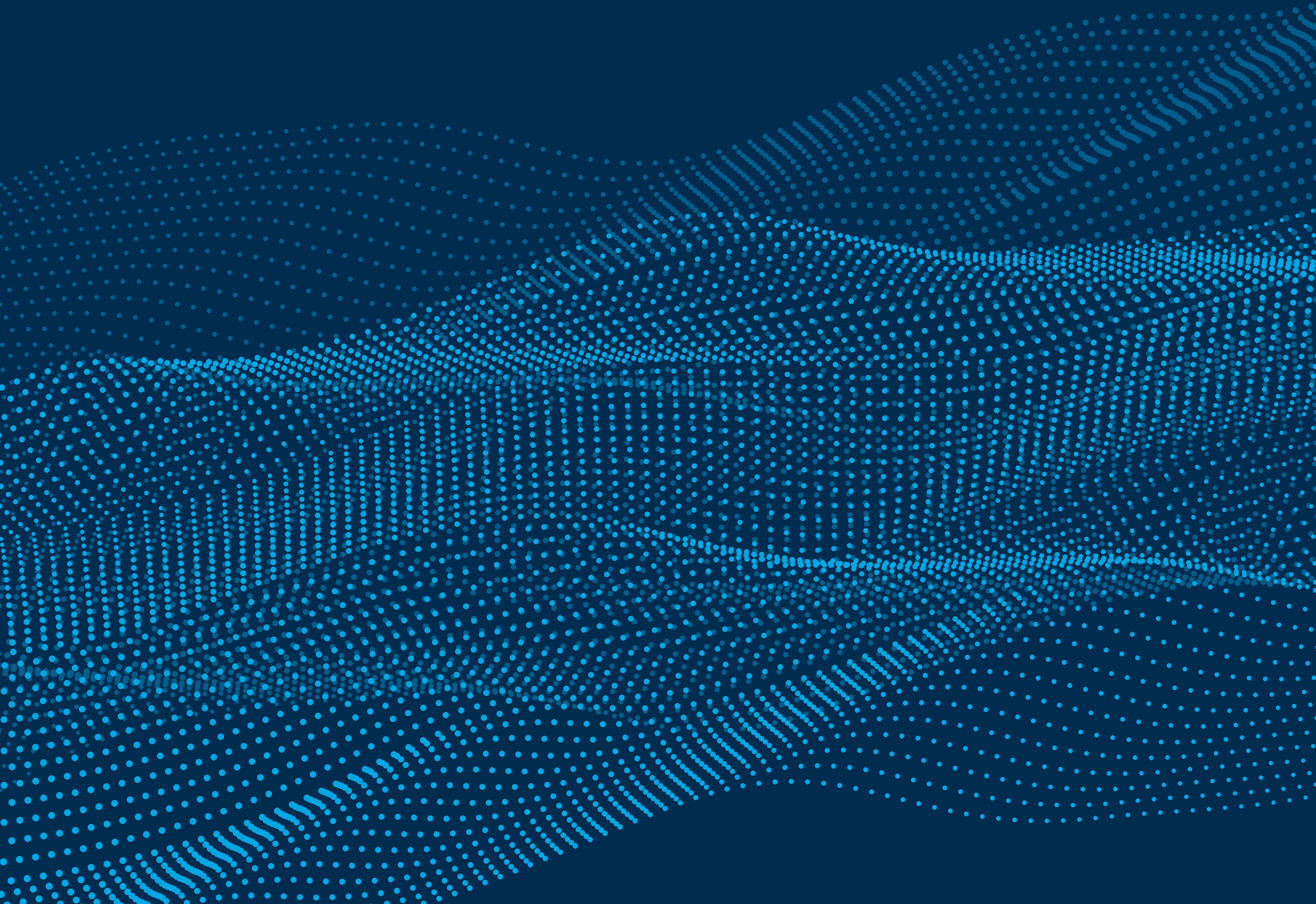
Center *for*
AI Safety



2023 IMPACT REPORT

In this Report

> “Preventing extreme risks from AI requires more than just technical work, so CAIS takes a multidisciplinary approach working across academic disciplines, public and private entities, and with the general public.”
>
> **– Dan Hendrycks** | *Executive Director, Center for AI Safety*

Center *for* AI Safety

The Center for AI Safety (CAIS) is on a mission to reduce societal-scale risks from artificial intelligence.

