FLI recommendations for the UK Global AI Safety Summit

Bletchley Park, 1-2 November 2023

Mark Brakel, Director of Policy
policy@futureoflife.org

"The time for saying that this is just pure research has long since passed [...] It's in no country's interest for any country to develop and release AI systems we cannot control. Insisting on sensible precautions is not anti-industry.

Chernobyl destroyed lives, but it also decimated the global nuclear industry. I'm an AI researcher. I do not want my field of research destroyed. Humanity has much to gain from AI, but also everything to lose."

Professor Stuart Russell
Founder of the Center for Human-Compatible AI at the University of California, Berkeley

Contents

The Future of Life Institute (FLI) works to promote the benefits of technology and reduce their associated risks. FLI has become one of the world's leading voices on the governance of artificial intelligence (AI) and created one of the earliest and most influential sets of governance principles, the Asilomar AI Principles. FLI maintains a large network among the world's top AI researchers in academia, civil society, and private industry.

In March of this year, we released an open letter that sparked a global debate on AI safety.

Introduction

Prime Minister Sunak,
Secretary of State Donelan,

The Future of Life Institute (FLI) is an independent non-profit organisation that works on reducing global catastrophic and existential risks from powerful technologies. Back in 2017, FLI organised a conference in Asilomar, California to formulate one of the earliest artificial intelligence (AI) governance instruments: the "Asilomar AI principles." The organisation has since become one of the leading voices on AI policy in Washington D.C. and Brussels, and is now the civil society champion for AI recommendations in the United Nations Secretary General's Digital Cooperation Roadmap.

In March, FLI - joined by over 30,000 leading AI researchers, professors, CEOs, engineers, and others - called for a pause of at least six months on the largest and riskiest AI experiments, to reduce the likelihood of catastrophic accidents. The letter sparked United States Senate hearings, a formal reply from the European Parliament, and a call from UNESCO to implement a global ethical framework for AI.

Despite this shift in the public conversation, we remain locked in a race that has only accelerated. No company has developed the shared safety protocols that we believe are necessary. In our letter, we also wrote: "*if such a pause cannot be enacted quickly, governments should step in*". The need for public sector involvement has never been clearer. As a result, we would like to thank you for your personal leadership in convening the world's first AI safety summit.

In our view, the Summit should achieve three things:

1. Establish a common understanding of the severity and urgency of AI risks;
2. Make the global nature of the AI challenge explicit, recognising that all of humanity has a stake in this issue and that some solutions require a unified global response, and;
3. Embrace the need for urgent government intervention, including hard law where appropriate.

With this document, we offer a draft outcome declaration, a number of recommendations to participating governments, and a roadmap for post-summit work. We imagine that summit preparations are well under way, but hope that this document can provide further inspiration as to what themes should be covered during the preparatory meetings and at the summit itself. Ultimately, we hope that the summit can kickstart the development of a new international architecture for AI regulation.

We wish you good luck with the preparations for the summit and stand ready to offer our expertise in support of effective global AI governance.

Sincerely,

Professor Max Tegmark
President

Professor Anthony Aguirre
Executive Director

Proposed Declaration on AI Safety

A. Increasingly powerful AI systems pose risks, through accidents, misuse, or structural problems, with potentially catastrophic consequences. The mitigation of these emerging risks must become a global priority.

B. The robust mitigation of AI risks demands leadership from the public sector. Advanced AI systems should, like other potentially dangerous technologies, be carefully regulated to ensure compliance with adequate safety measures.

C. Neither systems nor the risks they pose can be contained within the borders of one nation state. Adequate governance of advanced AI necessitates ongoing and intensive global coordination.

