ANNUAL REPORT 2025

GovAI

Table of Contents

A Note From Our Director

GovAI was founded ten years ago, based on a premise that was speculative at the time. We believed that, in the coming decades, progress in AI was likely to transform the world.

At the time, the most impressive AI systems were more or less toys. DeepMind, for example, had recently produced breakthrough systems that could beat the best human players at Go and certain Atari games. These kinds of systems were scientific marvels, but they had little practical significance. The question, back then, was whether this wave of AI progress would continue—and whether it would, past some point, spill out of laboratories and start to touch the real world.

Writing at the start of 2026, it's clear that the transformation has begun. Products like Claude Code have fundamentally changed how software development works: people without any technical background are suddenly spinning up complex software, while experienced software engineers are handing off more and more of their work to AI. Ordinary people now spend large parts of their days chatting with AI systems, to learn, to work, and sometimes to seek a friend. The security landscape is also beginning to shift: for example, this past year brought increasingly notable examples of AI being used both for cyberoffense and cyberdefense.

The key question is no longer whether AI will transform the world, but what decision-makers can do to positively shape this transformation as it unfolds.

The key question is no longer whether AI will transform the world, but what decision-makers can do to positively shape this transformation as it unfolds. How can they ensure that valuable AI systems are developed quickly and adopted widely? At the same time, how can they protect the public from potential harms, ranging from economic dislocation to threats to national security?

Unfortunately, decision-makers—across government, industry, and civil society—face many difficulties in trying to answer these questions. For one thing, they often lack essential information even about AI systems' current capabilities and impacts on the world. Right now, for example, there are live debates about how much AI is currently boosting productivity, about how American and Chinese AI companies currently compare, and whether AI is currently making any forms of terrorism meaningfully easier. For another

thing, decision-makers often lack a full picture of the policy options available to them and the trade-offs involved. Most fundamentally, expertise and staffing inside key institutions—especially government bodies—is normally uncomfortably scarce.

Our mission is to help decision-makers navigate the transition to a world with advanced AI.

This is where GovAI comes in. Our mission is to help decision-makers navigate the transition to a world with advanced AI. 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.

Responding to the unfolding transformation, 2025 was our most productive research year yet. Our researchers published analyses that helped answer questions such as "What data do governments need in order to understand AI's economic impacts?", "What AI capabilities do experts believe would genuinely heighten biological risks?", and "Why are US firms increasingly building datacenters in the UAE?" They also fielded a growing number of requests for expert input from a range of institutions, including both AI companies crafting their safety and security policies and government bodies crafting public policy.

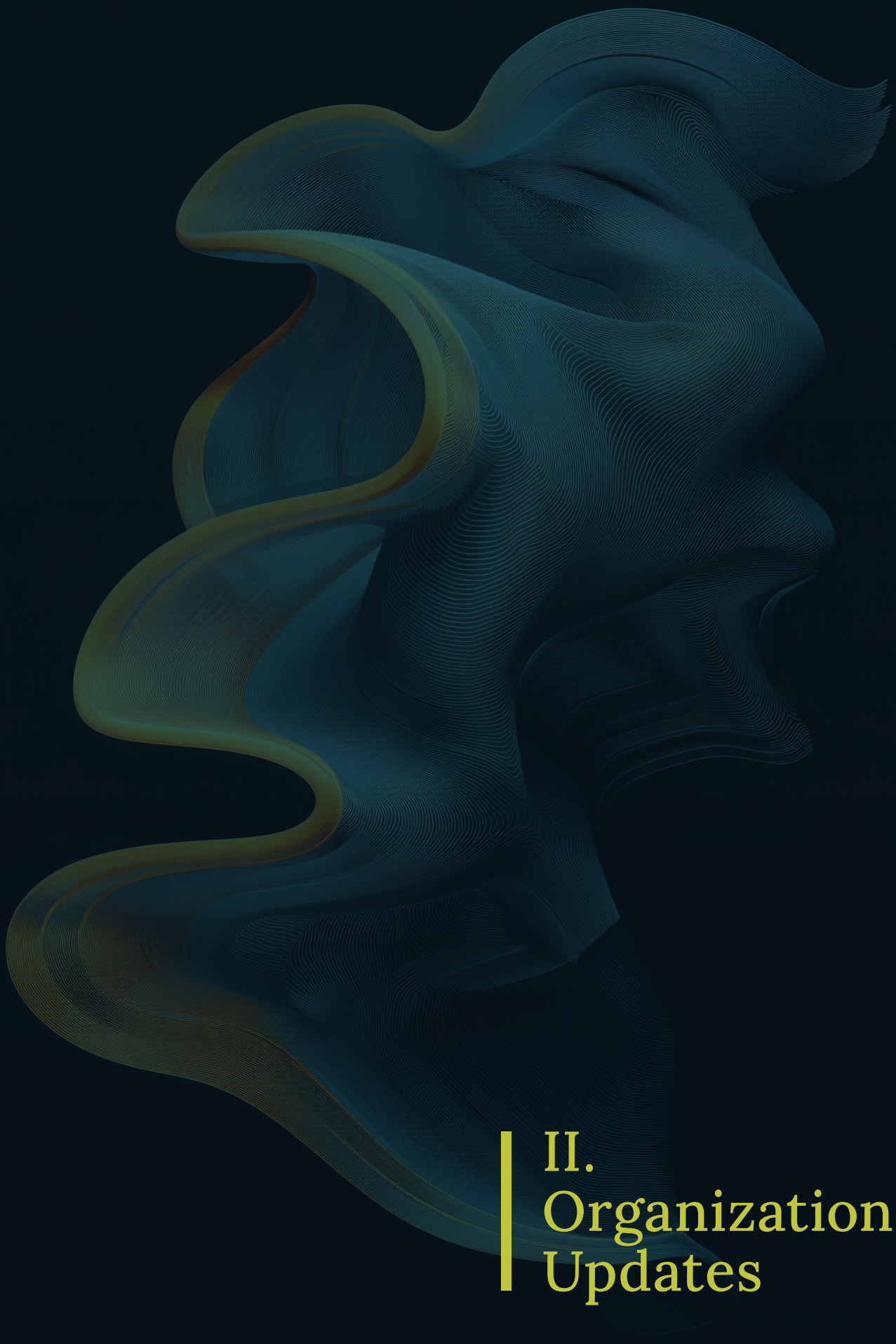
2025 was also an important step forward for our talent programs. A record seventy-four people participated in our full-time three-month-long fellowship program, and, for the first time, we hosted cohorts in both DC and London. We accepted our largest-ever cohort of year-long visitors, relaunched a US AI policy course, and experimented with offering co-working space and hosting networking events in a new London office. This year, alumni from our programs went on to destinations that included Google DeepMind, the US Senate, and the UK AI Security Institute.

