

Keeping Frontier AI Safe

→ Neil Chilson | Head of AI Policy

A GUIDE FOR POLICYMAKERS

Some worry that frontier AI could contribute to catastrophe through human misuse or failures of control. Policymakers should sort speculation from genuine risk, identify the specific mechanism of harm, and target interventions accordingly.

This brief covers catastrophic risks from developing, testing, and deploying frontier AI. Ordinary consumer harms fall outside its scope; existing consumer-protection and other laws still apply to them.

THE BOTTOM LINE

- 1. Start with the mechanism.** Before restricting AI, identify the concrete harm, the mechanism by which AI increases the risk over the status quo, and evidence that the increased risk is more than negligible.
- 2. Rely on liability.** Preserve common-law liability and add a narrowly-defined statutory duty, enforceable in court before injury, that requires reasonable mitigation and leaves the choice of methods to developers.
- 3. Count the costs.** Regulation can delay medical and cybersecurity benefits and raise entry barriers that favor incumbents. Weigh those costs against the harms a rule might prevent.
- 4. Keep model safety federal.** States should leave frontier model safety to [federal policymakers](#). Existing consumer-protection and other laws still apply to ordinary harms.

SEPARATE THREATS FROM MECHANISMS

Chemical, biological, radiological, and nuclear (CBRN) weapons and major cyberattacks are threats. AI could contribute to them through mechanisms such as human misuse and failures of control.

01

BIOLOGICAL RISK

Research tentatively shows that [AI boosts performance](#) on some biological tasks beyond internet-only access. But a [laboratory trial found](#) no significant improvement in its primary completion measure. It remains uncertain how much AI helps anyone build a biological weapon, which still requires materials, access, and practical skills. Biological findings cannot settle the other CBRN risks.

02

CYBER RISK

Attackers have already [used AI in cyberattacks](#), and defenders use it too. The open questions are how much risk AI adds and which side it favors.

03

CONTROL FAILURES

Reports document AI systems [taking unauthorized actions](#) in evaluations and internal deployments, sometimes knowingly violating task constraints. These incidents usually occur where developers have reduced safeguards. Even a system pursuing its [assigned objective can breach](#) operational limits. These findings warrant precautions, but they do not indicate a non-negligible probability of catastrophic loss of control.

SIX QUESTIONS TO ASK OF ANY AI SAFETY BILL

Preventing catastrophe matters. The question is how.

1. What concrete harm could occur?
2. By what reasonably concrete mechanism could AI cause it?
3. What evidence shows a non-negligible probability of harm through this mechanism?
4. What existing incentives, safeguards, and laws address this risk, and where do they fall short?
5. How would the proposed intervention close that gap and disrupt the harmful mechanism?
6. Would the expected reduction in harm justify the costs, including benefits forgone?

CHOOSE PRECAUTIONS THAT FIT THE MECHANISM

A

DANGEROUS INPUTS AND ACTIVITIES

Materials controls and screening can restrict misuse whether or not AI is involved. Drawn too broadly, they can burden legitimate research.

B

OPERATING ENVIRONMENTS

Sandboxes, access permissions, and monitoring can limit an unreliable model's actions. They require testing and maintenance, and they may restrict useful capabilities.

C

MODEL BEHAVIOR

Training can reduce harmful behavior but cannot anticipate every situation. Alignment may fail in unfamiliar circumstances, so it requires continuing evaluation.

These layers can work together, but the right mix depends on many factors. Instead of prescribing that mix, policymakers should allocate legal liability properly. Liability encourages developers to choose and improve precautions across all three.

WHAT POLICYMAKERS SHOULD DO

PRESERVE ACCOUNTABILITY

Assign responsibility to the actors best positioned to reasonably mitigate catastrophic risks; they often know the most about available precautions. Preserve common-law liability and add a narrowly defined statutory duty, enforceable in court before injury. Require reasonable mitigation, but let actors choose their methods. Audits or independent examiners can help courts assess care without granting immunity.

COUNT REGULATORY COSTS

Regulation can delay medical and cybersecurity benefits, raise entry barriers that favor incumbents, and constrain useful experimentation. Compliance can also displace work on effective safeguards. Policymakers must count those costs alongside the harms a rule might prevent.

**GET IN
TOUCH**

 abundance.institute
 [@abundanceinst](https://twitter.com/abundanceinst)
 [@abundanceinstitute](https://www.youtube.com/abundanceinstitute)

Neil Chilson
 Head of AI Policy
neil@abundance.institute

Digital Edition
 Scan for full sources
 and citations.