We believe that artificial intelligence has the potential to profoundly benefit the world, provided that we can develop and use it safely. However, in contrast to the dramatic progress in AI, many basic problems in AI safety have yet to be solved—leaving many risks unaddressed. **CAIS exists to mitigate these risks and ensure the safe development and deployment of AI.** To achieve this, we pursue three pillars of work: research, field-building, and advocacy.

Research

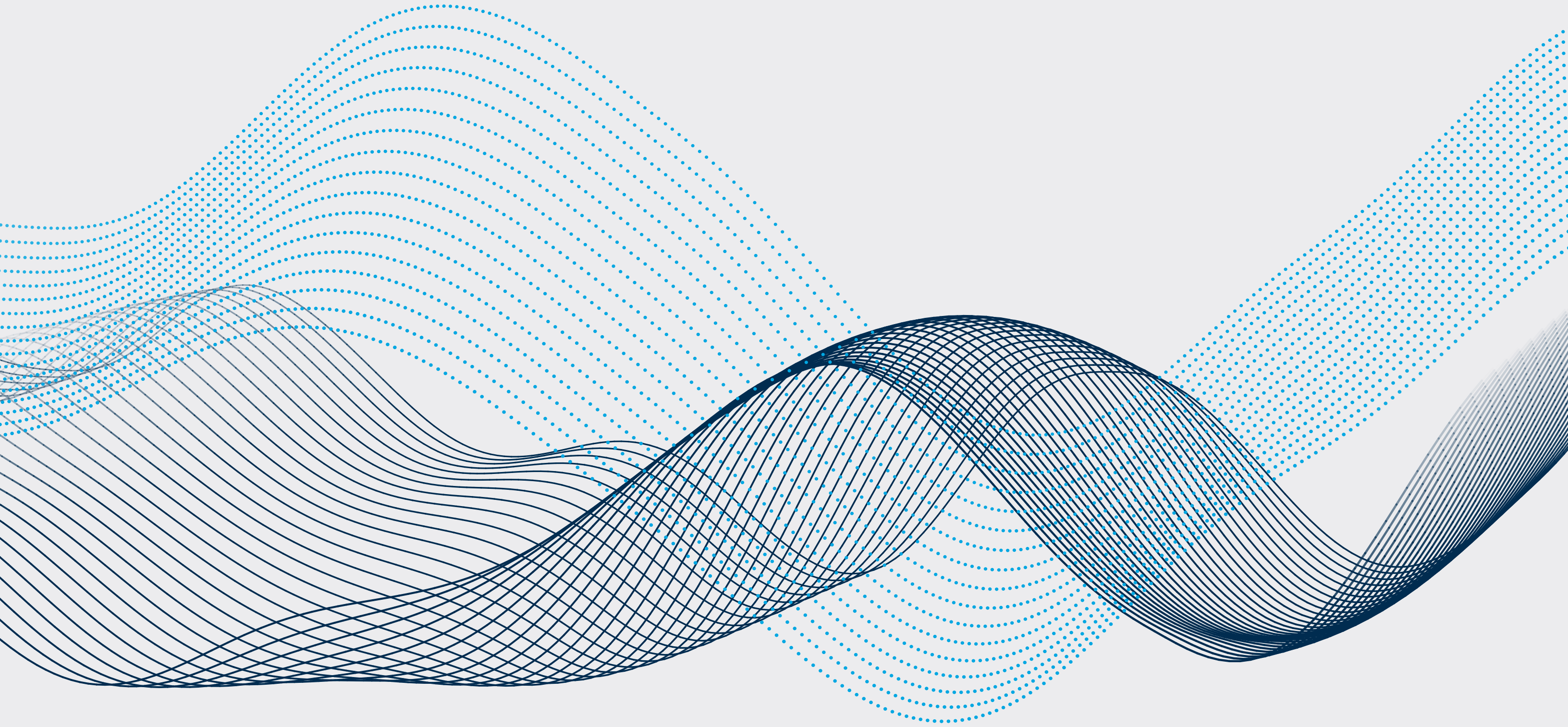
CAIS conducts research solely focused on improving the safety of AI systems. Through our research initiatives, we aim to identify and address AI safety issues before they become significant concerns.

Field-building

CAIS grows the AI safety research field through funding, research infrastructure, and educational resources. We aim to create a thriving research ecosystem that will drive progress towards safe AI.

Advocacy

CAIS works with industry leaders, policymakers, and other labs to bring AI safety research into the real world. We aim to build awareness and establish standards for safe development and deployment of AI.



Why AI Safety

AI is advancing at an alarming rate, and so are the associated risks

In just a few years, AI systems have learned to outperform 90% of lawyers at the Bar exam and discuss almost any topic fluently and at length. OpenAI, Google DeepMind, and other AI leaders anticipate the possibility of superhuman AI systems by the end of the decade.

