outcome, p90 is “9 years in 10 are better than this”, p95 is the near-worst case.

- **M/M/c**
= random (M), random handling times (M), c reviewers in parallel

Appendix C. Aiming at a graded scorecard with hard caps

e.g .25%

Effectiveness

High-severity issues caught before launch

e.g .15%

Post-launch control

Change re-review and production detection

e.g .10%

Triage

Fewer wrong exemptions

e.g .10%

Oversight

Human review on high tiers; fail-safe gates

e.g .10%

Proportionality

Review effort and cost match the risk

e.g .5%

Evidence

Key decisions leave a valid audit trail

e.g .15%

Efficiency

p90 lead time, cost per item, queue load

e.g .10%

Runtime impact

Risk reduction from the reference designs