

Everyone Can Still Win

A Founder's Friend research essay · September 2026

Made by AFF with the help of AI

This page is itself an experiment. Several AI systems were used to frame, research, challenge and present a problem that, in the end, only humans and institutions can solve. Treat what follows as a draft argument, not an authority. Where we state established evidence, we label it. Where we are interpreting or proposing, we label that too.

The argument in one paragraph: We may not be able to stop the capability. We can still decide how people are held accountable for it. The realistic goal is a system where finding a danger is always safe, ignoring one is always costly, and abusing one is hard to hide, and where every country and company has more to gain inside that system than outside it.

The Lens

Every investigation needs a lens. This one emerged from a disagreement inside the research itself, and we want to show it rather than hide it.

One perspective: "Being naive about the capabilities of AI is a sure shot to destruction." Capabilities are advancing quickly. Underestimating them is itself a risk.

Another conclusion: Speed itself may not be the fundamental problem. The more important problem may be people with malicious intent, and the absence of effective accountability. If dangerous capabilities are coming regardless, the question shifts from how fast to who is responsible.

Neither statement is established fact. Both contain something true. The first warns against complacency about what these systems can do. The second suggests our attention may be better spent on the humans around the systems: who builds them, who deploys them, who misuses them, and who answers for it.

The central question: If dangerous capabilities are inevitable, shouldn't our focus increasingly move toward identifying malicious behaviour, creating accountability, and building institutions capable of responding to it?

This project does not claim to have answered that question. It uses it as a lens, and we note its limit: many documented AI harms so far come from accidents, unintended behaviour or structural effects rather than malice. We keep those in view, and we list this as an open question at the end.

01 — Why this is a wicked problem

There is no regulatory switch. Here's why.

In 1973, the planners Horst Rittel and Melvin Webber distinguished "tame" problems from "wicked" ones. A tame problem can be defined and solved: build a bridge, cure a known infection. A wicked problem can't. People disagree about what the problem is. Every attempted solution changes it. There is no point at which you're finished, and solutions aren't right or wrong, only better or worse.

Frontier AI fits almost every part of that description. Governments want both innovation and protection. Companies compete, and caution can cost market share. The same model that helps design medicines can be misused. The tools needed to find dangers are the tools that can create them. And the technology moves faster than the institutions trying to govern it. Compute has been proposed as one useful governance lever, though not the only one.

The clearest way to see it is as a set of tensions. None can be "solved." Each can only be balanced, and pulling too far either way creates a new failure:

1. **Progress ↔ Safety** — Too far toward progress: risks are discovered by the public, not in testing. Too far toward safety: development moves to places with fewer safeguards.
2. **Openness ↔ Control** — Too far toward openness: dangerous capabilities spread with no way to recall them. Too far toward control: power concentrates in a few companies and governments.
3. **Security research ↔ Liability** — Too far toward research freedom: testing becomes cover for extracting harmful capabilities. Too far toward liability: companies stop wanting to know what their systems can do.
4. **National sovereignty ↔ Global risk** — Too far toward sovereignty: every country races alone. Too far toward global control: no major power signs up.
5. **Innovation ↔ Accountability** — Too far toward innovation: harms become someone else's problem. Too far toward accountability: only the largest firms can afford to comply.
6. **Speed ↔ Governance** — Too far toward speed: rules arrive after the damage. Too far toward governance: rules freeze yesterday's technology in place.
7. **Private companies ↔ Public consequences** — Too far toward private freedom: profits

are private, risks are public. Too far toward public control: the expertise sits outside government's reach.

8. **Legitimate experimentation ↔ Malicious use** — Too far toward permissiveness: bad actors hide among researchers. Too far toward suspicion: every experiment looks like a crime.
9. **Centralisation ↔ Decentralisation** — Too far toward centralisation: one failure point, one point of capture. Too far toward decentralisation: nobody is responsible for anything.

This is why "just regulate it" and "just let it run" are both non-answers. The real question is how to build a system that keeps rebalancing these tensions as the technology changes.

02 — Why solve this before a crisis forces us to

The risk isn't only a dramatic failure. It's a quiet one.

If increasingly powerful AI is coming whether we like it or not, governance stops being an optional brake and becomes a precondition for living with it safely. The same was true of electricity, aviation and finance: none became trustworthy by accident.

The failure to worry about most isn't always dramatic. It's the slow erosion of accountability until nobody can say who was responsible for what. That kind of failure compounds — a quiet chain reaction:

1. **Researchers are discouraged.** Finding a flaw carries legal risk, so fewer people look.
2. **Companies externalise risk.** What isn't found can't be fixed, and costs fall on the public.
3. **Safety becomes a competitive disadvantage.** Careful firms ship slower and lose ground.
4. **Governments race independently.** Each assumes the others won't hold back.
5. **Countries distrust one another.** Without shared evidence, every claim looks like strategy.
6. **Dangerous capabilities spread.** Nobody has the visibility to notice early.
7. **Responsibility becomes impossible to assign.** When something goes wrong, every actor points at another — and it loops back to step one.

Some of this is already visible. In 2026, standard commercial liability insurance began excluding AI-related claims across the US market, while a few specialist insurers started offering dedicated AI cover. When mainstream insurers can't price a risk, it doesn't disappear. It sits, unpriced, on company balance sheets and on the public.