For the first time, we hosted cohorts in both DC and London.

As 2026 brings new AI capabilities and new shifts in the world, we expect demand for our work to continue to grow. Beyond simply extending our work, a few key themes will be building up a larger on-the-ground presence in DC, adapting our talent programs to support career paths other than research and policy work, and launching new research workstreams to broaden our expertise and expand the ways we can be useful to decision-makers.

This report provides a summary of GovAI's work in 2025 and sketches an ambitious agenda for the year to come. Ten years on from our founding, our work has never been more important. I'm grateful to everyone who makes this work possible: our donors, our collaborators and partners, and, most of all, our dedicated staff.

BEN GARFINKEL
Executive Director



II. Organizational Updates

Research

PROGRESS IN 2025

More so than in previous years, our research covered a wide range of topics, including the economic impacts of AI, potential risks from frontier AI models, responsible frontier AI practices, regulatory design for frontier AI, and agent governance. We significantly expanded our research team this year, from 11 to 17. Further, we stood up two teams: the Risk Management Team led by Jonas Freund and the Threat Modelling Team initiated by Luca Righetti, who joined us in early 2025, and now led by John Halstead.

We welcomed 11 new Research Scholars working across our focus areas: James Coates, Aidan Homewood, John Lidiard, Amelia Michael, Peter McIntyre, Liam Patell, Kamile Lukosiute, Bhuvana Sudarshan, and most recently Rebecca Hersman, Tom Reed, and Zaheed Kara.

In addition, we offered former Research Scholars Matthew van der Merwe, Sophie Williams, and John Halstead more permanent roles on our team as Research Fellows. We also welcomed Elias Groll as Senior Research Editor in January 2026 and promoted Aquila Hassan to Research Manager.

Our research covered a wide range of topics, including the economic impacts of AI, potential risks from frontier AI models, responsible frontier AI practices, regulatory design for frontier AI, and agent governance.

Our work touched on many critical areas in AI governance. We primarily focused on the following workstreams.

Risk Management—Improving frontier AI practices: GovAI researchers developed frameworks for third-party compliance reviews for safety frameworks, examining how independent parties could assess whether frontier AI companies are adhering to their commitments. Our researchers analyzed the practice of assessing risk relative to competitors, identifying concerns with approaches that allow companies to lower mitigations based on competitors' model releases. GovAI team members also contributed to work on third-party flaw disclosure for general-purpose AI systems and surveyed experts on thresholds for advanced AI systems, finding general agreement that thresholds should be set by multiple stakeholders.

Threat Modeling—Assessing the potential impacts of frontier AI models: A GovAI researcher developed a framework for converting capability evaluations into risk assessments for dual-use biological capabilities. Our team also developed methods for forecasting LLM-enabled biorisk, recruiting subject-matter experts and forecasters to estimate how growing LLM capabilities affect

epidemic risk—finding that LLMs have already crossed capability thresholds experts thought wouldn't arrive until after 2030. GovAI researchers also proposed STREAM, a standard for transparently reporting evaluations of chemical and biological capabilities in AI model reports.

GovAI researchers expanded work aimed at informing US policymakers.

US Policy: GovAI researchers expanded work aimed at informing US policymakers, including by addressing foundational questions related to the labor impact of AI and the most effective ways to build out US AI infrastructure. This work included research related to export promotion, ways the Department of Labor could meet its mandate in the AI Action Plan to better measure AI's labor impacts, and whether it is cheaper to build data centers in the UAE than in the US.

Economics of AI: GovAI researchers extended prior work on labor market exposure to LLMs to the firm level, finding that companies in the US have an average of 17% of their workforce's tasks exposed to LLMs directly and 47% assuming partial integration of LLMs into software their workers use on the job. Another project team conducted a randomized controlled trial assessing how retrieval-augmented generation and AI reasoning models could affect legal work, finding significant productivity gains of 34–140% across most tasks tested.

Regulatory Design for Frontier AI: To inform training-compute threshold-based regulation, our researchers forecasted the number of frontier AI models that will exceed various training-compute thresholds through 2028. A GovAI researcher described regulatory supervision of frontier AI developers as a potentially good approach to manage risks while enabling innovation. GovAI team members examined the challenge of regulating downstream AI developers who fine-tune or modify foundation models. A comprehensive report on the UK's approach to AI regulation traced the evolution of UK AI policy and proposed establishing a flexible, principles-based regulator for advanced AI development.

Agent Governance: GovAI researchers helped develop the concept of agent infrastructure: technical systems and shared protocols that are external to agents and designed to mediate their interactions with their environments. Our researchers examined the role governments should play in providing such infrastructure, analyzed how Article 50 of the EU AI Act may require labelling of AI agent activity, proposed an incident analysis framework for agents, and examined how the shift toward inference scaling may affect AI governance.

