ANTHROP\C

# Anthropic's Advanced AI Framework

June 2026

This framework lays out what we think governments should do in the near term about the most serious risks from frontier AI.

The framework has two parts. The first is a set of obligations on frontier AI developers. These include testing the most capable models for catastrophic risks; engaging external evaluators; disclosing results, risks, and incidents on an ongoing basis; and being accountable to the government. These responsibilities are designed to prevent and mitigate potential catastrophic harms while preserving the conditions for continued innovation.

The second part is a set of cross-government and cross-sector investments in societal resilience, so that a biological or cyber attack is harder to carry out and easier to recover from, wherever the capability originates. Building that resilience is not something any one actor can do alone. It requires governments, industry, and civil society working in concert.

This framework is written primarily with the US federal government in mind, but many of the underlying principles and concepts apply more broadly. Where other countries or subnational governments (such as states) choose to act, not all of these recommendations will translate exactly as written. We encourage policymakers to tailor them to their jurisdiction's capacity and authority.

Frontier AI is moving fast, and the discipline of evaluating and governing it is young. We are more confident about some parts of this framework than others, and recognize it draws from both existing and novel concepts. Readers will not agree with every proposal, but we think the exercise of laying out concrete policy proposals is useful even where the specifics are contested.

PART 1

# Frontier developer obligations

The following sections set out proposed obligations for the companies building the most capable AI models. Under these obligations, frontier developers would test their models for a defined set of catastrophic risks, show how those risks have been mitigated, publish what they find, and answer to a government agency with the authority to act on the results.

## Scope: who and what is covered

Models are covered when they have strong enough capabilities to pose catastrophic risks, with revenue- and spending-based thresholds to capture very large developers, based on our current understanding of training requirements for potentially dangerous models.

The obligations in this part of the framework focus on four risk categories: biological weapons, offensive cyber operations, loss of control of AI systems, and automated research and development. In our view, these are the risks where the potential harm is severe, difficult to reverse, and plausibly enabled by frontier AI systems in the near term. We expect that the set of covered risks, and the requirements attached to them, may need to be adjusted over time as our understanding of frontier AI capabilities evolves.

RECOMMENDED SCOPE PROVISIONS

**Covered Developer.** The proposed obligations apply to AI developers that meet both of the following criteria:

- Develop AI models requiring more than $10^{25}$ training FLOP.
- Earn more than $500 million[1] in annual AI-derived revenue, or spend more than $1 billion per year on AI research and development.

**Capability threshold.** Over time, the FLOP needed to train dangerous models may fall, and it may make sense to introduce a threshold based on capabilities rather than simply on training costs.

1 Adjusted for inflation

**Catastrophic Risk.** Consistent with existing US state law, "Catastrophic Risk" means a foreseeable and material risk that a Covered Developer's development, storage, use, or deployment of a covered model will materially contribute to significant death, injury, or damage.[2]

**Critical Safety Incident.** Refers to any of the following: (a) unauthorized exfiltration of a covered model's weights, or unauthorized, deliberate, malicious modification of a covered model's weights; (b) harm resulting from the materialization of a Catastrophic Risk; (c) loss of control of a model causing death or bodily injury; (d) a model that uses deceptive techniques against the Covered Developer to subvert its controls or monitoring outside the context of an evaluation designed to elicit this behavior and in a manner that demonstrates materially increased Catastrophic Risk.

**Enumerated Risk categories.** The obligations in this framework prioritize catastrophic risks related to (a) biological weapons, (b) offensive cyber operations, (c) loss of control, and (d) automated research and development in key domains that could accelerate or amplify risks (a)–(c).

**Periodic review.** A government agency ("Agency") should review the criteria for Covered Developers at least annually, with input from industry, academia, governmental entities, and civil society. Agency rulemaking can also potentially play a broader role in keeping implementation guidance current as the frontier advances. That authority should be paired with safeguards—procedural, judicial, and structural—against overbroad discretion.

