GOV.UK

Department for Science, Innovation & Technology

Independent report

Frontier AI Taskforce: second progress report

Published 30 October 2023

Contents

This publication is available at https://www.gov.uk/government/publications/frontier-ai-taskforce-second-progress-report/frontier-ai-taskforce-second-progress-report

Today we are announcing that in the 7 weeks since our first progress report (https://www.gov.uk/government/publications/frontier-ai-taskforce-first-progress-report/frontier-ai-taskforce-first-progress-report) we have:

- Tripled the capacity of our research team, with 150 years of frontier AI research experience across the team.
- Recruited Jade Leung, a leading expert in safety protocols and governance of frontier AI systems, and Rumman Chowdhury, an expert in evaluating social harms from frontier models. Jade joins us from OpenAI; Rumman from Humane Intelligence.
- Cemented new partnerships with leading AI organisations. This brings our network of partner organisations to 11 and includes organisations expert in understanding AI risk across biosecurity, cybersecurity and deceptive behaviour.
- The Taskforce has supported the Department for Science, Innovation and Technology (DSIT) and the University of Bristol in establishing the UK's AI Research Resource (AIRR), Isambard-AI, an AI supercomputer on which we will be able to conduct more compute intensive safety research.
- Prepared a cutting-edge research programme; the initial results of which will be showcased on Day 1 of the Summit.

The Taskforce is a start-up inside government, delivering on the mission given to us by the Prime Minister: to build an AI research team that can evaluate risks at the frontier of AI. We are now 18 weeks old.

That frontier is moving very fast. On the current course, in the first half of 2024, we expect a small handful of companies to finish training models that could produce another significant jump in capabilities beyond state-of-the-art in 2023.

As these AI systems become more capable they may significantly augment risks. An AI system that advances towards expert ability at writing software could increase cybersecurity threats. An AI system that becomes more capable at advancing biology could escalate biosecurity threats.

We need to keep pace with those developments. We believe it is critical that frontier AI systems are developed safely and that the potential risks of new models are rigorously and independently assessed for harmful capabilities before and after they are deployed.

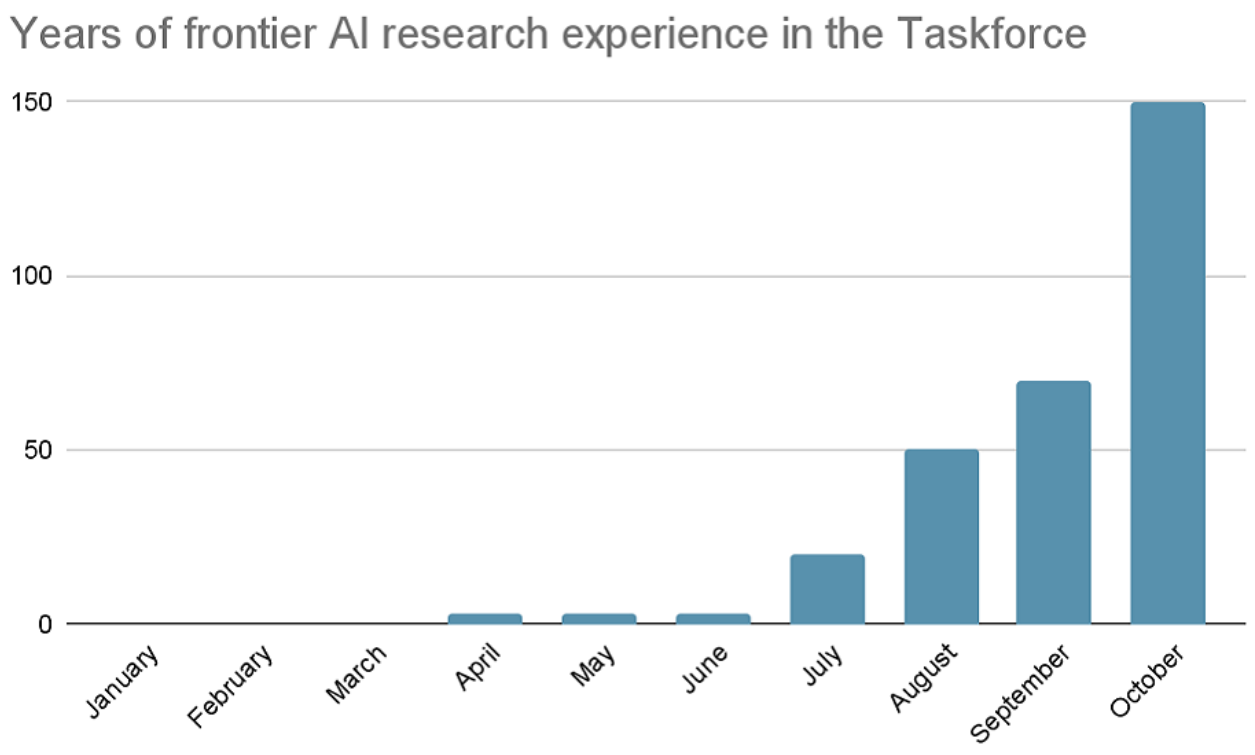
Effective start-ups send regular updates to investors. Since our first progress report (https://www.gov.uk/government/publications/frontier-ai-taskforce-first-progress-report/frontier-ai-taskforce-first-progress-report), published in early September, we've continued to make significant progress in building state capacity at the frontier of AI. This is our second progress report.

