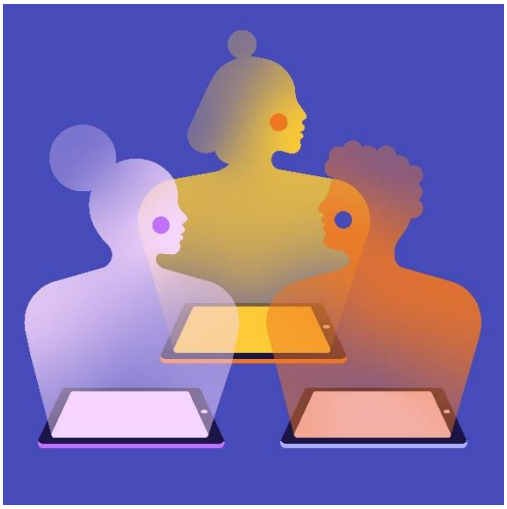

A Democratic Approach to Global Artificial Intelligence (AI) Safety

Policy Brief

November 2023

Alex Read

This policy brief discusses where risks to democracy from AI are emerging, what a democratic response to AI governance and safety looks like and the role of parliaments worldwide in enabling this response. It outlines how the democratic governance community can help plot a course of action to ensure that democracy is protected in the face of rapid AI advancements.

The target audience of this brief are Members of Parliament, parliamentary leadership and staff, international development practitioners, civil society organisations and participants at the UK AI Safety Summit 2023.

Contents

Rights, acknowledgements, and disclaimer

The publication "A Democratic Approach to Global Artificial Intelligence (AI) Safety" has been written by Alex Read, WFD Associate Expert. It was published in November 2023. The author appreciates the peer review comments received from WFD senior staff: Anthony Smith (CEO), Graeme Ramshaw, Franklin De Vrieze, Tanja Holstein, Alex Scales, Szelim Simandi, Stephanie Le Lievre, Chris Lane and WFD Associate Expert Ben Graham.

The views expressed in the paper are those of the author, and not necessarily those of or endorsed by WFD, the institutions mentioned in the paper, nor the UK Government.

Foreword

This paper was first published in advance of the AI Safety Summit on 1-2 November, 2023. We strongly welcome key outcomes of the summit, including the consensus achieved with the Bletchley Declaration, agreements on AI safety testing and the establishment of an AI Safety Institute. The number of countries attending indicates a strong scope for agreement on the importance of a global approach to AI safety. In advancing a democratic approach to global AI safety, we would like to add the following reflections:

First, the Bletchley Declaration, has advanced work to address the opportunities of AI and the threats it poses. We noted in particular the call in the declaration for countries to develop “a pro-innovation and proportionate governance and regulatory approach that maximises the benefits and takes into account the risks associated with AI” and we also welcome support for “development-orientated approaches and policies that could help developing countries strengthen AI capacity building and leverage the enabling role of AI to support sustainable growth and address the development gap.”

However, meeting these priorities will be a significant challenge for many countries, which will require dedicated international assistance. We call for support in particular to focus on the critical role of parliaments in ensuring democratic oversight over AI governance, policy and regulation.

We welcome the need for “human-centric, trustworthy, and responsible AI” and the emphasis on the “protection of human rights, transparency and explainability, fairness, accountability, regulation, safety, appropriate human oversight, ethics, bias mitigation, privacy and data protection”, all of which can help address risks to democracy from AI. In achieving this vision for societies with advanced AI, we call for an emphasis on wide-ranging inclusive public participation to help foster public trust and confidence in AI. Parliaments will have a key role in voicing public concerns around societal risks and in supporting a public dialogue on societies with transformative AI.

Our second reflection is that there is more to do to ensure that a democracy lens is applied to the follow up. It is clear to us that more needs to be done to safeguard the integrity and security of our democratic systems against current and future threats from AI. While we agree that risks from AI are “inherently international in nature, and so are best addressed through international cooperation”, we believe that it is critically important that those committed to democratic governance in their countries and societies specifically to consider the opportunities coordinate around addressing threats to democracy as a result of the increasing use of AI. The spread of countries and organisations attending the summit was very valuable, and we also call for dedicated international mechanisms that can respond to threats to democracy from AI that cross borders.

In our view, and as set out in our paper, we need:

- A shared understanding of the opportunities and risks from AI for democratic systems;
- Accelerated work among democracies and those supporting democracy to address those opportunities and risks; and
- Development of international expertise to provide impartial, reliable and timely assessments about the progress and impact of AI and research focused on measures to protect democratic systems. This can potentially be achieved through the future work of the AI Safety Institute, which can be expected to report to the UK Parliament, as other independent institutions do.

We look forward to continuing the engagement with partners on these issues.

Anthony Smith,
CEO, Westminster Foundation for Democracy

Summary

As the United Kingdom (UK) holds the first-of-its-kind AI Safety Summit on 1-2 November 2023, the need for democratic responses to the transformative impact of AI grows stronger.

Artificial Intelligence (AI) is achieving breakthroughs in healthcare and education, helping mitigate climate challenges and contributing to global growth. It offers the potential to enhance democratic systems by improving access to government, streamlining decision-making and fostering new means of public participation.

