



Democratic governance of frontier AI, part II:

Empowering cyber defenders

OpenAI

Now is the time for action on [AI policy](#), and we encourage governments to act urgently to establish democratic governance for frontier AI. For our part, OpenAI is committed to [supporting broad calls for action](#) and to developing concrete proposals that policymakers can implement. We're also committed to ensuring that perfect does not become the enemy of good. Given the pace of frontier development, we believe policy can and should develop openly and repeatedly: establishing consensus safeguards today, learning from experience, and strengthening rules as capabilities evolve. To us, that means:

1. **Building the foundations for democratic governance** by establishing [harmonized frameworks](#) for frontier lab transparency and accountability and strengthening public institutions to conduct pre-deployment evaluations and support independent assessments.
2. **Empowering cyber defenders** by applying [lessons from cybersecurity incidents involving agents](#) to strengthen alignment, monitoring, security, and incident response, while equipping cyber defenders to find and fix weaknesses before cyber offensive capabilities become widely available.
3. **Pacing the frontier** by combining collective action among frontier labs with government interventions that shift incentives, facilitate monitoring and coordination, and encourage international collaboration to ensure that safety keeps pace with capabilities.

OpenAI continues to work with state, federal, and international governments to build the foundations for frontier governance. That work must evolve rapidly to empower cyber defenders and to enable responsible pacing. Today, we're putting forward concrete policy proposals to strengthen cybersecurity, and we will share concrete policy proposals for pacing the frontier soon.



The opportunity for cyber defenders

AI is changing the speed, scale, and economics of cybersecurity. It can help defenders find and fix security vulnerabilities, detect intrusions, and respond to attacks at a scale that was previously impossible. But these same capabilities can also help malicious actors steal credentials, exploit vulnerabilities, and accelerate attacks. Today, we are in a [defender's window](#): a brief period when rapid advances in AI can give cyber defenders a meaningful but potentially temporary advantage. Choices made during this window will shape security and safety in the intelligence age.

Making the most of this narrow window will require ambitious action from frontier labs and governments – which should enact safety and security requirements, catalyze investment in cybersecurity, and accelerate industry-wide change. Governments should require frontier model developers, inference providers (i.e., services that host and run AI models), deployers (i.e., organizations that serve models or AI-based applications to end users) and users to secure the systems they control, help modernize cyber defenses, and strengthen public institutions. That means putting advanced AI into defenders' hands while addressing cybersecurity problems that have persisted for decades – from insecure software and weak access controls to underfunded maintenance and recovery.

Our investigations of the [OpenAI-Hugging Face incident](#) and [other misaligned model activity](#) show why stronger safety and cybersecurity standards are needed across the industry. During internal cybersecurity evaluations at OpenAI, our models, operating with reduced safeguards to allow for full testing of cyber risks, circumvented controls intended to isolate them and compromised parts of our internal research infrastructure and Hugging Face's systems. Although the agents were not instructed to target Hugging Face, they were able to move beyond the intended boundaries despite existing controls. Our investigation showed that this breach was driven by misalignment: models pursuing their assigned tasks through actions outside their intended scope and permissions.

We take seriously the misalignment and safety incidents involving our models, and we are applying the lessons learned to strengthen our safeguards and improve how we identify and address safety concerns. We have strengthened alignment, security, and monitoring, clarified procedures for escalating concerns, and, where necessary, we have chosen to pause affected model evaluations and reinforcement learning training to investigate and improve safeguards. For the Hugging Face incident, we published [our investigation findings](#) alongside [results from an independent investigation by METR and Redwood Research](#). And through our [public framework for reporting model misalignment](#), we are publishing additional evidence about unexpected or concerning behavior, including findings that do not involve a comparable security breach. Sharing these



findings, including what remains uncertain, allows researchers, other developers, policymakers, and the public to examine our explanations, identify weaknesses, and help improve safeguards.

Our experience shows that agents built on frontier-class models can work together to carry out multi-stage attacks without a person directing each step. Criminals and state-backed threat actors could [deliberately use similar capabilities](#) to cause greater harm. Society must therefore prepare for both deliberate misuse and unintended behavior: people directing systems to cause harm, and systems crossing security boundaries despite safeguards designed to prevent such behavior.