AI has the potential to profoundly benefit the world, but these advanced capabilities also bring considerable risks. AI systems can already spread misinformation or generate sophisticated phishing email campaigns. And their decision-making processes are too complex for even experts to decipher. As these systems quickly become more intelligent, the risks could escalate beyond our control.

Reducing risks from AI has emerged as a global priority, ranking alongside pandemics and nuclear war. Despite its importance, AI safety remains remarkably neglected, outpaced by the rapid rate of AI development.

Impact by the Numbers

Here are some highlights from our work in 2023

1 Global statement on AI Risk signed by 600+ leading AI researchers and public figures.

RESEARCH

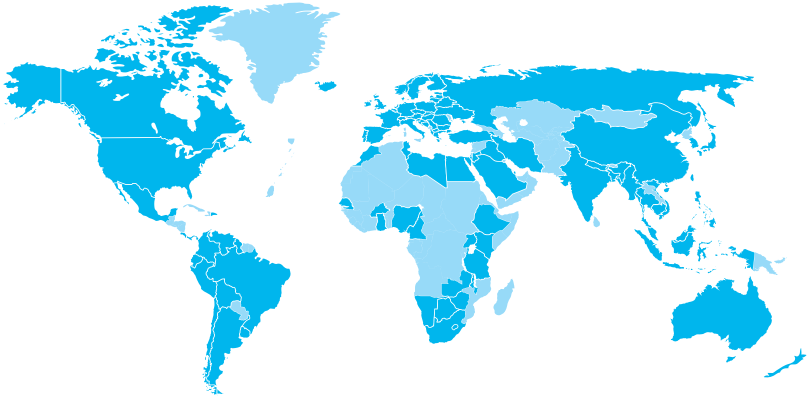
63	AI safety research projects enabled by the CAIS compute cluster.
9	AI safety research papers published.
55	Philosophy AI Safety papers produced.

ADVOCACY

100+	conversations with policymakers and government entities to advise them on technical AI safety.
600+	signatories on our Global Statement.
9000	subscribers to our AI safety newsletters.

FIELD-BUILDING

291	Students trained through our technical ML Safety Course.
25	Legal scholars and policy researchers convened to discuss AI safety, leading to the formation of a new AI safety & law consortium.
1000+	Machine learning researchers participated in our AI safety events at ML conferences.



123 COUNTRIES REACHED THROUGHOUT OUR ADVOCACY AND FIELD-BUILDING PROGRAMS

Research

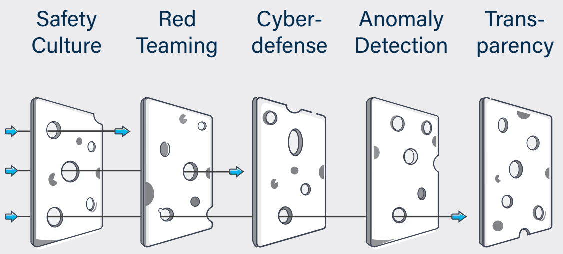
Conducting research solely focused on improving the safety of AI systems

Only 3% of technical research is focused on making AI systems safe, with the remaining efforts racing towards increasingly powerful AI. Experts believe this needs to be closer to 50/50.

To close this gap, CAIS conducts both Machine Learning (ML) and interdisciplinary research on AI safety.

In our ML research, we pursue multiple overlapping strategies which can be layered together to mitigate risk. In cybersecurity, this is known as “defense in depth” (Swiss cheese model). Though no individual technique brings risk to zero, we hope to build layered defenses which reduce risks to a negligible level.

We also examine AI safety from an interdisciplinary perspective, incorporating insights from safety engineering, complex systems, international relations, philosophy, and economics. Through this research, we create frameworks that aid in the understanding of the current challenges and publish papers that provide insight into the societal risks posed by current and future AI systems.



The Swiss cheese model shows how technical factors can improve organizational safety. Multiple layers of defense compensate for each other’s individual weaknesses, leading to a low overall level of risk.

Research Highlights

In 2023, CAIS published research papers addressing a spectrum of AI safety issues. These include techniques for interpreting an AI system's 'thoughts' and control its behavior, methods for measuring an AI system's ethical behavior, and methods to easily circumvent Large Language Model (LLM) safeguards. In total, CAIS's research team has completed nine papers which have all been published at or submitted to top conferences.

