ANNUAL REPORT 2024

Centre for the Governance of AI

Table of Contents

A Note From Our Director

As we begin 2025, the main theme in AI policy conversations is uncertainty.

Much of this uncertainty relates to the future of AI progress. I can now open my web browser, type in a short prompt, and have an AI system near-instantly write an undergrad-level research report or code a complex new video game. These are incredible technical feats, which AI systems could not manage a year ago. Now, people are putting forward very different visions of what next year will bring.

As we begin 2025, the main theme in AI policy conversations is uncertainty.

On the one hand, some commentators are excited by this progress and the benefits it could unlock — but concerned that they may be more fragile than they seem. Governments are increasingly betting that AI will support a renewal in economic growth, public service effectiveness, and leadership on the global stage. These commentators therefore worry that technical barriers or regulatory headwinds could prevent these optimistic visions from being realized. They are very conscious of past examples of technologies that failed to fulfil their apparent promise, from nuclear power to hypersonic flights.

At the same time, some commentators — often the same ones — are concerned that progress in AI could soon bring new capabilities that society is not ready for. More than one leading AI company has warned that their next generation of systems could plausibly facilitate biological weapons attacks. Significant labor market disruptions may also be on the horizon, with some professions like programming serving as canaries. There is also rising concern about AI-enabled cyberattacks, about the political speech implications of AI, about overreliance on AI, and about the possibility that we might ultimately struggle to control this new technology, among a wide range of other issues.

This uncertainty is compounded by shifts in the political environment. In the US, the UK, and many other countries, new cohorts of policymakers have been trying to work out their own distinctive approaches to AI policy. They have also been carefully watching each other, to try to anticipate what the others will do.

Despite this uncertainty, government bodies and AI companies must make increasingly consequential decisions. They must also often make these decisions while understaffed and short on key forms of expertise.

Our simple but critical mission is to help institutions make better AI governance decisions.

This is where GovAI comes in. Our simple but critical mission is to help institutions make better AI governance decisions. We achieve this mission in two ways. First, we produce analysis to increase clarity, decrease uncertainty, and ultimately inform these institutions. Second, we develop AI governance talent to help reduce the expertise shortages these institutions face.

While we have been doing this work since 2016, demand for it has never been greater. In 2024, our researchers' outputs and expertise were drawn on by a wide range of initiatives. These included the first authoritative international report on risks from AI, the code of practice that clarifies Europe's new general-purpose AI regulations, and leading AI companies' voluntary frameworks for making responsible development decisions. Our career development programs also attracted record-shattering numbers of applicants, with our flagship fellowship program demonstrating more than 100x growth in applications in the past three years. We have been proud to see program alumni go on to fill policy and governance roles at a wide range of government bodies, AI companies, and civil society organizations. This was by far the year where the need for our work was clearest.

This report summarizes all that we accomplished in 2024 and paints an ambitious vision of what we intend to accomplish in 2025. Highlights for the next year include launching and growing new research teams focused on threat assessment and risk management, piloting multiple new career development programs, expanding the portion of our work that is tailored to the needs of US policymakers, and moving into a new office that we intend to make a major AI governance hub.

I believe that GovAI's work is only becoming more urgent over time. I am grateful to everyone — to our donors, to our collaborators, and, most of all, to my brilliant and dedicated colleagues — who make this work possible.

BEN GARFINKEL
Director

II. Organizational Updates

2024 was GovAI's most productive year yet.

Research

PROGRESS IN 2024

With our growing team, we again published more research last year than in any prior year. Our work touched on multiple critical areas of AI governance and policy, appearing in venues such as Science, Lawfare, the International Conference on Machine Learning (ICML), and the ACM Conference on Fairness, Accountability, and Transparency (FAccT), while directly informing policymaking processes across the UK, Europe, and the US.

Core Research Areas

We continued to build on work from previous years in several core research areas, including compute governance, frontier AI risk management, and frontier AI regulation.

We published more research in 2024 than ever before.

Risk Management: GovAI researchers explored the case for internal audit functions for frontier AI developers in order to give their boards better insight into the organizations' risk-producing activities and designed a rubric to assess frontier AI companies' safety frameworks. In collaboration with the UK AI Security Institute, GovAI researchers also started exploring the role of safety cases in frontier AI governance; these are reports that make a structured argument, supported by evidence, that a system is sufficiently safe to deploy in a given context. This work included outlining a cyber inability argument, which shows how an AI developer could contend that a model is safe to deploy because its offensive cyber capabilities are not sufficiently advanced to significantly increase risk.

Compute Governance: GovAI researchers contributed to the publication of the first major report laying out the benefits and limitations of leveraging computing hardware and infrastructure as a tool for AI governance. GovAI researchers also studied the role of cloud providers in compute governance.