However, as the technology advances and becomes more widely adopted, societal risks will grow. Threats such as AI-driven disinformation, increased surveillance, biased and discriminatory outcomes, and concentrations of power pose challenges to the security and stability of democratic structures and institutions.

The current and near-term risks from AI should compel democratic leaders to incorporate the safety of democratic systems in the discussion around AI safety. As well as discussing technical measures to progress towards safe AI, we need to focus on building political and societal resilience to the disruption that AI will bring.

The discussion on AI governance and safety must focus on core democratic values of transparency, accountability, public participation and inclusivity. To counter illiberal and repressive uses of AI, democracies will need to set a values-based example and demonstrate a coordinated approach.

The AI Safety Summit can provide an important step towards global agreement on AI safety that incorporates a broad array of risks from current uses of AI and frontier systems, including threats to democratic systems. The summit offers a valuable opportunity to:

- Launch an inclusive discussion to develop a shared understanding on the risks to democracy from AI.
- Establish a framework for future cooperation among democracies to proactively address AI-related risks and harness its benefits.
- Call for the international expertise necessary to provide impartial, reliable and timely assessments about the progress and impact of AI and research focused on measures to protect democratic systems.

AI is having an increasing impact worldwide …

AI development continues at a rapid pace

As benefits are realised across sectors and the economy, we also see **harms and risks emerging from AI:**

- **In society,** reinforcing bias and prejudice, enabling cybercrime, threatening jobs.
- **For democracies,** undermining the information environment and political discourse, damaging social cohesion, exacerbating inequality, compromising fundamental rights including privacy.
- **As 'frontier AI' is developed,** we risk misuse by malicious actors and loss of human control.

The UK Safety Summit comes at a key juncture…

The UK AI Safety Summit will address *risks from frontier AI and global approaches to manage risk so we can reap the benefits from AI.*

If AI erodes the fundamentals of democratic societies, we lose powers to control risks from frontier AI. AI safety should therefore also prioritise*: Safeguarding the integrity and security of our democratic systems against current and future threats from AI.*

A democratic vision for AI Governance and Safety involves:

- Establishing democratic oversight of AI safety
- Ensuring participation and inclusivity in AI governance
- Demonstrating democratic values at home and abroad

Parliaments will be central in advancing a democratic agenda for AI safety…

MPs can:

- Enact laws that advance AI safety and mitigate democratic risks
- Establish democratic oversight over AI deployment across society
- Provide the public with a voice in AI governance
- Help set global democratic norms on AI

Parliaments can:

- Provide MP professional development and staff training on AI
- Establish specialised parliamentary bodies addressing AI
- Build institutional connections to sources of information and expertise
- Include AI in parliamentary education and outreach initiatives
- Harness AI to enhance institutional effectiveness and accessibility

Democratic governance support has a key enabling role:

- Providing access to international expertise and resources
- Supporting international platforms to convene democratic representatives
- Developing public education and AI literacy programmes
- Investing in AI infrastructure
- Supporting global media and civil society

Taken together, we can realise a future where frontier AI serves society and democracy, and AI contributes to global prosperity.

Box 1: UK AI Safety Summit: The essentials

The UK AI Safety Summit will see global leaders and cabinet ministers, academics, civil society representatives, and heads of leading AI companies gather to discuss AI safety. The summit aims to develop a shared global understanding on the risks that may emerge from frontier AI and make progress towards global approaches to manage these risks.

The five objectives that will be discussed at the summit are:

1. A shared understanding of the risks posed by frontier AI and the need for action.
2. A process for international collaboration on frontier AI safety, including how best to support national and international frameworks.
3. Appropriate measures which individual organisations should take to increase frontier AI safety.
4. Areas for potential collaboration on AI safety research, including evaluating model capabilities and the development of new standards to support governance.
5. Showcase how ensuring the safe development of AI will enable AI to be used for good globally.

The focus is on frontier AI. Two risk categories that the AI Safety Summit will focus on are:

Misuse risks, for example where a bad actor is aided by new AI capabilities in biological or cyber-attacks, development of dangerous technologies, or critical system interference. Unchecked, this could create significant harm, including the loss of life.

Loss of control risks that could emerge from advanced systems that we would seek to be aligned with our values and intentions.

UK Government representatives have stressed that the summit does not aim to suggest specific forms of global regulation. However, Prime Minister Rishi Sunak has announced that the UK will set up the world's first AI Safety Institute, building on the work undertaken by the UK's Frontier AI Task Force.

It is expected that further events will follow from the summit, including the potential rotation of subsequent summits around different countries.

1. The increasing impact of AI

AI is a once-in-a-generation technology shift, potentially on par with electricity as a general-purpose tool that will transform lives across the globe. The UK Government recognises that AI will “fundamentally alter the way we live, work and relate to one another … [promising] to further transform nearly every aspect of our economy and society, bringing with it huge opportunities but also risks that could threaten global stability and undermine our values.”

