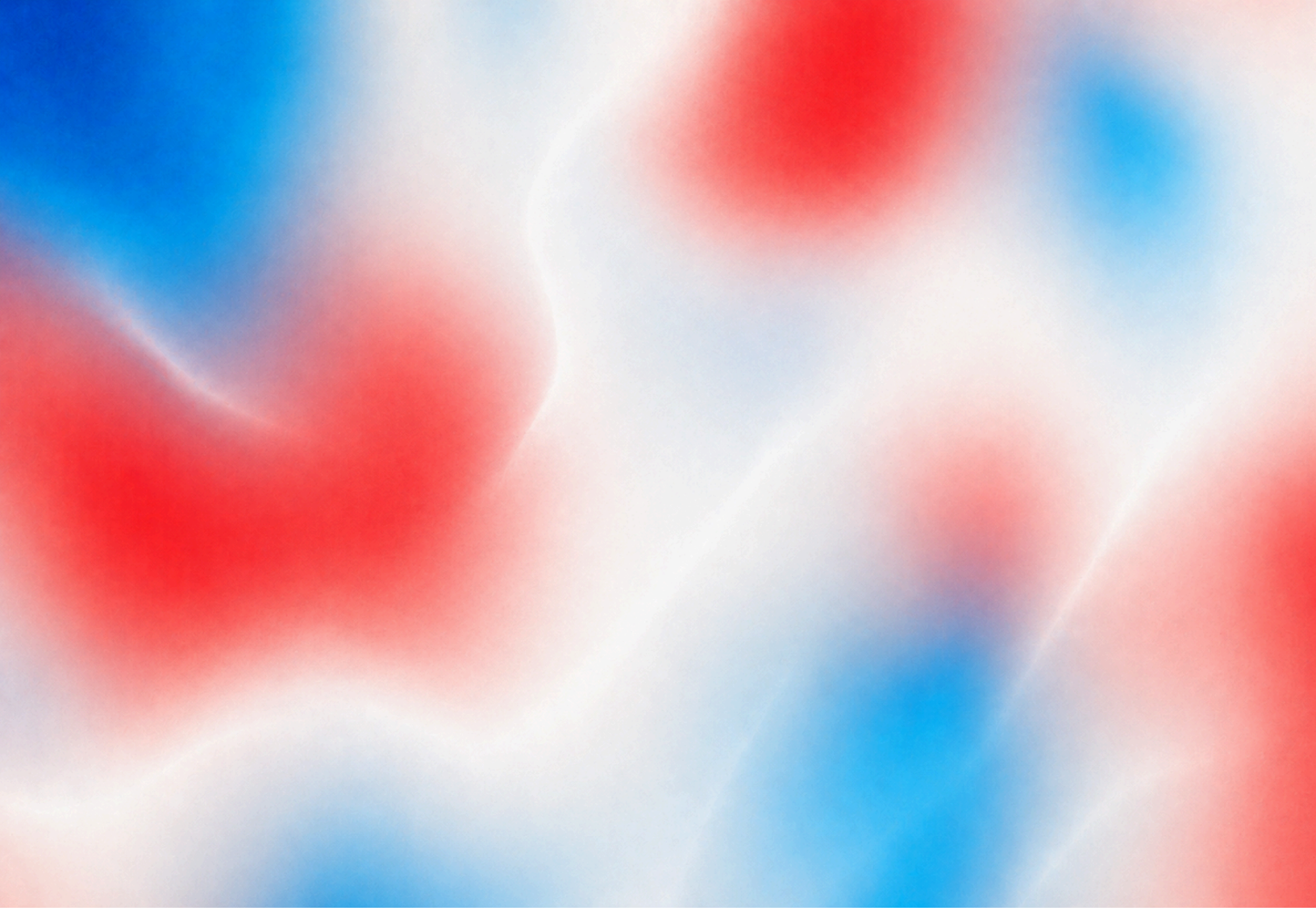

Cybersecurity in the Intelligence Age

An Action Plan for Democratizing AI-Powered Cyber Defense

April 2026

Artificial intelligence is reshaping cybersecurity. The same capabilities that help defenders identify vulnerabilities, automate remediation, and respond faster are also being used by malicious actors to scale attacks, lower barriers to entry, and increase sophistication. Recent incidents involving critical infrastructure disruption, large-scale ransomware, software supply-chain compromise, and increasingly sophisticated state-backed cyber activity have made clear that the threat environment is accelerating. AI intensifies that reality, for defenders and attackers alike.

That reality forces a strategic choice.