MACHINE LEARNING RESEARCH

Representation Engineering: Modern AI systems are known to be opaque and difficult to understand. We developed a general technique to directly read an AI system's 'thoughts' and control its behavior. Our experiments show **these techniques can detect when an AI is hallucinating and can make AI systems more truthful, power-averse, and ethical**. *This work was done in collaboration with researchers from organizations, including Carnegie Mellon University, UC Berkeley, and Stanford University.*

Large Language Model (LLM) Attacks: We showed that it was possible to automatically bypass the safety guardrails on GPT-4 and other AI systems, **causing the AIs to generate harmful content such as instructions for building a bomb or stealing another person's identity**. *This work was done in collaboration with researchers from Carnegie Mellon University, Bosch Center for AI, and Google DeepMind.* (*New York Times*)

DecodingTrust: We provided a thorough assessment of trustworthiness in GPT models, including toxicity, stereotype and bias, robustness, privacy, fairness, machine ethics, and so on. *This was part of a large research collaboration led by the University of Illinois Urbana-Champaign.* **It won the outstanding paper award at NeurIPS 2023.**

MACHIAVELLI Benchmark: We evaluated the tendency of AI systems to make ethical decisions in complex environments. The benchmark provides 13 measures of ethical behavior, including measures of whether the AI behaves deceptively, seeks power, and follows ethical rules. *This work was done in collaboration with researchers from UC Berkeley, Carnegie Mellon University, and Yale University.*

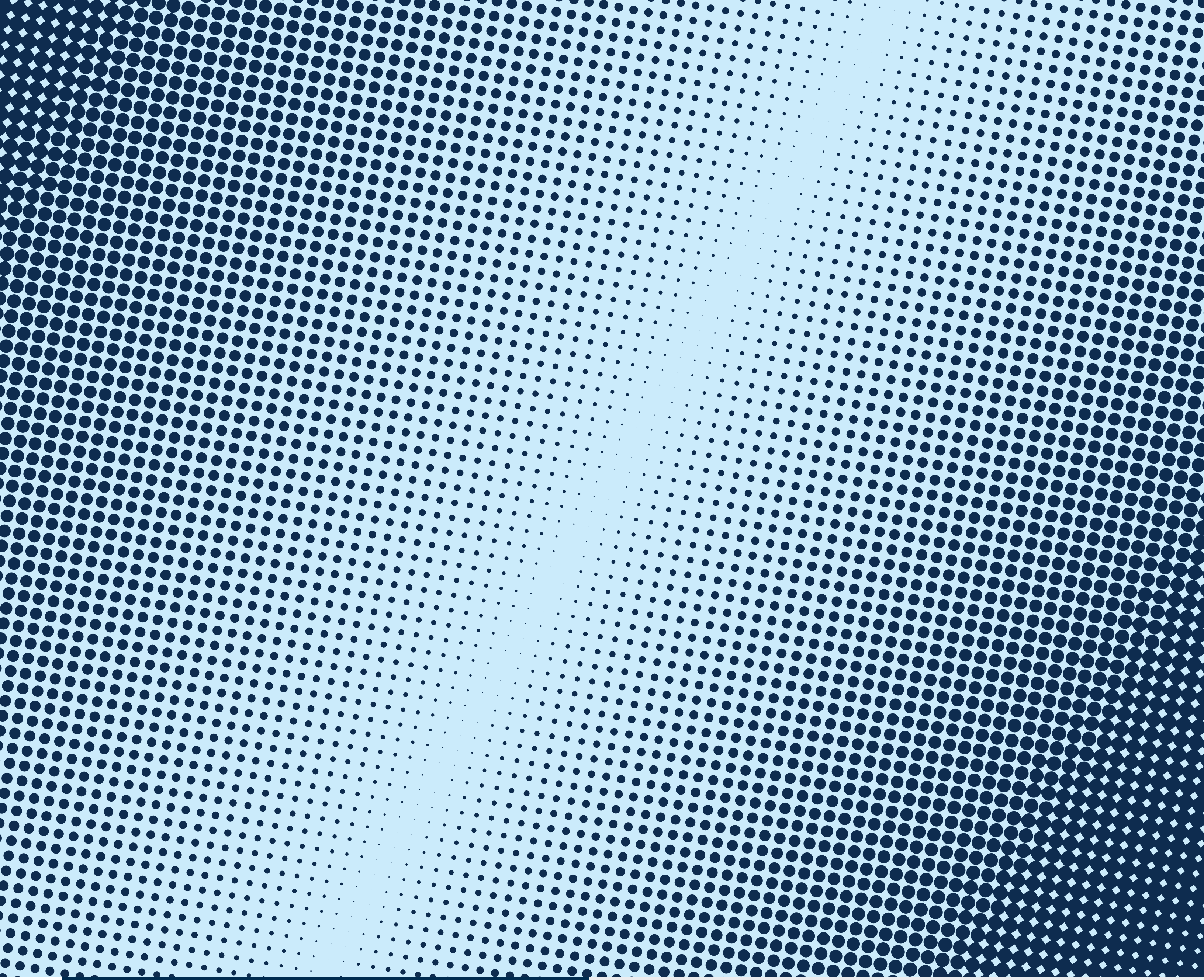
We also published research on rule-following for LLMs and unrestricted adversarial attacks.