The participating nations in the world's first global AI safety summit agree to:

1. Reconvene in six months, and every six months thereafter, to accelerate the creation of a robust AI governance regime;
2. Increase public funding for AI safety research to improve understanding of the risks and reduce the probability of accidents;
3. Develop national AI safety strategies consisting of the following measures:

   i. Standards for advanced AI, plus associated benchmarks and thresholds for dangerous capabilities,

   ii. Mandatory pre-deployment audits for potentially dangerous AI systems by independent third parties,

   iii. Monitoring of entities with large-scale AI compute concentrations,

   iv. Safety protocols to prevent systems with dangerous capabilities from being developed, deployed or stolen,

   v. Restrictions on open source AI based on capability thresholds to prevent the proliferation of powerful models amongst malicious actors,

   vi. Immediate enhancement of cybersecurity standards at leading AI companies,

   vii. Adaptation of national liability law to AI-specific challenges;

4. Establish a post-summit working group with a mandate to develop a blueprint for a new global agency that can coordinate the governance of advanced AI, advise on safety standards, and ensure global adherence;
5. Encourage leading companies to share information with the UK Foundation Model task force and welcome the UK's intention to put this entity at the disposal of the international community.

Recommendations in advance of the Summit

Recommendations for all participating governments:

- Do not let the perceived complexity of AI technology stand in the way of action. Many useful risk-mitigation measures are technology-neutral.
- Critically assess whether existing national AI strategies are sufficiently attuned to AI safety risk.

Recommendations for the UK hosts:

- Involve all major AI powers in the summit, including Brazil, China, the EU, India, Japan and the US. Instruct UK embassies in key capitals to convene informational sessions about AI safety.
- Ensure the summit takes a truly global perspective and avoid overly relying on the transatlantic relationship. Look beyond the US voluntary commitments for inspiration.
- Carefully balance private sector perspectives with those of independent experts from academia and civil society.
- Ask any private sector attendees to submit their safety plans ahead of time and make these documents available to other governments for scrutiny.
- Exclude autonomous weapons systems from the agenda. A treaty process is already emerging at other fora and the inclusion of too many issues could undermine progress on civilian AI safety.
- Amplify the UK's global leadership by bringing forward domestic AI legislation, recognising that societal-scale risks of AI cannot be managed by sector or through voluntary guidelines.

Recommendations for the People's Republic of China:

- Inform other governments about the recently enacted Interim Measures for the Management of Generative Artificial Intelligence Services.
- Engage in an open dialogue with the US, despite ongoing economic and strategic competition. Global threats from advanced AI, much like climate change, urgently demand cooperation even if this is not possible on most bilateral issues.

Recommendations for the European Union and its Member States:

- Inform other governments about the proposed EU AI Act and any key insights that have emerged during the drafting process, with a particular focus on the regime for more general AI systems.
- Where the AI Act falls short in mitigating risks identified at the summit, keep an open mind to adapting the overall EU regulatory framework.
- *[For the Spanish EU Presidency]* Inspire other governments by sharing information about the structure of the Spanish Agency for the Supervision of Artificial Intelligence (AESIA), the first specialized regulatory entity for AI in the EU.

Recommendations for the United States:

- Closely monitor compliance of key AI corporations with the recent voluntary commitments and apply appropriate pressure to ensure compliance.
- Do not leave the global regulatory conversation to Brazil, China and Europe; fast-track binding legislation.
- Engage in an open dialogue with China, despite ongoing economic and strategic competition. Global threats from advanced AI, much like climate change, urgently demand cooperation even if this is not possible on most bilateral issues.

Recommendations for the Summit programme

The UK government has set out five strong ambitions for the AI Safety Summit. Given how unfamiliar many government officials are with AI Safety, we would recommend that a final programme ensures all participants develop a shared understanding of the risks that we face. To this end, we would suggest involving independent experts to clearly articulate what risks the international community needs to address.

Existing large-scale harms

As the UK government frames the conversation, it may want to consider highlighting recent examples of large-scale harms caused by AI. The Australian Robodebt scheme and the Dutch childcare benefit scandal, for example, have shown how simple algorithms can already disrupt societies today.

Proposed speakers: Minister Alexandra van Huffelen (The Netherlands) and Royal Commissioner Catherine Holmes (Australia)

Catastrophic risks from accidents

A survey of 738 leading AI scientists found, in aggregate, that researchers believe that there is a 50% chance that we will develop systems surpassing human abilities in all domains before 2060. Currently, no robust mechanism exists to ensure that humans will stay in control of these incredibly powerful systems. Neither do we understand how to accurately align the objectives they pursue with our own. This session would lay out the risks from out-of-control AI systems.

