

PRINCIPLES for

Safe and Secure American AI

Frontier AI systems require constant oversight and vigilance at home and abroad. The United States should establish clear, transparent, enforceable safety standards, developed by federal experts at NIST in coordination with agencies such as DOE, DOD, and NSA. Covered models should not be released until they have undergone an independent audit confirming that they meet these safeguards and standards. It is past time for Congress to pass legislation establishing safeguards based on these principles:

Enforced Standards

SET STANDARDS FOR SAFE AI DEVELOPMENT AND DEPLOYMENT

- **Federal Standards:** NIST should develop clear, transparent, measurable, risk-based standards for AI systems that could cause catastrophic harm
- **Catastrophic Risks:** Standards must address catastrophic risks, including AI-enabled cyberattacks; chemical, biological, radiological, and nuclear threats; potential for loss of human control; and automated improvement and autonomous agents escaping secure testing environments
- **Risk Management and Controls:** Standards should establish proportionate safeguards and risk-management requirements, with heightened review and oversight for systems that exhibit dangerous capabilities or are capable of self-improvement that may bypass safety controls
- **Open-Source Protections:** Specialized standards should be developed for open-source AI models to reduce the risks of catastrophic misuse while providing developers with the clear guidance necessary to accelerate U.S. leadership of the open-source ecosystem
- **Keeping Pace:** Federal standards should be regularly updated as AI evolves

Independent Testing

VERIFY AI SAFETY THROUGH CONTINUAL REVIEW, CONSTANT OVERSIGHT

- **Continuous Testing:** Government and independent experts should provide constant oversight and continually scrutinize and stress test covered AI models against catastrophic risk standards, particularly where testing requires specialized expertise, sensitive information, or secure facilities
- **Built-In Safeguards:** AI systems must be stress-tested to ensure they halt self-modification that bypasses safety controls, monitoring, or human oversight
- **Independent Audits:** Audit firms should conduct comprehensive audits of AI developers, akin to outside auditors examining a company's financial statements, to ensure AI products are meeting applicable federal standards and safeguards

Transparency and Accountability

SAFETY REQUIRES DISCLOSURE, OVERSIGHT, AND CONSEQUENCES

- **Independent Access:** Developers should provide qualified independent testers and auditors with access and information necessary to assess catastrophic risks, investigate serious incidents, and verify that safety and security safeguards are functioning as intended
- **Public Disclosure:** Developers should provide clear, plain-language disclosures of material safety and security risks, and mitigation modeled on the risk disclosures public companies provide to investors
- **Incident Reporting:** Developers must promptly report material AI safety and security incidents, including loss of control over autonomous agents, dangerous cyber capabilities, failures of critical safeguards, and unsafe recursive self-improvement
- **Accountability:** Developers must remain liable under applicable civil and criminal law for foreseeable harms and unlawful conduct arising from their AI systems, including when their design, deployment, modification, or failure to implement reasonable safeguards materially contributes to that harm
- **Whistleblower Protection:** Employees and contractors who report safety, security, or compliance concerns should be protected from retaliation

Public Private Partnerships

GET THE BEST EXPERTS WORKING FOR THE PUBLIC GOOD

- **Defensive AI:** Government should work with industry to develop and deploy defensive AI that detects and contains malicious activity, identifies and patches cyber vulnerabilities, and protects critical infrastructure from increasingly automated attacks
- **Shared Capabilities:** Leading AI companies should contribute compute, technical expertise, and advanced defensive tools so small and medium sized businesses, critical infrastructure operators, and public institutions can benefit from capabilities they may not be able to build on their own
- **Public-Interest Compute:** Government and industry should expand access to advanced computing and AI tools for universities, researchers, and public-interest projects working on national security, scientific discovery, health, energy, and other national priorities
- **Local Resilience:** Federal agencies should equip state and local governments and private-sector operators with technical assistance, training, and resources needed to protect critical infrastructure and essential services

Defend Against Social Harm

MAKE AI WORK FOR PEOPLE, NOT AGAINST THEM

- **Online Child Safety:** AI products used by children should protect their safety and healthy development, rather than exacerbate social media harms such as addictive design, harmful content, or data exploitation
- **Worker Opportunity and Protections:** AI companies should help fund apprenticeships, education, and other worker training so that AI expands workers' skills, productivity, and earning power rather than displacing them, while guardrails protect the products of human creativity and incentivize employers to use AI to improve jobs rather than eliminate them
- **Human Judgment:** When AI informs consequential decisions about employment, health care, credit, or other essential services, it must be transparent and fair and be subject to meaningful human oversight—particularly where AI-assisted prior authorization, such as CMS's WISeR model, could delay or deny needed care
- **Recourse:** People should be able to understand when and how an AI system materially affected a consequential decision, challenge the outcome, and receive timely review by a qualified human decision-maker

Advance Global Cooperation

LEAD GLOBAL ADOPTION OF TRUSTED AMERICAN AI

- **Align Standards:** The United States should work with allies and partners to adopt common, rigorous AI safety and security standards that reduce conflicting requirements and expand global markets for trusted American AI
- **Standards Leadership:** The United States should lead in international standards-setting bodies, including ISO/IEC, to translate U.S. technical leadership in AI testing, risk measurement, and safeguards into widely adopted global standards
- **Shared Risk Information:** Allies should establish secure mechanisms to share information on serious AI incidents, emerging catastrophic capabilities, evaluation methods, and effective safeguards
- **U.S.–China Crisis Channel:** The U.S. should establish a secure channel with China like the Soviet era red phone for rapid communication when a significant AI incident or system-control failure could threaten either country or international security
- **Secure Access:** Export controls should ensure that America’s most advanced AI technologies are not made accessible to our adversaries and available only to countries that protect them from diversion and misuse and uphold U.S.-led safety and security standards