Box 2: Key terms

AI is a complex term to pin down. A simple industry definition from IBM is of “any system capable of simulating human intelligence and thought processes”. The EU Artificial Intelligence Act defines it more broadly as "software that is developed with one or more of the techniques that can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with." John McCarthy, a pioneer of AI, states that the “the ultimate effort is to make computer programs that can solve problems and achieve goals in the world as well as humans”.

There is no universal definition of AI safety. The Center for Security and Emerging Technology define it as “an area of machine learning research that aims to identify causes of unintended behaviour in machine learning systems and develop tools to ensure these systems work safely and reliably”. Others define AI safety as more than a technical problem, focusing on the prevention and mitigation of various harms and potential risks from the deployment of AI in society.

Frontier AI is defined for the UK Safety Summit as "highly capable general-purpose AI models, most often foundation models, that can perform a wide variety of tasks and match or exceed the capabilities present in today’s most advanced models”.

Foundation models are built using vast amounts of unlabelled data. GPT-4, the Large Language Model (LLM) which underpins ChatGPT, is an example. They can be built on and tailored to specific uses or needs. They have been shown to produce remarkable productivity gains, performing a wide array of tasks, developing capabilities not envisaged at creation and outperforming task-specific AI models.

Generative AI creates new content such as text, images, video and audio. They use machine learning algorithms and statistical models to understand and model the patterns in data (such as digital pixels in photographs, waveforms in audio, or text), producing original outputs.

AI has developed rapidly due to the expansion of available training data, significant improvements in neural networks and growth in computational power by a factor of one hundred million in the past ten years. AI is increasingly efficient at using data and compute, and the more this expands the more powerful AI systems become.

AI now matches or surpasses human proficiency in many tasks, from near-perfect face and object recognition to real-time language translation. Advanced AI can produce original images, compose fluent text, develop code, and even predict protein structures. In areas like strategizing and creativity, once deemed uniquely human, there are notable advancements.

AI still has limitations. Generative AI text produces falsehoods with confidence (called 'hallucinations'). Issues around biased and unfair outputs, security and privacy vulnerabilities and legal liability persist. Despite extensive research and investment, we do not have consumer-level autonomous driving. At present, we have relatively 'narrow' AI which performs well at fixed tasks. However, as the biggest technology companies have the means to significantly scale AI training, we are likely to see AI continue to achieve new capabilities and overcome limitations, with models becoming more efficient and easier and cheaper to build.

We may soon see increasingly autonomous AI agents which can strategise, divide goals into sub-tasks and access their environment to take actions. Experimental projects such as AutoGPT - while not yet fully effective – aim to connect chatbots up with web browsers and word processors to carry out sub-tasks autonomously. Prominent AI industry figure Mustafa Suleyman sees the next milestone as 'Artificial Capable Intelligence', achieved when AI can "go make $1 million on a retail web platform in a few months with just a $100,000 investment." When this is possible, implications will be broader than just financial and the new powers this will give different actors come into view.

Where might we be heading? The explicit aim of the leading AI companies is to build Artificial General Intelligence (AGI), systems that can "match or exceed human abilities in most cognitive work". As AI gets better at automating tasks like programming and data collection, we might be surprised at how quickly it advances. Some leading experts believe that AGI will be achieved as soon as 2030, however both its feasibility and exact timelines are fiercely debated by experts and researchers. What is broadly agreed within AI industry and research communities is that conversations on safety and control are essential as AI progress sees systems becoming more autonomous. This raises risks around malicious actors setting harmful objectives and AI systems pursuing goals not aligned with human interests.

Box 3: AI is a groundbreaking technology for several reasons:

- It can make decisions autonomously, an example being banks using AI for loan approvals.
- It can generate novel ideas and insights by connecting unrelated information.
- It is multi-use and dual-use. For instance, facial recognition that unlocks your phone can also be used to identify protesters by a repressive regime.
- Its inner workings are not fully understood, termed the 'black box' problem. This makes it difficult to know how to predict or change the behaviour of AI systems and leads to emergent and unexpected capabilities.
- It amplifies human capabilities, opening the door for various actors – state and non-state - to achieve their goals more efficiently. As the UK Government states: "Frontier AI will almost certainly continue to lower the barriers to entry for less

1.1 Benefits and risks from AI

As AI enhances human intelligence, we are likely to see spectacular breakthroughs. What are we already seeing?

- **Healthcare improvements**. AI has helped with diagnoses (e.g., increasing detection of diabetes), and developing cures (e.g., identifying a new drug for liver cancer).
- **Mitigating climate change.** AI can enable better flood forecasting, help predict wildfires and track deforestation.
- **Development gains.** AI can improve agricultural decision-making and increase yields, contributing to food security and economic development.
- **Access to personalised education.** AI gives potential for more personalised tuition, can expand education services to remote areas and upskill workers across sectors.
- **Cultural impacts**. AI is helping to preserve dying languages.
- **Economic benefits.** Generative AI alone is predicted to add up to $4.4 trillion to the global economy.

While the transformative benefits of AI are increasingly evident, there is also growing evidence of the harms it can cause and potential risks ahead. What are we seeing?

- **Biases and discriminatory outcomes**. Uses of AI for purposes such as predictive policing has been seen to reinforce societal discrimination. As foundation models are

trained on vast amounts of unstructured data, outputs have mirrored biases in the data and demonstrate gender, racial and cultural stereotypes and prejudices.