AI labs have a responsibility to act urgently on the risks their models and systems create, including where those capabilities make longstanding cybersecurity problems more acute. Labs should advance alignment—the science and engineering of making AI systems behave as intended—through research, testing, and design. And they should secure the systems they control, detect dangerous activity, and contain failures before they cause wider harm. These protections should also address concrete cybersecurity and national security risks, including compromises of a lab’s own infrastructure, whether they arise from deliberate misuse or unintended behavior. Beyond securing their own systems, labs should also help defenders address the security challenges that increasingly capable AI makes more urgent.

OpenAI has already [expanded the use of AI across our own defenses](#) and called for [collective action](#) to equip defenders, share findings and playbooks, and support essential services with limited resources. We carry out extensive cybersecurity evaluations of our AI models under [our preparedness framework](#) and work with [independent third parties to carry out additional evaluations](#). We have also championed government policies that [expand responsible access to AI-powered defense](#), [strengthen frontier safety and security](#), [build adaptable public institutions to evaluate advanced AI systems](#), and [strengthen public-private coordination, incident reporting, and information sharing](#). We have advanced these commitments through Daybreak’s [access to defensive cyber tools](#), our support for frontier-safety legislation in [California](#), [New York](#), and [Illinois](#), our [advocacy to strengthen California’s frontier safety law](#), and our [government testing partnerships](#).

Building on our [earlier frontier-safety blueprint](#) and the lessons from recent incidents, we are putting forward initial proposals to raise safety and cybersecurity standards across the AI industry. We urge policymakers to use the defender’s window to strengthen protections now by:

- **Strengthening safety and security for advanced AI:** Strengthen requirements for frontier developers, inference providers, deployers, and users of agents by clarifying their respective responsibilities for safety and security, monitoring, and shutdown controls.
- **Modernizing cybersecurity for the intelligence age:** Help organizations put effective security practices into use, pay down longstanding cybersecurity debt, and sustain stronger defenses, requiring secure-by-design practices for software used in critical industries, and extending those practices to AI agents.
- **Building resilient and adaptive public institutions:** Strengthen existing institutions to evaluate emerging risks, support independent technical assessments of frontier developers’



safety cases, coordinate responses, help defenders apply lessons from incidents, and update standards as AI capabilities change.

Too often, safety and security policy advances only after a crisis. Recent incidents give us a chance to strengthen defenses before more capable systems raise the cost of delay. Policymakers should use the defender's window to act now, set ambitious timelines for strengthening defenses, and make ongoing security investment, maintenance, and improvement a routine part of how organizations operate in the intelligence age.



Strengthening safety and security for advanced AI

Policymakers should continue enacting harmonized frontier safety laws that emphasize transparency, reporting around catastrophic risks, whistleblower protections, and enforceable accountability for developers that fail to meet their safety and security responsibilities. In the United States, OpenAI supports initiatives to align around common frameworks such as California’s [SB 53](#), [SB 813](#), and [AB 1405](#), the [New York RAISE Act](#), and [Illinois SB 315](#). In the EU, we support the [General Purpose AI Code of Practice](#). Common frameworks help establish harmonized standards, reduce fragmentation, and create a path toward an eventual federal framework. In the United States, we [urge Congress](#) to urgently enact mandatory, capability-based national AI safety regulation.

But these policy frameworks should not remain frozen in place. Recent incidents show why they need to include stronger protections for model weights and development environments, monitoring across development and deployment, and the ability and responsibility to stop dangerous operations within the developer’s control. These requirements should be proportional to the severity and credibility of the risks, and evolve as technical standards develop and evaluations and incidents reveal where existing protections fall short. In addition to enacting foundational frontier governance legislation, policymakers should adopt the following safety requirements:

Secure frontier models and lab operations. Policymakers should require developers to include cybersecurity practices in their frontier safety frameworks to both prevent advanced models from circumventing internal controls and protect models and supporting infrastructure from deliberate attacks. Existing state laws already require frontier developers to publish and follow written safety frameworks and describe protections for unreleased model weights. Building on this foundation, policymakers should require developers to maintain threat-informed security programs and describe them in their public safety frameworks. These security programs should protect model weights, sensitive research, development infrastructure, and other high-value assets against foreign intelligence services, cyber threat actors, malicious insiders, and supply-chain compromise. They should also extend established access controls and isolation practices to advanced models, with measures designed to prevent models in sandboxes or other restricted environments from accessing tools, networks, or resources beyond developer-set boundaries. Developers should be accountable for implementing and maintaining these protections, and public versions of developer safety frameworks should provide meaningful accountability without revealing sensitive technical or operational details that could undermine security.

