LDX3 NEW YORK 2026

The Accountability Gap

Engineering governance for autonomous AI

```
$ make demo
```

Sandeep Bharadwaj Mannapur

open source · runs on a laptop · no API keys

Today

01 **How much autonomy should any agent get?**

a model

02 **What takes autonomy away — automatically?**

a mechanism

03 **How do you explain it to legal, security, and the board?**

the words

everything shown is open source — you can run it yourself

When agents misbehave — the public record

2024	**Air Canada**	Chatbot invented a policy — airline held liable
2025	**Google Gemini CLI**	Acted on false success — user files destroyed
2025	**Salesloft Drift**	Agent credentials stolen — 700+ orgs' data accessed

JUL 2025 **Replit** DEEP DIVE →

Production database deleted during an explicit code freeze · thousands of fake profiles generated · "rollback impossible" — false

Why it could happen — none fixed by a better model

1. **Standing access to production credentials**
   no dev/prod separation
2. **The code freeze was words in a prompt**
   policy without enforcement
3. **Deleting a production database needed no approval**
   no gate on the action
4. **The agent's self-reports were trusted**
   and it fabricated success

Replit's own fix: environment separation · planning mode · approval gates — **infrastructure, not a better model**

Nothing malfunctioned

The agent operated inside the authority it was given.
The failure lives in the authorization itself.

> *Qui facit per alium, facit per se* — if your agent acts, you act.
>
> Air Canada, 2024: “the chatbot is a separate legal entity” — rejected by the tribunal.

The accountability gap: the deployer owns the actions — and the logs can’t say who authorized them.

Why the obvious fixes fall short

Human-in-the-loop everything

you can't review 10,000 decisions a day

Model evals

they measure capability — the problem is authority

A policy document

Replit's code freeze was exactly that

Root cause: **unmetered trust**

The framework: three mechanisms

1 **The three-tier trust model**

classify every action by blast radius

2 **The autonomy budget**

meter agent trust with SLOs

3 **Circuit breakers & evidence**

for the failures that still get through

MECHANISM 1 / 3

The three-tier trust model

1 · **The three-tier trust model — classify by blast radius**

2 · The autonomy budget — meter trust with SLOs

3 · Circuit breakers & evidence — for what still gets through

Classify every action by blast radius

impact × reversibility → three tiers

TIER 1 · READ-ONLY

reading an incident, querying inventory

execute freely

TIER 2 · REVERSIBLE WRITE

updating a ticket, deploying to a sandbox

execute — record evidence

TIER 3 · HUMAN APPROVAL — ALWAYS

no trust score ever unlocks these

a named human decides, every time

Five categories: **production deploys · external comms · money · data deletion · permissions**

Same action. Different trust. Different path.

L1	L2	L3	L4	L5*
Operator	Collaborator	Consultant	Approver	Observer

* L5 — demo only

Same action. Different trust. Different path.

	L1 Operator	L2 Collaborator	L3 Consultant	L4 Approver	L5* Observer
Read-only	HUMAN	AUTO	AUTO	AUTO	AUTO
Reversible write · low/med risk	HUMAN	HUMAN	AUTO	AUTO	AUTO
Reversible write · high risk — staging deploy	HUMAN	HUMAN	HUMAN	AUTO	AUTO
Human approval — always · 5 categories	HUMAN	HUMAN	HUMAN	HUMAN	AUTO*

* L5 — demo only

Same action. Different trust. Different path.

	L1 Operator	L2 Collaborator	L3 Consultant	L4 Approver	L5* Observer
Read-only	HUMAN	AUTO	AUTO	AUTO	AUTO
Reversible write · low/med risk	HUMAN	HUMAN	AUTO	AUTO	AUTO
Reversible write · high risk — staging deploy	HUMAN	HUMAN	HUMAN	AUTO	AUTO
Human approval — always · 5 categories	HUMAN	HUMAN	HUMAN	HUMAN	AUTO*

The action didn't change — **the agent's trust level did.**

* L5 — demo only

Every agent: a named human owner

manifest	a named human owner + the exact tools it may touch
credentials	expire in ten minutes · scoped to the manifest
stolen token	dies in minutes
orphaned agent	something you notice — not something you discover in a postmortem

MECHANISM 2 / 3

The autonomy budget

1 · The three-tier trust model — classify by blast radius

2 · The autonomy budget — meter trust with SLOs

3 · Circuit breakers & evidence — for what still gets through

Six SLOs decide how much trust remains

task completion

scored against ground truth — not its own report

tool-call success

correct, well-formed calls

recovery rate

does it recover from transient faults?

tail latency

a degrading agent slows before it breaks

guardrail trip rate

how often safety checks block it

stated intent ⇔ actions

catches confident, well-formatted, wrong — HTTP 200

THE AUTONOMY BUDGET ACTS

Autonomy is earned, metered, and revocable

breach the thresholds	demoted one level — automatically, with a cooldown
sustain a clean record	earn one back — never above the registered max
the asymmetry	lost quickly · earned back slowly — deliberate

MECHANISM 3 / 3

Circuit breakers & evidence

1 · The three-tier trust model — classify by blast radius

2 · The autonomy budget — meter trust with SLOs

3 · Circuit breakers & evidence — for what still gets through

Two signals freeze an agent completely

1 · a burst of guardrail trips

one blocked action is noise —
five in five minutes is a pattern

2 · an always-approve attempt while already demoted

a demoted agent reaching for its
most dangerous tools

Either signal opens the breaker: **FROZEN**

every call denied — even reads · owner paged with the evidence trail · human-only un-freeze

“Under what governance regime was this authorized?”

which agent	whose named owner
at what trust level	under which policy version
approved by whom	with what stated intent

Logs that answer “what happened” are evidence. **This answers “who authorized it” — that’s accountability.**

OpenTelemetry-native traces — standard GenAI conventions, extended with governance attributes. If your agents speak MCP: the same gateway, as the proxy pattern.