Frontier AI Regulation: Much of our work on frontier AI regulation focused on the role of training compute as well as risk thresholds in frontier AI regulation. A number of our researchers argued for a principles-based approach to frontier AI governance, describing conditions under which rules should become more prescriptive. Finally, GovAI researchers studied the limitations of frontier AI regulation, including exploring when interventions on AI capabilities to reduce misuse may be warranted and how frontier AI regulation may be insufficient as AI capabilities diffuse, necessitating societal adaptation to increasingly advanced AI.

Emerging Research Areas

We also expanded our work in several other areas, taking advantage of our increased capacity to respond to growing demand for policy-relevant research across multiple topics, including agent governance, the economics of AI, and technical and international governance.

We expanded our work to meet demand for policy-relevant research across multiple topics.

Agent Governance: GovAI researchers explored the governance challenges posed by AI agents, looking at how visibility into their use could be increased, for example via agent IDs.

Economics of AI: GovAI researchers contributed to a novel framework for estimating the potential labor impacts from AI systems and analyzed options for accelerating international access to AI's economic and societal benefits.

Technical Governance: GovAI researchers introduced and characterized technical AI governance — technical analysis and tools for supporting the effective governance of AI — through a paper presenting a taxonomy and an incomplete catalog of open problems in the field.

International Governance: GovAI researchers also worked more on international governance of AI, assessing what aspects of AI governance do or do not warrant intervention at the international level and offering a series of recommendations on the future of the AI Summit series.

Highlighted Research

Computing Power and the Governance of Artificial Intelligence

Girish Sastry, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K. Hadfield, et al.

Computing power, or "compute," is crucial for the development and deployment of AI capabilities. As a result, governments and companies have started to leverage compute as a means to govern AI and achieve policy objectives, such as ensuring the safety and beneficial use of AI. Compute is a particularly relevant intervention point because, unlike data and algorithms, it is detectable, excludable, and quantifiable, and it is produced via an extremely concentrated supply chain. However, while compute-based policies and technologies have the potential to assist in these areas, there is significant variation in their readiness for implementation. Furthermore, naïve or poorly scoped approaches to compute governance carry significant risks in areas like privacy, economic impacts, and centralization of power. This paper describes how compute may be used for AI governance and suggests guardrails to minimize the risks of such approaches.

From Principles to Rules: A Regulatory Approach for Frontier AI

Jonas Schuett, Markus Anderljung, Alexis Carlier, Leonie Koessler, Ben Garfinkel

Several jurisdictions are starting to regulate frontier AI systems. To reduce risks from these systems, regulators may require frontier AI developers to adopt safety measures. The requirements could be formulated as high-level principles (e.g. "AI systems should be safe and secure") or specific rules (e.g. "AI systems must be evaluated for dangerous model capabilities following the protocol set forth in…"). These regulatory approaches have distinct strengths and weaknesses but are not binary options. Policymakers must choose a point on the spectrum between them, recognizing that the right level of specificity may vary between requirements and may change over time. We provide specific recommendations on how policymakers should apply these two principles and evolve their approach over time.

Open Problems in Technical AI Governance

Anka Reuel, Ben Bucknall, et al.

AI progress is creating a growing range of risks and opportunities, but it is often unclear how they should be navigated. In many cases, the barriers and uncertainties faced are at least partly technical. Technical AI governance, referring to technical analysis and tools for supporting the effective governance of AI, seeks to address such challenges. In this paper, we explain what technical AI governance is and why it is important, and we present a taxonomy and incomplete catalogue of open problems in the field.

Visibility into AI Agents

Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, Markus Anderljung

Increased delegation of commercial, scientific, governmental, and personal activities to AI agents — systems capable of pursuing complex goals with limited supervision — may exacerbate existing societal risks and introduce new ones. Information about where, why, how, and by whom certain AI agents are used, which we refer to as *visibility*, is critical to understanding and mitigating these risks. In this paper, we assess three categories of measures to increase visibility into AI agents: agent identifiers, real-time monitoring, and activity logging. For each, we outline potential implementations.

We analyze how the measures apply across a spectrum of centralized and decentralized deployment contexts, accounting for various actors in the supply chain, including hardware and software service providers. Finally, we discuss the implications of our measures for privacy and concentration of power.

Safety Cases for Frontier AI

Marie Davidsen Buhl, Gaurav Sett, Leonie Koessler, Jonas Schuett, Markus Anderljung

As frontier AI systems become more capable, it becomes more important that developers can explain why their systems are sufficiently safe. One way to do so is via *safety cases*: reports that make a structured argument, supported by evidence, that a system is safe enough in a given operational context. In this paper, we explain why safety cases, already common in other safety-critical industries, may also be a useful tool in frontier AI governance. We then discuss the practicalities of safety cases, outlining how to produce a frontier AI safety case and discussing what still needs to happen before safety cases can substantially inform decisions.

