GOV.UK

Department for Science, Innovation & Technology

Independent report

Frontier AI Taskforce: first progress report

Published 7 September 2023

Contents

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

The Taskforce is a start-up inside government, delivering on the ambitious mission given to us by the Prime Minister: to build an AI research team that can evaluate risk at the frontier of AI. As AI systems become more capable they may significantly augment risks. An AI system that advances towards human ability at writing software could increase cybersecurity threats. An AI system that becomes more capable at modelling biology could escalate biosecurity threats. To manage this risk technical evaluations are critical - and these need to be developed by a neutral third party - otherwise we risk AI companies marking their own homework.

Given these potentially significant frontier risks, as of today, the Taskforce is being renamed to the Frontier AI Taskforce.

This is the Frontier AI Taskforce’s first progress report.

1. We have established an expert advisory board spanning AI Research and National Security

Given that a number of risks from frontier systems touch areas of national security, we have established an expert advisory board that bridges some of the

world's leading experts in AI research and safety as well as key figures from the UK's national security community. Our initial advisory board members are:

Yoshua Bengio. Yoshua is most known for his pioneering work in deep learning, earning him the 2018 A.M. Turing Award, "the Nobel Prize of Computing," with Geoffrey Hinton and Yann LeCun. He is a Full Professor at Université de Montréal, and the Founder and Scientific Director of Mila – Quebec AI Institute.

Paul Christiano. Paul is one of the leading researchers in the field of AI Alignment. He is co-founder of ARC, the Alignment Research Centre and previously ran the language model alignment team at OpenAI.

Matt Collins. Matt is the UK's Deputy National Security Adviser for Intelligence, Defence and Security. He has been in the civil service for 18 years, in a variety of roles covering national security, online harms and crime reduction.

Anne Keast-Butler. Anne is the director of GCHQ. Anne has an impressive track record at the heart of the UK's national security network, helping to counter threats posed by terrorists, cyber-criminals and malign foreign powers.

Alex van Someren. Alex is the UK's Chief Scientific Adviser for National Security. Alex was previously a venture capital investor and entrepreneur, focusing on investing in early stage 'deep technology' startups.

Helen Stokes-Lampard. Beyond national security and AI research expertise we are also excited to build an advisory board that can speak to critical uses of frontier AI on the frontlines of society. Helen is not only a practising General Practitioner observing how conversational AI tools can impact day to day medical diagnoses but also an incredibly experienced leader across the UK's medical community, having Chaired the Royal College of General Practitioners and currently Chair of the Academy of Medical Royal Colleges.

Matt Clifford. Matt is the Prime Minister's joint Representative for the AI Safety Summit, Chair of ARIA and co-founder of Entrepreneur First. His appointment as Expert Advisory Board Vice Chair demonstrates the level of coordination across UK initiatives around frontier AI - including the Taskforce and the AI Safety Summit.

We will announce other members of our advisory board in due course.

2. Recruitment of expert AI researchers

Eric Schmidt, former CEO of Google, recently captured the state of state capacity in frontier AI when he stated in May 2023: "The technology is

enormously complicated…there's no one in the government who could get it right[footnote 1]."

This cannot be true if society is to successfully navigate this rate of progress in AI.

Sam Altman, CEO of OpenAI, recently suggested that the public sector had a "lack of will" to lead on innovation, asking "Why don't you ask the government why they aren't doing these things, isn't that the horrible part?[footnote 2]"

We have an abundance of will to transform state capacity at the frontier of AI Safety. This is why we are hiring technical AI experts into government at start-up speed.

We are drawing on world-leading expertise:

Yarin Gal will join as Research Director of the Taskforce from Oxford where he is head of the Oxford Applied and Theoretical Machine Learning Group. Yarin is a globally recognised leader in Machine Learning, and will retain his position as Associate Professor at Oxford.

David Krueger will be working with the Taskforce as it scopes its research programme in the run up to the summit. David is an Assistant Professor at the University of Cambridge's Computational and Biological Learning lab, where he leads a research group focused on Deep Learning and AI Alignment.

The Taskforce is housed inside the UK's Department for Science Innovation and Technology (DSIT) which employs roughly 1500 civil servants. When I arrived in June (https://twitter.com/soundboy/status/1670343527723679744) there was just one frontier AI researcher employed by the department with 3 years of experience in frontier AI.

This lone researcher was Nitarshan Rajkumar who put his PhD on pause to join DSIT in April, and is a testament to what an earnest, incredibly hardworking technical expert can accomplish when they commit to public service. Michelle Donelan, Secretary of State for DSIT recruited Nitarshan and he has materially influenced many of the bold efforts that the UK has been making to invest in frontier AI Safety. We need more Nitarshans!

Thanks to a huge push by the Taskforce team we now have a growing team of AI researchers with over **50 years of collective experience** at the frontier of AI. If this is our metric for state capacity in frontier AI, we have managed to increase it by an order-of-magnitude in just 11 weeks. Our team now includes researchers with experience from DeepMind, Microsoft, Redwood Research, The Center for AI Safety and the Center for Human Compatible AI.

These are some of the hardest people to hire in the world given the huge wave of attention and money being thrown at AI right now. They have chosen to come into public service not because it's easy, but because it offers the opportunity to

fundamentally alter society's approach to tackling risks at the frontier of AI. These researchers and engineers will bring their skills towards giving the government the capability to work directly on frontier AI models and evaluate their risks – through model evaluations, red-teaming, and other aspects of safety infrastructure. I'm immensely grateful to these people for stepping up and performing this public service at a critical moment in the development of AI.