Highlighted Research

Infrastructure for AI Agents

Increasingly many AI systems can plan and execute interactions in open-ended environments, such as making phone calls or buying online goods. As developers grow the space of tasks that such AI agents can accomplish, we will need tools to both unlock their benefits and manage their risks. Current tools are largely insufficient because they are not designed to shape how agents interact with existing institutions (e.g. legal and economic systems) or actors (e.g. digital service providers, humans, other AI agents). For example, alignment techniques by nature do not assure counterparties that some human will be held accountable when a user instructs an agent to perform an illegal action. To fill this gap, we propose the concept of agent infrastructure: technical systems and shared protocols that are external to agents and designed to mediate and influence their interactions with and impacts on their environments. Agent infrastructure comprises both new tools and reconfigurations or extensions of existing tools. For example, to facilitate accountability, protocols that tie users to agents could build upon existing systems for user authentication, such as OpenID. Just as the Internet relies on infrastructure like HTTPS, we argue that agent infrastructure will be similarly indispensable to ecosystems of agents. We identify three functions for agent infrastructure: 1) attributing actions, properties, and other information to specific agents, their users, or other actors; 2) shaping agents' interactions; and 3) detecting and remedying harmful actions from agents. We propose infrastructure that could help achieve each function, explaining use cases, adoption, limitations, and open questions. Making progress on agent infrastructure can prepare society for the adoption of more advanced agents.

Alan Chan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K. Hadfield, Markus Anderljung | *January 17, 2025*

What Does the Public Think About AI?

Drawing from academic studies and public polling data, this report synthesizes public attitudes towards AI with a focus on the United Kingdom and United States. It discusses public views on issues such as concern about job loss, emerging trends, challenges, and future opportunities. The report also introduces and draws on a new resource to analyze emerging trends in AI public opinion surveys—the AI Survey Hub for Attitudes and Research Exchange (AI SHARE) database, which aggregates survey data from over 200 studies conducted between 2014 and 2023.

Noemi Dreksler, Harry Law, Chloe Ahn, Daniel S. Schiff, Kaylyn Jackson Schiff, Zachary Peskowitz | *January 22, 2025*

Extending "GPTs Are GPTs" to Firms

We extend Eloundou et al. (2024) to build firm-level measures of exposure to large language models (LLMs) with data from two sources: Eloundou et al. (2024) for occupation-level measures of LLM exposure and Revelio Labs for firm-level employee counts by occupation. The results indicate that companies with more technology workers and AI-skilled employees tend to have higher levels of LLM exposure. We also find that differences in LLM exposure are greater between exposure categories than within them, suggesting that integrating LLMs into corporate systems may lead to significant productivity gains.

Benjamin Labaschin, Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock | May 1, 2025

Dual-Use AI Capabilities and the Risk of Bioterrorism

Several frontier AI companies test their AI systems for dual-use biological capabilities that might be misused by threat actors. But what do these test results imply about the overall risk of bioterrorist attacks? There is much expert debate about how seriously to view such threats, especially from lone wolf actors. This report creates a framework for how to convert capability evaluations into risk assessments, using a simple model that draws on historical case studies, expert elicitation, and reference class forecasting. I conclude that if AI systems were to increase the number of STEM Bachelors able to synthesize pathogens as complex as influenza by 10 percentage points and also enable them to design concerning operational attack plans, then the annual probability of an epidemic caused by a lone wolf attack might increase from 0.15% to 1.0%. This is equivalent to 12,000 additional expected deaths per year, or ~$100B. Risk scenarios where AI or other tools also help discover novel viruses reach higher damages, whereas risk can also be significantly lowered if mitigations are put in place. A review of this report by six subject-matter experts and five superforecasters found similar medians, though all forecasts had high uncertainty. This work demonstrates a methodological approach for converting capability evaluations into risk assessments, whilst highlighting the continued need for better underlying evidence and expert discussion to refine assumptions.

Luca Righetti | December 12, 2025

Does the UAE Have an Advantage in Building Data Centers?

American hyperscalers are increasingly exploring building data centers in the UAE, but it is unclear whether doing so is intrinsically cheaper than building in the United States. This report compares the costs of building and operating a hypothetical 100-MW data center in the US versus the UAE. The analysis indicates that total costs are similar, although there are differences in specific costs. Given sparse public data—especially as no 100-MW+ data centers have yet been built in the UAE—these estimates are best treated as provisional. For example, in part due to data scarcity, the model does not account for differences in time to connect especially large data centers to the grid. However, if the total cost estimates are roughly correct, they suggest that current hyperscaler interest in the UAE is not driven by an intrinsic cost advantage. Interest may instead be driven by factors like subsidies, expectations of future cost declines, business opportunities unlocked by investments, or a desire to hedge against US delays. This report's estimates indicate that the US maintains structural advantages in data center construction, including (perhaps counterintuitively) cheaper energy, a more hospitable natural environment, and a robust domestic data center industry, with its primary disadvantages being higher building construction costs and permitting delays.

Amelia Michael | December 22, 2025

AMBITIONS FOR 2026

We expect to continue making significant investments in our existing workstreams, in particular risk management, threat modeling, US policy, and economic impacts of AI. We also aim to expand into more areas, including AI resilience (increasing society's ability to handle negative impacts from AI), state actors' use of AI (including for military purposes), the impacts of AI R&D automation, and how AI could increase innovation and economic growth more broadly.

We aim to expand into more areas this year, including AI resilience, state actors' use of AI, the impacts of AI R&D automation, and how AI could increase innovation and economic growth more broadly.

To enable the above, we aim to continue growing our research team by 25–50% over the course of 2026. We also expect to build out our Research Management Team further, with an additional Research Manager alongside Aquila Hassan and Elias Groll as Senior Research Editor.

Within our existing workstreams, we expect team members to focus on the following areas.

Risk management: The risk management team will continue to conduct research to support appropriate frontier AI safety framework design and implementation. Much of this work is likely to focus on how companies can comply with new legal requirements, such as those from California, New York, and the EU. This includes research on the auditing of frontier AI developers, how companies should assess and mitigate risks from harmful manipulation, and how companies should take into account the risk posed by competitor models.

Threat modeling: After the threat modeling team spent much of 2025 developing their approach to modeling threats from frontier and future AI systems (exemplified by their work on dual-use capabilities and bioterrorism), they intend to describe that methodology—establishing AI threat modeling as a field—while also publishing a number of threat models we worked on in 2025. These threat models include the use of AI to attack the electrical grid, conduct cybercrime, develop novel worm attacks, carry out chemical and explosive attacks, and more. Team members will also deepen their work on biorisk, e.g. to inform the choice of relevant thresholds.