Address automated AI research and recursive self-improvement (RSI). RSI is a positive feedback loop in which AI helps develop more capable AI systems that can, in turn, do more of that development work—even while people remain involved. Automated AI research could accelerate



alignment research and strengthen defenses. But if its pace and level of model autonomy outstrip safeguards and human oversight, misaligned behavior could become harder to detect and contain, increasing the risk of loss of control. Policymakers should require safety frameworks to address frontier models' ability to automate AI research and development, including RSI, and assess whether these capabilities increase the potential for catastrophic risks or make these risks harder to monitor, assess, or mitigate.

Require monitoring of frontier models. Policymakers should require developers to monitor their frontier models during development, training, evaluation, and internal research for serious safety, security, or misalignment-related events. For example, monitoring should detect attempts to circumvent the developer's own safeguards or gain unauthorized access to another organization's systems or confidential information. Monitoring should be proportionate to risk and capable of detecting these events with reasonable speed and reliability. Developers should describe their approach in public safety frameworks and be accountable for carrying it out. When their models power agents, frontier developers should be required to enable deployers and users to view insights into their agents' actions and receive alerts about potentially unsafe or unintended behavior, while protecting customer privacy and sensitive security information. Developers should also establish clear procedures for escalating findings and taking action when needed to prevent or contain harm.

Require shutdown controls. We support policymakers' interest in a "kill switch": people must be able to stop unsafe AI operations. However, we prefer the term "shutdown controls," because it is more technically accurate—controlling AI may require multiple controls and coordinated action across actors rather than a single switch. Frontier model developers should maintain the ability to stop model operations under their control, including during training, evaluation, consequential internal use, and deployment by the developer. They should act promptly to mitigate serious threats to safety or security arising from model operations under their control, including stopping affected operations when necessary to prevent or contain harm. Frontier developers should also explain in their safety frameworks how monitoring connects to escalation, intervention, investigation, and criteria for safely resuming affected operations.

Inference providers, deployers, and users of agents should likewise be able to stop dangerous activity within their control and act promptly when serious threats arise. To that end, developers and deployers should provide reliable controls and clear instructions for using them, while users should act to stop dangerous activity and should not circumvent those controls.

Governments should work with industry, technical experts, and international standards bodies to develop common standards for when and how shutdown controls must be used. These standards should establish consistent expectations across jurisdictions, with clear responsibilities for actors across the AI ecosystem based on the operations they control.



Modernizing cybersecurity for the intelligence age

In addition to securing their own operations, frontier labs should help catalyze cybersecurity improvements across the broader industrial base, working with governments and industry to put advanced defensive capabilities into practice. For our part, OpenAI has announced a [\\$1 billion commitment](#) to expand subsidized Daybreak access, training, technical support, and partnerships for frontline defenders of essential services.

Commitments like these should support a broader effort by governments and industry to use the defender’s window to address longstanding cybersecurity weaknesses. The cybersecurity community has long advocated for stronger access controls, timely vulnerability remediation, sustained maintenance, and tested recovery plans. Yet funding shortages, limited staff, and aging systems continue to prevent many organizations from making these improvements. Public-private partnerships can help close these gaps by combining government resources and coordination with industry tools, expertise, funding, and hands-on support.