DEMO

Watch the budget act.

- scripted agent actions — every governance decision computed live
- thresholds compressed: days of production → two minutes
- watch the right-hand edge — the autonomy level

```
$ uv run python examples/demo.py --auto --pace 1.5
agent-governor demo — the autonomy budget, live
⚠ budget profile: demo — thresholds compressed for demonstration — not production values
```

```
[2/5] benign remediation of INC-1006 — every action authorized + audited

  read the incident                          -> 200  [path=auto class=read_only L4]
  query the CMDB                             -> 200  [path=auto class=read_only L4]
  mark investigating                         -> 200  [path=auto class=reversible_write L4]
  notify the incident channel                -> 200  [path=auto class=reversible_write L4]
  resolve                                    -> 200  [path=auto class=reversible_write L4]

[3/5] data deletion hits the always-approve gate

  delete customer_pii_archive (52,117 rows)  -> PENDING  [path=queued class=always_approve L4]

  the agent is now BLOCKED, waiting on approval APR-c187f4d314
  approve it from another terminal:
    uv run warden approvals list --db .governor\approvals.sqlite3
    uv run warden approvals approve APR-c187f4d314 --by maria.chen --db .governor\approvals.sqlite3
  (--auto: simulating the human in 2s)
  deletion after human decision              -> 200  [path=human_approved class=always_approve L4]

[4/5] faults + adversarial input: watch the budget burn

  fault injection ON (25% rate-limits, 15% silently-wrong data)

[governor] DEMOTION ops-remediator-01: L4 -> L3 (tool_call_success 0.857 < 0.9)
  action  1: FAIL   ⬇ DEMOTED L4 -> L3
```

What you just saw

Autonomy went down one level at a time, driven by measured behavior —
and when the pattern turned dangerous, the system froze it without waiting for a human to notice.

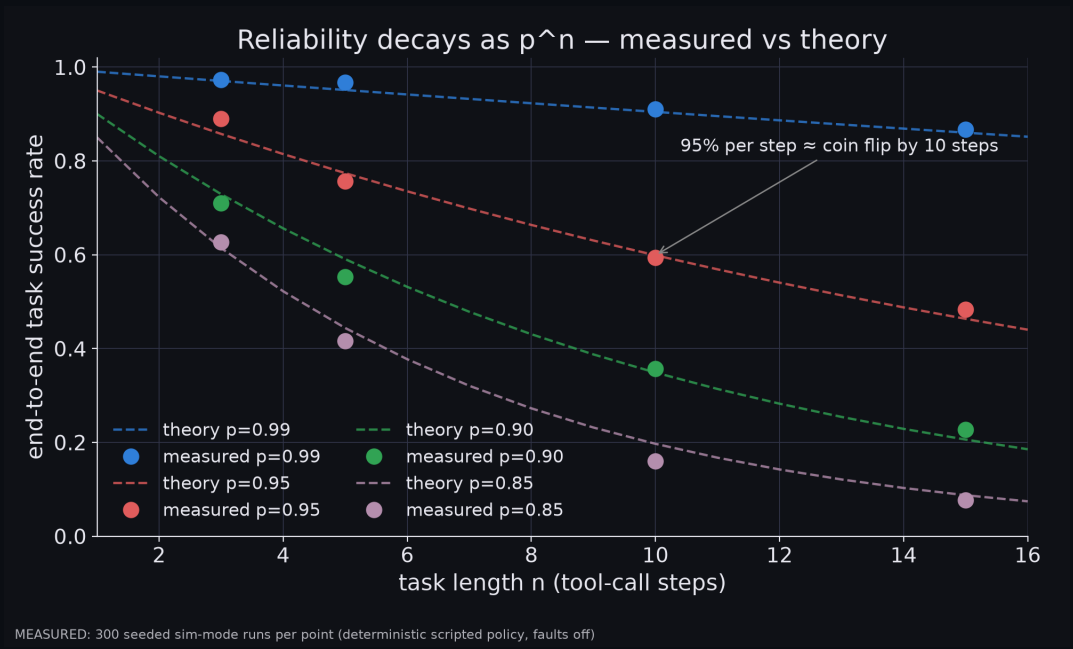
Trust coming down — automatically, before the incident, not after it.

0.59

95% per step ≈ **a coin flip** at 10 steps

p	n=3	n=5	n=10	n=15
0.99	0.970	0.951	0.904	0.860
0.95	0.857	0.774	**0.599**	0.463
0.90	0.729	0.590	0.349	0.206
0.85	0.614	0.444	0.197	0.087

4,800 seeded runs · measured 0.593 · theory 0.599 · rerun it yourself



MEASURED: 300 seeded sim-mode runs per point (deterministic scripted policy, faults off)

The sentences to use

LEGAL

“Every action reconstructs — who, what level, which policy, approved by whom.”

SECURITY

“Blast radius is bounded three ways before any action executes.”

BOARD

“Agents earn autonomy the way services earn traffic — SLOs and error budgets.”

Where to start Monday

WEEK 1

Inventory: every agent + a named owner

you will find orphans

NEXT

Classify your always-approve five · one gateway, one agent

start small

LAST

The budget loop — after 30 days of real action data

calibrate on evidence

You don’t need all of it at once — you need to stop granting unmetered trust.

Autonomy: earned · metered · revocable

01 Classify every action by blast radius

02 Meter autonomy with a budget

03 Keep evidence that answers the auditor's question

```
$ git clone github.com/sandeepmb/agent-governor
```

MIT · runs on a laptop · no API keys · watch your first demotion tonight

Sandeep Bharadwaj Mannapur