Further, we are excited to see the leadership of the threat modeling team transition from Luca Righetti to John Halstead, with Luca moving to more hands-on biosecurity work. John has extensive experience in threat modeling and running research teams. We're excited to see what he and the team will accomplish.

US policy: Team members intend to continue working to inform the implementation of the AI Action Plan, examine workforce policy in light of the opportunities and challenges posed by AI, and conduct analysis relevant to US competitiveness in AI development and adoption.

Economic impacts of AI: We plan to grow the team of researchers working on AI's economic impacts at GovAI while building on researchers' previous empirical and conceptual work related to this topic. This includes developing methods to track labor market impacts and understand drivers of worker resilience to disruptions posed by AI.

As always, researchers' priorities may substantially shift over the course of the year, as new opportunities arise and AI develops in unexpected directions.

Talent Development

PROGRESS IN 2025

In 2025, we established a wider Talent Development Team around Valerie Belu.

In 2025, we established a wider Talent Development Team around Valerie Belu, whom we promoted from Head of Talent Development to Director of Talent in June. We expanded the team to four members with a new DC Fellowship Coordinator, Martin Fukui, in April and two new Research Management Associates focused on our UK programs—Jannis Hamida and Verena Heusser—joining in July and August, respectively.

This growth enabled us to expand and improve our flagship talent development program, the Seasonal Fellowships, which included launching a new edition in Washington, DC. We also maintained our Research Scholar program at a comparable size to the previous year. Lastly, we ran the second iteration of our Policy Program, piloted a Women in AI Governance Mentorship program, and hosted our first alumni reunion event.

Seasonal Fellowships are three-month opportunities designed to launch or accelerate impactful careers in AI governance and policy. Participants conduct a supervised research project of their choice, while also participating in an expert Q&A and seminar series aimed at developing a broad understanding of the field, and spending time forging connections with other researchers and practitioners. In 2025, we increased the number of cohorts we welcome per year from two to three. We hosted two cohorts of our UK-based fellowship—one from January and one from July—in our new London office. In addition, we launched our first ever US-based fellowship in October in Washington, DC. In total, this enabled us to host 81 Seasonal Fellows, a 125% increase compared to the previous year. Our application numbers also continued to rise, nearly doubling from 9,700 in 2024 to 18,000 in 2025.

About 60% of our alumni have settled into a longer-term role in AI governance a year after their fellowship ended.

Many of our 2025 graduates have already gone on to influential roles in AI governance, including at government institutions like the UK's AI Security Institute; at think tanks such as the Foundation for American Innovation, the Institute for Law and AI, and the SafeAI Forum; and in industry, at Google DeepMind, Anthropic, and the Frontier Model Forum. Recent data shows that about 60% of our alumni have settled into a longer-term role in AI governance a year after their fellowship ended; the remainder are split among predominately graduate education, temporary roles, or work in other fields (including AI safety more broadly).

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 2025, we supported 15 Research Scholars to work on a wide range of topics, including UK, US, and EU AI policy, national security implications of AI, risk management, technical governance, and threat modeling.

We continue to support secondments for Research Scholars who hope to gain direct policy experience. Over the past year, we have increased the number of staff members who were seconded to relevant institutions, including the UK's Department of Science, Innovation and Technology 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 2025, we saw Scholars move on to work at the UK AI Security Institute, the Centre for Long-Term Resilience, and Microsoft. We are also increasingly offering some Scholars the chance to stay on at GovAI as Research Fellows.

In addition to our established programs, we also ran an improved version of our remote Policy Program first trialed in 2023. This program is a twelve-week part-time program structured around guided self-study, workshops, and seminars on AI policy. The material is designed to distill key context on the current AI landscape and give cohort members hands-on experience engaging with policy questions, while also accommodating busy academic or work schedules. Sixteen people completed the program, and in the post-program survey, 90% of respondents reported that they would strongly recommend it to others. We also piloted a mentorship program for women interested in entering the field and hosted our first alumni reunion event, with more to follow in 2026.

In 2026, we will focus on continuing to expand our talent development work in the US, while also experimenting with new ways of meeting the field's urgent demand for non-research focused talent.

AMBITIONS FOR 2026

In 2026, we will focus on continuing to expand our talent development work in the US, while also experimenting with new ways of meeting the field's urgent demand for non-research focused talent.

We will host another, larger fellowship in Washington, DC, this time in summer 2026, concurrently with our 2026 UK Summer Fellowship. In combination with a slight increase in the expected size of our UK cohorts, this will allow us to host around 100 Seasonal Fellows across our UK and DC fellowship programs, a nearly 25% increase compared to 2025. We also expect to run two additional iterations of our remote, part-time Policy Program: one more closed round from March to May and a second, open round aimed at US congressional

In response to the increasing demand for non-research-focused talent, we are trialing a new Applied Track for the 2026 UK Summer Fellowship.

and executive staffers, as well as think tank researchers, from September to November. We plan to expand our Research Scholar program in 2026 to keep in step with our overall growth plans for our research and policy workstreams, aiming to support approximately 18–20 of them this year.

In response to the increasing demand for non-research-focused talent such as operations and communications staff, research and program managers, and policy engagement professionals, we are trialing a new Applied Track for the 2026 UK Summer Fellowship. We are looking for fellows with prior experience in communications, policy, issue advocacy, events, research management, program management, operations, and fundraising in other fields, and will invite them to spend three months at our offices, completing a project in an area other than traditional research while improving their familiarity with AI governance through seminars and expert Q&As and building a network in the field.

We also plan to make ongoing improvements to our existing programs. For example, we have constructed our first comprehensive database of our alumni's post-program career trajectories and will use this to refine our outreach strategy, selection processes, and the career support we offer our fellows. We are also currently hiring for a Head of Community and Partnerships, whose portfolio of responsibilities will include significantly expanding our alumni engagement.