Societal Adaptation to Advanced AI

Jamie Bernardi, Gabriel Mukobi, Hilary Greaves, Lennart Heim, Markus Anderljung

Existing strategies for managing risks from advanced AI systems often focus on affecting what AI systems are developed and how they diffuse. However, this approach becomes less feasible as the number of advanced AI developers grows, and it hampers beneficial use cases alongside harmful ones. In response, we propose a complementary approach: increasing societal adaptation to advanced AI. We introduce a conceptual framework that helps identify adaptive interventions that avoid, defend against, and remedy harmful uses of AI, with examples from election interference, cyberterrorism, and AI control failures. We outline a three-step cycle for society and provide concrete recommendations for governments, industry, and third parties.

AMBITIONS FOR 2025

We expect to continue our ongoing research workstreams in 2025, with particular emphasis on three priority areas:

We are establishing new research teams focused on risk management and threat analysis.

- **Informing risk management policies and safety frameworks of frontier AI developers**: Many frontier AI developers now have published frontier AI safety frameworks as part of the Frontier AI Safety Commitments they made at the AI Summit in Seoul. In these frameworks, developers define the policies they will follow to keep the risk of the systems they're developing to acceptable levels. Our work will contribute to the development, refinement, and evaluation of these policies to improve their effectiveness.
- **Frontier AI regulation:** We will continue researching the design and limitations of frontier AI regulation. This includes exploring how (and whether) such regimes address risks that may be caused by downstream developers, the extent to which such rules may lead to regulatory flight, and how lessons from financial regulation may apply.
- **Threat assessments**: This new workstream at GovAI will seek to empirically ground and assess claims about AI risk, including how frontier AI developers should set thresholds for dangerous capabilities and when certain mitigation measures to address those risks are warranted.

We also expect to continue supporting work in a number of other areas, including AI agents and agent governance, the economics of AI and its impact on labor markets, and international governance.

However, our plans are subject to change; we often update our plans in response to new developments in AI and new opportunities that arise.

Talent Development

We are heartened by the rate at which interest in the field continues to grow.

PROGRESS IN 2024

In 2024, we continued to grow and improve our flagship talent-development program, the Seasonal Fellowships. We have also maintained our Research Scholar program at a comparable size to the previous year.

Our Seasonal Fellowships bring a cohort of early-stage researchers or individuals new to the field of AI governance to Oxford to spend three months working on an AI governance research project, learning about the field, and making connections with other researchers and practitioners. In 2024, we hosted 36 Seasonal Fellows, a 50% increase compared to the previous year.

Considering the rapid rise in demand for AI governance talent, we are heartened by the rate at which interest in the field continues to grow. In 2024, this was most clearly evidenced by the applications to our Seasonal Fellowships. Not only did the number of applicants jump more than 14-fold compared to 2023, from 670 to 9700, but the profile of our applicants also continued to evolve, with more mid- to late-career professionals looking to transition into AI governance from adjacent fields.

The growing skills and expertise of our Seasonal Fellows — and the quality of the work they do while at GovAI — enabled us to significantly expand our external supervision capacity, with positive implications for our ability to continue scaling up the program going forward. We also continued to refine both the selection process and the structure of the program in response to feedback from fellows and supervisors, and to the evolving needs of the field.

Our efforts are bearing fruit: In our first comprehensive alumni survey, the vast majority of respondents credited the three-month program not only with accelerating their career progression, but also with significantly increasing their future impact in the field. The average respondent estimated a one-year acceleration and a 50% increase in impact.

Many of our 2024 graduates have already gone on to influential roles in AI governance, including in government institutions such as the UK's Department for Science, Innovation and Technology, the EU AI Office, and the US Department of Commerce; at think tanks such as RAND, the Future of Life Institute, and the French Center for AI Safety; and in industry, at Google DeepMind and the Frontier Model Forum.

The program aims to give these researchers the freedom and support to build their expertise and networks while doing impactful work.

We also continued to run our Research Scholar program, which offers promising AI governance researchers one-year visiting positions at GovAI. The program aims to give these researchers the freedom and support to build their expertise and networks while doing impactful work. In 2024, we supported 14 Research Scholars to work on a wide range of topics, including UK and EU AI policy, international governance, and technical governance.

We are also more frequently supporting secondments for Research Scholars who hope to gain direct policy experience. Over the past year, we have increased the number of staff members who are seconded to relevant institutions, including the UK's DSIT and the OECD, on a part- or full-time basis. We continue to see our Scholars move on to exciting roles after their year at GovAI. In 2024, we saw Scholars move on to work at the EU AI Office, Demos, and Google DeepMind. We are also increasingly offering some Scholars the chance to stay on at GovAI as Research Fellows.

AMBITIONS FOR 2025

In 2025, we will continue to expand our efforts to fill the most urgent talent gaps in AI governance.