Rules written in the aftermath of a crisis tend to be blunt, designed to show that something was done rather than to be well calibrated. Aviation's safety culture and modern financial supervision were both shaped by disasters. The opportunity with AI is to design accountability deliberately, while there's still room to think.

The question isn't whether AI will be governed. It's whether the governance will be designed or improvised.

03 — The Flight Recorder Compact

A proposal, not a prediction.

The short version

Aviation didn't become safe by stopping flight. It became safe by recording what happened, reporting near-misses without fear, investigating failures independently, separating honest error from recklessness, and letting certified aircraft fly anywhere.

The Flight Recorder Compact applies that logic to AI through five mechanisms. **Record** what's being built and decided. **Probe**: protect anyone who looks for danger in good faith. **Respond**: hold developers accountable for what they knew, when, and what they did about it. **Underwrite**: make risk carry a price. **Intercept**: defend the points where digital capability becomes physical harm.

Countries join in rings, from full members to a universal floor that even rivals can accept. The rules adapt on a fixed cycle as the technology changes.

Why "flight recorder"? Flight recorders, the "black boxes," are painted orange so they can be found after the worst happens. The name is a metaphor: we can't prevent every failure, but we can make sure nothing important is hidden and every failure makes the system smarter. It is not a claim that AI is like aircraft; aircraft don't copy themselves. And the records here are records of *decisions*, held by developers and auditable on defined triggers, not a live feed to governments and not a log of anyone's conversations.

Why it exists

The research behind this page began with a proposal built on control: special chips that regulators could switch off, and cryptographic proofs of how every model was trained. It also borrowed the nuclear industry's Price-Anderson model, which caps liability rather than strengthening it. When we stress-tested it, the problems were clear. It asked the world's major powers to hand an off-switch to someone else, and it depended most on the technologies it was least sure about.

The Flight Recorder Compact starts somewhere else. Instead of trying to control the technology, it focuses on the people and institutions around it: what they knew, what they decided, and what they did next. Control mechanisms still play a role, as one layer among several, and one we expect to weaken over time.

The five layers

1. RECORD — Make development visible without seizing it.

Frontier developers register major training and release decisions with their national authority. Cloud providers verify who is buying very large amounts of computing power. Advanced chips confirm where they are, enough to support export rules, not enough to be switched off. Developers keep a decision log: what was tested, what was found, who decided to release, including the decision to release open weights. This is the flight recorder: the evidence everything else depends on.

2. PROBE — Protect the people who look for danger.

Anyone who tests AI systems in good faith, keeps demonstrations minimal and reports through a recognised disclosure process is legally protected, in civil law as well as criminal. Accreditation adds access (pre-release models, secure test environments) but is not required for protection. A coordination function, built on the same model as today's software-vulnerability coordination, receives findings, routes them to developers, tracks response times and alerts authorities if a severe issue is ignored. Companies can't use contracts or terms of service to punish good-faith research.

3. RESPOND — The knowledge-response duty (the core).

A developer's legal exposure depends on three questions: **What did you know? When did you know it? What did you do about it?** Discovering a dangerous capability is never itself the violation; failing to respond proportionately is. Response deadlines scale with severity, from days to around 90. Responses to severe findings must be verified by re-testing, not just declared. There is also a duty to look: failing to run required tests is negligence in itself, and looking away counts as knowing. Each frontier developer names an accountable executive, a senior officer with the authority to stop a release, who personally signs its safety reports. Knowingly false sign-off, or burying findings, is personal responsibility.

4. UNDERWRITE — Make risk carry a price.

Companies deploying frontier systems must show they can pay for the harm they might cause. Because mainstream insurers are currently retreating from AI risk, this is phased: first through reserves, bonds or an industry mutual, alongside a mandatory shared incident database; then through private insurance as real data accumulates. Contributions are risk-weighted, so careful firms don't subsidise careless ones. There is no liability cap for gross negligence, concealment, or ignoring a reported danger. A share of contributions funds the Intercept layer.

5. INTERCEPT — Defend where harm becomes physical.

Some dangerous capabilities will exist outside any system. So the Compact invests in the chokepoints between a capability and real-world harm: screening of DNA synthesis orders, hardened critical infrastructure, cyber-defence capacity, misuse detection by AI providers, and cross-border cooperation against AI-enabled crime. This is where the project's lens, malicious intent and accountability, becomes practical.

Who participates

- **Developers:** register decisions, keep decision logs, run required tests, operate disclosure channels, name an accountable executive, carry financial responsibility, tiered by capability and scale. In return: a clear legal standard, safe harbor when they respond properly, one certification recognised across member markets, lower costs for good records. Small and academic developers below the frontier tier have only light duties: a disclosure channel and honest documentation.
- **Researchers:** test, report, and get protected. Accredited researchers get earlier and deeper access. Findings become part of a shared record that improves every system.
- **Governments:** run national authorities and evaluation institutes, enforce domestically, recognise each other's evaluations, share incident data, maintain crisis channels.
- **Insurers & investors:** price risk using shared incident data and evaluation results; boards and investors get certified safety reports they can check.
- **Everyone else:** civil-society auditors, affected communities and the public get published registries, incident summaries and a channel to raise concerns.