Building an AI research team inside government

The hardest challenge we have faced in building the Taskforce is persuading leading AI researchers to join the government. Compensation for machine learning researchers has ballooned in recent years. Beyond money, the prestige and learning opportunities from working at leading AI organisations are a huge draw for researchers.

We can't compete on compensation, but we can compete on mission. We are building the first team inside a G7 government that can evaluate the risks of frontier AI models. This is a crucial step towards meaningful accountability and governance of frontier AI companies, informed by the science and motivated by the public interest.

In our first progress report we said we would hold ourselves accountable for progress by tracking the total years of experience in our research team. When I arrived as Chair in June, there was just one frontier AI researcher employed full time by the department with 3 years of experience. We managed to increase that to 50 years of experience in our first 11 weeks of operation. Today that has grown to 150 years of collective experience in frontier AI research.



Our research team and partners have published hundreds of papers at top conferences. Some of their recent publications span:

- algorithms for AI systems to search and improve (https://arxiv.org/abs/2310.13032)
- how large language models can fail to generalise semantics (https://owainevans.github.io/reversal_curse.pdf)
- publicly sourced constitutions for AI alignment (https://cip.org/blog/ccai)

We are delighted to welcome **Jade Leung** to our team. Jade joins us from OpenAI where she led the firm's work on AGI governance, with a particular focus on frontier AI regulation, international governance, and safety protocols for upcoming generations of AI systems. If you reach the end of this blog with interest and energy to spare, read Jade's PhD thesis (https://ora.ox.ac.uk/objects/uuid:ea3c7cb8-2464-45f1-a47c-c7b568f27665/download_file?file_format=pdf&safe_filename=JADE+LEUNG+-+DPHIL+THESIS+-+Sep19.pdf&type_of_work=Thesis) on approaching the hardest problems in AI governance.

Furthermore, we are excited to welcome **Rumman Chowdhury**, who will be working with the Taskforce to develop its work on safety infrastructure, as well as its work on evaluating societal impacts from AI. Rumman is the CEO and co-founder of Humane Intelligence, and led efforts for the largest generative AI public red-teaming event at DEFCON this year. She is also a Responsible AI fellow at the Harvard Berkman Klein Center, and previously led the META (ML Ethics, Transparency, and Accountability) team at Twitter.

> We are continuing to scale up our research and engineering team. Please consider applying to the EOI form (https://docs.google.com/forms/d/1HKVnrV_rBHF3w4StWUs0XNhxTtKkTHs5CpMnH6PM0pQ/viewform?edit_requested=true) or via these job postings for Senior Research Engineers (https://www.linkedin.com/authwall?trk=bf&trkInfo=AQF3x16JvjEiZAAAAYtwT2uYugw1rn2iSL7K2FV-9mH01L2CIMHWzZ_6BxqngJgmKUo3Jq_Kt3eKworXD1mg_TT-mm2BtUdNI-OKfn0BBkkb3oj3SrAAr433DeD22xAP-SnyH4I=&original_referer=&sessionRedirect=https%3A%2F%2Fwww.linkedin.com%2Fjobs%2Fview%2F3749739500) and Senior Software Engineers (https://www.linkedin.com/jobs/view/3749671437).

Partnering with leading technical organisations

Leading on AI safety does not mean starting from scratch or working alone – we are building on and supporting the work conducted by a range of cutting-edge organisations. In our first progress report and recent update on GOV.UK we announced partnerships with ARC Evals, Advai, The Centre for AI Safety, Collective Intelligence Project, Faculty, Gryphon Scientific, RAND, Redwood Research and Trail of Bits. By partnering with leading organisations we are helping to coordinate research on frontier AI safety and feed it into the government's efforts.