Proposed speaker: Stuart Russell (University of California, Berkeley)

Catastrophic risks from misuse and proliferation

Through cybertheft or voluntary open-sourcing, very powerful AI systems can end up in the hands of malicious actors and be used to cause significant harm to the public. Once the compute-intensive training phase has been completed, consumer hardware can be sufficient to fine-tune AI models for destructive behaviour (e.g. automated cyberattacks that disable critical infrastructure or the creation of pathogens that cause catastrophic pandemics).

Proposed speaker: Professor Yoshua Bengio (University of Montreal)

Post-summit roadmap

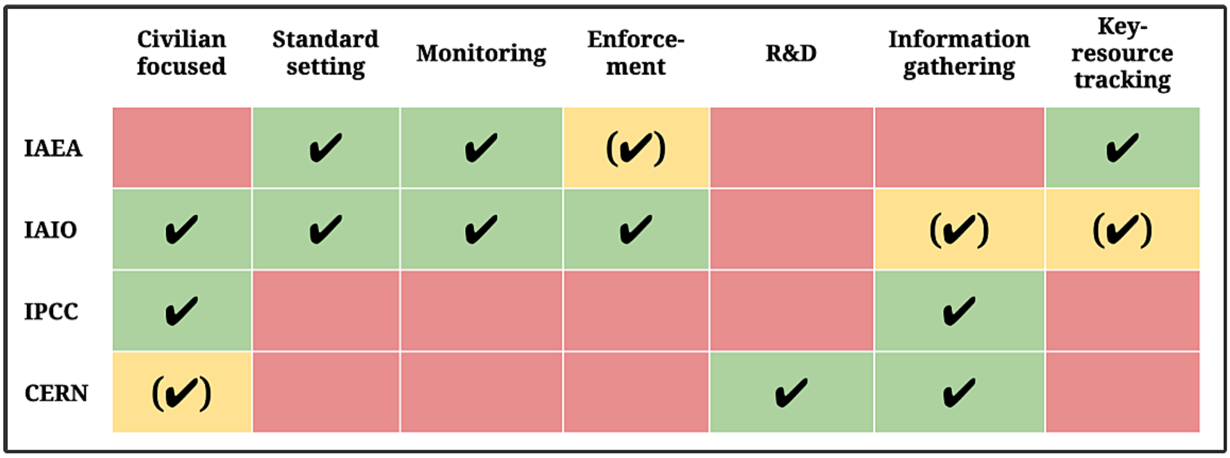
Building on initial agreements at fora like the G7, the Bletchley Summit should be the start of a process, rather than a one-off event. Ahead of a successor Summit, for which FLI would suggest May 2024, we would suggest the following roadmap.

For the post-summit working group:

The proposed working group would have a mandate to develop the blueprint for a new global agency that can coordinate the governance of advanced AI, advise on safety standards, and ensure global adherence.

Functions of the agency (or associated entities) would need to include i) risk identification, ii) promoting agreement on governance standards, such as thresholds that risky capabilities ought not to exceed, and iii) assistance with implementation and enforcement.

No perfect template exists for dealing with the challenges that AI will bring. In a recent working paper, Trager et al. look at the features of relevant analogous institutions, show which relevant functions they fulfil, and propose a design for an International AI Organisation (IAIO):



	Civilian focused	Standard setting	Monitoring	Enforce-ment	R&D	Information gathering	Key-resource tracking
IAEA		✔	✔	(✔)			✔
IAIO	✔	✔	✔	✔		(✔)	(✔)
IPCC	✔					✔	
CERN	(✔)				✔	✔	

Note: Green indicates that the model fulfills this function; red indicates that it does not. Yellow means that there is some ambiguity; for instance, the IAEA only refers violations to the Security Council which then potentially takes action, a process that could be counted as enforcement. Similarly, tracking of key AI inputs could be part of the IAIO model but is optional. In the case of CERN, despite its civilian focus, the research could be classified as dual-use to a degree. These institutions were chosen for comparison because they represent commonly discussed models for international AI governance.[4] The IAIO is based on the ICAO, IMO, and FATF models, and thus these are not listed because they share similar characteristics.