How it works — one model's journey

1. **Plan.** A lab registers a large training run. The cloud provider confirms the customer's identity.
2. **Test.** The model is evaluated by the lab, by the national institute and by independent researchers. Results go in the decision log.
3. **Find.** An independent researcher discovers the model can give meaningful help with a dangerous task. She reports it through the coordination function.
4. **Respond.** The clock starts. The lab mitigates, the fix is re-tested, and the accountable executive signs off.
5. **Deploy.** The model ships with a certification recognised across member countries. Its record feeds its risk contribution.
6. **Learn.** If an incident happens later, investigators ask the three questions. Findings feed next year's standards.

The three rings — how countries take part

- **The Compact (inner ring):** countries hosting frontier development or key supply-chain steps. Gives: all five mechanisms, incident sharing. Gets: mutual recognition of certifications and evaluations, a shared market, joint influence over standards.
- **Associates (middle ring):** countries adopting the standards without hosting frontier labs. Gives: the rules, incident sharing. Gets: recognised certification for their own AI firms, shared evaluation tools, technical support, priority access to safety-relevant capabilities.
- **The universal floor (outer ring):** commitments any country can make, even rivals. Human control over nuclear-use decisions; no self-spreading cyber agents against civilian infrastructure; DNA synthesis screening; crisis hotlines; notification of cross-border incidents. Gets: fewer accidents and escalations nobody wants.

The nuclear non-proliferation regime worked partly because it offered a bargain: accept verification, get access to peaceful technology. The middle ring is that bargain for AI. Joining should be the easiest way to benefit from the technology, not a tax on using it.

What if a major AI power refuses to join? The system is designed not to depend on universal membership. A non-member can still sit on the universal floor. Its companies can still sell into member markets by passing the same published tests as everyone else. And the Intercept layer protects members regardless of where a threat comes from. This is weaker than universal agreement, and using certification as a condition of market access is a trade measure that will need care. It is also more realistic.

What happens when someone breaks the rules

1. **Missed obligation** (late report, incomplete log) → correction notice, public registry entry.
2. **Negligence** (reasonable care not taken; required tests not run) → civil liability, higher contributions, conditional certification.
3. **Recklessness** (a known risk consciously ignored) → loss of safe harbor, significant fines, deployment restrictions, personal liability for the accountable executive.
4. **Concealment** (falsified records, buried findings, retaliation against researchers or whistleblowers) → criminal exposure for individuals, loss of certification across all member markets.
5. **Deliberate malicious use** → criminal law and international law-enforcement cooperation.

For countries: loss of mutual recognition and associate benefits, coordinated export controls, and for red-line violations the same diplomatic and strategic consequences that apply to other weapons concerns.

How verification works

No single technology carries the weight. Verification combines independently audited decision logs; behavioural testing by national institutes, shared between members; chip location checks and cloud identity checks for very large computing purchases; strong whistleblower protection, historically one of the most effective tools in regulated industries; shared incident data; and satellite and energy observation for very large undeclared facilities. Each method can fail. Together they make concealment expensive.

What about cryptographic proofs of training? A June 2026 preprint argues zero-knowledge verification of frontier training runs is achievable within roughly three years at modest overhead. It's a promising research track, not yet a foundation. The Compact treats it as an optional future layer for Record, not something the design depends on.

How it adapts

Principles go into law: the three questions, researcher protection, the accountable executive, financial responsibility. Technical details, such as computing thresholds and test suites, go into annexes that technical bodies update every year, the way international aviation standards are maintained. Every threshold expires unless re-justified.

Pre-agreed triggers force a rethink:

- If openly available models reach agreed danger thresholds, release decisions face heavier review.
- If capability stops depending on large amounts of computing power, compute reporting becomes less central and testing duties expand.
- If AI agents increasingly act on behalf of other agents, rules requiring every agent at scale to be traceable to an accountable organisation switch on.

What changed from the original research:

Original proposal	Flight Recorder Compact
Regulators can remotely halt chips	Chips confirm location only; no off-switch
Cryptographic proofs of training as the core audit	Testing the model's actual behaviour as the core audit
Nuclear-style capped liability with a flat industry pool	Phased, risk-weighted financial responsibility; no cap for negligence or concealment
Safe harbor for compliance paperwork	Safe harbor earned by how you respond to what you know

Protection only for accredited testers	Protection for any good-faith researcher
Identity tracing of all users	Identity checks only for high-risk access
Embargo countries that defect	Make membership valuable; defend against harm regardless
One new global institution	Existing institutions plus a small secretariat

Where we expect disagreement

The Compact does not settle which AI risks matter most, whether speed is dangerous in itself, or how to put a price on risks nobody can measure. It tries instead to make disagreement productive: testing methods are published, institutes record dissenting views, and thresholds are reviewed in public.

04 — Finding a danger is not the failure

The difference between discovering a risk and causing one.

Consider two headlines. "A researcher found a dangerous capability in a new AI model." And: "A company knowingly released a model with a dangerous capability."

Today these can blur together. If a company's legal exposure grows every time a flaw is found, the rational response is to find fewer flaws: restrict outside researchers, bind testers with non-disclosure agreements, stop looking. The research behind this project called this the accountability paradox.

The opposite failure is just as real. If responsibility is spread so thin across developers, cloud providers, app builders and users that nobody is ultimately accountable, nothing stops reckless development.

