LAB L2 LIVE AI SAFETY & GOVERNANCE

Trustworthy AI Toolkit

Red-team - Eval - Lineage - EU AI Act

Bias auditing, red-teaming and policy guardrails baked into the model lifecycle.

Thesis: If safety is a checklist at release, you've already lost. It must be a CI gate on every commit.

AUDIT-READY RELEASES	RED-TEAM COVERAGE (MITRE ATLAS)	TIME TO REGULATOR DOSSIER
100%	+ 4.1	9d 0d

MANIFESTO

Why this lab exists

Trustworthy AI cannot be a slide deck. The Toolkit ships continuous red-teaming, bias and toxicity evals, model lineage and signed releases - all wired into CI so a model that fails audit literally cannot reach production. Compliance becomes a build artefact.

KPIS

Outcomes we measure

- Audit-ready releases: 100%
- Red-team coverage (MITRE ATLAS): + 4.1
- Time to regulator dossier: 9d 0d

ACTIVE EXPERIMENTS

What the lab is testing now

> Continuous prompt-injection red-team
Adversarial agents probe production endpoints daily across MITRE ATLAS coverage.

> Bias suites per protected class
Statistical parity, equalised odds, calibration - measured on every model card revision.

> Lineage proofs
Cryptographically signed chain from data features model deploy response.

> Policy-as-code guardrails
Rego rules enforced at the model gateway: residency, consent, retention, exit-list filtering.

SHIPPABLE ARTEFACTS

Everything that ships

> Eval harness
Quality, bias, toxicity, jailbreak and ATLAS suites running on every commit.

> Red-team agents
Continuous adversarial probes with severity scoring and auto-tickets to detection-as-code.

> Lineage plane
Signed graph of every artefact in the model lifecycle, queryable for audit.

> Regulator dossier
Auto-generated EU AI Act, NIST AI RMF and ISO 42001 evidence packs, updated continuously.

> Policy gateway
Rego rules at the model gateway: residency, consent, retention, response filters.

LAB TEAM

Who you'll work with

- Responsible AI Principal
- Adversarial ML Lead
- Policy & Legal Engineer
- Lineage / MLOps Lead

ENGAGEMENT TIMELINE

Weeks 1-6 - first regulator dossier signed off by week 4

1. **Weeks 1-2 - Baseline + lineage**
   Wire eval harness, lineage capture, model cards across the production fleet.
2. **Weeks 2-4 - Red-team + dossier**
   Continuous adversarial probes, severity scoring, first regulator dossier signed.
3. **Weeks 4-6 - CI gates live**
   Quality, bias and ATLAS gates fail-build on regressions; exec scorecard live.

FLAGSHIP PODS

Squads that productionise this lab

- EU AI Act Readiness Pod
- Red-Team Continuous Pod
- Bias & Fairness Pod
- Model Lineage Pod

PARTNERS

Who we collaborate with

NIST AI RMF - MITRE ATLAS - Hugging Face - OpenAI - Anthropic - Microsoft Responsible AI

PUBLICATIONS

Receipts

Continuous red-teaming reduces jailbreak success by 78% at iso-cost
USENIX Security Workshop - 2025

From policy doc to Rego: making AI Act controls executable
AXP Internal Whitepaper - 2026

FAQS

What partners actually ask

Q. Is this a one-off audit?
A. No - it's a continuous control plane. Every commit re-runs evals, red-team and lineage; the dossier auto-updates.

Q. EU AI Act ready?
A. Yes - high-risk and limited-risk obligations are mapped, with Annex IV evidence pack auto-generated.

Q. Does it slow shipping?
A. The opposite. Failing fast in CI is far cheaper than failing in front of a regulator.

Q. Can we use our own evals?
A. Yes - the harness is plug-in. Bring HELM, Big-Bench Hard, internal golden sets or domain suites.

Partner with Trustworthy AI Toolkit

Outcome-priced. Sovereign by default. Refund-backed if the contracted KPI isn't hit.

Apply: alfaxprienz.com/labs/trustworthy-ai-toolkit#partner