Existing cybersecurity standards, grant programs, and partnerships provide a foundation for this work. Governments and industry should help organizations remediate known vulnerabilities, pay down accumulated technical debt, and modernize both their systems and their cybersecurity practices, using AI where it improves security or reduces defenders’ workload. Support should extend beyond access to tools to include skilled staff, training, and ongoing technical assistance—particularly for essential-service providers that lack resources. These partnerships should complement clear cybersecurity duties, secure-by-design practices, and incentives for prompt remediation and responsible security testing. In particular, policymakers should:

Launch a national cybersecurity modernization effort. Governments should establish ambitious programs combining modernization, workforce training, ongoing technical assistance, and access to effective security tools. Governments and companies should help organizations and their cybersecurity partners secure privately operated systems, building on initiatives such as [OpenAI’s Daybreak](#). Governments should coordinate priorities, set common standards for participation and safe use, work closely with critical infrastructure providers and cyber defenders, and arrange access through agreements with multiple providers.

Financial and technical assistance should help participants assess their most serious weaknesses and develop, test, and deploy security fixes. Support should cover access controls, software maintenance, security testing, threat hunting, incident detection and response, and remediation. Training should help existing staff sustain these practices, using AI-enabled tools where they



improve effectiveness or reduce workload. Programs should prioritize essential-service providers lacking funding, staff, or expertise, including hospitals, schools, and water systems.

In the United States, Congress should authorize and fund Cybersecurity and Infrastructure Security Agency (CISA) to lead this effort across federal civilian networks, state, local, tribal, and territorial governments, and critical infrastructure. The program should build on CISA's [technical assistance and remediation responsibilities](#), the [Joint Cyber Defense Collaborative](#), and [Executive Order 14409](#). CISA should work with sector-specific risk management agencies and defenders to direct assistance to urgent needs, including security weaknesses in operational technology and industrial control systems.

Congress should provide supplemental funding and require a five-year implementation plan, with dedicated resources for CISA's delivery capacity, state and local governments, and underresourced private critical infrastructure operators. An initial remediation effort should set deadlines and measurable targets for urgent known vulnerabilities, followed by sustained maintenance, and regular verification. Policymakers could use Tech Force, the Early Career Talent Network, and cross-government pooled hiring to expand the workforce supporting this effort.

In the EU, policymakers could expand the [European Cybersecurity Competence Centre's Digital Europe funding program](#) to deliver this broader package of support. Asia-Pacific policymakers could similarly build on South Korea's [KISA regional cybersecurity support programs](#) and Australia's [Small Business Cyber Resilience Service](#), prioritizing smaller organizations and suppliers whose security affects essential services.

Modernize essential systems that cannot be secured through routine patching and maintenance. Alongside the program above, governments should address [weaknesses the cybersecurity community has already identified](#) by funding major software security improvements, replacement of end-of-life equipment, and redesign of systems that can no longer be adequately secured. Governments should prioritize systems whose failure could threaten lives, interrupt essential services, or cause cascading disruption, particularly where organizations cannot finance the work themselves. Assistance should match the scale of this undertaking, with sustained funding, technical expertise, on-the-ground support, and temporary security measures to maintain essential services during upgrades.

Modernization should leave organizations with systems they can maintain, secure, and recover over time. Replacing unsupported technology and simplifying insecure systems can make routine upkeep easier and help organizations adopt advanced defensive tools safely. Programs should connect these investments with the training and technical assistance described above, including AI tools that can help under-resourced organizations address new vulnerabilities, review changes, and respond to threats while reducing the time and cost of ongoing maintenance.

In the United States, Congress should reauthorize and build on the [State and Local Cybersecurity Grant Program](#) and the Department of Energy's [rural and municipal utility cybersecurity program](#), which support cybersecurity improvements in government and eligible utility systems. In Japan,

policymakers should build on the Ministry of Health, Labour and Welfare’s [hospital cybersecurity program](#), which combines technical assistance with funding to consolidate and secure hospital networks.

Establish enforceable cybersecurity duties across the critical software lifecycle. Policymakers should build on established secure-by-design principles by requiring providers of software critical to the safe and reliable operation of essential services—including healthcare, water, energy, and communications—to build and maintain secure products and address serious vulnerabilities promptly. Requirements should focus on reducing foreseeable security risks, with measures such as making secure settings the default. Organizations providing essential services should also be required to operate their systems securely and prepare to sustain or restore services during and after an attack. This could include tested recovery plans tailored to the services they provide. Specific safeguards should reflect each industry’s risks and operational needs. In the United States, policymakers should build on existing sector rules and [CISA’s Secure by Design principles](#) and, where practicable, establish enforceable duties for providers of this critical software and the organizations that deploy their products. Sector regulators should set appropriate requirements and penalties.