Policy Engagement

We worked with decision-makers across the UK, the EU, the US, and frontier AI companies.

PROGRESS IN 2025

2025 was our most active year to date in terms of policy engagement as we worked with decision-makers across the UK, the EU, the US, and frontier AI companies.

We increased the scope of US policy engagement at GovAI compared to 2024, through the first US edition of our Seasonal Fellowship in Washington, DC, and through hiring two Research Scholars focused on informing US AI policy decisions. We engaged with policymakers on issues including export promotion, how the Department of Labor could improve our understanding of AI's labor impacts, and the cost of building data centers in the US.

Another focus of 2025 was providing expertise to frontier AI companies on the design and implementation of their 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, GovAI researchers offered independent research, provided feedback on frameworks, and engaged with policymakers seeking to understand such frameworks.

In the UK, GovAI researchers continued to provide information and analysis on domains including frontier AI regulation, international governance, and agent governance. We hosted a series of events at our new London office, covering topics from CBRN risk evaluation to lessons from two years of post-Bletchley AI governance. We also partnered with the British Foreign Policy Group on a roundtable exploring the UK's role in the international AI landscape and potential for US-China mediation.

As the EU has started implementing its 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 served 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 offered expert opinions on questions around the AI Act's scope via the Commission's Joint Research Centre. Alan Chan has also started engaging with the EU's forthcoming Code of Practice regarding Article 50, which focuses on ensuring citizens are informed that they are engaging with AI systems when applicable.

On the international level, GovAI researchers attended the 2025 Paris AI Action Summit—including hosting an event there—and engaged with the hosts of the India AI Impact Summit scheduled for February 2026. Further, Jonas Freund led a survey on thresholds for advanced AI systems in collaboration with the OECD, and we also hosted two secondees to the organization who supported its Strategic Foresight Unit.

We hosted a series of events at our new London office, covering topics from CBRN risk evaluation to lessons from two years of post-Bletchley AI governance.

Notable External Engagements

Reception on the Role of Third Parties in Safety Frameworks
GovAI hosted a closed reception in Paris on 7 February 2025, alongside the AI Action Summit. The event, held under the Chatham House Rule, brought together experts from academia and research organizations to discuss improving the quality and implementation of safety frameworks. Many leading AI developers have adopted safety frameworks as a formalized approach to keeping risks from frontier AI systems to a tolerable level.

Frontier AI and CBRN Risk Seminar
On 31 July 2025, we hosted an event on frontier AI and CBRN risk at our London office. The event featured presentations from Luca Righetti on forecasting LLM biorisk, Tegan McCaslin and Tom Reed on establishing rigorous grading standards for CBRN safety evaluations, and Matthew van der Merwe and John Halstead on determining the appropriate scope of safety frameworks beyond viral threats, followed by a moderated discussion.

Two Years Since Bletchley: Progress, Challenges, and What's Next?
Two years ago, the Bletchley AI Safety Summit marked the first major intergovernmental convening on frontier AI, resulting in the Bletchley Declaration and announcement of the UK AI Safety Institute. On 10 November 2025, we hosted an event at our London office inviting participants to reflect on these two years and look ahead for the field. The event featured talks from Kanishka Narayan MP (Parliamentary Under-Secretary of State for AI and Online Safety), Victoria Krakovna (Senior Research Scientist at Google DeepMind), Rishi Bommasani (Society Lead at the Stanford Center for Research on Foundation Models), and Jade Leung (Prime Minister's AI adviser and CTO of the AI Security Institute), alongside a panel discussion with Nitarshan Rajkumar (International Policy at Anthropic), Marc Warner (CEO of Faculty AI), and Henry de Zoete (former adviser to the Prime Minister on AI).

Roundtable on the UK's Position in the International AI Landscape
GovAI and the British Foreign Policy Group (BFPG) co-hosted a closed roundtable discussion on 11 September 2025 covering the UK's position in the international AI landscape. The session brought together 10–15 experts under the Chatham House Rule to inform research on the UK's international AI partnerships and its potential role in facilitating US-China cooperation. The roundtable explored how the UK can strengthen its position as a leading nation in AI development and governance, debated specific partnership strategies with the US, middle powers, and China, and examined key challenges to achieving a stable AI governance landscape, including the UK's potential for US-China mediation.

AMBITIONS FOR 2026

We have ambitions to expand the portion of our team's research that is relevant to and tailored for US policymakers, as well as to grow our US-focused talent programs.

We aim to continue enabling work providing technical assistance to policymakers in the UK, EU, and US, looking to ensure that these economies capture the potential upside benefits of AI systems while mitigating downside risks. We also hope to considerably increase the quality of our team's independent research and advice 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 team's 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 support and lead these efforts.

Organizational Capacity

PROGRESS IN 2025

GovAI has grown substantially over the course of 2025: we ended the year with 40 full-time staff members across both permanent and visiting positions, up from around 30 at the start of the year, and have six additional team members joining in January 2026.

In June 2025, Abraham Rowe joined GovAI as Director of Operations on our leadership team, succeeding Ryan Fugate.

We significantly increased our operational capacity throughout the year to enable continued growth in 2026.

We significantly increased our operational capacity throughout the year to enable continued growth in 2026, including building out larger recruitment and people operations teams, and introducing additional structure within the growing team. We promoted Arianna Bosio to lead the newly established People Operations Team as Manager. We also added several new operations staff, including Isha Paik (Operations Specialist), Leyla Ava (Team Executive Assistant), Robyn Seabrook (Executive Assistant to the Director), Anastacia Button-Gentry (Operations Associate), Nathan T. (Office Manager), John Drummond (Talent Systems Specialist), Natasha Martemianova (Hiring Specialist), Maddie Ruwitch (Hiring Specialist), Romina Wojcik (Executive Assistant and General Ops), Victoria Lloyd (Executive Assistant to the Director of Policy and Research), and most recently Emily Moore as General Counsel.