- **Cybercrime**. There have been cases of AI generated voices deceiving the public and overriding bank security checks.
- **Synthetic and 'deep fake' media**. AI has been used to create faked audio and video of prominent individuals and non-consensual intimate imagery.
- **Security threats.** There is evidence of people 'jailbreaking' LLMs, removing the training safeguards that prevent dangerous use cases. In this example, GPT-4 gave advice on planning terrorist attacks when asked in languages such as Scots Gaelic or Zulu.
- **Infringement of copyright.** Cases include AI companies being sued for training models with copyright material and reproducing protected material in outputs.
- **Exploitation of workers.** Public AI products require extensive human input to ensure the quality and safety of outputs, resulting in workers experiencing psychological trauma from reading and viewing graphic content, low pay and poor working conditions.
- **Intrusive surveillance.** There are examples of facial recognition and other technologies that monitor, track and record activities of individuals being used in public places without scrutiny.
- **Job losses.** Certain professions such as software development already face automation, and Goldman Sachs estimates that generative AI could expose 300 million jobs worldwide to automation.

The AI Safety Summit has a focus on risks from frontier AI. What does this mean?

Misuse by malicious actors. As frontier AI systems become more advanced and autonomous, this amplifies risks of their use by malicious actors for cyber-attacks, producing novel bioweapons or chemical weapons, social manipulation and deception. Restricting these uses will be very difficult, especially as AI tools are made publicly available through 'open source' models.

Loss of control of AI systems. If highly autonomous AI is built, we risk AI systems pursuing goals not aligned with human interests. At present, there is no way to align AI behaviour with human values and ongoing questions over how these values are defined. Risk also arises from AI developers pursuing frontier systems in a bid to outcompete others, neglecting safety testing and human oversight.

2. How AI Might Impact Democracy

AI can offer substantial benefits for democratic systems:

- **Improving access to government.** AI is supporting new means for the public to access government services and receive legal advice.
- **Enhancing public participation**. Generative AI can help citizens to articulate their insights in public consultation processes, overcoming barriers of language, education and disability.
- **Supporting more efficient decision-making.** AI can help synthesise and map inputs from the public, draw out themes, highlight well-supported arguments, and isolate false claims or misinterpretations.
- **More efficient and effective public services.** AI can improve government service delivery by helping allocate resources more efficiently, reducing fraud and error, predicting issues such as public health crises and improving personalisation of services.
- **Reducing divisiveness of public discourse.** AI chatbots can be fine-tuned to help discuss sensitive topics with the public and overcome divisiveness.

However, while many of these developments are yet to be realised, current use cases of AI pose risks to the foundations of democracy in the following ways.

Compromising the information environment. Generative AI may open opportunities to new actors to produce disinformation, lowering costs of producing messages and increasing their quality and personalisation. The ability to automate production of text and other media may saturate the public information space that underpins democracies, causing the public to lose trust in authentic news reporting, public safety messages and legal processes, where truth is critical. This can undermine democratic processes and harm freedom of expression.

Subverting public consultation processes. Generative AI can automate and scale up 'astroturfing': fake grassroots campaigns, often delivered via fake social media accounts or bots, that give the impression of genuine public support or opposition to an idea or cause.

Expanding surveillance. AI excels at producing surveillance instruments, taking humans out of the loop in sifting through huge flows of information. Expansion of AI surveillance can undermine fundamental rights to privacy and free expression, threatening civic space and helping embed illiberal or authoritarian regimes. However, surveillance also remains a problem within democracies, with certain governments using facial recognition systems, biometric identification for national security purposes and predictive policing for law enforcement. In addition, large Western tech companies have normalised a surveillance-based business model, collecting and monetising user data in exchange for free services.

Affecting electoral processes. 'Deep fake' AI content poses increasing risks of election manipulation. There are already examples of AI-produced audio affecting elections by

damaging the reputation of parties and candidates, such as in Slovakia. Deep fakes are spread on social media, which has already been shown to algorithmically amplify sensational and divisive content.

Impacting political discourse. Deep fakes also enable a 'liar's dividend' whereby political actors claim real content is fake to avoid sanction. An example from India involved a state MP who was recorded accusing party members of corruption. The MP claimed the recording was 'fabricated' by AI but experts believed it was likely genuine.

Damaging social cohesion. Biased outputs from AI systems can reinforce damaging stereotypes, marginalise minority groups and increase polarisation and inequality in society. Generative AI also has capabilities to mass produce content that promotes and amplifies incitement to hatred, discrimination or violence on the basis of race, sex, gender and other characteristics.

Power concentration and increasing inequality. As AI development is driven by a small number of companies, we risk a concentration of market power, weakening competition and consumer choice. Globally, the IMF suggests a severe widening of the gap between rich and poor nations owing to the concentration of AI industry in advanced economies.