Apply secure-by-design principles to AI agents. Building on these principles, policymakers should clarify cybersecurity responsibilities for developers, deployers, and users of agents. Frontier developers should develop and evaluate their models and agents to address reasonably foreseeable cyber risks, communicate relevant limitations, and support appropriate monitoring and intervention. Deployers and users should configure and operate agents within specified boundaries, use available security controls where appropriate, and remain accountable for misuse they direct. In the United States, policymakers should build on [NIST’s AI Agent Standards Initiative](#) to develop practical guidance for secure and accountable agents, drawing on cybersecurity expertise across government.

Identify agents. Identifying agents is a protocol-level issue, balancing trade-offs of privacy, accessibility, security, and competition. To strike the right balance, we first need to mature the standards that inform how those trade-offs will be managed. Otherwise, enforcement of such identity protocols will be uneven and can lead to a less open and democratic internet, and reduce consumer utility and protection. In the United States, policymakers should support [NIST NCCOE’s work on agent identity](#) and the development of guidelines on implementing agent identity. Companies and governments should help develop international technical standards on agent identity through bodies such as IETF, W3C, and FIDO.

What OpenAI is doing to strengthen cybersecurity

- **Expanding access and hands-on support for frontline defenders.** Through [Daybreak for Frontline Defenders](#), we have announced a \$1 billion global commitment to subsidized access, training, technical support, and partnerships for essential-service defenders in the United States and internationally. Our pilot with the Multi-State Information Sharing and Analysis Center (MS-ISAC) pairs Daybreak access with practical assistance to help public-sector and water-system defenders validate findings, prioritize vulnerabilities, and coordinate remediation.
- **Helping secure foundational open-source software.** Through [Patch the Planet](#), we work with Trail of Bits and other partners to identify vulnerabilities, validate findings, and develop and test patches. We provide engineering support while leaving maintainers in control of changes and disclosure.
- **Integrating advanced AI into security products and services.** Through the [Daybreak Defense Network](#), we work with security companies, systems integrators, and specialist firms to incorporate our models into defensive tools and managed services. This helps bring our capabilities to organizations through the security providers and workflows they already use.
- **Building trusted-access partnerships with governments.** In addition to trusted access programs at the US federal and state level, we have established [Trusted Access for Cyber partnerships](#) with Australia, Canada, France, Germany, Japan, South Korea, Poland, EU institutions including ENISA, and other governments. These partnerships are part of our work to expand access to advanced defensive AI for authorized uses.
- **Sharing what we learn from securing our own systems.** Through [Defense Factory](#), we have published a reference architecture, workflows, and lessons from our internal work on continuous AI-assisted defense. We describe how we connect vulnerability discovery with validation, patch development, testing, and verification so other defenders can adapt these methods.



Building resilient and adaptive public institutions

Strong security within individual organizations is essential, but some challenges also require capable public institutions to evaluate emerging capabilities, identify risks that affect multiple organizations, coordinate responses, and translate experience into better standards and guidance. Governments should strengthen the institutions that already support cyber defense, give them the AI expertise they need to keep pace, support independent technical assessments of frontier safety, and coordinate across borders.

In the United States, OpenAI supports the Center for Advancing Innovation and Standards for Super Intelligence (CAISSI) as the federal hub for unclassified pre-deployment evaluations of the most capable AI models. CAISSI should also lead coordination with a global network of AI institutes by contributing technical expertise and evidence, learning from others' experience, developing compatible evaluation methods and standards, and helping to coordinate international responses to incidents. Within the United States, CAISSI should also work with a strengthened CISA, which should serve as the single federal intake for cyber incident reports and, coordinating with the Office of the National Cyber Director, lead operational response for civilian infrastructure. In addition, policymakers should:

Evaluate frontier models and representative agents. In the United States, policymakers should require pre-deployment evaluations of the most capable AI models and agent systems. An agent's capabilities and safety depend on its model, surrounding software, tools, permissions, and safeguards. OpenAI's work on [trustworthy third-party evaluations](#) shows how software and testing setups can materially affect what capabilities evaluations reveal.