Given the exponential growth in AI capabilities and the corresponding urgency of mitigating risk, the blueprint should be ready for discussion at the next summit. As with other international organisations, FLI recommends that the UK hosts act as a temporary secretariat in developing the agency until such a time when the agency can itself support national governments.

At national level:

Following the summit, governments should revise their national AI strategies. Whereas these strategies[1] previously focused almost exclusively on economic competitiveness, recalibration is required to account for AI safety risks.

[1] See the OECD AI Policy Observatory for an overview.

Firstly, governments need to establish safety standards for the responsible design, development, and deployment of powerful AI systems. These standards should regularly be updated as technology progresses, and include:

1. Comprehensive pre-deployment risk assessments informed by internal and independent third party model audits. These audits should test for dangerous capabilities, controllability, and ethical alignment.
2. Standardised protocols for permissible deployment options for AI systems. These should range from fully open sourcing a model to not deploying it at all. If a system fails to pass an audit successfully, deployment should be prohibited.
3. Post-deployment monitoring requirements that can trigger 1) repeated risk assessments if post-deployment enhancement techniques significantly alter system capabilities and 2) immediate termination of model deployment if unacceptably dangerous behaviour is detected.
4. Categories of AI capabilities, such as automated cyberattacks or fraud, that should be restricted to prevent large harms to public safety.
5. Strong measures to prevent and track model leaks. These should include robust cybersecurity standards as well as safeguards against threats from outside the relevant companies.

Alongside standards, robust **enforcement mechanisms** should be enshrined in national legislation to ensure leading AI corporations comply with appropriate safety standards. To enable adequate enforcement, governments should **create national AI agencies** with the authority to initiate enforcement action. National authorities also have a role in mandating private, third-party actors to **audit the most capable AI systems** and to put arrangements in place that minimise conflicts of interest. Moreover, the adaptation of **national liability law** to AI can help dissuade corporate leaders from taking excessive risks.

Governments should also improve their institutional understanding of the key risks. On the one hand, and especially in countries with leading AI corporations, **mandatory information-sharing regimes** should be put in place. Continual information-sharing will provide governments with insight into development processes, compute usage, and model capabilities and grant governments early access to models for testing purposes. Furthermore, a global AI incident database should be created to monitor recorded harms.

On the other hand, all governments should **expand academic research on AI safety**. Additional funding needs both to increase the number of scientists working on the problem and to expand the computational resources available to safety researchers. This will empower publicly-funded academics to conduct the type of safety research (on large scale models) that has recently become the exclusive preserve of the private sector. The establishment of an international research institution for AI safety similar to CERN should be seriously considered.

Finally, governments should establish **hardware governance regimes**. Giant data centers with several thousand cutting-edge AI chips are needed to develop the most capable systems. This physical infrastructure marks the most amenable bottleneck for government intervention. In a first step, large domestic AI compute concentrations and their highly centralised global supply chains need to be mapped. Additionally, reporting requirements for large training runs should be instantiated for monitoring purposes. To substantially reduce the risk of catastrophic accidents, licensing regimes for large-scale training runs must be developed to ensure requesting entities can demonstrate compliance with the required safety precautions.

Recommended Experts

Tackling these new challenges will require governments to build considerable expertise. Below is a list of suggested experts to involve in preparatory meetings for the summit at expert level, and in eventual post-summit working groups.

INTERNATIONAL INSTITUTIONS FOR THE GOVERNANCE OF ADVANCED AI

- Professor Robert Trager (University of California, Los Angeles),
- Professor Duncan Snidal (University of Oxford),
- Dr. Allan Dafoe (Google DeepMind),
- Mustafa Suleyman (Inflection AI)

AUDITING REGIMES FOR HIGH-RISK AI SYSTEMS

- Professor Ellen P. Goodman (Rutgers Law School),
- Dr. Paul Christiano (Alignment Research Center),
- Markus Anderljung (Center for the Governance of AI),
- Elizabeth Barnes (ARC Evals)

HARDWARE GOVERNANCE

- Professor Anthony Aguirre (University of California, Santa Cruz),
- Dr. Jess Whittlestone (Centre for Long-Term Resilience),
- Lennart Heim (Center for the Governance of AI),
- Dr. Shahar Avin (Centre for the Study of Existential Risk, University of Cambridge)