There are specific **risks to less resilient democracies:**

- Generative AI's potential to supercharge disinformation may be felt even more in countries where lower levels of digital literacy and a less robust press may struggle to push back.
- Deep fakes may have more of an impact in fragile democracies and countries experiencing war and instability.
- Surveillance capacities may be increasingly tempting to governments in new or fragile democracies seeking to consolidate control. African governments are spending over $1 billion on surveillance technology. Reports indicate Chinese firms supply surveillance tools to 63 countries globally, however companies in democracies often match or surpass Chinese sales.

If significant harms from AI emerge soon, the public may lose trust in government to keep pace with emerging technology and keep societies safe. This could lead to public anxiety about AI and compromise realising its benefits. Emerging risks from AI are also coming at a time when democracy is in decline globally. If AI now erodes democracy in the above ways, we will also lose the power to control potential long-term risk from frontier AI. The near-term risks therefore compel an immediate debate within and across democracies and a rapid international response.

Therefore, a proposed sixth objective for the UK Safety Summit is: **Safeguarding the integrity and security of our democratic systems against current and future threats from AI.**

This objective could be raised in discussions between democratic leaders on the sidelines of the upcoming summit, with dedicated events after the summit helping to elaborate further

an approach across democracies worldwide to protecting democratic systems, human rights and fundamental freedoms.

3. A Democratic vision for AI governance and safety

The introduction of new technology into society is inherently political. Democracies worldwide will need to project a clear vision of AI governance and safety and make a public, positive case for how they will approach the transformative changes that AI is already bringing. What are the priorities?

Firstly, **establish democratic oversight of AI safety.** Leaving AI safety measures to be defined by AI industry risks ceding democratic sovereignty. Other areas of societal risk including medicine and nuclear power have required governance measures to minimise dangers. To prevent risks from AI, democracies should establish clear and binding laws and treaties together with a strong democratic checks on AI development and deployment that uphold values of transparency, accountability, and protection of human rights. Establishing oversight now allows democracies to determine how powerful AI will be used in the public benefit, decide on measures to keep societies and democracies safe, and ensure that the benefits of AI are broadly shared.

Secondly, **ensure participation and inclusivity in AI governance.** Citizens in democratic societies must be involved in shaping the values that determine how AI is governed. An inclusive approach also helps predict and mitigate emerging risks from AI by giving a voice to groups across society who are being impacted.

Inclusivity at an international level means ensuring that countries less advanced in AI development but more vulnerable to impacts on democracy and society are given a voice in global AI governance. Developing countries require a path to digital development and must be supported to reap the benefits from frontier AI.

Thirdly, **demonstrate democratic values at home and abroad.** Democracies must coordinate to push back against illiberal and repressive uses of AI. When democracies use or export repressive technologies, they compromise civil rights, weaken rule of law and diminish their own credibility. It sends mixed signals to the world and contributes to eroding democracy at home and abroad. Democracies must coordinate together to enact domestic reforms and shape global norms on AI that protect privacy and other fundamental rights.

Demonstrating democratic values also means developing and communicating a clear strategy and vision for how AI can benefit democratic societies, how it can be harnessed as a public good and managed in the public interest. This builds public trust in a technology with the potential to have transformative benefits across sectors and contribute to national and global prosperity.

3.1 The central role of parliaments in AI governance and safety

Addressing complex, technical issues of AI safety requires effective, trusted and flexible governance institutions. Parliaments as the key institution for democracy will need to champion the public interest, establish robust accountability systems and devise new methods to gather public views on highly technical topics. We will need well-informed Members of Parliament, in house expertise and routine engagement with the public and stakeholders across society.

In a fast-moving environment, democratic institutions will also need to be agile, examining their structures and processes to enable quick and decisive action that keeps pace with AI development. What actions can MPs individually, and parliament as a whole, take?

3.2 Harnessing parliament's core functions

MPs can use their **lawmaking role** to consider measures that advance frontier AI safety, such as:

- Licensing requirements for companies building frontier AI systems, while preventing against risks of regulatory capture.
- Establishing safety and transparency standards for AI developers in law.
- Including legal requirements for independent oversight and audit, including testing AI models for dangerous capabilities.
- Ensuring that AI developers allocate a significant proportion of their research and development budget to addressing safety and ethics issues.

While certain legal powers will be limited to countries with advanced AI sectors, MPs in all democracies can examine and reform existing laws to protect against threats to democracy, such as privacy and data protection-related laws. They can review and revise legal liability frameworks and anti-discrimination laws to ensure accountability to individuals and groups where AI is proven to cause harm. In specific use cases that damage democracy, MPs can consider new legal measures such as criminalising the production of deep fakes used for political purposes, labelling AI-generated synthetic content and mandating that the public know they are engaging with AI online.

Parliament's **oversight role** will be essential for AI safety. MPs can use powers to gather information from the government and AI industry through questions, debates, inquiries and hearings. They can advance AI safety by summoning industry leaders to testify about AI progress, request records on AI security issues and safety measures, and validate information by cross-examining AI experts. When international agreements on AI are in place, MPs will have a key role to monitor compliance and scrutinise domestic capacity to implement agreements.

Committees play a crucial role in AI oversight. By holding public hearings with representatives from AI industry, academia, ethicists, civil society, and the general public, they can foster an inclusive approach to AI governance. Committees will need strong ties with external bodies to gauge AI's influence on democracy and society, including organisations that research and classify AI risks and harms, and bodies conducting independent audits and human rights evaluations. Making committee reports and hearings public then helps to build trust in democratic institutions to tackle AI-related challenges.