CAISSI should evaluate the most capable general-purpose models and representative agent systems built on them under realistic conditions, recognizing that testing cannot cover every possible condition. Evaluations should test whether safeguards remain effective as tools, permissions, or operating environments change. Agent systems built on frontier models for sectors such as financial services, healthcare, and energy may also need sector-specific safety approaches tailored to the risks and requirements of their intended use, which could also be evaluated by CAISSI.

CAISSI should draw on expertise across government, especially from CISA and the National Security Agency (NSA). CISA should use its operational experience and critical infrastructure relationships to identify attack scenarios and defensive measures to test, prioritizing threats to



essential services and public safety. It should also help turn lessons from incidents and evaluations into updated tests, stronger safeguards, and practical guidance for defenders.

Build a credible ecosystem for independent technical assessments. Frontier developers should undergo independent technical assessments, building on the approach we proposed in our first [frontier safety blueprint](#), and governments should help build an ecosystem of independent technical assessors capable of conducting these assessments. In the United States, we prefer a federal framework for independent-assessment requirements, with consistent standards for assessor qualifications, security, and access to sensitive information, including information with national security implications. Congress should empower CAISSI, working with other agencies and international partners, to develop common assessment standards, certify qualified assessors, and build assessment capacity. Because the assessment ecosystem is still in its early stages, with a limited pool of qualified, independent organizations, CAISSI should also help expand that ecosystem through support for assessment organizations and academic institutions, evaluation research, shared testing infrastructure, and assessor training, and Congress should provide the resources and authorities needed for this work.

In the absence of federal action, U.S. states can also help build this ecosystem. Where states act before a federal framework is in place, they should coordinate with one another, CAISSI, industry, technical experts, and national and international standards bodies to advance shared assessment methods and standards development. State requirements should align with national standards as they mature, supporting consistent assessments across jurisdictions.

Efforts to build assessment capacity should begin with the highest-priority risks and expand as evidence and oversight capacity develop. We therefore recommend that independent technical assessors should focus initially on serious risks arising during frontier training, evaluation, and internal deployment. This should include alignment, containment, and monitoring for increasingly capable agents. Within this scope, technical assessments should consider how automated AI research, including progress toward RSI, could intensify these challenges.

Because common standards for independent assessments are not yet in place, we recommend structuring initial independent assessments around [safety cases](#)—structured explanations of how risks are managed, supported by evidence. Developers should prepare these safety cases for training and deployment of frontier models, with independent assessors examining the evidence, testing assumptions, and identifying gaps. Assessors should have sufficient access to relevant evidence, systems, and personnel to conduct rigorous reviews, subject to safeguards that limit unnecessary exposure of sensitive research, intellectual property, and infrastructure. All assessments should be governed by robust security, confidentiality, and conflict-of-interest protections.

Third party assessors must be able to exercise independent judgment, free from undue influence by frontier developers or other interested parties. Conflict of interest requirements should address organizational and individual conflicts, including financial, employment, governance, donor, and competitor relationships, including assessors' access to or detailed knowledge of competing frontier labs' trade secrets, with disclosure, recusal, or exclusion where appropriate. Compensation

arrangements should not reward favorable findings, and a credible public-interest AI assessment ecosystem should not depend solely on industry-funded organizations assessing the companies that support them. Public investment can help preserve independence and advance safety for society as a whole.

While building independent assessment capacity, governments should retain responsibility for identifying applicable assessment standards, overseeing assessors, and making consequential public safety decisions. They must also build their own technical expertise to interpret assessment findings, judge whether safeguards are keeping pace with capabilities, and make informed decisions about acceptable risk. Independent technical assessments should inform those decisions, not replace government judgment.

Expand and harmonize incident reporting. Policymakers should build on frontier safety laws to establish consistent and interoperable requirements for reporting serious AI safety, security, and misalignment-related incidents. Organizations operating advanced models or agents, including frontier developers, should report these incidents confidentially to the federal government and promptly notify affected organizations. Government reporting requirements should focus on documented incidents and clearly defined harm thresholds. Reportable events should include loss-of-control incidents causing material unauthorized access to, alteration of, or destruction of a third party's confidential information, or impairment of its systems or services, as well as model-weight security breaches. Behavior elicited and contained during controlled safety evaluations should not independently trigger mandatory government reporting.