One approach is to treat these systems as too dangerous for broad defensive use and limit them to a very small number of approved partners.

We believe that misses the central challenge. Attackers will not wait. Existing models are already useful for many cyber workflows and capabilities will continue to advance. Criminal groups will adopt whatever tools are available. The best way to reduce national risk is to responsibly equip and accelerate trusted defenders faster than adversaries can adapt. The task now is to move with purpose and urgency, not panic – to act decisively and put advanced capabilities in the hands of trusted defenders at every level, paired with the safeguards and oversight needed to use them responsibly.

The United States and its allies face a rapidly changing cyber threat environment, and private-sector innovators have an important responsibility to help meet that challenge. OpenAI takes that responsibility seriously, and this work builds on efforts we have been advancing for some time to support defenders, strengthen safeguards, and partner closely with governments and industry. This Action Plan has been informed by conversations with cybersecurity and national security experts across federal and state government and major commercial entities. It consists of five pillars:

1. Democratizing cyber defense
2. Coordinate across government and industry
3. Strengthening security around frontier cyber capabilities
4. Preserving visibility and control in deployment
5. Enabling users to protect themselves

The plan describes how we will deepen our existing commitment by building the infrastructure needed to support cybersecurity defenders, organized around democratizing access to the defensive tools that trusted actors across society should be able to use. Building resilience in the Intelligence Age will require both working through democratic institutions and processes, and broadening access to the technologies that can help protect communities, critical systems, and our national security.

Sasha Baker
Head of National Security Policy
OpenAI

New Era for Cybersecurity

We are entering a world where frontier cyber capability is unlikely to remain concentrated within a small number of leading American labs for long. Advanced AI capabilities tend to diffuse quickly: techniques spread, competitors adapt, research gaps narrow, and open-source alternatives improve over time.

The question is not whether advanced cyber-capable AI will become globally available—it almost certainly will—but whether democratic societies can collectively maintain today's temporary capability lead into lasting defensive advantage before malicious actors catch up.

At the same time, the near-term threat environment is already evolving. Malicious actors are using AI to improve phishing, automate reconnaissance, accelerate malware development, evade detection, and increase the scale of cyber operations. These groups don't need the most advanced frontier models to cause real harm; even capable mid-tier systems can provide meaningful operational advantage. They are also operating against a digital environment already burdened by longstanding weaknesses: aging and end-of-life systems, inconsistent patching, insecure-by-design software, and vulnerabilities across widely used open-source dependencies. Taken together, these realities suggest the right framework is neither unrestricted release nor a model where frontier cyber capabilities are confined to a small number of selected organizations.

Cybersecurity is a team effort.

AI capability has reached a new level. That creates a new responsibility to ensure the operating model keeps pace, with clear rules, strong safety practices, and broad access so more people, customers, and developers can benefit. National resilience depends on a broad constellation of defenders—including government agencies, critical infrastructure operators, financial institutions, software providers, cloud platforms, security companies, and state and local governments. The right approach is controlled acceleration: moving quickly enough to put advanced capabilities in the hands of trusted defenders across that ecosystem, while preserving the safeguards, monitoring, and intervention tools needed to prevent misuse, respond to emerging threats, and continuously strengthen security over time.

More broadly, cybersecurity is only one example of how advances in AI will require communities, institutions, and industries to work together in new ways to manage risk and capture public benefit.

Five Pillars for National Cyber Defense

1. Democratizing cyber defense.

A central part of our strategy is to get the best available cyber-capable models into the hands of trusted defenders quickly, responsibly, and with controls that scale with risk.

Trusted Access for Cyber (TAC) is our mechanism for doing that. It creates a pathway for legitimate cyber experts to access more capable and more permissive models for defensive work, while preserving safeguards against misuse. Trusted Access uses graduated tiers based on trust, mission need, and defensive impact—from standard users who are interested in hardening their personal code, to organizations capable of protecting others at scale.

The core principle is simple: the more powerful or permissive the capability, the stronger the vetting, security commitments, monitoring, and use-case requirements. The goal is to reduce unnecessary friction for legitimate defensive work while preventing destructive, disruptive, or malicious cyber activity.