**Preemption.** In the US context, Congress should not preempt state law unless it enacts a rigorous federal regime that meets or exceeds the strongest measures proposed in this framework for addressing catastrophic AI risks. Even then, preemption should be limited to the specific frontier governance functions Congress has expressly chosen to occupy, such as catastrophic-risk testing of covered models, evaluator licensing, or closely related reporting obligations. Outside those functions, federal law should not be read to occupy the field of AI regulation, to displace state law by implication, or to preempt state statutory or common-law claims. Any preemption should be construed narrowly, with ambiguity resolved in favor of preserving state authority outside the specifically occupied federal function. Compliance with a federal regime should not

2 We recommend a definition of Catastrophic Risk that is in line with the definition provided in California SB 53

itself confer immunity, a safe harbor, or a presumption against liability under otherwise applicable state law. Absent a strong federal law, states should retain full authority to legislate.

## Transparency: what developers should publish

While transparency alone is not sufficient for advanced AI, it gives governments, evaluators, and the public a record of what a developer's models can do and how risk is being managed. Developers should test their models against each of the four Enumerated Risk categories and publish a summary of the results. They should publish a safety framework explaining how they evaluate those risks, a system card whenever they release a covered model or add a major new capability, and a risk report at least every six months describing their overall posture across Enumerated Risks. They should also report Critical Safety Incidents to the government soon after they occur.

RECOMMENDED TRANSPARENCY PROVISIONS

**Develop and publish a safety framework.** Require Covered Developers to develop, publish, and follow a safety framework that sets out how they evaluate risks and how they make decisions about AI development and deployment. The safety framework should, at minimum:

- Identify the covered model(s) to which it applies.
- Describe how the developer evaluates each Enumerated Risk, including the standards relied on, capability evaluations performed, and mitigations applied.
- Identify the corporate officer primarily accountable for the framework's implementation.
- Describe the developer's process for modifying the safety framework.

**Annual certification.** Require Covered Developers to certify annual compliance with their safety framework to the Agency, with civil penalties for material misrepresentation.

**Risk reports.** Require Covered Developers to publish a risk report at least every six months that offers an overall assessment of each of the Enumerated Risks posed by the developer's models. As noted below, this assessment is also updated on a per-model basis via system cards. Risk reports should, at minimum, include:

- A summary of relevant capabilities across the models the developer deploys, covering each Enumerated Risk category and discussing material changes in capabilities since the prior report. This summary should give a representative picture of what the developer knows about its models' Enumerated Risk-relevant capabilities. This report should cover all models the developer serves externally, as well as any substantially riskier models the developer deploys internally.
- The threat models being tracked, how observed capabilities map to them, and the mitigations addressing each.
- An assessment of the risk in each Enumerated Risk category that remains after accounting for the safety mitigations the developer has in place. Risks from internal model deployments should be considered, as appropriate, in addition to risks from external deployments. It should be possible to understand most of the reasoning behind the risk assessment from public information, and to reach a similar overall conclusion about the level of risk as one would reach with access to all the information the developer has.
- Risk reports should be required every six months to start, but a higher frequency may be needed in the future if AI progress accelerates.

**System cards.** Require Covered Developers to publish a system card or equivalent documentation when deploying a covered model for general access that, in any Enumerated Risk category, (a) is materially more capable than any covered model it has previously deployed; or (b) is deployed under materially weaker safeguards for an Enumerated Risk than a previously deployed covered model of comparable or greater capability. This document should include a summary of:

- Testing methodologies and results.
- Model capabilities and limitations, as well as intended and observed model behaviors.
- How, if at all, the model's internal or external deployment changes the risk assessment compared to prior public analyses in system cards and risk reports.

**Critical Safety Incidents.** Require Covered Developers to report Critical Safety Incidents to the designated Agency within 15 days of the Covered Developer discovering the Critical Safety Incident or facts that would lead the Covered Developer to have a reasonable belief that a Critical Safety Incident has occurred. Reports should be shared with relevant federal government agencies and national laboratories. Information shared should be exempt from public records disclosure laws, consistent with existing state law.

