Center for
AI Policy

2024 AI Action Plan

Artificial Intelligence: The Risks

Artificial Intelligence is rapidly becoming more powerful. It can already debug code, write poetry, translate languages, beat humans at chess, and more. At its best, AI will drive economic growth, improve medical diagnoses, and make educational resources more accessible. At its worst, AI could cause widespread harm due to misuse by malicious actors or ‘misalignment’ with human values.

For example, terrorists could use AI to create and deploy biological weapons. Rogue states may leverage AI to scale cyberattacks on critical infrastructure such as hospitals,reservoirs, and telecommunications. Elections may be influenced by the proliferation of highly realistic deepfake videos. As AI becomes increasingly integrated into our lives, the potential for harm grows.

What can be done now

Given the rate of technological development and the scale of these risks, we need to implement guardrails as soon as possible to ensure safe innovation. The Center for AI Policy would like to propose three actions that can make a difference now.

1. Whistleblower protections for employees of AI companies

Whistleblowers are one of the strongest incentives for companies to adhere to safe AI processes. Yet employees of AI companies are only tenuously covered by existing whistleblower laws. While there are protections for meat safety, aviation, and securities, there are no formal protections for AI employees. If we want employees to risk their careers by speaking up, we need to protect them.

2. “Wargaming” AI catastrophes for government preparedness

With the proliferation of AI, critical US systems and infrastructure face increased risk of attack. Government agencies already conduct wargames focused on natural disasters, physical attacks, and cyberattacks. However, they need additional resources to simulate AI-specific catastrophes. Only through practice can we better understand our vulnerabilities and refine our responses.

3. Reporting of cybersecurity standards

As AI becomes more powerful, it is critical to stop malicious actors from accessing unrestricted versions of models. AI companies need to ensure they have sufficient cybersecurity to protect these models both for public safety and their intellectual property. DoD has existing private sector cybersecurity tiers[1] against which AI firms should be required to assess and report their standards. Reporting their current security tier to the government (and any plans for improving it) is a low-touch incentive for AI companies to shore up security.

[1] Cybersecurity Maturity Model Certification (CMMC) program

Center for
AI Policy

How you can help

We need support from you to work towards a safer future. There is an opportunity to introduce these policies ahead of the upcoming election. Please reach out to info@aipolicy.us to start the conversation – if you have concerns, we want to hear about them, and if you are ready to endorse these policies, then we want to get you involved.

About us

The Center for AI Policy (CAIP)

The Center for AI Policy's mission is to ensure safer AI now and going forwards. We are a nonpartisan research organization dedicated to mitigating the catastrophic risks of AI through policy development and advocacy.

We're working with Congress and federal agencies to help them understand advanced AI development and effectively prepare for it. We share policy proposals, draft model legislation, and give feedback on others' policies.

AI poses many policy challenges including those related to privacy, discrimination, intellectual property, economic displacement, and climate. While we care deeply about these issues, our focus is predominantly on catastrophic risks.

AI policy briefings

We regularly host virtual and in-person briefings on key AI policy issues. Upcoming briefings include *Autonomous Weapons & Human Control* in July, *AI & Education* in September, and *AI & Intellectual Property* in November. Previous briefings focused on *AI & Privacy* and *AI, Automation, & the Workforce*.

Weekly Newsletter

The Center for AI Policy publishes a newsletter called **AI Policy Weekly**. Each week, we share clear and detailed analyses of three important developments that AI policy professionals should know about, especially those working on US federal policy.

Visit aipolicyus.substack.com or scan the QR code below to read a sample issue.