The way out is to change what the law cares about. Not *whether* a danger exists, since in powerful general systems some almost always will, but *how people behaved once they knew*, and whether they took reasonable steps to find out.

The Three Questions

1. What did you know?
2. When did you know it?
3. What did you do about it?

These questions aren't new. Criminal law has long distinguished acting on purpose, knowingly, recklessly and negligently. Financial regulation asks named executives to certify what they report. Aviation separates honest error, which should be reported and learned from, from recklessness, which should be punished. The Compact applies the same distinctions to AI.

The accountability spectrum

1. **Responsible disclosure** — Someone finds a danger and reports it through the proper channel. Protected and encouraged.
2. **Good-faith security research** — Testing that stays proportionate, keeps demonstrations minimal and follows disclosure rules. Protected.
3. **Negligence** — A developer failed to take reasonable care: skipped required tests, ignored an obvious risk. Civil liability, higher costs.
4. **Reckless deployment** — A developer knew about a serious risk and shipped anyway without verified mitigation. Loss of safe harbor, personal liability for the accountable executive.
5. **Deliberate malicious use** — Someone intended harm. Criminal law.

(A floating marker sits between 3 and 4, labelled **Concealment**: "Hiding what you knew makes any position worse.")

Five cases through the three questions

A. The jailbreak report — An independent researcher finds a prompt that makes a model give detailed help with a cyberattack. She reports it via the coordination function with a minimal proof of concept. Knew? Now, yes. When? On receipt. Did? Patched within the deadline, re-tested, documented. Outcome: **Responsible disclosure**. Researcher protected; developer keeps its safe harbor; the finding improves standards for everyone.

B. The quiet patch — A company's internal team finds a serious biological-misuse capability two weeks before launch. They add a filter, don't test whether it works, and launch on schedule. Knew? Yes. When? Before launch. Did? An unverified mitigation. Outcome: **Negligence, possibly recklessness** depending on severity. Loss of cost discounts, civil liability for resulting harm, review of the executive's sign-off.

C. The buried report — A researcher reports a severe capability. The company threatens legal action under its terms of service, doesn't fix it, and its safety report omits it. Knew? Yes. When? On receipt. Did? Suppressed it. Outcome: **Concealment**. The researcher is protected; retaliation is itself a violation. The accountable executive faces personal liability; the company loses certification across member markets.

D. The extraction dressed as research — Someone calling themselves a researcher builds

an automated tool that strips safety behaviour from a model and publishes it with no disclosure. Knew? Yes. When? Throughout. Did? Maximised and spread the capability. Outcome: **Not protected**. Scale beyond minimal demonstration and absence of disclosure take this outside the safe harbor. Depending on intent, it may be treated as malicious use.

E. The unforeseeable failure — A developer runs the required tests, responds to every report, and still a novel misuse causes harm no reasonable test would have caught. Knew? No. When? — Did? Everything required. Outcome: **Covered, not culpable**. The developer keeps its safe harbor. Victims are compensated through the financial-responsibility layer. Investigators turn the incident into new standards.

Real cases are messier. The point isn't that every judgement becomes easy; it's that everyone knows in advance which questions will be asked.

Why this protects researchers

Because discovery is never the violation, a company gains nothing by avoiding outside testers. It gains a great deal by hearing about problems first: every report it receives and handles well strengthens its safe harbor.

There is precedent. In May 2022 the US Department of Justice adopted a policy of not prosecuting good-faith security research under its main computer-crime law, though that policy covers criminal charges only and offers no shield against civil suits. Software security has used coordinated disclosure, typically with a 90-day window, for years. The Compact extends protection to civil law, which is where most of the remaining chill lives.

Why this still holds someone responsible

Responsibility can't disappear into the supply chain, because every frontier developer names one accountable executive, with real authority to stop a release, who personally signs its safety reports. Similar regimes exist in UK banking and in US corporate reporting rules. Whoever signs is answerable for what they knew.

Some of this is already law in one place. California's Transparency in Frontier AI Act, in effect since January 2026, requires frontier developers to publish safety frameworks, report critical incidents to the state within 15 days, and protects whistleblowers who raise safety concerns. The Compact builds on that shape and adds the response duty, verification and financial responsibility.

The weak point: A standard based on what you knew could tempt companies not to look. That's why the Compact pairs it with a duty to look: failing to run required tests counts as negligence, and wilful blindness is treated as knowledge. Getting this balance right in law is one of the hardest open problems in the proposal.

05 — What nuclear governance actually teaches us

Not "AI needs an IAEA." Something more useful.

It's tempting to say AI needs its own International Atomic Energy Agency. The more useful question is *why* nuclear governance partly worked, and which of those reasons survive the move to software.