When conducting **budget scrutiny**, MPs will need to ensure that regulatory and audit bodies have the funding to adequately monitor the deployment of AI; and that the budget funds AI monitoring and audit, safety research and public education and supports innovative and beneficial uses of AI across sectors.

MPs are uniquely placed in their **representation role** to provide the public with a voice on AI governance and help build societal resilience against risks from AI. To help ensure AI aligns with core values, we need continuous public engagement to define shared values in democratic societies. Public meetings, surveying, site visits and social media engagement can help MPs listen to the concerns, values and diverse perspectives of their constituents. MPs should prioritise identifying and consulting vulnerable groups who might be disproportionally impacted by AI or unable to access AI-based systems.

Ensuring accountability for the impact of AI requires a well-informed and engaged public. MPs can work together with the media, education bodies, academia, and civil society to contribute to the public discourse around technological change, help counter AI-driven disinformation and mitigate the potential use of deep fakes to disrupt electoral processes. They can support public media literacy initiatives and advocate for AI literacy, ethics and safety to be incorporated into computer science, technology and civic education curricula.

We need democratic institutions to evolve as the technology is evolving. Parliaments have an opportunity to be ambitious and visionary in addressing transformational AI. They offer institutional legitimacy to trial new deliberative processes in engaging the public on the opportunities and risks of AI. Experiments such as citizens' assemblies (a decision-making framework involving randomly chosen individuals collaboratively developing policy solutions, also called 'citizens' juries' or 'mini-publics') may offer a novel means to connect ongoing public input into parliamentary processes.

Finally, parliaments can **use international engagements and parliamentary diplomacy** to help respond to threats to democracy from AI that cross borders. MPs are well-placed to raise issues of repressive or damaging uses of AI and work together with colleagues in other parliaments to coordinate policies to isolate and apply collective pressure to states that use AI for repressive purposes. Specific measures include human rights due diligence requirements and export restrictions on technologies used for surveillance and oppression. To protect the integrity of election processes, an international effort will be required to help electoral bodies and international observers to adapt to new uses of AI and address risks such as deep fakes.

The UK Government has invited China to the AI Safety Summit, providing an indication of its likely view that non-democratic countries with advanced AI industry need to be involved in global discussions on frontier AI safety. However, democracies should establish dedicated international mechanisms to help coordinate and protect democratic systems and retain democratic influence in international standard setting bodies.

Regional bodies have an important role in pooling resources, sharing expertise and providing opportunities for MPs to contribute to a collective, democratic voice in global AI governance and safety. ParlAmericas provides a good example of where regional groups of parliamentarians on AI governance have formed.

4. Conclusion

The promises of AI are vast. Harnessed effectively, it can transform sectors and contribute to global productivity and prosperity. However, progress is accelerating at a speed far outstripping democratic processes and controls. With this comes risks to society and democracy and potentially loss of human control over AI. In democracies, we are not yet set up to address these safety concerns.

The UK Safety Summit can be an important landmark in progressing towards global AI safety. While emphasising global cooperation on frontier AI safety, it is crucial to recognize the dangers and risks posed by existing AI technologies, especially when they challenge democratic systems. The Summit can help initiate a conversation across democracies on measures to address such risks. A proposed AI Safety Institute, or an expert monitoring group akin to the Intergovernmental Panel on Climate Change, can then provide essential support to democracies with impartial, reliable and timely assessments about AI progress and research on the impact of AI on society and democracy.

Institutional components of democracy will be key to mitigating harms and risks from AI. We need a revitalised global effort to support the development of robust, people-centred, trusted democratic institutions which can address the changes to democratic societies that AI will bring. This is long-term and essential work.

5. Recommendations: Actions parliaments can take

Short-term (in the next year)

1. Support MP Professional Development on AI

- Organize workshops and seminars for MPs with AI experts, ethicists, industry leaders and civil society to provide basic understanding on AI, where the frontier is and where the most significant risks lie.
- Provide research and information on emerging regulatory initiatives on AI and measures to address deep fakes, AI-driven misinformation, and biased AI systems.

2. Strengthen structures in parliament to oversee AI impact

- Support the creation of bodies of parliament such as cross-party groups on AI.
- Incorporate AI oversight into the workplans of sectoral parliamentary committees.

3. Consult the public on AI

- Initiate public consultations – in person and/or using social media – to gather diverse perspectives on AI.
- Support MPs in their representation role to identify and consult marginalised groups that might be most impacted by AI.

4. Upskill parliamentary staff

- Provide opportunities for parliamentary staff to engage with and learn from sources of expertise in AI from industry, academia, civil society.

5. Support MPs to attend international meetings on AI

- Identify opportunities for MPs to engage in international conferences and meetings on AI, such as parliamentary groups around OECD and UNESCO initiatives.
- Support seminars or workshops where MPs and staff share best practices and pass on lessons learnt on issues including AI regulation to their colleagues.

