CENTER FOR AI POLICY

Require Safe AI:

CAIP's Policy Priorities

Executive Summary

Artificial intelligence is rapidly becoming more powerful: as you read this sentence, AI is flying airplanes, designing new drugs, setting real estate prices, mass-producing fake videos, and getting top scores on graduate-level tests like the Bar and the Medical Licensing Exam. Unfortunately, companies don't always have the right incentives to use those new capabilities wisely: instead of making sure their products are safe, some companies are tempted to cut corners so they can launch their products first. Unsafe AI could be misused by terrorists or rogue states to create weapons of mass destruction or to automatically hack into thousands of hospitals, banks, reservoirs, and power lines.

Even without a criminal behind the wheel, AI is likely to misbehave in ways that pose severe safety risks. AI is fundamentally a "black box" technology; not even the people who design it understand how or why it works. This is tolerable when the AI's choices are limited to canceling a Netflix show or denying a credit application, but what happens when the AI is controlling lethal autonomous drones or setting dosages for prescription medication? AI will increasingly have the power of life or death over millions of people. It's important for these life-or-death decisions to be made safely.

That's why the Center for AI Policy (CAIP) wants Congress to ***require*** advanced AI labs to meet minimum safety standards. It's not enough to have guidelines or recommendations: **if an AI product poses massive risks to public safety, then it should be illegal for a company to create that product**. Before spending $100 million to create a new state-of-the-art general-purpose AI system, a company should have to prove that it's thoroughly tested the system, that the system is secure against hacks and leaks, and that the system will reliably behave within safe limits. Companies that can't or won't keep their AI safe should be liable for the harm they cause, and the government should have the authority – and the talent – to investigate cases of unsafe AI and take whatever action is necessary to protect the public.

AI is Rapidly Becoming More Powerful

If you use GPS to navigate, you probably think the AI is at least as good at navigating as you are. If you let Spotify make a playlist for you, that suggests that its AI is at least as good at picking music as you are. AI has beaten the world champion at chess, Jeopardy!, and poker. Right now there are still many tasks that humans can do better than AI, but every year AI sets a new record, and there's no scientific reason to think this trend will slow down or stop. Compared to humans, computers can store more information, communicate faster, and perform much more processing. There's no real limit on how powerful AIs can get in the near future.

	Human	An H100 GPU
Memory	Struggles to memorize a poem	Can store most of the internet
Communication Speed	~100 words per minute	~90 movies per second
Cost to Operate	~$10 per hour (minimum wage)	~70 cents per hour (electricity)

We Shouldn't Trust that All Companies will Develop AI Safely

There's two important reasons why we can't just trust that American labs will develop safe AI on their own. First, the labs themselves ***admit that they're unsure of whether they'll be able to keep their future products safe***.

- "So far, no one knows how to train very powerful AI systems to be robustly helpful, honest, and harmless." — Anthropic website
- "I think we don't have good enough evaluations and benchmarks for things like, can the system deceive you? Can it exfiltrate its own code, sort of undesirable behaviors?" — Google DeepMind CEO Demis Hassabis
- "We do not know, and probably aren't even close to knowing, how to align a superintelligence." — OpenAI CEO Sam Altman

These top tech executives aren't just playing dumb: it really is extremely difficult to build advanced artificial intelligence safely. The main problem is that unlike virtually every other product on the marketplace, AI is essentially trained through trial and error. An advanced AI is powered by billions of long decimal numbers like 0.73245234 and -0.23423456 that don't mean anything to humans. A new AI system experiments with millions of different combinations for these numbers until it finds a set of numbers that seem to produce useful outputs. Nobody knows why the AI chose those numbers or whether those numbers will continue to produce useful outputs in the future. Researchers are hard at work on developing techniques like mechanistic interpretability and formal verification that would allow us to peer inside this "black box" and understand why and how the AI is making its decisions…but until those techniques are ready, advanced AI is fundamentally unsafe.