We expect this larger team to allow us to add up to 30 additional staff over the coming years, support over 100 seasonal fellows, and run our office as an AI governance hub for visitors and guests in London.

OUR NEW OFFICE IN LONDON

We moved our main office to central London expanding our ability to engage with and support key stakeholders in the UK and beyond.

In April 2025, we moved our UK office from Oxford to central London. This relocation marked a key milestone and has expanded our ability to engage with and support key stakeholders in the UK and beyond. Being situated in King's Cross, we are able to serve increasingly as a central node in the field of AI governance, holding events and establishing a visitors program. On 5 June 2025, we officially celebrated our move to London with an evening event that brought together our alumni, collaborators, and the broader AI governance community. The launch welcomed guests from across academia, civil society, industry, and government to our new space. The evening featured brief remarks from our leadership team outlining our vision for GovAI's presence in London, followed by extended opportunities for conversation among some of the city's leading AI governance researchers and practitioners.

FINANCES

We at GovAI are extremely grateful to our funders, who continued to generously support our mission in 2025. During the year, we received commitments and donations from Coefficient Giving (formerly Open Philanthropy), the Survival and Flourishing Fund, Longview Philanthropy, along with many donors on the Giving What We Can and Every.org platforms.

We are extremely grateful to our funders, who continued to generously support our mission in 2025.

In our first full year as an independent organization, we established a wide spectrum of financial operations to support GovAI's expansion and completed our first audit related to our 2024 accounts.

In 2025, our total expenses for the year came to $11.9 million USD, growing by almost 100% compared to 2024. Of this, approximately 60% went to staff costs, including salary, visas, taxes, and other payroll costs for our researchers, operations team, seasonal fellows, and management. Approximately 20% was spent on facilities and equipment, primarily rent and office management fees, including one-off costs for the fit out of our new London office. The remaining roughly 20% was deployed on operational costs, including contract services (legal, financial, and other contractors), events, and team travel.

You can donate to us at www.every.org/govai, or if you already have a setup with Giving What We Can, at www.givingwhatwecan.org/charities/govai. 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 development@governance.ai.

AMBITIONS FOR 2026

In 2026, we expect to continue our growth trajectory. We are establishing a new office in Washington, DC, to host our DC fellowships along with our growing cohort of long-term DC researchers. With the office in London, we have significantly more capacity to host events, run workshops, and run larger fellowship cohorts. We'll continue growing these programs and workstreams further. Finally, in 2026, we intend to update our branding, website, and public communication pipelines to more effectively reach decision-makers and other stakeholders with our research.

People

SENIOR LEADERSHIP TEAM

BEN GARFINKEL

Executive 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 role of economic growth in promoting security. 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 the Executive 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 leads GovAI's research and advisory work to support the safe and beneficial development of frontier AI systems. His work focuses on the regulation of the most capable AI models, AI risk assessment, national security implications, and compute governance as a policy tool.

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. Markus has previously served as a vice-chair drafting the EU Code of Practice for General-Purpose AI and was seconded to the UK Cabinet Office as a Senior AI Policy Specialist.

VALERIE BELU

Director of Talent

Valerie 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.

ABRAHAM ROWE

Director of Operations

Abraham leads GovAI's operations and other internal functions. His past experience includes co-founding and running the scientific grantmaking organization Wild Animal Initiative, serving as the COO of the think tank Rethink Priorities, and managing a nonprofit operations consultancy and service provider.

BOARD

JEFFREY DING

Jeff is an Assistant Professor of Political Science at George Washington University.

TASHA MCCAULEY

Tasha is a technology entrepreneur, former CEO of GeoSim Systems, and a co-founder of Fellow Robots.

SEÁN Ó HÉIGEARTAIGH

Seán is the Director of the AI: Futures and Responsibility Programme, part of the Centre for the Future of Intelligence at the University of Cambridge, and a Research Professor at Cambridge.

TOBY ORD

Toby is a Senior Researcher at the Oxford Martin AI Governance Initiative.

HELEN TONER

Helen is the Interim Executive Director 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, Director of Research and Policy, and Paul Harding, Operations Manager.