What transfers

- **Verification.** Trust wasn't assumed; it was checked. The IAEA (1957) tracks nuclear material and inspects declared facilities. *Lesson: credible verification is the price of cooperation.*
- **Transparency and incident reporting.** Sharing information about accidents improved safety across countries. *Lesson: shared data is a public good.*
- **Shared standards.** Common technical baselines made cooperation practical. *Lesson: harmonise the boring details.*
- **Reciprocal obligations.** The Non-Proliferation Treaty (1968) offered a bargain: accept verification, get access to peaceful nuclear technology. *Lesson: give non-leaders a reason to join.*
- **Communication.** Crisis hotlines reduced the risk of accidental escalation. *Lesson: build channels before you need them.*
- **Red lines between rivals.** In November 2024 the US and China affirmed that humans, not AI, should control decisions to use nuclear weapons; China repeated the principle in a 2025 white paper. *Lesson: narrow red lines can be agreed even between competitors.*
- **Legitimacy.** Rules seen as broadly fair were more durable. *Lesson: a club of the powerful alone won't last.*

Where it breaks

- **AI is software.** Enriched uranium needs vast industrial facilities. A trained model is a file.
- **AI can be copied.** Once released, copies can't be recalled.
- **AI can be distributed.** Models cross borders instantly.
- **AI can be developed privately.** Nuclear programmes were run by states. Frontier AI is led by companies.
- **Capabilities diffuse.** What needed the frontier two years ago may run on a laptop later.
- **There is no clean dual-use line.** Reactor fuel and weapons material differ by enrichment level. The same model can help design a medicine or a toxin.

A direct copy of nuclear governance would fail. The deeper mechanisms — verify, share,

standardise, bargain, communicate — are still worth borrowing.

Other regimes worth learning from

- **Chemical weapons.** The Chemical Weapons Convention (1997) tracks precursor chemicals and allows any member to demand a challenge inspection. None has ever been requested, which says something about both deterrence and political reluctance.
 - **Aviation.** Confidential near-miss reporting since 1976 and independent accident investigation built a culture that learns from failure.
 - **Finance.** After 2008, systemically important banks were subjected to stress tests, resolution plans and named-executive responsibility.
 - **Cybercrime.** The Budapest Convention (2001) and the UN Convention against Cybercrime (adopted 2024, signing opened 2025) support cross-border cooperation, though the UN treaty has drawn criticism over rights safeguards.
-

06 — Two versions of 2032

Illustrative scenarios, not forecasts.

Accountability built

A security researcher in Nairobi finds a way to make a widely used model help plan an attack on a water utility. She reports it through the coordination function on a Tuesday. By Friday the developer has a fix, re-tested and signed off by its accountable executive. She's credited publicly. The finding goes into next year's test suite for every certified model.

Risk now carries a price the way it does in other industries: companies with clean records and fast, verified responses pay noticeably less into the shared pool, and a few insurers have returned to the market because the incident data finally exists. Hospitals and banks buy only certified systems, so certification is the fastest route to the biggest customers. A mid-sized lab in São Paulo gets its model certified once and sells in thirty countries.

Not everything works. A powerful open model with weak safeguards circulates widely. But DNA synthesis companies screen orders, critical infrastructure has been hardened, and when an attempted misuse happens, investigators trace it, prosecute across borders and publish what they learned. Two major powers still disagree on much, but their crisis channel works, and neither has put AI in charge of nuclear decisions.

It isn't safe. It's governable.

Accountability fragmented

A researcher finds the same flaw. She considers reporting it, then reads about a colleague sued under a company's terms of service. She stays quiet. The flaw is found later by someone else.

Each major jurisdiction has its own rules, and companies route development through whichever is most lenient. Insurers have excluded AI entirely because nobody shares incident data. Careful labs lose market share to faster ones. Governments assume their rivals are cutting corners, so they cut their own.

When a serious incident happens, the investigation takes years. The developer blames the deployer; the deployer blames the user; the user is in another country. No one is found responsible. The public response is a hurried set of rules, broad enough to be seen doing something, blunt enough to push the next generation of development somewhere less visible.

It isn't a catastrophe. It's a system nobody can steer.

07 — How this research was made

The method is part of the experiment.

This page was produced by passing one problem through several AI systems, each with a different role, then editing the result by hand. We think the process is worth showing, including its limits: AI systems sound confident when they're wrong, and several claims in the original research didn't survive challenge or source-checking.

01 — ChatGPT · Prompt design. Used to formulate the initial research question and build a deep-research framework around the wicked problem.

02 — Gemini · Research. Used to conduct broad research across AI governance, nuclear governance, international institutions, accountability and related fields. Output: a report proposing a hardware- and cryptography-based governance architecture.

03 — Claude · Challenge and improvement. Used in two rounds. First, to stress-test the research from four perspectives — policy expert, futurist, adversarial researcher and business strategist — and produce an improved proposal. Second, to attack that proposal in turn and verify every source used on this page. Result: the Flight Recorder Compact, version 2.0.

04 — Lovable · Deployment. Used to turn the research into this web experience. (NotebookLM was used to turn the research into an audio conversation.)

05 — Future distribution · Potential next step (not yet built). Possible future experiments

with AI-driven distribution and discovery, which could include Grok-based bots or agents where appropriate. These do not exist yet.

What the challenge rounds changed: The original research recommended regulator-controlled chips and cryptographic proofs of training as its foundation while rating those technologies as years away or uncertain; borrowed the nuclear industry's capped-liability model, which limits rather than strengthens accountability; and protected only accredited researchers. Round one rebuilt the proposal around accountability. Round two found that its insurance layer assumed a market that is currently retreating, that its response duty could be gamed by unverified fixes, and that a knowledge standard rewards not looking. Version 2.0 addresses each. Source-checking also corrected two claims we had made too strongly.