**Redactions.** For public-facing documents, Covered Developers may redact information that constitutes the developer's trade secrets, that would materially compromise public safety, model security, or national security if disclosed, or that should be withheld to comply with applicable law. Public-facing materials should be redacted only as necessary, and developers should note the nature of the redaction in any published version.

## Independent evaluation

Self-assessment is not enough. At the same time, we recognize that a mature independent evaluation ecosystem does not yet exist. This framework aims to bridge that gap by requiring developers to run their own evaluations and risk assessments on the Enumerated Risks and, within six months of the enactment of the regulation, to regularly engage at least one qualified evaluator for an independent, external point of view on those risk assessments.

RECOMMENDED INDEPENDENT EVALUATION PROVISIONS

**Require independent evaluation.** Within six months of the enactment of the regulation, require Covered Developers to regularly engage at least one qualified independent evaluator to examine the relevant covered models, gather information about safeguards, and provide their own assessment of Enumerated Risks.

- The evaluator should be independent in the sense of not having a financial interest in the developer and being free of major conflicts of interest with respect to the individuals involved in conducting the review.
- The evaluator should receive access to an *unredacted* version of the developer's most recent risk report and system cards, access to the developer's most capable models, and the opportunity to ask and receive

reasonable responses to relevant questions about models, likelihood of catastrophic risks, and safeguards.

- The evaluator should publish a review of the developer's most recent risk report, addressing the adequacy of the report's information; its analytical rigor; the appropriateness and materiality of any redactions made to the public version; and whether the evaluator disagrees with any of the report's key claims, especially its overall assessment of the level of risk for each Enumerated Risk category.
- Evaluators should be bound by obligations not to copy, retain, or disclose confidential information (e.g., confidential intellectual property, matters of national security, or proprietary details such as model architecture, cost, or size), but beyond that should generally not be restricted in what they can publish, including concerns about the risk report or the developer's conduct in connection with the review process.

**Independent evaluator ecosystem.** Take steps to grow an independent evaluation ecosystem, including developing and publishing standards for evaluators; exploring a licensing system to qualify evaluators; providing government funding or arranging pooled funding so that evaluators can do their work while remaining financially independent of any given developer; and providing resources and funding for nascent organizations seeking to become evaluators.

**Evaluator independence safeguards.** Require evaluators to certify their independence from AI developers as a condition of qualification or licensure. This may include providing information regarding funding and remuneration, conflict-of-interest and equity-ownership disclosures, board participation, and employment. Additional safeguards could include post-employment cooling-off periods or pooled funding through a common industry fund.

**Avoid "evaluator shopping."** There is a risk that companies seek out whichever evaluator(s) will ask the least of them and be most generous in their assessments. To mitigate this, relevant government agencies could rate evaluators based on predefined criteria such as rigor of public reasoning. Highly rated evaluators could be randomly assigned to AI developers, particularly in high-stakes cases.

**Build government capacity to fulfill the independent evaluator function.** Ideally, a government agency would be able to fulfill the functions above—including capability evaluations as well as holistic risk assessment that takes threat models, model capabilities, and developer safeguards into account. A government evaluation function could eventually supplement or even substitute for an independent evaluator ecosystem.

## Security

Frontier model weights and training infrastructure are high-value targets, including for sophisticated state actors, and a model that is safe to deploy is not safe if it can be stolen or quietly copied. Covered Developers should maintain a security program across the full development environment, robust to both external and insider threats and scaled to the consequences of compromise. They should describe that program publicly at a general level and give the Agency detailed documentation on request. They should also monitor for unauthorized distillation attacks on their models and run regular penetration testing of their own defenses (against external and insider threats), reporting findings to the government.

RECOMMENDED SECURITY PROVISIONS

**Security program.** Require Covered Developers to maintain a security program covering the full AI development environment (model weights, training and inference infrastructure, trusted partner access controls, and internal development processes), robust to external and insider threats and scaled to the consequences of compromise.