In January, we promoted our Research Manager, Valerie Belu, previously responsible for organizing and running the Seasonal Fellowships, to Head of Talent Development. In her new role, she will oversee, build up, and develop the overall strategy for our AI governance talent pipeline programs. We are also currently hiring for Research Management Associates and/or Research Managers to support our talent-development work. With this expanded capacity, we will continue to scale up our existing talent programs, improve our alumni engagement, and pilot (and evaluate) several new programs to determine the best strategy for GovAI's talent-development efforts going forward.

In 2025, we will host more than 60 Seasonal Fellows, a nearly 70% increase on 2024. To support this expansion, we will implement new processes for selecting supervisors and managing Seasonal Fellows. We also plan to expand our Research Scholar program to keep in step with our overall growth plans for our research and policy workstreams.

Our 2024 alumni survey highlighted that GovAI's role in expanding our alumni's professional networks made a large contribution to their subsequent careers. We will thus double down on this path to impact by deepening our alumni engagement through events and other initiatives aimed at building connections between our current fellows, alumni, and the field at large.

We are planning to pilot an ambitious slate of new programs over the coming year.

Lastly, we are planning to pilot an ambitious slate of new programs over the coming year. These will include a DC-based version of our Seasonal Fellowship, designed specifically for people pursuing US AI policy careers; an improved version of the remote, part-time Policy Program first trialed in 2023; and a Senior Visiting Fellowship to continue to smooth the path for established professionals into the field.

Policy Engagement

PROGRESS IN 2024

2024 was our most active year to date in terms of policy engagement as we worked with policymakers across the UK, the EU, the US; frontier AI companies; civil society organizations; and intergovernmental organizations.

In the UK, we continued to provide information and analysis on domains including frontier AI regulation, international governance, and agent governance. Some of our team members also spent time seconded into the Department for Science, Innovation and Technology, as well as the AI Security Institute. We deepened our research collaboration with the AI Security Institute, publishing papers on safety cases. We also actively engaged with broader civil society by organizing roundtables with the Tony Blair Institute on the UK's role in international AI governance and with Chatham House on compute governance, as well as a workshop with the Centre for Long-Term Resilience on UK AI regulation.

We continued to provide information and analysis on domains including frontier AI regulation, international governance, and agent governance.

As the EU has started implementing the AI Act, a number of our researchers have provided research and expertise, in particular with regard to general-purpose AI with systemic risk. Markus Anderljung began serving as a vice-chair in the drawing up of the EU's Code of Practice for general-purpose AI systems with systemic risk, detailing how companies could, in practice, meet their obligations under the AI Act. Other staff started offering expert opinions on questions around the AI Act's scope via the Commission's Joint Research Centre.

We spent comparatively less time engaging with policymakers in the US. As in previous years, we continued engaging on questions related to compute governance, in particular export controls. Our main engagement was with the National Institute of Standards and Technology as it started to codify standards and best practices in frontier AI risk management through, for example, the AI Safety Institute Consortium and three RFI responses.

On the international level, we engaged deeply with the AI Summit series, offering support to the International AI Safety Report, attending the AI Seoul Summit, and writing a number of pieces offering guidance on the series' future. Jonas Schuett led a survey on thresholds for advanced AI systems in collaboration with the OECD, and we also started hosting two secondees to the OECD, who support the organization's Strategic Foresight Unit.

Another focus of 2024 was providing expertise to frontier AI companies in the wake of their commitment to devise and implement frontier AI safety frameworks. These frameworks establish thresholds beyond which AI systems would pose unacceptable risks and include policies to keep below those thresholds. Doing so well is a considerable challenge and will likely warrant significant scientific study. To support this work, we offered independent research, provided feedback to some of these companies on their frameworks, and co-organized an academic conference with the UK's AI Security Institute in November 2024 to advance the development and implementation of frontier AI safety frameworks ahead of France's AI Action Summit in February 2025.

Notable External Engagements

Roundtable on Technical Governance

GovAI and Chatham House jointly hosted a roundtable on May 2, 2024, bringing together stakeholders from various sectors to discuss computing power's role in AI governance. The event followed GovAI's publication of two papers on the topic: "Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation" and "Computing Power and the Governance of Artificial Intelligence." The discussion explored "hardware governance" as a new point of policy intervention and the role of compute governance.

Roundtable on the UK's role in international AI governance

GovAI hosted an event on June 24, 2024, bringing together a small group of relevant policymakers and experts to discuss critical issues in international AI governance. The event focused on two questions: What should the key goals for international governance efforts be? And how can the UK best contribute toward these goals in the coming years?

Workshop on the AI Regulation Bill

GovAI and the Centre for Long-Term Resilience jointly hosted a workshop on the forthcoming proposals for frontier AI legislation on October 10, 2024. The workshop featured a high-level discussion of possibilities for the bill and deeper policy discussions on key issues.