INTERDISCIPLINARY RESEARCH

An Overview of Catastrophic AI Risks provides a structured introduction to the various types of catastrophic risks from AI. (*Wall Street Journal*)

Natural Selection Favors AIs over Humans argues that AI development will be shaped by natural selection, which will lead to selfish AI systems which prioritize their own proliferation over human goals. (*TIME* op-ed)

AI Deception: A Survey of Examples, Risks, and Potential Solutions provides examples of existing AI systems being deceptive, discusses potential risks from AI deception, and proposes technical and policy solutions.



Field-Building

Empowering the research community

AI safety is a challenge that is beyond any single research lab or academic discipline. We run a variety of initiatives to support and empower the existing research community while lowering barriers to entry and further expanding the community. Our efforts include providing infrastructure and resources for the AI safety research ecosystem, initiating multi-disciplinary projects to explore the societal effects of AI from new, expert perspectives, and creating educational resources to encourage newcomers to join. We aim to create a thriving and diverse research ecosystem that will drive progress towards safe AI.

CAIS Compute Cluster

PROBLEM:

Conducting useful AI safety research often requires working with cutting-edge models, but running large-scale models is expensive and often cumbersome to implement. As a result, many researchers are unable to pursue advanced AI safety research.

SOLUTION:

To address this issue, CAIS launched a large compute cluster in February 2023. This compute cluster has enabled a diverse portfolio of AI safety research on large-scale models and incentivized talented researchers to pursue AI safety.

63	AI safety research projects run on the cluster.
36	AI safety research papers under review or published.
200+	AI safety researchers using the cluster.
100%	of survey respondents report the compute cluster significantly supports their AI safety research — 70% noted their research project would not have been possible in its current scope without the cluster; the other 30% responded that the cluster significantly accelerates their research progress.

IN THE WORDS OF OUR RESEARCH COMMUNITY MEMBERS

"A lot of the work we want to do on LLM safety and security requires A100/H100's, and we can't afford to buy them or to rent access at the level needed for our research. So there are a bunch of research projects we just wouldn't be able to do, if we didn't have access to the CAIS cluster."

– David Wagner | *Professor, UC Berkeley*

"It [the cluster] has allowed us to aim to scale our methods to LLMs, when we would not have been able to otherwise. This greatly expands the relevance and impact of our work."

– Scott Niekum | *Associate Professor, UMass Amherst*

"The CAIS Compute Cluster has enabled me to do many experiments to submit [two recent papers]. Without the CAIS Compute Cluster, [...] these would not have been possible."

– Yizheng Chen | *Assistant Professor, University of Maryland*

Field-Building Highlights

In 2023, CAIS supported hundreds of researchers across multiple disciplines through workshops, competitions, social events, fellowships, and educational resources.

AI SAFETY TEXTBOOK

Previously, there was no comprehensive textbook that covered the concepts of AI safety for non-ML researchers or professors, but there are many other academic disciplines that have the potential to play an important role in reducing societal-scale risk from AI. ***Introduction to AI Safety, Ethics, and Society*** is a new textbook that aims to provide an accessible and comprehensive introduction to AI safety that draws on safety engineering, economics, philosophy, and other disciplines.

PHILOSOPHY FELLOWSHIP

CAIS hosted a dozen academic philosophers for a seven month research fellowship. The Fellows **produced 21 papers** on AI safety on topics such as power-seeking AI and the moral status of AI agents. **They've also initiated three workshops at leading philosophy conferences, two books,** and **a special issue of a top journal**, all focused on producing philosophy research on AI safety.