**Disclosure.** Require Covered Developers to describe their security program at a general level in their public safety framework and to provide detailed documentation to the Agency on request.

**Monitoring for distillation attacks.** Create channels for Covered Developers to report known model extraction and distillation attacks as well as detection efforts and prevention measures to the Agency and to other Covered Developers.

**Penetration testing.** Require Covered Developers to conduct regular red teaming and penetration testing covering model weights, algorithmic secrets, training infrastructure, and insider threats. Require Covered Developers to inform the

Agency of the categories of findings and the status of remediation. As the field matures, support trials of penetration testing and red teaming run by, or jointly with, government agencies.

**Security of privileged-access parties.** Require independent evaluators, and any third party granted privileged or pre-deployment access to covered models in a way that could be misused (e.g., access to versions with safety mitigations reduced or removed), to maintain security protections commensurate with the sensitivity of that access and scaled to the consequences of compromise.

## Enforcement and regulatory authority

Transparency and independent evaluation make it easier to assess which models are safe. They need to be paired with real enforcement: developers who lie about compliance should face meaningful penalties, and employees and contractors should be able to raise concerns without retaliation. The provisions in the first box below should hold under any enforcement design a jurisdiction chooses.

Ultimately, there should also be a way to block or deter deployment of models that pose significant catastrophic risks. There is much room for debate on how best to accomplish this, while avoiding overly broad or heavy-handed regulatory power. Recognizing the difficulty of this topic, we lay out a range of options for policymakers. Policymakers could begin with a lighter-touch model and revisit that choice as model capabilities advance and the independent evaluation ecosystem matures.

RECOMMENDED ENFORCEMENT PROVISIONS

*Applicable under any enforcement model*

**False statements.** Prohibit intentionally false or materially misleading statements related to safety framework compliance, required evaluations, or required disclosures.

**Civil penalties.** Authorize the Agency to seek civil penalties for failure to conduct required evaluations, publish a safety framework, or report incidents. Penalties should escalate with repeated violations and scale with global annual revenue.

**Whistleblower protections.** Require Covered Developers to maintain anonymous internal channels for employees and contractors to flag compliance concerns or activities that pose serious public safety risks. Prohibit retaliation against persons who report in good faith to the Agency, federal authorities, or other appropriate recipients, and prohibit contractual restrictions on such reporting. Existing federal and state whistleblower protections apply.

POSSIBLE APPROACHES TO ESTABLISHING AUTHORITY TO BLOCK OR DETER UNSAFE MODELS

**Government agency review.** Empower an Agency to review the information available in system cards, risk reports, and reports from independent evaluations, discussed above, and identify cases where one of the following violations has occurred:

1. The required system cards, risk reports, and/or independent evaluations have not been completed and published as required.
2. The independent evaluation was not done by a sufficiently disinterested or qualified evaluator. Determining whether an evaluator is sufficiently qualified could involve requiring licensing for evaluators, setting standards for evaluator qualifications, establishing an evaluator rating system, or simply reviewing evaluator qualifications qualitatively.
3. The independent evaluator did not have sufficient time and/or access to conduct a high-integrity evaluation.
4. The risk assessment and/or independent evaluation found that the developer's covered models, as deployed, pose a significant risk of catastrophic harm in an Enumerated Risk category, even taking active safeguards into account.

**Remedies.** Where a violation has occurred, the Agency may pursue remedies including:

- Fines for deployment of models with inadequately mitigated risks.
- Prohibitions on deployment of further covered models until the developer corrects the violations.
- In extreme cases, requirements to restrict usage of, and access to, already-deployed models as needed to reduce catastrophic risks.