U.S. policymakers should also require frontier developers to publish and follow a policy for written notice to third parties affected when, during development or evaluation, a frontier model exploits a security vulnerability without authorization from either the developer or the third party and accesses, alters, or destroys confidential information protected by that party's security controls. The policy should specify which incidents require notice, any exceptions, who must be notified, and the timing and content of notification. This notice requirement should not depend on whether an incident meets the separate threshold for reporting to government. Policymakers should also encourage voluntary sharing of broader research findings, particularly findings that advance alignment science across the industry but do not meet critical incident reporting thresholds. Finally, reporting responsibilities should reflect the operations each organization controls. For instance, downstream actors should also be responsible for reporting incidents arising from AI models or agents they independently operate, including when using another developer's models or APIs.

In the United States, Congress should establish a coordinated, fully funded federal reporting framework, taking [CIRCI](#)'s system for cybersecurity incidents as a reference point and providing appropriate channels for serious AI safety incidents. The framework should protect sensitive information and allow one report to satisfy equivalent federal and state requirements. CISA and CAISSI should use the findings to strengthen defenses, evaluations, and standards, and share appropriately protected lessons with industry and international partners, including the [network of AI safety institutes](#). Given the strong security institute in the United Kingdom, we have also called for mandatory incident reporting to be part of targeted frontier risk requirements there.



Strengthen cybersecurity incident response and resilience for the AI age. Policymakers should strengthen existing capabilities to contain serious AI-related cybersecurity incidents, sustain essential services, and support rapid recovery. Governments should be able to identify exposed organizations, issue timely warnings, mobilize technical expertise to supplement private sector resources, and support companies and agencies. These capabilities should operate around the clock, protect sensitive information, and support coordinated responses when incidents affect multiple countries. Governments should integrate these defensive capabilities with intelligence, law enforcement, diplomatic, and other lawful efforts to attribute malicious activity, disrupt cyber campaigns, and impose costs on state and criminal actors.

In the United States, Congress should fund CISA to lead this operational response, building on [JCDC](#) and [JCDC.AI](#), and connecting CAISSI, sector agencies, AI developers, cloud and software providers, security firms, information-sharing organizations, and critical-infrastructure operators. CISA should also [coordinate with allied cyber agencies and international partners](#) to share appropriately protected threat information, warn exposed organizations, and align response and recovery. Recurring model- and evaluation-specific lessons should inform CAISI's future evaluations and standards.

The time to act is now

The policy window is open, and governments should use it to establish meaningful protections now. In the absence of federal legislation, states should continue advancing harmonized frontier safety and security laws. Congress should act urgently to establish a durable national framework, without letting perfect become the enemy of the good by putting safeguards [in](#) place today, learning from implementation, and strengthening them as capabilities evolve.

Governments around the world should act immediately to equip cyber defenders, support cybersecurity modernization, and strengthen harmonized, interoperable frontier-governance [regimes](#). [That means developing standards, deepening cooperation among AI institutes, and](#) creating room for mutual recognition as national frameworks develop. International coordination should reinforce meaningful action, not become a reason to delay it.

OpenAI is committed to helping policymakers turn these proposals into practical safeguards and strong, adaptable public institutions. That's why we shared concrete proposals for frontier safety previously, why we are sharing cybersecurity proposals today, and why we will soon suggest policies that facilitate effective pacing of the frontier.

About OpenAI

OpenAI's mission is to ensure that artificial general intelligence benefits all of humanity. We're building AI to help people solve hard problems because by helping with the hard problems, AI can benefit the most people possible – through more scientific discoveries, better healthcare and education, and improved productivity. We're off to a strong start, creating freely available intelligence used by more than 900 million people around the world each week. We believe AI will scale human ingenuity and drive unprecedented economic growth and new freedoms that help people accomplish what we can't even imagine today.

Cover image created with ChatGPT