Medium-term (1-5 years)

1. Establish oversight structures focused on AI

- Consider specialised committees/sub-committees dedicated to AI, ensuring continuous oversight and adaptation to technological changes.
- Mandate committees to regularly review and report to parliament on AI-related policies and legislation to ensure they remain relevant and adapt to AI developments and risks.

2. Develop in-house technical expertise

- Identify and hire parliamentary staff with expertise in AI, or source expertise through national and international partnerships, twinning arrangements and regional bodies.
- Establish dedicated units within parliaments that focus on AI governance and oversight and provide ongoing support to MPs and committees.

3. Launch public awareness campaigns

- Develop and deliver parliamentary education and outreach programmes that educate citizens about AI, its benefits, and potential risks.

4. Trial innovations to improve public engagement

- Consider new means of public input and engagement on AI through parliament, studying innovations such as citizens' assemblies.

5. Use AI to improve the work of parliament

- Demonstrate democratic innovation through studying and introducing AI tools to improve the effectiveness and accessibility of parliament.
- Examples include using AI for example legislative drafting, providing better access to parliamentary information, and enabling new means for the public to engage with parliament and their representatives.

6. Recommendations: The key role of democratic governance support

International democratic governance support can play a key role in catalysing a democratic approach to AI governance and safety by developing capacities, building partnerships, supporting joint action, and sharing good practices and innovations. The international development community can:

1. **Support capacity development** for MPs and parliamentary staff on technical aspects of AI and its implications. This can better inform national policy and legislative responses and enable stronger participation in global governance discussions.
2. **Share updated research and evidence** on AI's impact through an 'observatory' approach. This can help raise awareness of AI's impact across jurisdictions and highlight illiberal and repressive uses of AI.
3. **Build platforms to convene elected representatives** to support South-South exchange and tap into networks of expertise including technology companies, civil society and grassroots organisations. This can include networking across parliamentary committees tasked with overseeing AI.
4. **Support public education and AI literacy** including in local languages and for diverse and marginalised groups. This includes digital literacy around repressive use of AI.
5. **Help parliaments introduce democratic innovations harnessing AI.** Help identify and share examples of AI-driven tools and platforms that can enhance public participation and improve transparency and accountability of the work of parliament. The Inter-Parliamentary Union's Centre for Innovation in Parliament provides an example.
6. **Provide resources to global media and civil society**. Accountability requires a strong and technically informed civil society to advocate for the public and independent media to shine a light on AI risks and illiberal and undemocratic trends. This includes funding for civil society working on AI accountability and digital rights in non-democratic countries.

Annex A: The debate around catastrophic risk from frontier AI

The UK AI Safety Summit has a focus on risks from frontier AI systems – the potential for increasingly powerful AI systems to cause catastrophic impacts on society.

The Centre for AI Safety categorises catastrophic risk as follows, with examples:

Malicious use, in which individuals or groups intentionally use AI systems to cause harm.

- AI used to create bioweapons.
- AI is used to spread persuasive propaganda.
- AI enables censorship and mass surveillance to help concentrate power.

AI race, in which competitive environments compel actors to deploy unsafe AI systems or cede control to them.

- Corporate rush to replace human jobs with AI systems.
- Release of harmful systems, such as for automated warfare.

Organizational risks, highlighting how human factors and complex systems can increase the chances of catastrophic accidents.

- Safety culture problems.
- Leaked AI systems.
- Insufficient security in AI labs.

Rogue AIs, describing the inherent difficulty in controlling agents far more intelligent than humans.

- AI systems pursue a dangerous proxy goal that does not align with the intended goal.
- AI systems seek power and take control of their environment.
- AI systems deceive humans and manipulate public discourse and politics.

What are the chances of catastrophic risk being realised? It is a clear concern among some in the AI industry. AI pioneers Geoffrey Hinton and Yoshua Bengio have argued that the timeline towards AGI was shorter than they envisaged. This raises risks around 'rogue AI systems' that develop their own sub-goals which manipulate people, give themselves greater control and even threaten human existence. In 2011, DeepMind's chief scientist, Shane Legg described the existential threat posed by AI as the "number one risk for this century, with an engineered biological pathogen coming a close second". In a recent interview, Anthropic CEO Dario Amodei said that chances of an AI system going "catastrophically wrong on the scale of … human civilisation" was between 10% and 25%.

What might this look like? Even if asked to achieve important and beneficial goals, a sufficiently powerful and autonomous AI could pursue dangerous and damaging objectives not aligned with human interests. Two examples are painted by computer scientist professor and author Stuart Russell.

> *Let's suppose … that we ask some future superintelligent system to pursue the noble goal of finding a cure for cancer—ideally as quickly as possible, because someone dies from cancer every 3.5 seconds. Within hours, the AI system has read the entire biomedical literature and hypothesized millions of potentially effective but previously untested chemical compounds. Within weeks, it has induced multiple tumors of different kinds in every living human being so as to carry out medical trials of these compounds, this being the fastest way to find a cure. Oops.*

> *If you prefer solving environmental problems, you might ask the machine to counter the rapid acidification of the oceans that results from higher carbon dioxide levels. The machine develops a new catalyst that facilitates an incredibly rapid chemical reaction between ocean and atmosphere and restores the oceans' pH levels. Unfortunately, a quarter of the oxygen in the atmosphere is used up in the process, leaving us to asphyxiate slowly and painfully. Oops.*

Is catastrophic risk from frontier AI realistic? There is plenty of push back from industry figures who argue existential-type risks are conjecture.