Conference on Frontier AI Safety Commitments (FAISC)

In November 2024, GovAI co-hosted an academic-style conference alongside the UK AI Security Institute. The Conference on Frontier AI Safety Commitments saw 100 representatives of signatory companies, academia, and civil society come together on the road to the AI Action Summit in France, to discuss progress and challenges in developing and implementing frontier AI safety frameworks. With presentations given on 51 accepted papers, the conference resulted in debate, several areas of consensus, and directions for future work that emerged from the discussion.

AMBITIONS FOR 2025

We aim to continue our efforts to provide technical assistance and engage with policymakers in the UK and the EU — in particular, but not exclusively, advising on the design of appropriate requirements on developers of frontier AI systems. We also hope to considerably increase the quality of our independent research and analysis for AI companies in their design and implementation of frontier AI frameworks, for example via expanding our workstreams for risk management and threat assessments.

In addition, we have ambitions to expand the portion of our research that is relevant to and tailored for US policymakers, as well as to grow our US-focused talent programs. We are hoping to bring on new team members to lead these efforts.

We are also likely to increase our policy engagement by organizing events, particularly as we move our offices to London.

We are also likely to increase our policy engagement by organizing more events.

Organizational Capacity

PROGRESS IN 2024

GovAI has grown substantially over the course of 2024: we ended the year with 30 full-time staff members across both permanent and visiting positions, up from about 20 at the start of the year, with six additional team members anticipated in the early months of 2025. We also completed our spin-out from our prior fiscal sponsor, Effective Ventures, which is discussed in more detail below.

We ended the year with 30 full-time staff members across both permanent and visiting positions, up from about 20 at the start of the year.

GovAI's Operations team more than doubled in size in 2024, including hiring Ryan Fugate as our new Director of Operations, as well as Leia Wang, Arianna Bosio, and William Ehlers as Operations Associates in the first half of the year. This group joined the team to take over the operational functions previously managed by Effective Ventures and to support the growing demand for recruiting coordination and program support.

We have also hired further administrative capacity across the organization, prompted by the departure of Gina Moss, former Executive Assistant to the Director. Joining GovAI in an executive assistant capacity across the organization are Rakhsana Ramzan, Robyn Seabrook, and Leyla Ava. We have also hired Isha Paik as an Operations Specialist, Anastacia Button-Gentry as an Operations Associate, and John Drummond as a Talent Systems Specialist. We expect these additional hires will significantly increase our capacity to scale our existing programs and experiment with new paths to impact.

We started growing and formalizing our research management structures in 2024, expanding Markus Anderljung's responsibilities to Director of Policy and Research, beyond specifically managing the GovAI Policy Team. We have restructured what was formerly called the Policy Team and moved to a more flexible workstream-based research model.

Stephen Clare also joined the team as a Research Manager, focused on supporting our cohorts of Research Scholars, leading our research dissemination efforts, and providing research support for our core research team.

Spin-out

In 2024, GovAI successfully completed its spin-out from Effective Ventures, which previously provided fiscal sponsorship for the organization. This was an important project with a lot of moving pieces, and we consider it a major milestone to have completed a smooth spin-out with no disruptions to our programs. GovAI is now an entirely independent organization, with entities incorporated in both the United States and the United Kingdom, each with its own activities, staff, and board of directors. GovAI's new US entity is a section 501(c)(3) organization, and its UK entity (a non-profit company limited by guarantee) is a wholly controlled subsidiary of the US entity, ensuring that both are organized and operated exclusively for 501(c)(3) charitable, scientific, and educational purposes. As part of this spin-out, GovAI assumed full responsibility and control over its operations, strengthened its governance mechanisms, and built strong foundations for the growth we anticipate in the coming years.

GovAI assumed full responsibility and control over its operations, strengthened its governance mechanisms, and built strong foundations for the growth we anticipate in the coming years.

With GovAI spinning out of our fiscal sponsor and establishing independent entities, our former advisory board transitioned into a formal board of directors in 2024. As part of that transition, we have been preparing to expand the set of directors serving on the board throughout the year.

With the establishment of the formal board, Allan Dafoe and Ajeya Cotra have departed. We are grateful to both for their dedicated service and leadership throughout the past three years while GovAI established itself as an organization independent of the University of Oxford. Their expertise and advice have been instrumental in guiding us through that phase, helping to shape the new organization's direction and success.

To ensure that the board retains the capacity necessary to oversee and steer GovAI's activities, we have run a thorough process to identify additional board members and vet dozens of candidates. As a result of this process, the board recently invited Jeffrey Ding and Seán Ó hÉigeartaigh to join its ranks.

Jeff is an Associate Professor of Political Science at George Washington University, where he focuses on emerging technologies and international politics. He recently wrote a book investigating how technological revolutions affect the rise and fall of great powers. By analyzing historical cases of industrial revolutions that sparked power transitions and conducting statistical analysis on cross-country technology adoption, he develops insights into how emerging technologies like AI could affect the US-China power balance.