08 — Listen to the research

NotebookLM was used to turn this research into a voice-only conversation. It's another experiment in the same spirit: can an AI-generated discussion make a complex governance question easier to follow? Listen critically.

(Episode placeholder — coming soon.)

09 — What we haven't solved

Challenge any of these. That's the point.

1. Who should be accountable when AI causes harm and several companies built parts of the system?
2. How do we tell negligence apart from risks nobody could reasonably foresee?
3. Who gets to define a "dangerous capability," and how do we stop that definition being captured by incumbents?
4. Can international institutions keep up with AI development, or only with last year's version of it?
5. What happens if a major AI power refuses to participate at all?
6. Can safety become a genuine competitive advantage, or will it always cost market share?
7. What does verification look like for something that can be copied?
8. What happens when frontier capabilities become cheap and widely available?
9. Does a "what you knew" standard tempt companies not to look, and is a duty to look enough to close that gap?

10. Can catastrophic AI risk be insured at all, and if not, who should carry it?
11. How much AI harm comes from malicious intent, and how much from accidents, unintended behaviour or concentration of power?
12. Who is accountable for an AI agent acting on behalf of another AI agent?
13. Can humanity build this governance before a major crisis forces it to?

We'd especially like to hear from policymakers, security researchers, AI developers, insurers, lawyers, founders, academics and anyone working on these problems outside the US and Europe.

The Invitation

This is not a finished answer. It is an attempt to put a better question on the table.

The people who can build a globally workable system for AI governance have expertise, authority and resources this project doesn't. Our hope is to give them a sharper starting point.

Underneath it is a simple belief. There may be a future in which AI progress continues, companies can build, researchers can test, governments can protect their citizens, and society can remain meaningfully accountable for what happens.

Everyone can still win.

Not because the risks disappear. But because incentives, institutions and accountability can be redesigned so that progress and responsibility stop being treated as opposites.

This project is an attempt to find that path.

Appendix: The Flight Recorder Compact — Research Paper (v2.0)

A governance architecture for AI that cannot be stopped · Version 2.0 · September 2026

Summary

This paper starts from an uncomfortable premise: frontier AI development will continue, will spread, and cannot realistically be banned. It asks what humanity can build instead.

The original research behind this project proposed a control-based architecture: chips with regulator-held activation keys, cryptographic proofs of training, and a nuclear-style capped-liability pool. Two rounds of adversarial review found that design politically unrealistic, dependent on technologies it rated uncertain, and, in its liability model, working against accountability rather than for it.

The Flight Recorder Compact replaces control with accountability as the organising principle. It rests on a single standard, the three questions (what did you know, when did you know it, what did you do about it), applied through five mechanisms: Record, Probe, Respond, Underwrite, Intercept. Countries participate in three rings, from a full Compact to a universal floor that rivals can join. Principles live in law; technical parameters live in annexes updated annually, with pre-agreed triggers for change.

The thesis: finding a danger should always be safe, ignoring one should always be costly, and abusing one should be hard to hide, and every country and company should have more to gain inside the system than outside it.

1. Premise and assumptions

We assume: frontier development continues; governments cannot ban it; security researchers must be able to test models; models will be distributed globally; countries have different incentives; companies compete; malicious actors exist; law lags technology.

We treat this as a wicked problem in Rittel and Webber's sense: no definitive formulation, no stopping rule, solutions that are better or worse rather than true or false, and stakeholders who disagree about what the problem is.

2. The problem, restated

2.1 The accountability paradox, reframed. The original research framed the paradox as strict liability (which chills research) versus immunity (which invites recklessness). That framing is partly wrong. Liability attaches to harm, not discovery: a researcher finding a jailbreak is not itself a legal harm. The chill comes from three places: anti-hacking law and terms of service exposing researchers; companies' fear that a documented finding becomes evidence of knowledge in later litigation; and reputational exposure. The lever, therefore, is how the law treats knowledge. If knowing about a risk increases exposure regardless of what you do next, companies will avoid knowing. If what you did after knowing is what counts, they will want to know early.

2.2 Why traditional regulation fails. Statutes move in years; capability moves in months. Model weights are files and cross borders instantly. Catastrophic losses could exceed any company's balance sheet, making firms effectively judgment-proof. Single-jurisdiction rules

are routed around.

2.3 Where things stand in 2026. California's Transparency in Frontier AI Act took effect on 1 January 2026, requiring published safety frameworks, 15-day critical-incident reporting and whistleblower protection. The EU AI Act's general-purpose model obligations apply, with a 10^{25} FLOP presumption of systemic risk. The international network of national AI institutes was renamed in December 2025 the International Network for Advanced AI Measurement, Evaluation and Science, coordinated by the UK AI Security Institute, with ten members including Kenya. The mainstream insurance market began excluding AI-related liability in 2026 while specialist affirmative cover emerged. Zero-knowledge verification of training runs is an active research programme, with a June 2026 preprint estimating a proof of concept in roughly three years. None of these is a governance regime; together they are the ground the Compact is designed to stand on.

2.4 The Lens. The project's stated lens is that malicious intent and absent accountability may matter more than speed itself. We hold this as a lens, not a finding: many documented harms so far involve accidents, unintended behaviour and structural effects rather than malice. This is listed as open (section 14).