**Safeguards against overreach.** A number of options are available for avoiding overbroad or inappropriate use of Agency authority, including:

- **Court enforcement.** Rather than having the authority to impose remedies directly, the Agency may pursue them only by initiating a lawsuit seeking the appropriate remedies. In cases of imminent catastrophic risks, provisional remedies could be permitted (reversed after a set period if a court does not uphold them), or lawsuits could be fast-tracked.
- **Cabined discretion.** The Agency may not pursue remedies based on its own assessment of risk, but only based on the specific violations listed above.
- **Consistent treatment and judicial review.** The Agency should apply a fact-based review process consistently across developers, holding all covered models of equivalent capabilities to the same standards. No developer may be advantaged or disadvantaged on grounds unrelated to the model's evaluation record or capabilities related to the Enumerated Risk categories, and a developer may challenge remedies through an expedited judicial review process.

The remedies and safeguards specified above can be mixed and matched to achieve different tradeoffs between the risks of inadequately safeguarded models and the risks of overbroad regulation. The specifics of these enforcement and regulatory approaches may vary depending on the unique legal and jurisdictional context.

PART 2

# Societal resilience measures

The preceding sections address the conduct of AI developers and how to build the right governance framework to shape that conduct. The following sections address a different question: how society can prepare to withstand the threats—particularly biological and cyber—that advancing AI capabilities may accelerate or enable. This distinction matters because the most consequential resilience investments (e.g., surveillance systems, stockpiles, hardened infrastructure, and response capacity) take years to build and cannot be stood up in a crisis. The recommendations that follow are about building them now, and many are worth making regardless of how AI develops.

These are not new policy areas. Biological and cyber resilience have deep bodies of work, established institutions, and live debates that long predate frontier AI. The recommendations below are meant to build on that work, not restart it. We offer them as priorities and directions rather than finished policy designs; detailed proposals should be shaped with the experts and communities already active in these fields.

This part focuses on biological and cyber threats, where the resilience interventions are clearest and most tractable today. We believe analogous resilience measures exist for additional emerging risks where the agenda is less developed, including loss of control and automated R&D, and will share findings and proposals as we learn more.

## Biological resilience

Biological threats are unique in that a released agent replicates, spreads person-to-person, and evolves under selection pressure. A small breach can become a large one without further action by the attacker, and the agent can adapt to evade countermeasures developed against it. The developer obligations in Part 1 are the upstream end of that defense: testing frontier models for biological-weapons uplift, securing model weights, and giving an Agency authority to act on what evaluations find. The recommendations below are the societal complement, organized as layers of defense. *Prevention* raises

the cost of acquiring the materials and expertise needed to build a biological weapon; *detection* builds the surveillance needed to spot a threat early enough to act; and *preparedness* puts protective equipment, medical countermeasures, a more resilient built environment, and hardened public health and response infrastructure in place before they are needed. These investments pay off regardless of where a threat originates—they protect against natural outbreaks with pandemic potential and conventional bioterrorism, as well as against AI-enabled attacks.

RECOMMENDED BIOLOGICAL RESILIENCE MEASURES

**PREVENTION**

**Modernizing biosafety and biosecurity standards.** Update and unify national biosafety and biosecurity rules such as lab safety protocols—many of which predate synthetic biology, directed evolution, and AI-assisted design tools—into a coherent, enforceable framework that reflects today's threat landscape.

**Closing oversight gaps.** Extend enforceable biosafety and biosecurity standards to privately funded research, strengthen institutional review committees as a front line of local oversight, and require personnel vetting for access to the most dangerous pathogens.

**Gene synthesis screening.** Require gene synthesis providers and benchtop synthesizer manufacturers to screen requested sequences for known or predicted hazards and verify customers seeking sequences of concern.

**Threat monitoring and intelligence sharing.** Dedicate new intelligence and law enforcement resources to monitoring and disrupting actors pursuing biological weapons and to deterring and attributing state-level bioweapons activity. Establish structured, two-way channels for governments and AI developers to share CBRN threat intelligence, supported by legal safe harbors, information-sharing protections, and antitrust carve-outs. Allow AI and biotechnology companies to share threat intelligence with each other and with the government, and provide for independent stress testing of biosecurity safeguards.

**DETECTION**

**Pathogen-agnostic biosurveillance.** Establish, fund, and maintain pathogen-agnostic biosurveillance systems at the national, subnational, and international levels, designed to provide actionable early warning and characterization of biological threats.