Seán is the Director of the AI: Futures and Responsibility Programme at the University of Cambridge. Seán's work focuses on foresight and governance of frontier AI, as well as the geopolitical implications of advanced AI. Previously, Seán was the founding Executive Director of CSER, where he led its establishment alongside the founders from 2013 to 2015 and directed major research projects from 2015 to 2020.

We are excited to welcome Jeff and Seán. Their commitment to GovAI's mission, important work in the field, and new perspectives promise to enrich discussions and enhance the organization's strategic vision in our next phase of growth and development.

> We are excited to welcome Jeff and Seán to the board.

Board

- **Jeffrey Ding**
  Jeff is an Associate Professor of Political Science at George Washington University.
- **Tasha McCauley**
  Tasha is an Adjunct Senior Management Scientist at RAND Corporation.
- **Seán Ó hÉigeartaigh**
  Seán is the Director of AI: Futures and Responsibility Programme at the University of Cambridge.
- **Toby Ord**
  Toby is a Senior Researcher at the Oxford Martin AI Governance Initiative.
- **Helen Toner**
  Helen is the Director of Strategy at the Center for Security and Emerging Technology.

The board of our UK subsidiary comprises Toby Ord and two members of the GovAI team: Markus Anderljung, our Director of Research and Policy, and Paul Harding, our Operations Manager.

FINANCES

The entire GovAI team remains grateful to all our funders, whose trust in us and generosity continue to enable us to achieve our mission. In the past year, we received large donations from several supporters, including but not limited to support from Open Philanthropy, Longview Philanthropy, The Waking Up Foundation, Frank Batten IV, and several anonymous donors.

The entire GovAI team remains grateful to all our funders, whose trust in us and generosity continue to enable us to achieve our mission.

In 2024, our expenses totaled US$6.07 million. Approximately 60% of this amount went to our staff costs, including salaries and taxes for our researchers, seasonal fellows, operations team, and management team, as well as visa costs. 17% went to operational costs, including events and travel, and 13% went to contract services, including legal costs. The remaining 10% went to facilities and equipment, primarily rent.

You can donate to us through Giving What We Can. If you are interested in supporting our work and would like to develop a more detailed understanding of our funding needs and specific opportunities, please get in touch at contact@governance.ai.

AMBITIONS FOR 2025

Our mission is more urgent than ever, and we anticipate that 2025 will be a year of continued growth. Following our spin-out from Effective Ventures in 2024, we expect to invest time and energy in the first half of the year building robust operational systems and processes to streamline our work, reduce the operational burden on our researchers, and ensure that GovAI can continue to scale effectively.

In mid-2025, GovAI's UK presence will relocate from Oxford to London.

To accommodate growth, we are implementing more formal structures on the research side and establishing teams around specific workstreams, notably Risk Management (led by Jonas Schuett) and Threat Assessments (led by Luca Righetti).

In mid-2025, GovAI's UK presence will relocate from Oxford to London in order to more completely realize the benefits of a larger talent pool, closer proximity to policymakers, and stronger connections to other civil society organizations. We have settled on a larger office in central London and anticipate using the space to accommodate growing fellowship cohorts, invite visiting experts to work alongside our team, and host seminars and roundtable discussions on AI governance and policy. Relocating to a new city will naturally cause some disruption for our staff, but we expect that London as an international hub will confer important benefits on our talent programs and policy work in the years to come.

We also remain committed to further diversifying our sources of funding. In 2025, we expect to hire a dedicated professional fundraiser with the express intention of cultivating a wider range of donors and improving our organizational resilience by reducing our reliance on the relatively small number of donors who currently support our work.

Finally, we will continue to expand our efforts to create highly visible summaries and compelling presentations of our work across multiple channels. Translating our research into direct impact requires us to help key audiences quickly grasp our findings and recommendations. In 2025, we anticipate expanding and refining our online presence, including through our website, social media channels, and direct media engagement.

Senior Leadership Team

BEN GARFINKEL
Director

Ben leads GovAI and is responsible for setting the direction of the organization and making key decisions. His own research has focused on the security implications of AI, the causes of war, and the methodological challenge of forecasting risks from technology. He earned a BS in Intensive Physics and in Mathematics and Philosophy from Yale University before receiving a DPhil in International Relations at the University of Oxford.

GEORG ARNDT
Chief of Staff

Georg supports GovAI's Director in day-to-day decision-making, maintains a high-level overview of GovAI programs, and manages GovAI's non-research staff. He has previously worked as an economic consultant for NERA; as Project Manager for the Future of Humanity Institute, University of Oxford; and as Chief Executive of the Future of Humanity Foundation.