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

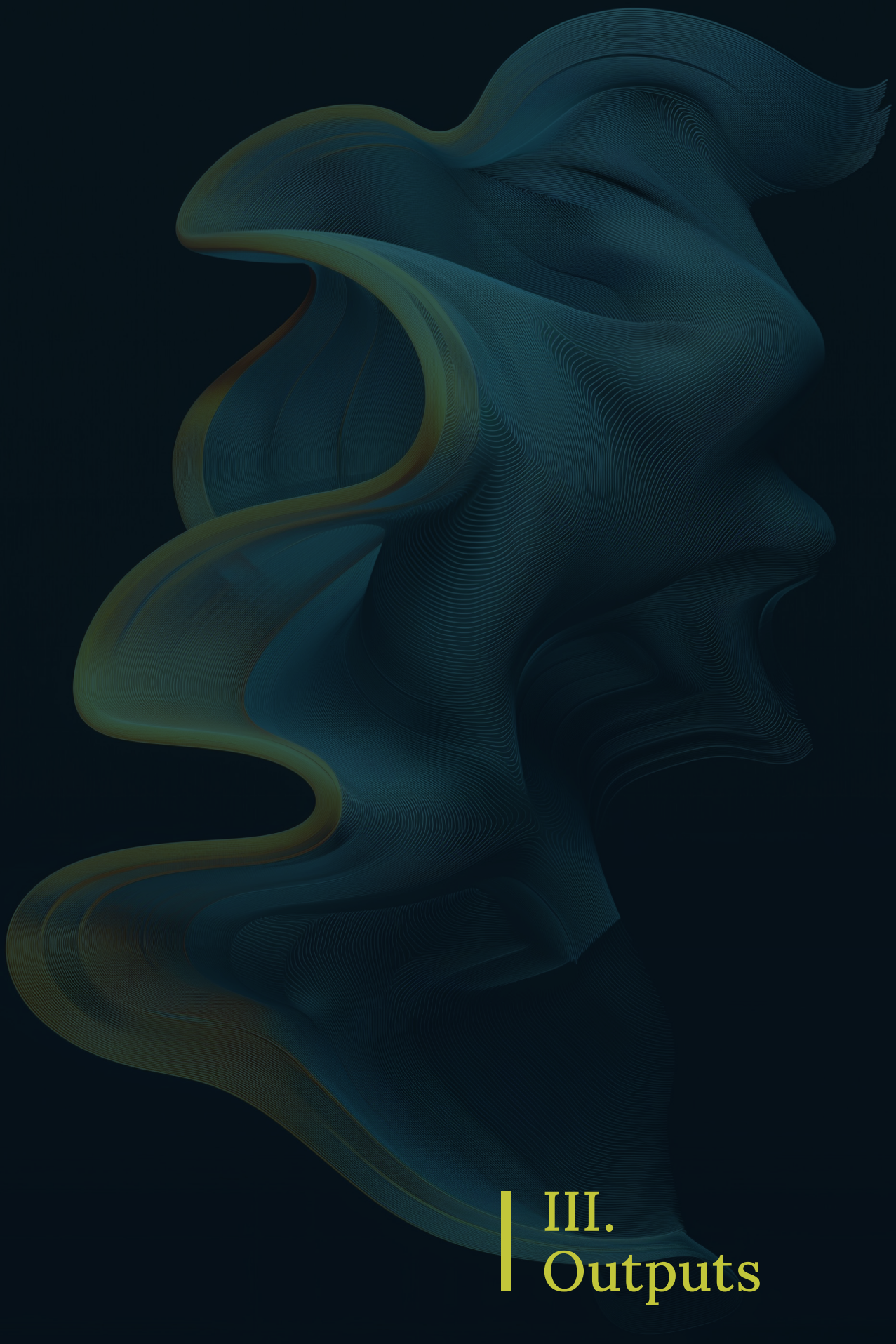
In Summary

Ten years after our founding, the premise which motivated GovAI's creation—that AI might transform the world—is no longer speculative. AI is transforming how people work, communicate, and navigate the world—and the key question has become what decision-makers can do to positively shape that transformation as it unfolds.

In 2025, we responded to this moment by producing our most substantial body of research yet, expanding our talent programs to new geographies and formats, and deepening our engagement with decision-makers across government, industry, and civil society. We grew our team significantly, moved into a new London office that is becoming a hub for the AI governance community, and laid the groundwork for a growing presence in Washington, DC.

We are determined to meet the challenges arising from the transition to a world with advanced AI systems.

In 2026, we will build on this foundation. We plan to launch new research workstreams, scale our fellowship programs, open a new DC office, and find new ways to support the talent the field increasingly needs. As AI capabilities continue to advance, we expect demand for rigorous, independent analysis and well-prepared professionals to only grow. We are determined to meet the challenges arising from the transition to a world with advanced AI systems.



III. Outputs

Publications, Reports, and Working Papers

- **Permission Manifests for Web Agents**
  Samuele Marro, Alan Chan, Xinxing Ren, Lewis Hammond, Jesse Wright, Gurjyot Wanga et al.

- **Does the UAE Have an Advantage in Building Data Centers?**
  Amelia Michael

- **Dual-Use AI Capabilities and the Risk of Bioterrorism**
  Luca Righetti

- **Labeling of AI Agent Activity in Article 50 of the EU AI Act**
  Alan Chan

- **Assessing Risk Relative to Competitors: An Analysis of Current AI Company Policies**
  Sophie Williams, Noemi Dreksler, Aidan Homewood, Markus Anderljung, Jonas Freund

- **Inference Scaling and AI Governance**
  Toby Ord

- **International AI Safety Report 2025: First Key Update: Capabilities and Risk Implications**
  Yoshua Bengio et al., including Sam Manning

- **STREAM (ChemBio): A Standard for Transparently Reporting Evaluations in AI Model Reports**
  Tegan McCaslin, Jide Alaga, Samira Nedungadi, Seth Donoughe, Tom Reed, Rishi Bommasani, Chris Painter, Luca Righetti

- **Survey on Thresholds for Advanced AI Systems**
  Jonas Freund, Eunseo Choi, Kasumi Sugimoto, Bosco Hung, Robert Trager, Karine Perset

- **Incident Analysis for AI Agents**
  Carson Ezell, Xavier Roberts-Gaal, Alan Chan

- **From Turing to Tomorrow: The UK's Approach to AI Regulation**
  Oliver Ritchie, Markus Anderljung, Tom Rachman

- **Forecasting LLM-Enabled Biorisk and the Efficacy of Safeguards**
  Bridget Williams, Luca Righetti, Josh Rosenberg, Rebecca Ceppas de Castro, Otto Kuusela, Rhiannon Britt, Emily Soice, Alvaro Morales, Jon Sanders, Seth Donoughe, James Black, Ezra Karger, Philip E. Tetlock

- **Third-Party Compliance Reviews for Frontier AI Safety Frameworks**
  Aidan Homewood et al., including Sophie Williams, Noemi Dreksler, John Lidiard, Ben Garfinkel, Jonas Freund

- **Extending "GPTs Are GPTs" to Firms**
  Benjamin Labaschin, Tyna Eloundou, Sam Manning, Pamela Mishkin, Daniel Rock

- **Trends in Frontier AI Model Count: A Forecast to 2028**
  Iyngkarran Kumar, Sam Manning

- **Trends in AI Supercomputers**
  Konstantin F. Pilz, James Sanders, Robi Rahman, Lennart Heim

- **n-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI**
  Shayne Longpre et al., including Markus Anderljung

- **On Regulating Downstream AI Developers**
  Sophie Williams, Jonas Freund, Markus Anderljung

- **AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice**
  Daniel Schwarcz, Sam Manning, Patrick Barry, David R. Cleveland, JJ Prescott, Beverly Rich

- **Regulatory Supervision of Frontier AI Developers**
  Peter Wills

- **Multi-Agent Risks from Advanced AI**
  Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, et al. (44 authors total including Carson Ezell, Anka Reuel)

