A progressive realist approach to

Countering the manipulative use of AI

Ben Graham Jones
Franklin De Vrieze

2025

WFD

Foreign, Commonwealth & Development Office

About the authors

Ben Graham Jones is a consultant specialising in emerging challenges to election integrity. He has served on more than 30 election observation, assistance, and advisory missions, including as head of mission, media and social media analyst, and online campaign analyst. Ben has contributed to WFD's work since 2015. Ben has worked on building resilience to the manipulative use of AI in places including Taiwan, Kenya, and the United Kingdom, and has contributed to recent publications on AI resilience by the National Endowment for Democracy and WFD. He holds degrees from the University of Cambridge, and King's College London, where he researched the impact of autocratic disinformation on elections. He is a Churchill Fellow.

Franklin De Vrieze is Head of Practice Accountability at WFD. He leads WFD's approach to accountability, with a focus on financial scrutiny and public debt accountability, anti-corruption, post-legislative scrutiny and institutional performance of parliament. He supports WFD's programme on Democratic Resilience in a Digital World (DRDW) with a focus on the digital transformation of parliaments. He co-edited the WFD Guidelines for AI in Parliaments. He advises WFD country programmes on strengthening parliaments and independent institutions, providing strategic direction, technical guidance, and thought leadership. This includes developing methodologies, leading communities of practice, and co-producing global standards with partner organizations. His work centres on ensuring democratic institutions are equipped to deliver meaningful oversight, adapt to complex challenges, and serve the public with transparency and accountability.

Disclaimer

This policy brief has been written by Ben Graham Jones and Franklin De Vrieze and published in 2025. The authors appreciate the peer-review comments by Anthony Smith, Graeme Ramshaw, Tanja Holstein, Charlotte Egan, Alex Scales, and Ravio Patra from WFD, by Dr Fotis Fitsilis of the Hellenic Parliament, and James Pinnell from the Commonwealth Parliamentary Association (CPA).

The research has been made possible through WFD's Grant-in-Aid funding received from the UK's Foreign, Commonwealth and Development Office (FCDO).

Contents

Acronyms

AGI	Artificial General Intelligence
ASI	Artificial Super Intelligence
AI	Artificial Intelligence
CPA	Commonwealth Parliamentary Association
CSIS	Centre for Strategic and International Studies
CSO	Civil Society Organisation
DISARM	Disinformation Analysis and Risk Management (an open source framework for reporting on disinformation incidents and responses)
EU	European Union
FCDO	Foreign, Commonwealth and Development Office (UK)
FIMI	Foreign Information Manipulation and Interference
G7	Group of Seven (most advanced economies)
GRU	Glavnoye Razvedyvatel'noye Upravlenie (Russia's military intelligence agency)
ICCPR	International Covenant on Civil and Political Rights
LLM	Large Language Model
OAS	Organization of American States
OECD	Organization for Economic Co-operation and Development
OSCE	Organization for Security and Co-operation in Europe
PRC	People's Republic of China
STIX	Structured Threat Information Expression (a language for sharing cyber threat intelligence)
TTPs	Tactics, Techniques, and Procedures
UK	United Kingdom of Great Britain and Northern Ireland
UN	United Nations
UNGA	United Nations General Assembly
USAID	United States Agency for International Development
WFD	Westminster Foundation for Democracy

Summary

Artificial intelligence (AI) offers unprecedented opportunities for democratic empowerment and societal progress. However, the manipulative use of AI threatens a wide range of democratic rights, including the rights to freedom of expression, freedom of thought, and genuine elections. It also presents grave national security threats. This policy brief contains recommendations for democratic policymakers who seek to build resilience to the manipulative use of AI.

The recommendations present a progressive realist approach, which involves using realist means to pursue progressive ends. A progressive realist approach involves three behaviours: identifying incentives, forming coalitions of purpose, and mitigating risk.

1. What is a progressive realist approach to the manipulative use of AI?

Uncertainty over the ultimate future of AI has created a new arena of power competition. Conscious of the enormous opportunities AI offers, state and private actors are competing to shape this arena in favour of their political, economic, and security interests.

Some actors pursue these interests in ways that threaten democratic processes and national security. These include autocratic states for whom the information ecosystem is not a venue for open debate but an opportunity to exert social control. This brief is not primarily focused on the many unintentional risks associated with AI, such as unintended hallucination, but on the manipulative use of AI as a deliberate matter of policy.