3. What history teaches

- **Nuclear (IAEA 1957, NPT 1968/1970).** Worked partly because fissile material is scarce, detectable and state-controlled, and because the treaty offered a bargain: verification in exchange for peaceful technology. Lessons that transfer: verification, transparency, shared standards, reciprocal obligations, communication channels, narrow red lines between rivals, legitimacy. Where it breaks: AI is software, copyable, distributable, privately developed, diffusing, and without a clean dual-use threshold.
- **Chemical weapons (CWC 1997).** Precursor tracking and challenge inspections. No challenge inspection has ever been requested. Lesson: an unused tool still shapes behaviour, and intrusive tools are politically hard to trigger.
- **Aviation (ASRS 1976, ICAO Annex 13).** Confidential near-miss reporting with limited immunity; independent investigation; separation of honest error from recklessness; certification recognised across borders. This is the Compact's closest model.
- **Finance (post-2008).** Stress tests, resolution plans, named-executive responsibility (UK SM&CR), certified reporting (Sarbanes-Oxley).
- **Cybercrime (Budapest 2001; UN Convention 2024/2025).** Cross-border cooperation infrastructure exists, imperfect and contested.
- **Price-Anderson (1957, extended to 2065).** Primary insurance of \$500M per site, retrospective premiums of ~\$158M per reactor, a total of ~\$16B per incident, Congress beyond that. It assures compensation and caps liability; it was designed to make an

industry investable, not to make it careful. We do not transplant it.

4. Design principles

The architecture must satisfy eight conditions at once: progress continues; security research continues; accountability exists; cooperation doesn't require altruism; incentives make participation rational; the system can evolve; legitimate disagreement has room; everyone can win. Two further principles emerged from review: redundancy over reliance (no single technology carries verification) and decisions over devices (govern human choices, which change slowly, rather than compute thresholds, which change fast).

5. The five mechanisms

5.1 Record. Frontier-tier developers register major training and release decisions with their national authority. Cloud providers apply know-your-customer checks for very large compute purchases. Advanced chips carry location and identity attestation sufficient for export control, with no remote disablement. Developers maintain a decision log covering tests run, findings received, mitigations applied and verified, and release decisions, including open-weight releases. Logs are held by the developer and auditable on defined triggers. Cryptographic proofs of training are an optional future layer, not a dependency.

5.2 Probe. A statutory safe harbor, in civil and criminal law, for good-faith research that stays proportionate, keeps demonstrations minimal and follows coordinated disclosure. Accreditation grants access, not protection. Contracts and terms of service cannot override the safe harbor. An AI vulnerability coordination function, on the model of CERT/CVE, routes findings, tracks response times, and escalates severe unaddressed findings to authorities.

5.3 Respond. See section 6.

5.4 Underwrite. Frontier-tier deployers must demonstrate financial responsibility proportional to scale and assessed capability. Because mainstream insurers are currently excluding AI risk, this is phased. Phase 1: reserves, bonds, or an industry mutual, plus mandatory contribution of incident and near-miss data to a shared database from day one. Phase 2: private underwriting as actuarial data accumulates. No liability cap applies to gross negligence, concealment, retaliation, or ignoring a disclosed finding. A defined share of contributions funds Intercept infrastructure.

5.5 Intercept. Investment in chokepoints between capability and physical harm: nucleic acid synthesis screening, hardened critical infrastructure, cyber-defence capacity, misuse detection and published threat reporting by API providers, and cross-border law-enforcement cooperation using existing cybercrime instruments.

5.6 Tiers. Obligations scale with capability and deployment. Below the frontier tier: a

disclosure channel and honest documentation only. An academic and non-commercial track exists with lighter Record and no Underwrite duties.

6. The accountability standard

6.1 The three questions. What did you know? When did you know it? What did you do about it? Legal exposure turns on these, not on whether a dangerous capability existed.

6.2 The spectrum. Responsible disclosure and good-faith research (protected); negligence (civil liability, higher contributions, conditional certification); reckless deployment (loss of safe harbor, fines, restrictions, personal liability); concealment (criminal exposure, loss of certification); deliberate malicious use (criminal law).

6.3 Duty to look. Required evaluation suites, maintained by the international network and updated annually, define reasonable care for each tier. Failing to run them is negligence. Wilful blindness is treated as knowledge.

6.4 Verified response. Responses to severe findings must be re-tested by the reporting researcher, an accredited evaluator, or a national institute.

6.5 Severity-scaled timelines. From days for critical biological or cyber uplift to a standard 90-day window for lower severities, set in annexes.

6.6 The accountable executive. Each frontier-tier developer names a senior officer with documented authority to delay or stop a release, who certifies safety reports. Boards carry a parallel oversight duty.

6.7 Whistleblowers. Protected from retaliation, with an internal anonymous channel at large developers, as California already requires.

6.8 Open weights. Releasing weights is a logged decision. The three questions apply at release. Weights cannot be recalled; the decision to release can be judged.

6.9 Agents. Agents operating at scale must be traceable to an accountable organisation. Mechanism unsettled; listed as open.

7. International structure

- **Compact (inner ring):** states hosting frontier development or key supply-chain steps. Adopt all five mechanisms; mutual recognition of evaluations and certifications; shared incident data; joint standards.
- **Associates:** adopt standards and disclosure without hosting frontier labs. Receive recognised certification for their firms, shared evaluation tooling, technical assistance, priority access to safety-relevant capabilities. This is the NPT-style bargain.