This is not to diminish the incredible broader team at DSIT but part of an active effort supercharged (https://www.ft.com/content/0e103268-f641-4be4-ba5c-91dcf134ac01) by our Secretary of State to expand the diversity of people inside government to include more AI researchers, so we can together tackle this challenge.

We are rapidly expanding this team and are looking for researchers with an interest in catalysing state capacity in AI Safety. **We plan to scale up the team by another order of magnitude**. Please consider applying to join us here (https://docs.google.com/forms/d/1HKVnrV_rBHF3w4StWUs0XNhxTtKkTHs5CpMnH6PM0pQ/edit). With the first ever AI Safety Summit in the UK on 1 and 2 November, this is a critical moment to influence AI Safety. We are particularly focused on AI researchers with an interest in technical risk assessments of frontier models.

3. 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. We are excited to announce our initial set of partnerships with:

ARC Evals is a non-profit that works on assessing catastrophic risks from frontier AI systems, and have previously worked with OpenAI and Anthropic on evaluating the "autonomous replication and adaptation" capabilities of their systems before release. We'll be working closely with the ARC Evals team to assess risks just beyond the frontier in the lead up to the UK's AI Safety Summit. We'll also be engaging with the team at **Redwood Research**, and with Jeff Alstott, Christopher Mouton and their team at non-profit **RAND** in driving forward this agenda.

Trail of Bits is a leading cybersecurity research and consulting firm that has helped secure some of the world's most targeted organisations. We are kicking off a deep collaboration to understand risks at the intersection of cybersecurity and frontier AI systems. This work will be led by Heidy Khlaaf (https://twitter.com/heidykhlaaf?lang=en), who specialises in software evaluation, specification, and verification for safety-critical systems, and who also led the safety evaluation of Codex while at OpenAI.

The Collective Intelligence Project is a non-profit that incubates new governance models for transformative technology, with a mission to direct technological development towards the collective good. Co-founders Divya Siddarth and Saffron Huang will join us on secondment to help us develop a range of social evaluations for frontier models.

The Center for AI Safety is a non-profit that works to reduce societal-scale risks from AI, through fundamental safety research, research infrastructure, and technical expertise to support policymakers. We'll be working with Dan Hendrycks and his team in the lead up to the summit to interface with and enable the broader scientific community.

If you think your organisation can step up and join these teams and contribute to our efforts, please let us know here (https://docs.google.com/forms/d/1HKVnrV_rBHF3w4StWUs0XNhxTtKkTHs5CpMnH6PM0pQ/edit).

4. Building the technical foundations for AI research inside government

A core goal of the Taskforce is to give AI researchers inside the government the same resources to work on AI Safety that they would find at leading companies like Anthropic, DeepMind, or OpenAI. As the Prime Minister announced, these companies have already committed to giving us deep model access so that researchers in the Taskforce are not constrained in their ability to work on model evaluations. We're also working in close collaboration with No10 Data Science ('10DS') so that our researchers and engineers have the compute infrastructure they need to hit the ground running, for model fine-tuning, interpretability research, and more.

5. Moving fast matters

Getting this much done in 11 weeks in government from a standing start has taken a forceful effort from an incredible team of dedicated and brilliant civil servants. Building that team has been as important as the technical team mentioned above.

This is why we've brought in Ollie Ilott as the Director for the Taskforce. Ollie joins us from Downing Street, where he led the Prime Minister's domestic private office, in a critical role known as "Deputy Principal Private Secretary" (I'm still learning how to speak 'civil service'). He is known 'across the piece' for his ability to recruit and shape best-in-class teams. Before joining the Prime

Minister's office, Ollie ran the Cabinet Office's COVID strategy team in the first year of the pandemic and led teams involved in Brexit negotiations and passing Brexit legislation. Ollie's leadership of the talented civil service team will be critical to making the Frontier AI Taskforce a success. We are grateful to the Prime Minister for releasing such a key member of the team at No.10.

I'm so grateful for the 'General Groves' energy that Ollie embodies, and the determination and enthusiasm so many in the Taskforce, No.10 and the Department of Science, Innovation and Technology have brought to this challenge.

We are moving fast but time is of the essence. The UK is hosting the first ever AI Safety Summit (https://www.gov.uk/government/news/uk-government-sets-out-ai-safety-summit-ambitions) on 1 and November, just 8 weeks away. We need more technical experts and more leading technical organisations to come and support us. The evaluations for frontier models we develop will help set a standard for the world and ensure we can safely capture AI's potential benefits. Please apply here (https://docs.google.com/forms/d/1HKVnrV_rBHF3w4StWUs0XNhxTtKkTHs5CpMnH6PM0pQ/edit)!

Background

The Taskforce is a start-up that will move fast, but it plays by the same rules as the rest of government. The Taskforce is functionally a part of the Department for Science, Innovation and Technology (DSIT), with the Permanent Secretary serving as the Accounting Officer and DSIT Ministers accountable to Parliament. The Taskforce's expenditure will therefore be accounted for in the usual DSIT annual report and accounts and subject to the same HM Treasury (HMT) controls as other departmental expenditure. All appointments will comply with the normal DSIT conflicts of interest policy and will be subject to the standard business appointment rules when their term comes to an end. Ian Hogarth, as Taskforce chair, reports jointly to the Prime Minister and to the DSIT Secretary of State.

---

1. https://www.nbcnews.com/politics/congress/washington-struggling-catch-artificial-intelligence-rcna84489 (https://www.nbcnews.com/politics/congress/washington-struggling-catch-artificial-intelligence-rcna84489)
2. https://www.ft.com/content/1d1cb2b3-391c-4dc8-ba5b-fedd379b7fb0 (https://www.ft.com/content/1d1cb2b3-391c-4dc8-ba5b-fedd379b7fb0)

↑ Back to top