> "The fellowship definitely changed my overall trajectory. By giving me time to produce two drafts in AI safety, and giving the opportunity to build connections with other AI safety philosophers, I can now better market myself as a philosopher in AI, which makes it more likely for me to get hired as an AI philosopher, which makes it more likely that I can continue to do this kind of work." – *Philosophy Fellow, 2023*

LAW & AI SAFETY WORKSHOP

The world is in a critical period for influencing AI policy, safety standards, and regulations. Legal expertise is essential for developing policies that will survive constitutional and political challenges, but there is limited legal expertise focused on risks from AI. To orient more legal work towards the most pressing AI safety policy needs, CAIS ran a Law and AI Safety workshop in August. Twenty-five legal scholars and policy researchers convened to discuss the most urgent legal questions related to AI safety policy. **100% of researchers came away with more research ideas; 91% reported that they found the workshop very useful for meeting research collaborators.** Additionally, a group of attendees are **launching a new AI Safety Law Institute to continue this important work**.

INTRO TO ML SAFETY COURSE

The aim of this online course is to increase the number of ML or computer science undergrads who understand and pursue empirical AI safety. We ran this course three times in 2023, with 291 students participating. **87% of participants would recommend this course to a friend or colleague**, and **93% of students who were not already working in AI safety reported they were likely to actively take steps towards exploring a career in AI safety within the year.**

> "The Intro to ML Safety course provided the confidence and community connections necessary to take steps towards a career in ML Safety research. The mentorship I received was invaluable, playing a key role in my decision to specialize in technical ML Safety for my master's thesis. I appreciate this wonderful opportunity and the direction it has given to my career." – *Student, Fall 2023*

STUDENT SCHOLARSHIPS

CAIS awarded 36 students a $2,000 scholarship to pursue a research project related to AI safety. We also provide these students with a mentor to guide their research, and **90% of participants report that their experience made them more likely to pursue a career in research on AI Safety**.

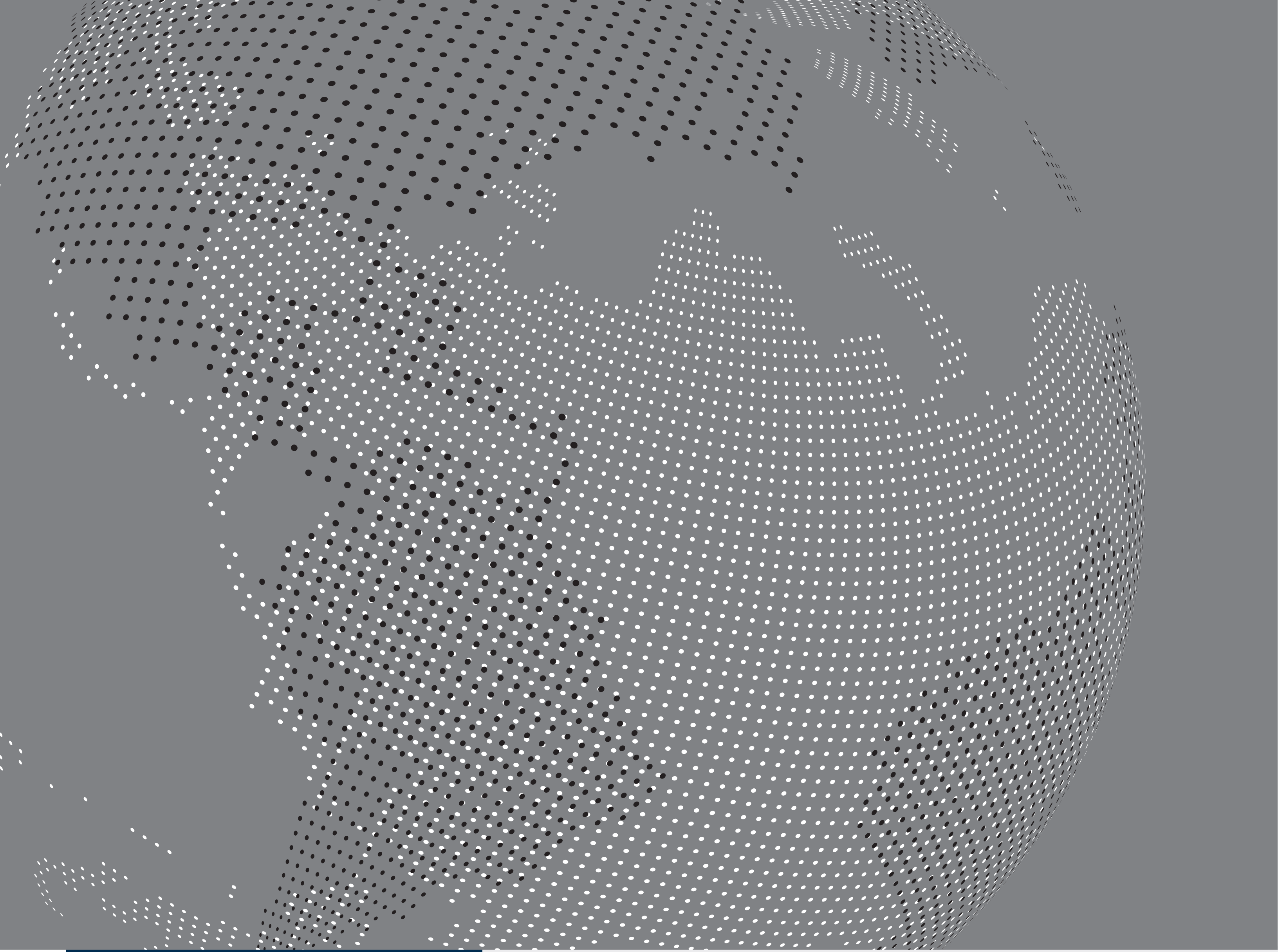
> "The mentorship and guidance from the team was especially valuable. I also enjoyed the autonomy to learn and contribute by exploring new avenues for what to try next. Being able to guide other members on AI safety projects was also a rewarding experience."
> – *Scholarship Student, 2023*