"The advent of AI is ushering in profound changes to competition and conflict"

Rand Corporation

In the face of this challenge, many democratic states and civil society organisations (CSOs) have regarded mechanisms such as politically impartial fact-checking, media literacy, and limiting the impact of bot accounts as limited but necessary measures to preserve a free, open, and inclusive sphere of democratic debate. Such measures have so far played an important role in mitigating and building resilience against the manipulative use of AI.

Fostering lasting progress amid differing emphases will require a clear-eyed view of incentives. Many democratic actors have argued that, just as AI can have unintended negative consequences, attempts to limit the manipulative use of AI may also have unintended consequences such as restricting innovation or giving unchecked power to social media companies. However, amid an often-polarised discussion, there are points of consensus between democratic actors.

For example, most democratic actors are likely to reject the highest-risk applications of AI, such as social credit scoring systems; to oppose artificial general intelligence (AGI) or artificial super intelligence (ASI) under the attempted custodianship of unaccountable companies or governments; to question the undermining of companies subject to democratic oversight by autocratic competitors working on sensitive tech; and to shun the prospect of elections being purchased by whomever acquires the most powerful AI capabilities to influence public opinion.

These points of agreement may have the potential to bring together lasting coalitions for sustainable action against the manipulative use of AI. The centrality of incentives is reflected in the UK government's overall approach to foreign policy, which it describes as 'progressive realism'. This means recognising that states pursue their perceived self-interests and then working with that reality to pursue just ends.

It is precisely because progressive realism is ambitious in its aims that it is realistic about the need to collaborate to achieve them. Progressive realists regard building partnerships with like-minded allies as crucial to securing policy objectives. They are equally realistic about the extent to which those actors who use AI for manipulation, or who lack appropriate safeguards, can be trusted to collaborate in the building of a secure and democratic future for AI.

All manner of institutions across society have a role to play in mitigating risks by building resilience. However, states do have unique levers to address the problem, including by fostering collaboration across society at large, and governments' scope for action is the focus of this brief.

Recent manipulative uses by a vast patchwork of entities, from states to lone actors to private companies, have included:

- conducting astroturfing operations to influence political campaigns
- launching AI assistant apps which censor factual content and store user data in autocratic jurisdictions
- hacking critics across the EU
- promoting AI-generated content masquerading as local news sources
- facilitating surveillance of host populations in countries like Serbia and Hungary
- using biometric data of Zimbabweans to build a facial recognition database capable of distinguishing between skin tones
- developing databases for AI-facilitated cognitive warfare against Taiwan
- generating deepfakes to spread misogynistic disinformation against female candidates
- creating non-existent jobs to facilitate money-laundering to finance information operations[11]
- using concealment tactics to hide deepfake audios in Lithuania[12]
- debugging code for propaganda websites and networks[13]
- deploying FIMI operations to shift foreign public opinion towards autocratic policy perspectives[14]
- generating supportive comments to influence Ghana's 2024 election[15]

The breadth of these recently deployed tactics underscores the need to engage a broad swathe of actors as part of the response. Equally, the fact that many of these incidents are attributed to states, state proxies, and companies based in autocratic states underscores the threat of AI being used for manipulation as a matter of policy.

However, the risks of AI manipulation are growing, not only in breadth but also in magnitude, as technological progress places further risks on the horizon. These include:

- new data-driven capabilities to produce increasingly individualised video and audio disinformation, enabling granular microtargeting at scale
- advances in text-to-image and text-to-video technologies that make it easier for non-experts to produce compelling disinformation
- improvements in the hardware underpinning AI, such as higher-capability semiconductors, which make AI capabilities cheaper, more accessible, and more powerful
- emerging advances in the stability and scale of quantum computing systems, which may pose AI-enhanced risks such as threats to the cryptographic security of election systems and unprecedented optimization of AI-driven microtargeting

These developments underscore the need for the broad range of actors working to build resilience to collaborate in ensuring that insight on the very latest challenges is adequately shared. Concerted action can play a role in building resilience against the manipulative use of AI both now and in the future.