**Microbial forensics and attribution.** Build, fund, and exercise federal capabilities for rapid attribution of biological incidents to enable accountability for state and non-state actors and underwrite deterrence-by-denial. Maintain reference databases, accredited laboratories, standing rapid-deployment forensic teams, and international channels for cross-border evidence sharing and joint attribution.

#### PREPAREDNESS AND RESPONSE

**Public health infrastructure and planning.** Maintain and harden public health and emergency response infrastructure, workforce, and logistics at all levels of government. Test and strengthen plans for protecting critical infrastructure and population centers during a severe biological event, accounting for worst-case outcomes enabled by advances in biotechnology, and exercise them in live operational drills.

**Protective equipment for the essential workforce.** Stockpile pandemic-grade respiratory protection, especially reusable respirators, for the essential workforce at the national, subnational, and international levels, and maintain surge manufacturing capacity.

**Airborne transmission suppression.** Support standard-setting, research and development, incentives, and investment in measures that make the built environment more resistant to respiratory pathogen transmission in critical infrastructure (e.g., public utilities, schools, hospitals, transit hubs, and government facilities).

**AI-accelerated countermeasure development.** Leverage AI-accelerated drug discovery and protein design to shorten development timelines for novel countermeasures, and fund partnerships between AI developers and biodefense research institutions to build this capacity before it is needed.

**Broad-spectrum antivirals.** Increase investment in research and development of broad-spectrum antiviral medicines, which have the potential to provide resilience by reducing the pressure to rapidly characterize a pathogen and develop novel countermeasures.

**Responsive medical countermeasures.** Invest in adaptable platforms and manufacturing capacity to rapidly produce vaccines, therapeutics, and diagnostics against novel or engineered pathogens. Prioritize countermeasures that are readily scalable, broad spectrum, and transmission suppressing. Stockpile precursors and maintain pre-established clinical trial protocols.

**After-action institutionalization.** Require binding after-action reviews following any nationally significant biological incident, with public reporting, statutory deadlines for corrective action, and independent oversight for unimplemented findings. Codify lessons-learned cycles to prevent the recurrent erosion of institutional memory between events.

## Cyber resilience

Frontier AI is shifting the economics of cyber offense, accelerating vulnerability discovery and exploit development at a scale that will spread well beyond the handful of developers covered by Part 1 of this framework. The recommendations below aim to make sure the same technology that lowers the cost of attack also lowers the cost of defense. *Prevention* hardens the open-source and legacy systems the internet depends on and supports the operators least able to defend themselves; *measurement and situational awareness* builds the government's ability to track how frontier cyber capabilities are advancing and who is misusing them; *defender advantage* puts AI-enabled defensive tools and faster patching in the hands of defenders; and *security modernization* updates policies built for a slower era, from replacing legacy systems in critical infrastructure to funding experimental research into new security paradigms.

RECOMMENDED CYBER RESILIENCE MEASURES

**PREVENTION**

**Open-source and legacy software security.** Fund sustained maintenance, security auditing, AI-assisted vulnerability remediation, and migration to more secure software development practices (e.g., memory-safe languages) for open-source and legacy software underpinning critical infrastructure and the internet.

**Phishing-resistant identity and provenance.** Accelerate ecosystem-wide adoption of phishing-resistant authentication and content provenance standards.

**Forward-deployed support for under-resourced operators.** Fund forward-deployed engineers, shared regional security operations centers, and managed security services for critical-infrastructure operators that lack the resources to defend themselves (e.g., water and wastewater utilities, municipal governments,

school districts, regional hospitals, and rural electric cooperatives). This work should prioritize the basics with the highest payoff against known-but-unpatched vulnerabilities: maintaining asset inventories, deploying patches promptly, and reducing internet exposure of operational technology (via data diodes).