MARKUS ANDERLJUNG
Director of Policy and Research

Markus's work aims to produce rigorous policy analysis for governments and AI companies. His research focuses on frontier AI regulation, responsible cutting-edge development, national security implications of AI, and compute governance. He is an Adjunct Fellow at the Center for a New American Security, and a member of the OECD AI Policy Observatory's Expert Group on AI Futures. He was previously seconded to the UK Cabinet Office as a Senior Policy Specialist after roles as GovAI's Deputy Director and Senior Consultant at EY Sweden.

RYAN FUGATE

Director of Operations

Ryan leads GovAI's Operations team and is responsible for the efficient day-to-day functioning of the organization, as well as ensuring that GovAI's operating model continuously and effectively supports its mission. He has previously worked in both the private and non-profit sectors and has experience leading teams focused on strategy, operations, product management, and corporate development.

VALERIE BELU

Head of Talent Development

Valerie is the Head of Talent Development at GovAI. She is responsible for overseeing, building up, and strategizing for our AI governance talent pipeline programs. Before joining GovAI, Valerie was a Fellow at the LSE's European Institute and a Stipendiary Lecturer at St. Hilda's College, Oxford. She holds a DPhil in Politics, an MPhil in Comparative Government, and a BA in Philosophy, Politics, and Economics, all from the University of Oxford.

Our full team, including researchers, operations personnel, and affiliates, can be found on our website.

In Summary

2025 will be a year of exciting progress and evolution for GovAI across all of our priority areas. We will grow and develop our team, accelerate our research and policy work, and mature our organizational infrastructure. The AI governance decisions made in 2025 may turn out to shape the lives of billions of people for decades to come. We are committed to doing everything we can to help these decisions be made well.

The AI governance decisions made in 2025 may turn out to shape the lives of billions of people for decades to come.

III. Outputs

Since our last update, GovAI researchers have authored or contributed to over 60 research outputs.

Publications, Reports, and Working Papers

- **Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation.**
  Lennart Heim, Tim Fist, Janet Egan, Sihao Huang, Stephen Zekany, Robert Trager, Michael A Osborne, Noa Zilberman

- **Computing Power and the Governance of Artificial Intelligence.**
  Girish Sastry et al., including Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K. Hadfield

- **Visibility into AI Agents.**
  Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, Markus Anderljung

- **Responsible Reporting for Frontier AI Development.**
  Noam Kolt, Markus Anderljung, Joslyn Barnhart, Asher Brass, Kevin Esvelt, Gillian K. Hadfield, Lennart Heim, Mikel Rodriguez, Jonas B. Sandbrink, Thomas Woodside

- **Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090.**
  RAND Working Paper Series. Gabriel Kulp, Daniel Gonzales, Everett Smith, Lennart Heim, Prateek Puri, Michael J. D. Vermeer, Zev Winkelman

- **How to Design an AI Ethics Board.**
  Jonas Schuett, Anka Reuel, Alexis Carlier

- **From Principles to Rules: A Regulatory Approach for Frontier AI.**
  Jonas Schuett, Markus Anderljung, Alexis Carlier, Leonie Koessler, Ben Garfinkel

- **Risk Thresholds for Frontier AI.**
  Leonie Koessler, Jonas Schuett, Markus Anderljung

- **GPTs are GPTs: Labor Market Impact Potential of LLMs.**
  Tyna Eloundou, Sam Manning, Pamela Mishkin, Daniel Rock

- **Societal Adaptation to Advanced AI.**
  Jamie Bernardi, Gabriel Mukobi, Hilary Greaves, Lennart Heim, Markus Anderljung

- **IDs for AI Systems.**
  Alan Chan, Noam Kolt, Peter Wills, Usman Anwar, Christian Schroeder de Witt, Nitarshan Rajkumar, Lewis Hammond, David Krueger, Lennart Heim, Markus Anderljung

- **Beyond Static AI Evaluations: Advancing Human Interaction Evaluations for LLM Harms and Risks.**
  Lujain Ibrahim, Saffron Huang, Lama Ahmad, Markus Anderljung

- **Foundational Challenges in Assuring Alignment and Safety of Large Language Models.**
  Usman Anwar et al. including Anton Korinek, Alan Chan, Markus Anderljung, Lewis Hammond, Atoosa Kasirzadeh

- **Responsible Reporting for Frontier AI Development.**
  Noam Kolt, Markus Anderljung, Joslyn Barnhart, Asher Brass, Kevin Esvelt, Gillian K. Hadfield, Lennart Heim, Mikel Rodriguez, Jonas B. Sandbrink, Thomas Woodside

- **Black-Box Access is Insufficient for Rigorous AI Audits.**
  Stephen Casper et al. including Benjamin Bucknall, Alan Chan