TROJAN DETECTION CHALLENGE

CAIS ran its second Trojan Detection Challenge at NeurIPS 2023. This competition aims to advance the understanding and development of methods for detecting hidden functionality in LLMs. We had 307 participants, a 60% increase over the previous year, and 3,432 submissions in total.

ML SAFETY SOCIALS

Socials at top conferences play an important role in supporting topic-specific discussions and for newcomers to learn about a subfield. CAIS organized socials on ML Safety at ICML and NeurIPS, two top AI conferences, **convening an estimated 700+ researchers to discuss AI safety**.



Advocacy

Advancing global conversation and action on AI safety

AI safety became a priority for global governments and the general public in 2023. The White House issued a landmark Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, the UK hosted an inaugural AI Safety Summit with global leaders. And there is growing public concern about artificial intelligence, with 52% of Americans reporting they feel more concerned than excited about the increased use of artificial intelligence.

CAIS advises industry leaders, policymakers, and other labs to bring AI safety research into the real-world. We aim to build awareness and establish guidelines for the safe and responsible development and deployment of AI. Relevant activities include: raising public awareness of AI risks and safety, providing technical expertise to inform policymaking at governmental bodies, and advising industry leaders on structures and practices to prioritize AI safety.

STATEMENT ON AI RISK

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

CAIS published the joint statement on AI extinction risk which notably raised public and government awareness of the scale and importance of AI risks. Crucially, by publishing this statement, CAIS placed AI risk firmly within the Overton Window of acceptable public disclosure. The statement, which received 600+ signatures from AI scientists and notable public figures, **has significantly affected the thinking of top leaders globally**.

"These scientists and experts have called on the world to act, declaring AI an existential threat to humanity on par with the risk of nuclear war ... And the loudest are the developers who designed it ... We must take those warnings seriously."

– Antonio Guterres | *Secretary General of the United Nations*

"It [AI] is moving faster than even its developers anticipated ... we have a narrowing window of opportunity to guide this technology responsibly."

– Urusla Von Leyen | *President of European Commission*

THE STATEMENT DOMINATED THE GLOBAL NEWS CYCLE AND WAS COVERED BY NUMEROUS MEDIA OUTLETS, INCLUDING:

The New York Times | THE WALL STREET JOURNAL. | BBC | CNN

The Guardian | REUTERS | The Washington Post | Bloomberg

Advocacy Highlights

In 2023, CAIS advocated for safe AI by advising policymakers around the world, educating readers through our newsletters, and promoting AI safety through leading media outlets.

TECHNICAL ADVISING

We served as a **technical advisor for the UK AI Safety Summit**. We worked with the UK Task force to plan and prepare for the science track.

We partnered with the National Science Foundation to **develop a $20M grantmaking round for AI safety research**.

We raised the profile of AI catastrophic risks at relevant governance events, such as the UN's Governing AI for Humanity event and the World Economic Forum's San Francisco AI Governance Alliance working group.

We responded to the National Telecommunications and Information Administration's (NTIA) Request For Information with a proposed regulatory framework for AI, and to the President's Council of Advisors on Science and Technology (PCAST) Request For Information with recommendations on how to reduce and detect model-generated falsehoods.