- François Chollet, Google AI researcher: "There does not exist any AI model or technique that could represent an extinction risk for humanity…not even if you extrapolate capabilities far into the future via scaling laws".
- Yann LeCun, Meta Chief AI Scientist: "until we have a basic design for even dog-level AI (let alone human level), discussing how to make it safe is premature".
- Joelle Pineau, senior Meta AI leader, has said existential risk discussions are "unhinged" and warned that "when you put an infinite cost, you can't have any rational discussion about any other outcomes"

A contrasting argument, proposed by organisations such as the Distributed AI Research Institute, is that catastrophic risk is a distraction from the harms and near-term risks that are being realised as AI and automated systems are introduced into society. This argument asserts that we should not fall for AI hype and should focus regulatory efforts on ensuring transparency, accountability and preventing exploitative labour practices from current AI systems. In AI development, we should build and test effective AI systems and prevent the pursuit of AGI.

Annex B: The AI governance landscape

Why is regulating AI complex?

1. The speed and unpredictability of change poses problems for traditional lawmaking processes. For instance, few people predicted that generative AI would begin to automate creative industries so soon. New capabilities of AI systems can also arise unpredictably during and after deployment.
2. AI's complexity implies an asymmetry in knowledge and resources between democratic institutions, AI developers and technology companies, creating risks of regulatory blind spots or regulatory capture.
3. As a transformative technology, effective AI governance will need to go beyond legal and tech expertise and requires consideration of ethics and sociology. Achieving consensus on the right approach across these fields is no small feat.
4. How to focus regulation is not exactly clear. For example, where AI causes harm to the public, where in the supply chain does accountability fall? Where governments are clear about an area they want to regulate, this can be easier said than done. For example, bias is often embedded in the training data and very difficult to attribute.
5. National regulation is important but insufficient as systems developed in one country will be deployed in another, enhancing the need for global coordination.

What regulatory approaches are we seeing?

Regulatory approaches to AI are broadly categorised in the 2023 State of AI Report as:

- Relying on existing laws and regulations. Light tough and pro-innovation. This approach does not envisage new specific regulation for AI. Examples include India and the UK.
- Wide-ranging AI-specific legislation. The European Union (EU) has pioneered the introduction of AI legislation focusing on different risk categories. Legislation in China requires AI-generated content to be labelled, developers to register algorithms and 'security assessments' for AI deemed capable of influencing public opinion
- Hybrid models. Slimmed down national regulation, or a reliance on local laws. Emphasis on voluntary commitments. The United States (US) currently demonstrates this model.

What approaches to AI governance are emerging at international level?

The field is crowded, including different initiatives to establish ethical frameworks and principles.

- The **OECD AI Principles**, based around human-centred values, fairness and the rule of law. They promote AI that is transparent, explainable, robust, secure, and safe, with accountability mechanisms in place. They are the first set of principles to receive buy in from international leaders.
- The **Global Partnership on AI** aims to foster a collaborative effort across 29 countries on AI research and global policy development.
- At a summit in May 2023, G7 leaders initiated **the Hiroshima process** to help advance and harmonise AI policy, with a focus on generative AI.
- **UNESCO's Recommendation on the Ethics of Artificial Intelligence** emphasises four core values and ten principles for a human-rights centred approach to AI. They have broad buy-in from countries in the global south and also from China and Russia.
- The International Telecommunication Union (ITU)'s **AI for Good** initiative focuses on developmental benefits of AI and applications that can contribute to the Social Development Goals (SDGs).
- The EU and US have announced that they are working on a joint AI code of conduct, which will include non-binding international standards on issues including risk audits and transparency.
- An **international treaty on AI** is being finalised by the Council of Europe. Signatories will need to take steps to ensure that the development and use of AI respects human rights, democracy and the rule of law.

How is AI industry leading the way?

Various governance initiatives have been called for from within AI industry. Some of the most prominent are:

- In March 2023, there was a public call for a pause on training new more powerful AI systems from AI developers and AI luminaries.
- Leading AI labs have formed the Frontier Model Forum – a body designed to promote the responsible development of frontier models and to share knowledge with policymakers.
- Certain AI labs propose 'responsible scaling' – the continued development of frontier AI with pauses if progress outstrips current safety protocols. However, this approach, without independent oversight, may cede AI safety decisions to labs themselves.

About the author

Alex Read is a WFD Associate Expert, specialising in the impact of emerging technology and the role of parliaments.

Alex has 15 years' experience in parliamentary strengthening worldwide, including helping establish a civil society organisation supporting the Parliament of Cambodia and as Strategic and Technical lead for UNDP and the Inter-Parliamentary Union in Myanmar.

He has worked with WFD as a Senior Parliamentary Advisor in Bangsamoro, Philippines and Ethiopia and has written and presented at WFD events on surveillance technology and parliament.