- **Libel via Language Models.**
  Peter Wills

- **Care for Chatbots.**
  Peter Wills

- **Position Paper: Technical Research and Talent is Needed for Effective AI Governance.**
  Anka Reuel, Lisa Soder, Benjamin Bucknall, Trond Arne Undheim

- **Position: Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data.**
  Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, Marius Hobbhahn

- **Training Compute Thresholds: Features and Functions in AI Governance.**
  Lennart Heim, Leonie Koessler

- **Open Problems in Technical AI Governance.**
  Anka Reuel, Ben Bucknall, et al. including Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, and Robert Trager

- **A Grading Rubric for AI Safety Frameworks.**
  Jide Alaga, Jonas Schuett, Markus Anderljung

- **Voice and Access in AI: Global AI Majority Participation in Artificial Intelligence Development and Governance.**
  Sumaya N. Adan, Robert Trager, Kayla Blomquist, Claire Dennis, Gemma Edom, Lucia Velasco, Cecil Abungu, Ben Garfinkel, Julian Jacobs, Chinasa T. Okolo, Boxi Wu, Jai Vipra

- **Safety Cases for Frontier AI.**
  Marie Davidsen Buhl, Gaurav Sett, Leonie Koessler, Jonas Schuett, Markus Anderljung

- **Frontier AI Developers Need an Internal Audit Function.**
  Jonas Schuett

- **Safety Case Template for Frontier AI: A Cyber Inability Argument.**
  Arthur Goemans, Marie Davidsen Buhl, Jonas Schuett, Tomek Korbak, Jessica Wang, Benjamin Hilton, Geoffrey Irving

- **What Should Be Internationalised in AI Governance?**
  Claire Dennis, Stephen Clare, et al. including Robert Trager, Markus Anderljung, Malcolm Murray, and Lennart Heim

- **Protecting Society From AI Misuse: When Are Restrictions on Capabilities Warranted?**
  Markus Anderljung, Julian Hazell, Moritz von Knebel (Originally published as a pre-print in 2023)

- **Introduction to AI Safety, Ethics, and Society (Chapter 8: Governance).**
  Jonas Schuett, Robert Trager, Lennart Heim, et al.

- **The Future of International Scientific Assessments of AI's Risks.**
  Claire Dennis, et al.

Short Analysis

- **What Increasing Compute Efficiency Means for the Proliferation of Dangerous Capabilities.**
  Lennart Heim, Konstantin Pilz

- **Goals for the Second AI Safety Summit.**
  Ben Garfinkel, Markus Anderljung, Lennart Heim, Robert Trager, Ben Clifford, Elizabeth Seger

- **Computing Power and the Governance of AI.**
  Lennart Heim, Markus Anderljung, Emma Bluemke, Robert Trager

- **Predicting AI's Impact on Work.**
  Sam Manning

- **Evaluating Predictions of Model Behaviour.**
  Alan Chan

- **Visibility into AI Agents.**
  Alan Chan

- **Managing Risks from AI-Enabled Biological Tools.**
  John Halstead

- **The AI Summit Series: What Should Its Niche Be?**
  Lucia Velasco

- **Effective Mitigations for Systemic Risks from General-Purpose AI.**
  Risto Uuk, Annemieke Brouwer, Tim Schrier, Noemi Dreksler, Valeria Puglanino, Rishi Bommasani

Opinion Articles

- **Frontier AI Regulation: Safeguards Amid Rapid Progress.**
  Markus Anderljung, Anton Korinek

- **To Govern AI, We Must Govern Compute.**
  Lennart Heim, Markus Anderljung, Haydn Belfield

- **Tort Law and Frontier AI Governance.**
  Matthew van der Merwe, Ketan Ramakrishnan, Markus Anderljung

- **Predistribution over Redistribution: Beyond the Windfall Clause.**
  Sam Manning, Saffron Huang

- **Frontier AI Regulation in the UK.**
  Markus Anderljung

- **AI's Impact on Income Inequality in the US.**
  Sam Manning

Policy Advice

The policy recommendations and advice presented in this section reflect the views of individual GovAI researchers, rather than the views of the organization.

- **Response to the RFI Related to NIST‘s Assignments Under the Executive Order Concerning AI.**
  Jonas Schuett, Leonie Koessler, Markus Anderljung

- **Accessing Controlled AI Chips via Infrastructure-as-a-Service (IaaS): Implications for Export Controls.**
  Lennart Heim, Janet Egan

- **Comments on NIST’s Draft Profile on Generative AI.**
  Malcolm Murray, Jonas Schuett, Sam Manning, Alan Chan, Leonie Koessler, Markus Anderljung

- **Managing Misuse Risk for Dual-Use Foundation Models: Comments on the Initial Public Draft of NIST AI 800-1.**
  Jonas Schuett, Sophie Williams, Marie Buhl, Alan Chan, Markus Anderljung

- **Comment on the Bureau of Industry and Security Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters.**
  Sam Manning, Markus Anderljung