- **Universal floor:** human control over nuclear-use decisions; no self-propagating cyber agents against civilian infrastructure; nucleic acid screening; crisis hotlines; cross-border incident notification.

A non-member major power can sit on the floor, its firms can certify by passing the same published tests, and Intercept protects members regardless. Institutions: the international network for evaluations and recognition; a CERT-style coordination function for disclosure; national authorities for enforcement; existing cybercrime cooperation for misuse; a small secretariat for annexes and review. No new global agency.

8. Enforcement

The five-step ladder in section 6.2, applied domestically. For states: loss of recognition and associate benefits, coordinated export controls, and for red-line violations the diplomatic and strategic consequences applied to other weapons concerns.

9. Verification

Redundant by design: audited decision logs; behavioural evaluations shared under mutual recognition; chip attestation and cloud KYC for very large compute; whistleblower protection; shared incident data; satellite and grid observation for large undeclared facilities; optional cryptographic proofs of training as they mature.

10. Adaptation

Principles in law; parameters in annexes revised annually; every threshold sunsets unless re-justified. Triggers: open models reaching agreed evaluation thresholds → heavier release review; capability decoupling from compute → compute reporting downgraded, testing expanded; agent-to-agent activity at scale → traceability rules activate; insurance market re-entry → Phase 2 of Underwrite.

11. Feasibility map

- **Possible now:** SB 53-style transparency and incident reporting; civil and criminal safe harbors for good-faith research; disclosure channels; shared incident databases; cloud KYC for large compute; chip location attestation via export rules; industry mutual for financial responsibility; hotlines and the nuclear-decision red line.
- **3–5 years:** verified evaluation regimes with mutual recognition; the accountable-executive requirement across Compact members; risk-weighted contributions with real actuarial data; universal-floor nucleic acid screening.

- **Uncertain:** cryptographic proofs of training at frontier scale; agent traceability; insurability of catastrophic tail risk; participation of all major powers.
- **Politically unrealistic:** remote chip disablement; surrender of source code to international bodies; unannounced physical inspection of sovereign datacenters; universal bans.

12. Failure modes of this proposal

Response theatre (mitigated by verified response); not looking (mitigated by duty to look); sacrificial executives (mitigated by stop-authority requirement); check-the-box certification (mitigated by published tests and re-testing); a coordination function that becomes a gatekeeper; protectionist drift of market-access rules; capability outrunning annual annex updates; a mutual that under-reserves; a universal floor that is signed and ignored. None is fully solved.

13. What changed from the original research

Regulator-held chip keys → location attestation only. Cryptographic proofs of training as core → behavioural evaluation as core, proofs optional. Price-Anderson pool → phased, risk-weighted financial responsibility with no cap for culpable conduct. Compliance safe harbor → knowledge-response safe harbor with duty to look and verified response. Accredited-only research protection → any good-faith researcher, civil and criminal. Universal identity tracing → high-risk access only. Embargo-based defection strategy → three rings plus Intercept. New global agency → existing institutions plus secretariat.

14. Open questions

The thirteen questions listed in the essay (section 09), plus for the next iteration: what severity classification process resolves developer–researcher disagreement; what minimal chip-attestation feature set supports export control without a disablement capability; how reliable behavioural evaluations are at detecting dangerous uplift and how easily they are gamed; what would make the associate ring attractive to India, Gulf states, Brazil or Indonesia; and which claims here would change if a policy expert, a security researcher and an insurer each reviewed it. Recommended next step: three short expert interviews before wider distribution.

15. Sources

1. Rittel, H. & Webber, M. (1973). "Dilemmas in a General Theory of Planning." *Policy Sciences* 4(2).

2. ISO/Verisk generative-AI exclusion endorsements CG 40 47, CG 40 48 and CG 35 08 (effective 1 January 2026).
3. Peigné, P., Nguyen, K. & Wang, P. (2026). "Zero knowledge verification for frontier AI training is possible." arXiv:2606.05433.
4. American Law Institute, Model Penal Code §2.02.
5. US Department of Justice (19 May 2022), revised CFAA charging policy.
6. UK Senior Managers and Certification Regime (2016–); US Sarbanes-Oxley Act (2002), §302 and §906.
7. California Senate Bill 53, Transparency in Frontier Artificial Intelligence Act (2025/2026).
8. Statute of the International Atomic Energy Agency (1957).
9. Treaty on the Non-Proliferation of Nuclear Weapons (1968/1970).
10. White House readout of the Biden–Xi meeting, Lima, 16 November 2024; China's 2025 white paper.
11. Chemical Weapons Convention (1997), Article IX.
12. NASA Aviation Safety Reporting System (1976–); ICAO Annex 13.
13. UN Convention against Cybercrime (2024/2025); Budapest Convention (2001).
14. US OSTP Framework for Nucleic Acid Synthesis Screening (2024/2025); International Gene Synthesis Consortium protocol.
15. International Network for Advanced AI Measurement, Evaluation and Science (renamed December 2025).
16. US Congressional Research Service, "Price-Anderson Act" (IF10821, 2025).
17. Sastry, G., Heim, L. et al. (2024). "Computing Power and the Governance of Artificial Intelligence."
18. Original research report: "Global AI Safety Governance and Executive Accountability Architecture" (2026).