#### MEASUREMENT AND SITUATIONAL AWARENESS

**Tracking frontier-model cyber capability.** Direct the national AI safety institute or equivalent body to develop and maintain cyber capability evaluations, in partnership with the intelligence community and national-laboratory operational-technology testbeds, and to issue capability guidance to government, AI developers, and critical-infrastructure sectors as thresholds are crossed.

**Threat intelligence on AI-enabled cyber operations.** Establish a dedicated threat-intelligence function, in coordination with the intelligence community, that fuses developer abuse monitoring, government reporting, and incident-response telemetry into a single picture of which actors are misusing or distilling frontier cyber capabilities. Encourage short-term data retention to enable this function.

**Safe harbor for defensive coordination.** Provide legal safe harbor and information protections for AI developers to share defensive research, vetting practices, jailbreak data, distillation defenses, and threat intelligence with each other, with the government, and with the broader ecosystem.

#### DEFENDER ADVANTAGE

**Designing for breach.** Direct the government to distribute best practices on hardware-root-of-trust network isolation that prevent attackers from moving laterally inside of networks.

**AI-enabled breach remediation.** Direct the government to distribute best practices on utilizing ML and AI systems to detect and remediate breaches within minutes or hours instead of days or weeks.

**Safeguards for broad release of defensive capability.** Direct the government to assist AI developers in co-developing safeguards (e.g., vetting, user verification, misuse and distillation safeguards, and threat intelligence sharing) that enable the responsible release of cyber capabilities to large numbers of defending organizations.

**Strategic reserve of operational-technology hardware.** Use existing industrial-base and defense-production authorities to build a strategic reserve of long-lead-time operational-technology hardware (e.g., protective relays, remote terminal units, and programmable logic controllers) for grid and water operators whose equipment a destructive attack would otherwise render inoperable.

**Software supply chain mapping.** Analyze and identify the most critical components in the software supply chain. Fund the national cybersecurity agency to map out the most critical and wide-reaching AI and software supply chain nodes. Use this mapping to inform and prioritize coordinated vulnerability disclosure at scale.

**Coordinated vulnerability remediation at scale.** Fund the national cybersecurity agency and coordinated defense bodies to handle 10–100 times the current disclosure volume with sub-seven-day turnaround, as frontier models accelerate vulnerability discovery beyond what the current pipeline can absorb. Direct the relevant standards body to revisit disclosure timeline norms for the AI era.

**Patching at scale.** Establish binding patch-deployment frequency for critical-infrastructure operators calibrated to sector-specific operational constraints. Fund research, development, and broad deployment of AI-assisted patch generation, backporting, and validation tooling so that faster timelines are achievable rather than aspirational.

#### SECURITY MODERNIZATION

**End-of-life and legacy system replacement.** Require vendors of software and connected devices used in critical infrastructure to publish support lifecycles and end-of-life dates; fund operators to inventory and isolate systems that can no longer receive security updates; and establish a funded replacement program—modeled on prior national-security-driven equipment replacement efforts—prioritized by exposure and consequence.

**Experimental research into new security paradigms.** Fund research that transcends find-and-patch (e.g., formal verification, runtime patching, and polymorphic defense), with a requirement that it demonstrate value beyond what frontier models can already do, rather than reproducing their capabilities.

## Loss of control and automated R&D

Part 1 of this framework addresses risks from loss of control and AI R&D on the developer side, through capability testing, incident reporting, and security requirements. The societal resilience agenda for loss of control and automated R&D is less mature than it is for biological and cyber risks, and we believe it needs much more active work across the field, from governments, researchers, and industry alike. Promising directions include the capacity to detect and respond to AI systems acting outside their developers' control, and infrastructure for containing or shutting down such systems. We continue to conduct research in this space and will share more as our understanding matures.

## Conclusion

The measures proposed in this framework—imposing clear obligations on developers of the most capable models and strengthening society's resilience against the threats those models may accelerate—are necessary first steps. There will be more work ahead as the technology matures. But the cost of waiting for perfect policy is having none during a critical period: when the most severe risks of AI are in view, but not yet realized. We invite feedback on these proposals, and we are ready to work with policymakers and experts to put them into practice.