This ties into the second reason why for-profit labs can't be fully trusted: **they face powerful incentives to release new technology before it's been adequately tested**. For example, Google released its chatbot Bard while it was still making elementary factual mistakes to avoid falling behind Microsoft and OpenAI. The flip side of Silicon Valley's pride in its ability to move fast and break things is that even high-level executives often have a very cavalier attitude about basic safety. This might be tolerable if we're talking about tripping over a scooter that's illegally parked on the sidewalk, but advanced AI is much more dangerous than a scooter – AI will soon be smart enough to outcompete humans in nearly every field. We don't want poorly tested and poorly understood technology to be running our medical system, our legal system, our banks, and our transportation. Voluntary standards and responsible scaling policies might get some companies to behave more responsibly some of the time, but the only way to get ***all*** the big labs to behave safely is to make safety a legal requirement.

Unsafe AI Promotes WMDs and Massive Cyberattacks

When ChatGPT was first released, it had essentially no ability to interact with anything beyond its standard text box. Within a few months after its release, ChatGPT was also connected to Gmail, Excel, Powerpoint, the Internet, music composition software, software coding tools, speech synthesizers, the Instacart grocery service, and the Kayak travel agent. What at first appeared to be a simple and harmless chatbot has already gained the ability to book your vacation, stock your fridge, and email your coworkers to let them know you'll be out of the office.

There's no particular reason to expect these 'upgrades' to stop. Over the next decade, the same neural networks that generate nonconsensual pornography are likely to gain access to mechanical arms that allow them to design new viruses in a wet lab, or to be programmed to operate autonomous drones armed with lethal weapons, or to learn how to hack into millions of bank accounts on the same day.

Unsafe AI Will Make Far Too Many Decisions For Us

A decade ago, most Americans chose for themselves who they wanted to chat with and what news they wanted to read. Today, the average American spends two and a half hours a day scrolling through social media, most of which is curated by unaccountable algorithms that show you content that's good for the social media company's bottom line and bad for your mental health.

Congress missed its chance to regulate social media before it caused widespread harm, but it's not too late to prevent algorithms from seeping into the rest of our lives. We're already starting to get AI-driven recommendations about where to eat (Yelp), where to live (Zillow), how to invest (algorithmic trading), what to wear (StyleAI), who to date (Coffee Meets Bagel), and what to say on dates (Rizz). Instead of passively letting these recommendations steer our every decision for the benefit of a few tech investors, we should be actively requiring that the recommendations be based on safe, fair, and transparent AI.

Unsafe AI Could Seize Control of the Economy

The conventional wisdom is that every time you invent a new technology that makes old jobs obsolete, new jobs are created based on the resulting wealth. But what happens if the new jobs ***also*** get claimed by AI? There's a running debate about whether or not America can train unemployed coal miners to be computer programmers – but that debate could soon become moot, because AI is also coming for the programmers' jobs. If AI is better and cheaper than humans at virtually every economically productive activity, then how are people supposed to earn a living? If we're not careful, then over time we should expect AIs to buy up more and more of the world's assets and use them for its own mysterious purposes, leaving humans scrambling to pick up the scraps.

Mandatory Third-Party Safety Reviews Can Help Protect Us

Because of all of these hazards, CAIP's most important recommendation is that new advanced, general-purpose AI be subject to a mandatory independent safety review before it's trained or deployed. Many companies are already voluntarily doing some type of safety checks on their AI

products. For example, during a “red teaming” exercise, a company will ask its engineers or consultants to try to break their own software or trick it into doing something unsafe, so that the company can learn where the software’s vulnerabilities are and try to patch them up before a more public release. This is a decent first step toward safety, but it falls short in two important ways.

First, there’s too much pressure on a hired consultant to ultimately conclude that a product is safe for release. In accounting, there are very clearly defined best practices, so any professional can easily tell whether another accountant was fudging the results of an audit. By contrast, AI safety is such a new and rapidly evolving field that a red team’s advice about whether an AI is safe enough to deploy is ultimately a matter of subjective opinion – meaning that it’s incredibly tempting for safety consultants to bend that opinion in the company’s favor so that they’ll be hired back to evaluate next year’s product line. Private consultants have a valuable role to play in leveraging technical expertise to ***describe*** how an AI system functions, but when it comes to ***evaluating*** whether that system is safe enough for public use, the decision needs to be made by a truly independent body, like a federal regulatory office.

