

Imagination With Levers: A Foresight Lab for Enforceable AI Governance Futures

Joël Naoki Christoph · 10Billion.org · joel.christoph@proton.me

An imagined future lasts only when someone gains from keeping it and loses from breaching it. **This lab builds that test into the foresight session itself.**

Why foresight needs an enforcement test

Foresight produces vivid images of AI futures. Most go no further than a report, while the rules that decide which future arrives get written in budgets and regulation. This 90-minute format sends each imagined future out of the room as a draft institution, with the incentives that would keep it alive already attached. It keeps the open, judgement-free spirit of a Futures Literacy Lab (Miller, 2018) and adds a closing stage that the lab tradition has left out.

The lab in three stages

1

Imagine

Participants write one short future of AI in 2040. Myths and fears count as much as hopes. Critique is switched off.

Artefact: a one-page future



2

Translate

The group identifies who would bring that future about and who would resist it, and where each side draws its power.

Artefact: an actor map



3

Test

The group attaches a rule, an incentive to follow it, and a way to verify compliance, then hunts for the point where self-interest would defeat the future.

Artefact: an instrument sketch, a register of assumptions, three early-warning indicators

Limits

Critique is banned until stage 3 because early testing kills fragile ideas. A mechanism can be coherent and still serve a poor future, so the future itself stays open to challenge until the end. Calibrations appear as comparative scenarios, never as forecasts.

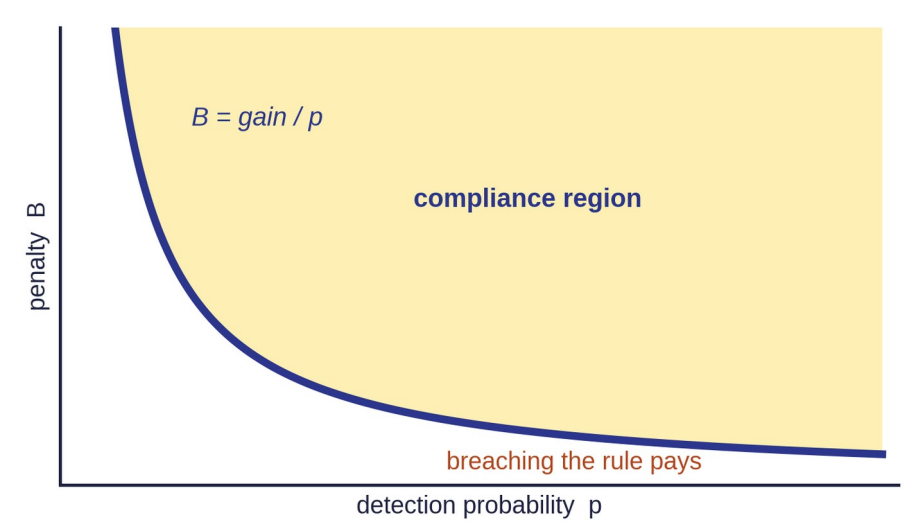
Worked example: a licensed frontier

The future, as written in stage 1

“By 2040, training a frontier AI model needs a permit, the way flying an aircraft needs certification. Permits are scarce, priced, and public.”

Stage 2 makes it a design. A regulator sets a period cap on high-risk compute. Permits trade within the cap, so compute goes to the projects that value it most. Audits check what developers report.

Stage 3 tests it against self-interest. A developer complies when the expected penalty exceeds the gain from breaching the rule: $B \geq \text{gain} / p$, where B is the penalty and p the chance that an audit detects the breach.



The design task becomes choosing a point on this curve. Heavier penalties can compensate for weaker audits, up to the point where a penalty stops being credible, so audit capacity sets the floor of any enforceable future.

Two futures, one anatomy

	Licensed frontier	Reviewable decisions
Future, 2040	Frontier training needs a permit	Anyone can demand a human review of an automated public decision
Rule	Cap on high-risk compute	Per-decision fee paid into an appeals fund
Incentive	Permits are scarce and tradable	Reported decision counts set the fee
Verifier	Metering and random audits	Sampled re-audits of decisions
Weakest point	Hidden or offshore training	Under-reporting decision counts
Fix	Penalty large relative to detection chance	Penalty on under-reporting; same inequality

The same inequality governs both. That is what the format offers: an enforceability test that works on any imagined future, whatever its subject.

Run the lab. Ninety minutes, 8 to 25 participants, paper and a wall. The materials are open. Write to joel.christoph@proton.me to host a session or to read the mechanism papers behind both examples.

References. Miller, R. (2018) Transforming the Future: Anticipation in the 21st Century, UNESCO. · Sastry, G., Heim, L., Belfield, H. et al. (2024) Computing Power and the Governance of AI, Centre for the Governance of AI. · Petrie, J., Aarne, O., Dalrymple, D. (2024) Mechanisms for Flexible Hardware-Enabled Guarantees. · Scher, A. and Thiergart, L. (2024) Mechanisms to Verify International Agreements About AI Development, MIRI. · Christoph, J. (2026) Risk-Weighted Compute Permit Markets under Imperfect Monitoring, SSRN Working Paper 6361077.