In the coming days, OpenAI plans to expand the Trusted Access program by:

- Expanding access for government defenders at every level. Cybersecurity responsibilities span federal, state, and local governments—from national security missions and threat response to public health systems, emergency management, benefits delivery, and local critical infrastructure. We are creating pathways for governments at all levels of scope and expertise to responsibly access these capabilities, including by expanding our existing Trusted Access for Cyber program to government users, alongside technical resources and support tailored to mission needs.
- Scaling through industry. We will prioritize industry actors whose defensive work can protect thousands or millions of downstream users, including major security platforms, hyperscalers, infrastructure providers, internet-facing technologies, critical infrastructure operators, and software-supply-chain defenders. We will also prioritize key sectors starting with the financial sector, one of the prime targets of sophisticated cyber threats, and one of the most critical to maintaining stability across the globe.
- Reaching smaller critical infrastructure providers. Many smaller hospitals, school districts, water utilities, municipalities, and local infrastructure providers lack the capacity to operate frontier cyber models directly. We intend to reach them through trusted intermediaries, including MSSPs, sector-specific organizations, major security vendors, and CISA-supported programs.

- Coordinating allied access. Cyber defense is a transnational challenge, and many critical systems are transnational. We intend to expand Trusted Access over time, working with trusted democratic allies and partners.

2. Coordinate across government and industry.

While the Trusted Access program helps get advanced cyber capabilities into the hands of the right defenders, access alone is not enough. To make this operational at ecosystem scale, we also need coordination mechanisms that allow government, industry, and frontier AI labs to act on the same threat picture, prioritize the most important defensive use cases, and share information quickly when abuse or emerging threats appear.

This work can build upon the work that governments already have underway to defend critical infrastructure, and draw lessons from established information-sharing across government and industry. In the era of highly capable cyber AI models, several near-term priorities stand out for further work:

- Aligning on the threat model. We want to work with governments to validate the long-term risk that competitors in adversarial jurisdictions will fast-follow these capabilities, and the near-term risk that malicious actors are already using multiple models, accounts, and platforms to scale cyber activity. A shared threat model will help align decisions about access, safeguards, monitoring, and escalation.
- Sharing operational threat intelligence faster. We need faster, more actionable sharing between government and industry about threat actors, infrastructure, tooling, tradecraft, targeting patterns, and safeguard-evasion techniques. That information will help us detect misuse earlier, tune safeguards more effectively, and ensure trusted defenders are using these capabilities against the highest-priority threats.
- Prioritizing sectors and use cases. Government can help identify the federal missions, critical infrastructure sectors, state and local systems, and software supply-chain risks where frontier models can have the greatest defensive impact.
- Coordinating through existing government channels. We want to plug into the structures that government already uses for cyber defense and incident response—including existing cyber defense, intelligence-sharing, and incident response channels. The government can take the lead in establishing a real-time coordination hub for AI-enabled cyber defense, allowing AI companies, cloud providers, and other key stakeholders to share threat information and deploy urgent mitigations more quickly.
- Strengthening cross-lab coordination. No single lab will have the full picture. We want to support faster cross-lab sharing through the Frontier Model Forum or similar trusted mechanisms, especially for abuse patterns, indicators, infrastructure, tactics, and emerging threat activity.

Trusted Access is how we get the best tools into the hands of defenders and other responsible actors. Coordination is what turns that into a national cyber-defense capacity. We are taking steps to work with government and industry partners to build the channels needed to move from individual access decisions to a coordinated defensive ecosystem.

3. Strengthening security around frontier cyber capabilities.

One of the most important safety controls for frontier cyber-capable AI is preventing unauthorized access to the model, its weights, and the operational knowledge surrounding it. Responsible deployment therefore requires more than user-facing safeguards. It also requires robust internal security measures designed to reduce the risk of leaks, theft, unauthorized replication, or distillation by malicious actors.

We are continuing to raise the security baseline across our research and production environments. That work includes tighter access controls, stronger segmentation of sensitive environments, enhanced monitoring, software and hardware supply chain security, and more rigorous protections for high-value assets.

We are also partnering with leading organizations to independently stress test and evaluate our security posture. Outside expertise is essential to ensuring our defenses keep pace with a rapidly evolving threat landscape. As one example, we recently announced an expanded partnership with Microsoft focused on collective defense efforts to help protect our infrastructure and disrupt threat actors seeking to misuse our technology.

Insider risk is another critical dimension of frontier model security. We are strengthening need-to-know access, anomaly detection, auditability, privileged-access governance, and investigation-ready telemetry across critical surfaces. As frontier capabilities become more strategically significant, protecting against insider compromise must be treated with the same seriousness as defending against external intrusion.

We are increasingly using our own technologies to augment defenders internally, including helping identify vulnerabilities in code, surface suspicious network activity, accelerate defensive workflows, and disrupt malicious uses of AI. Frontier cyber capabilities need to be deployed for both securing customers externally and strengthening our own resilience as well.