- **Measuring AI Agent Autonomy: Towards a Scalable Approach with Code Inspection**
  Peter Cihon, Michael Stein, Ganesh Bansal, Sam Manning, Kevin Xu

- **What Does the Public Think About AI?**
  Noemi Dreksler, Harry Law, Chloe Ahn, Daniel S. Schiff, Kaylyn Jackson Schiff, Zachary Peskowitz

- **Infrastructure for AI Agents**
  Alan Chan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K. Hadfield, Markus Anderljung

- **Local US Officials' Views on the Impacts and Governance of AI: Evidence from 2022 and 2023 Survey Waves**
  Sophia Hatz, Noemi Dreksler, Kevin Wei, Baobao Zhang

- **Options and Motivations for International AI Benefit Sharing**
  Claire Dennis, Sam Manning, Stephen Clare, Boxi Wu, Jake Okechukwu Effoduh, Chinasa T. Okolo, Lennart Heim, Katya Klinova

- **International AI Safety Report 2025**
  Yoshua Bengio et al. (including GovAI researchers Sam Manning and Ben Garfinkel as writers)

- **Authenticated Delegation and Authorized AI Agents**
  Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, Alex Pentland

Other Outputs

- **What Role Should Governments Play in Providing AI Agent Infrastructure**
  Alan Chan

- **The Case for Supervising Frontier AI Developers**
  Peter Wills

- **Options and Motivations for International AI Benefit Sharing**
  Claire Dennis, Sam Manning, Stephen Clare

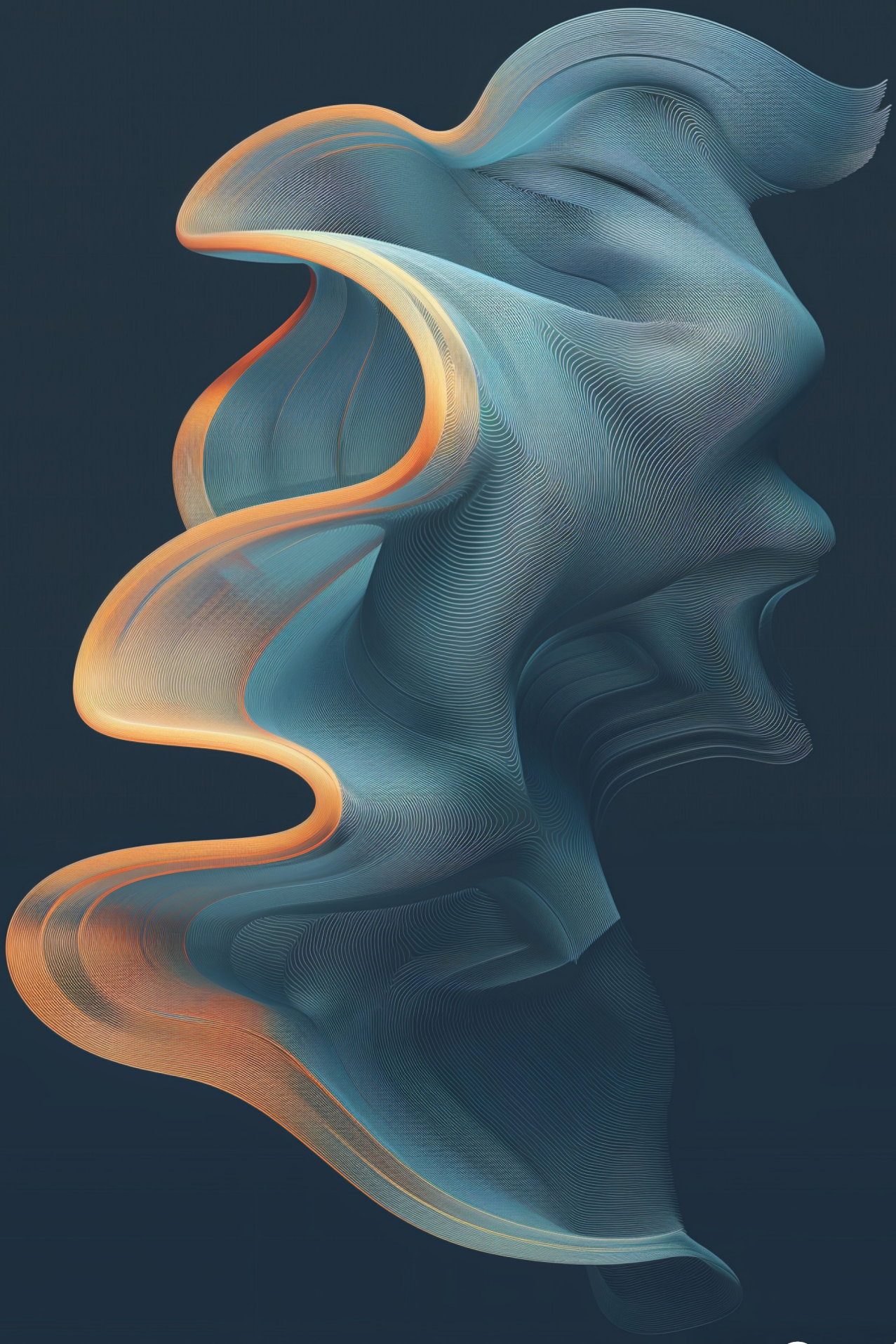
- **What Success Looks Like for the French AI Action Summit**
  Claire Dennis, Ben Clifford, Ben Garfinkel, Markus Anderljung, Robert Trager

- **Export Controls and Export Promotion**
  Sam Manning (The Republic Journal - June 2025)

- **Here's How To Share AI's Future Wealth**
  Saffron Huang, Sam Manning (Noema Magazine - April 2025)

- **Addressing the U.S. Labor Market Impacts of Advanced AI**
  Sam Manning

- **Understanding AI's Labor Market Impacts: Opportunities for the Department of Labor's AI Workforce Research Hub**
  Sam Manning (Foundation for American Innovation – 2025



GovAI