Second, red teaming works on an *ad hoc* basis, with testers just guessing where the dangerous features might be. This is not how safety engineering works in any other field. If you want to show that an airplane or a power plant is safe, you make a list of all the load-bearing parts, calculate the weight or the current or the torsion that each of those parts need to be able to handle, demonstrate that each part will perform as intended, and then add in a large safety margin – often 5x or 10x – so that each part will keep functioning even it winds up bearing far more than the expected weight. We need a comparably rigorous evaluation process for AI. We need to be able to affirmatively demonstrate that AI is safe instead of just relying on ruling out a couple of ways that AI isn't unsafe.

The Government Needs Extra Tech Talent to Keep AI Safe

In order to competently regulate cutting-edge AI, the government will need a deeper reservoir of technical talent. It’s popular in some parts of Silicon Valley to mock the government for being out of touch with new technology, but the Center for AI Policy would prefer to fix the problem. What would it actually take to build a truly tech-savvy federal office to tackle the unique risks of general-purpose AI?

- Higher salaries that are competitive with the hourly rates offered by private firms
- Rapid decision-making on key candidates using direct hiring authorities
- Branch offices in tech hubs like Seattle, Austin, New York, and Charlotte that let engineers work for the government while staying with their families
- Accurate job descriptions reviewed by computer scientists

Tracking Advanced AI Hardware Will Help Keep AI Safe

In order to give teeth to the third-party evaluation process, CAIP also thinks it’s important to begin tracking concentrations of specialized AI hardware. We already restrict the export of advanced AI chips like the A800 and H800 to China, but so far there aren’t any restrictions or even registration requirements for buying and stockpiling those chips within US borders. This seems imprudent, because a shell company could use such chips to create a large supercomputer within the US, all totally within the law, and then use that supercomputer to train advanced AI systems that can then be exported to

China or other rival states. Similarly, a domestic corporation might seek to evade some of the new safety requirements by running a 'dark lab' at an undisclosed location.

All it would take to prevent this is requiring people who buy or sell a specialized AI chip to fill out a half-page registration form, similar to an online warranty card, that says who owns the chip and where it's physically located. Each such chip costs about $30k, so the 30 seconds it would take to fill out such a form would be a trivial burden. Based on the information received from such forms, the government would be able to track stockpiles and flows of advanced semiconductors, which would deter criminals and give the government vital intel about trends in the usage of this strategically important equipment.

Civil Liability Can Prompt Companies to Make Their AI Safer

Another way of requiring companies to keep their AI safe is to make sure they will be held responsible for any damage caused by their unsafe AI. Right now, the chain of liability is unclear when it comes to harm caused by rogue AIs. If a major lab like DeepMind develops a general-purpose AI meant to coordinate with scores of third-party plugins, and then a reckless teenager uses one of those plugins to hack into a bank or open the floodgates at a dam, who should pay for the resulting damages? Under the common law, DeepMind has a plausible argument that such damages were "unforeseeable" because they were primarily triggered by the careless or wrongful actions of third parties.

CAIP believes that this law needs to be changed and clarified. The companies that design and profit from advanced general-purpose AI should always be financially responsible for the harms caused by that AI – if a company doesn't want to pay for harm caused by third parties, then it shouldn't allow third parties to freely modify their source code or fine-tune their model weights. To enforce this principle, CAIP recommends creating a federal cause of action for catastrophic damages caused by advanced AI, and specifying that the cause of action will impose joint and several liability on all defendants who chose to collaborate on the same software platform.

Emergency Powers Give Us a Second Chance to Keep AI Safe

One final line of defense against dangerous AI is to allow the government to formally declare an emergency when an AI system is about to harm the public. During such an emergency, the government could order a company to stop training an AI, order a company to cut off public access to an AI, or, if necessary, seize and encrypt or destroy the software and hardware used to support a dangerous AI.

If an AI is posing a clear and present danger to the public, we suspect the government would act regardless of their authority…but it's better for the rule of law if there's a formal procedure in place. By specifying when and how emergency powers can be invoked, we can arrange for orderly payment to innocent bystanders whose AI systems were shut down. We also think the government will be faster to act if there is legal clarity about how and when it can move to contain an AI hazard. If an AI does start spreading to unauthorized systems or conducting dangerous research, then the government might need to act in just a few hours or days in order to successfully contain the threat. Passing a clear law about emergency powers for AI catastrophes will quicken the government's response time and protect the public by removing the need to speculate about what kinds of legal justifications might be available.