Finally, internal security extends beyond our own perimeter. The modern digital ecosystem depends heavily on open-source software, and vulnerabilities in widely used libraries can create systemic risk. We are investing in efforts to help secure the open-source supply chain by supporting maintainers, improving vulnerability detection and remediation, and helping defenders everywhere benefit from stronger, more accessible security tools. A more secure, resilient ecosystem is in everyone's interest.

4. Preserving visibility and control in deployment.

Broadening access to advanced cyber-capable AI must be paired with the visibility and control needed to detect misuse, enforce safeguards, and respond as the threat landscape evolves. Deployment is not binary; responsible deployment includes controlling who has access to what capabilities and maintaining oversight to ensure they are not abused.

For general users, we will continue investing in robust default safeguards, such as model behavior (i.e. what an AI model does and refuses to do) and automated system protections (i.e. classifier-based detection of potential abuse). For higher-trust and mission-relevant users who are given access to more permissive and capable AI models for cyber tasks, access should be tiered based on identity, use case, security posture, and defensive impact. As models become more capable or permissive in dual-use cyber areas, obligations should increase accordingly, including identity verification, legal attestations, baseline security commitments, abuse reporting, and monitoring designed to reduce misuse while enabling legitimate defensive work.

Beyond real-time safeguards, we maintain additional layers of detection through offline monitoring and threat-intelligence enrichment. As our models become more capable, we are partnering across the industry and with our customers that have high volumes of cybersecurity-relevant workflows to ensure we collectively have insight into activity that may involve cyber abuse. This includes monitoring and evaluation of potential misuse against trusted intelligence sources to identify potential threat actors or compromised accounts. This provides a second line of defense beyond front-end controls and helps inform enforcement actions, partner notifications, and future improvements to models or policies. Our goal is a risk-based framework that preserves privacy where appropriate while ensuring misuse can still be effectively detected and disrupted.

Finally, responsible deployment requires maintaining post-launch levers to respond as the threat environment evolves. If risk increases, we must adapt our default configurations, applying more restrictive blocking, introducing account-level friction, reducing quotas, requiring reauthentication, downgrading access tiers, or removing access altogether where abuse is detected. Static safeguards are not enough in a dynamic threat environment. Effective safeguards require the ability to adapt quickly, proportionately, and based on evidence.

5. Enabling users to protect themselves.

Every person deserves access to tools that help them stay safe online. The greatest public benefit of advanced cybersecurity tools will come not from helping a handful of chosen users, but from raising the baseline of security across society. Cybersecurity is no longer only an enterprise or government challenge. Individuals and families are increasingly having to deal with the potential consequences of cyber misuse—targeted by phishing, fraud, identity theft, account compromise, and increasingly

sophisticated scams enabled by AI. National cyber resilience therefore depends on widening access to defensive benefits, not concentrating it.

AI can help close that gap by making expert-grade security guidance available to ordinary users in moments that matter. ChatGPT can help users identify suspicious messages, understand potential scams, secure their accounts, adopt stronger passwords and multifactor authentication, respond to breaches, and recover more quickly from fraud or compromise. ChatGPT users are already using it to protect themselves from digital threats—they send over 15 million messages per month to ChatGPT asking it to check if something is a scam—and we can continue to build on this. For many people, the hardest part of cybersecurity is not willingness—it is knowing what to do, quickly and confidently, when something goes wrong. AI can help translate complex security problems into clear, practical next steps.

We also see an opportunity to better protect households, parents, seniors, and small businesses that often lack dedicated security resources but face real and growing risk. In the coming days, we will be introducing additional security features to help users better secure their ChatGPT accounts, alongside continued investments in tools and guidance that make personal cyber hygiene simpler and more accessible. A safer digital ecosystem depends not only on protecting critical infrastructure, but on helping millions of people become harder targets and more capable digital citizens.

The Strategic Imperative

AI will reshape cybersecurity whether institutions are ready or not. This moment is a wake-up call, and we have a limited window of time to get it right.

That means ensuring frontier capabilities reach trusted defenders who can protect others at scale; strengthening coordination across the public and private sectors so threats can be identified and disrupted faster; securing frontier models, systems, and sensitive knowledge against theft or misuse; preserving visibility into higher-risk deployments so abuse can be detected and addressed; and widening access to cyber defense so individuals, small businesses, and underserved organizations are not left behind.

We are confident that advanced AI can shift the balance toward defense rather than offense: faster patching, stronger resilience, smarter detection, more secure infrastructure, and a broader community of empowered defenders. If we move quickly and responsibly, the United States and its allies can turn today's lead in capabilities into a real and lasting cyber-defense advantage.

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