Today we are announcing that the Taskforce has entered new partnerships with:

Apollo Research is an AI safety organisation that works with large language models to understand their behaviour and interpret their inner workings. They aim to evaluate models' high-risk failure modes, such as deceptive alignment. We have been working with Apollo to better understand the risks associated with potential loss of human control over AI systems.

OpenMined is a global non-profit building open source AI governance infrastructure - helping the AI ecosystem manage access to its data, compute, talent, and models. We are working with OpenMined to develop and deploy technical infrastructure that will facilitate AI safety research across governments and AI research organisations

Building the foundations for AI safety research

In June this year, several large AI companies committed to giving the UK government, via the Frontier AI Taskforce, early and deeper access to their models. Since then, we have been working in collaboration with these leading AI companies to secure this world first model access.

But model access is only one part of the picture. For too long researchers in industry have had access to much greater computing resources than those in academia and the public-sector, creating a so-called 'compute divide'. Having the compute infrastructure to conduct cutting-edge research is pivotal for building state capacity in AI safety - for example being able to run large-scale interpretability experiments.

To tackle this over the last months, the Taskforce has supported DSIT and the University of Bristol to help launch major investments in compute. The University of Bristol will soon host the first component of the UK's AI Research Resource, Isambard-AI (https://www.gov.uk/government/news/bristol-set-to-host-uks-most-powerful-supercomputer-to-turbocharge-ai-innovation), which will itself be one of the most powerful supercomputers in Europe when built. It will have thousands of state-of-the-art GPUs and will vastly increase our public-sector AI compute capacity. There is more to do, but these great strides fundamentally change the kind of projects researchers can take on inside the Taskforce.

Preparing for the first global summit on AI Safety

Since Prime Minister Rishi Sunak announced the UK would host the world's first global AI Safety Summit, the Taskforce and Summit teams in the Department for Science, Innovation and Technology have visited and engaged at Ministerial and/or senior level with over 30 countries across the world. This has driven a global conversation on AI safety ahead of the Summit's launch in just a week's time.

Our growing research team has been working at pace with our partners to develop a cutting-edge research programme; the initial results of which will be showcased on Day 1 of the Summit.

Here, our team will present 10-minute demonstrations, focused on 4 key areas of risk:

- misuse
- societal harm

- loss of human control
- unpredictable progress

We believe these demonstrations will be the most compelling and nuanced presentations of frontier AI risks done by any government to date. Our hope is that these demonstrations will raise awareness of frontier AI risk and the need for coordinated action before new - more capable - systems are developed and deployed.

To the emerging network

AI is a general purpose and dual-use technology. We need a clear-eyed commitment to empirically understanding and mitigating the risks of AI so we can enjoy the benefits. In 1955 Von Neumann, pioneer of nuclear weapons and computing wrote a crisp and prescient essay “Can we survive technology?” in which he considered the implications of accelerating technological progress. His final conclusion rings true today:

> “ Any attempt to find automatically safe channels for the present explosive variety of progress must lead to frustration. The only safety possible is relative, and it lies in an intelligent exercise of day-to-day judgement”

We must do our best and contribute to that empirical foundation for day-to-day judgement.

One thing that struck me on a recent trip to the Bay Area to meet AI companies and researchers was the number of people who, unbeknown to us, have been rooting for the Taskforce behind the scenes. There have been AI researchers telling their network to join and fighting for us in private. There have been people spending their goodwill with company leaders to promote the importance of our work. To the emerging international network of people who take this technology seriously - we are very grateful for your support, and can’t wait to tell you more about what we’ve been working towards in just a few days time at the AI Safety Summit.

A home for AI Safety research

Last week the Prime Minister announced (https://www.gov.uk/government/speeches/prime-ministers-speech-on-ai-26-october-2023) that he is putting the UK’s work on AI safety on a longer term basis by creating an AI Safety Institute in which our work will continue. The AI Safety Institute is the first state-backed organisation focused on frontier AI safety for

the public interest. Its mission is to minimise surprise to the UK and humanity from rapid and unexpected advances in AI, and will work towards this by developing the sociotechnical infrastructure needed to understand the risks of advanced AI and support its governance. AI has the power to revolutionise industries, enhance our lives and address complex global challenges, but we must also confront the global risks. The future of AI is safe AI.

↑ Back to top