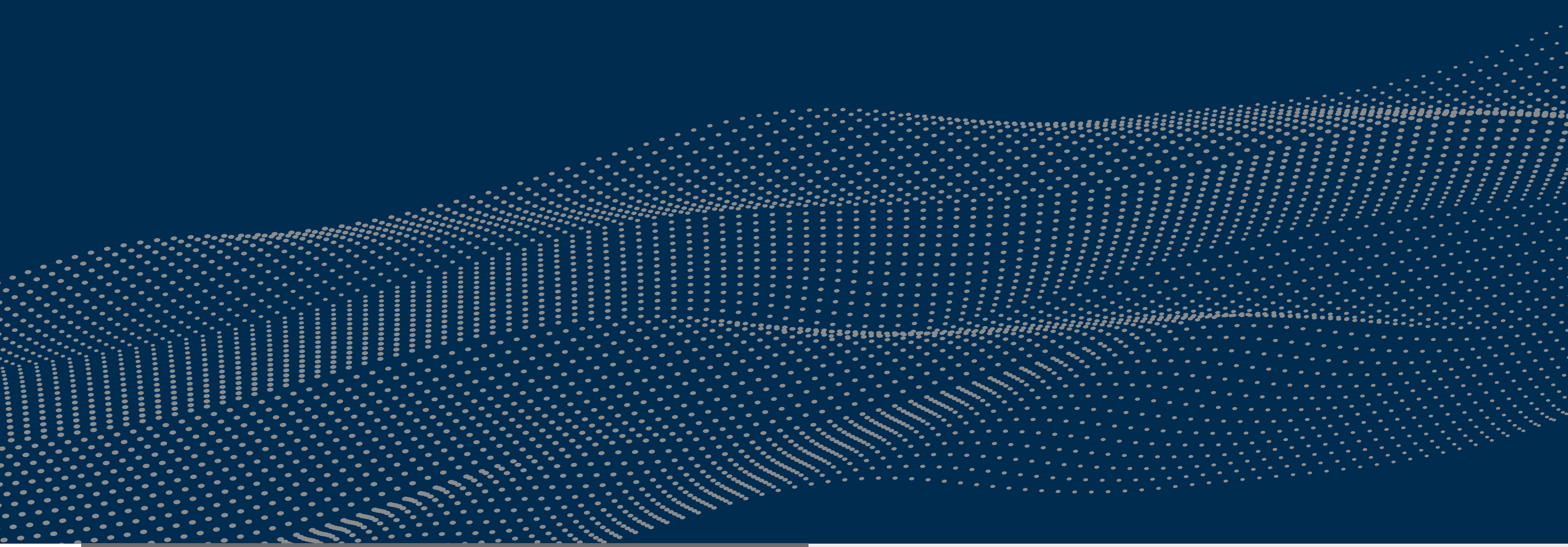
COMMUNICATIONS & MEDIA

CAIS Director Dan Hendrycks was named one of ***Time's* 100 Most Influential People in AI**.

We **made the case for AI safety** with articles that were featured in leading media outlets, such as *TIME* and the *Wall Street Journal*.

We educated 9,000 readers with relevant and timely AI safety information through our weekly AI Safety Newsletter and monthly ML Safety Newsletter.

CAIS advanced public discourse on AI safety **with over 70 interviews across global news outlets, podcasts, and documentaries**. Our perspectives were covered by Fox News, BBC, CNN, The Washington Post, Financial Times, *Wired*, Al Jazeera, and many more.



Looking Ahead

We have more work to do

In 2023, our small team of 12 employees, with the generosity of our early supporters, spurred global discussion on AI safety with our Statement on AI risk, had our research featured at top Machine Learning conferences and by leading media outlets, and trained hundreds of researchers on AI safety. **But our work is only just getting started.**

CAIS is dedicated to accelerating the study of AI risk and the implementation of real-world solutions. In 2024, we will continue to focus on solving the most important problems in AI safety across our three pillars, including:

Research

Removing hazardous knowledge from AI systems and systematically studying AI deception and honesty.

Field-building

Motivating hundreds of researchers to work on AI safety, supporting them with computing power, educational resources, and mentorship.

Advocacy

Advising the general public, frontier AI developers, and global governments on the most pressing AI risks and mitigation strategies.

As a non-profit, CAIS relies on the generosity of individuals and foundations to accomplish this important work. If you are interested in learning more about our priorities and how you can accelerate AI safety, please contact us at contact@safe.ai.

Center *for* AI Safety

www.safe.ai | contact@safe.ai