2. Towards collective resilience: Identify incentives, mitigate risks, forge coalitions

Across the globe, a wide range of steps are already being taken to shield democratic processes from the manipulative use of AI:

- Finland's Parliament has used AI to debug code to increase resilience to potential cyberattacks[16]
- Latin American politicians share insight into best-practice legislative responses to manipulative AI through the ParlAmericas initiative[17]
- Kenyan and Nigerian workers' rights advocates founded the Content Moderators' Union to protect workers checking that decisions made by AI are free from manipulation[18]
- Indian political campaigns have developed rapid response systems to combat false accusations that legitimate content was AI-generated, addressing a phenomenon known as the Liar's Dividend[19]
- United States security agencies have identified and sanctioned actors linked to Russia's GRU implicated in the manipulative use of AI and declassified useable insight[20]
- the Philippines' electoral management body COMELEC issued guidelines requiring political parties to disclose any use of AI in their campaigning materials[21]
- in Brazil, organisations are leveraging AI to aggregate unstructured data, enabling new insights into elected representatives' usage of public expenditure[22]

There is therefore a wealth of best practice from which like-minded changemakers can draw. Faced with common risks, diverse actors have found a shared need to build resilience. However, converting this shared interest into collective action on the international stage faces risks. These include:[23]

- limiting the development of high-risk AI may limit AI-for-profit, generating commercial and political resistance
- not all democratic actors share the same view of threat actors
- lacking institutional knowledge, capacity, and accountability mechanisms, states may be unable to mitigate risks effectively and responsibly, risking collaborations that produce unintended harmful consequences
- threat actors seek to actively compromise and influence the decision-making of democratic actors
- inadequate incentivisation may lead states to seek relative security advantages through noncompliance with collective mechanisms
- political actors who benefit from AI technologies in their campaigns may be disincentivised to support limits on the use of AI

Mitigating these limitations will require a clear assessment of incentives. For example, it may be unrealistic to assume that actors with a track record of using AI to violate human rights will voluntarily decide to comply with global accords that contravene their economic and security interests. Building collective resilience is not the same thing as building the broadest possible coalitions. However, strong partnerships which carefully assess risks can help reduce the costs of working together.

Positively, there are already many examples of meaningful collaboration to mitigate the manipulative use of AI. These successful initiatives have often benefitted from at least three features, which are identified below. This is not to say that all such interventions must include these features, but these aspects are chosen for being present across multiple examples of emerging best practice, and are also features that have been identified in other work and reviews examining the evidence on what works to manage AI risks.

A. Focus on incentives

Work to bridge different views on the challenge posed by the manipulative use of AI has benefitted from identifying and acting on shared incentives. Work from the Mozilla Foundation has identified the importance of balancing regulation with incentivising "trustworthy AI" but notes that too often governments face capacity barriers to doing so.[24] When such barriers have been overcome and incentives have been navigated, tangible progress is being made in enhancing resilience.

Illustrative example:

Despite facing information ecosystems specific to their geographies, countries have found that like-minded partners are similarly incentivised in the face of common threats. Actors including the UK, USA, EU, and Japan each established their own AI safety institutes but found that these institutes share common aims which can be pursued through one international consortium.[25]

Examination of the early work of this mode of collaboration has underscored the role of incentives. The Centre for Strategic and International Studies (CSIS) has noted that, unlike private initiatives where AI safety concerns conflict with incentives to maximise profit, this form of broad collaboration benefits from being free of such financial self-interests.[26] CSIS has further noted that this form of collaboration has the potential to enhance gains for the full diversity of actors involved, with governments benefitting from greater economies of scale in AI safety work, regulators benefitting from greater interoperability through common standards and means, and private companies benefitting from consistency in the requirements their AI products must meet.[27]

B. Mitigate risk

The rapid evolution of AI-enhanced threats ensures that risk-based approaches are essential. Risk assessment has been widely recognised as a core feature of best practice in building resilience to the manipulative use of AI. The OECD is working on common standards for reporting AI incidents to improve the overall evidence base and has established a database collecting AI incidents. It has noted that effective identification of risks can help bolster preparedness and help prioritise finite resources.

Illustrative example:

As the scale and power of AI applications grew in Europe, it became clear that for any regulatory action to be manageable and fair, it would need to be prioritised based on precise metrics. The transnational character of the problem ensured that this prioritisation would be best implemented at EU level Consequently, states supported legislation at EU level, built around a risk-based approach, through the AI Act 2024.[28] The legislation defined unacceptable-risk AI, high-risk AI, and minimal-risk AI, with specific obligations pertaining to each category.

"Artificial intelligence promises tremendous benefits but also carries real risks. Some of these risks are already materialising into harms to people and societies: bias and discrimination, polarisation of opinions, privacy infringements, and security and safety issues"

OECD[29]

C. Act through coalitions of purpose

Coalitions are groupings of some permanence that amplify the benefits of action. As the International Forum for Democratic Studies notes, "Coalitions bring together diverse skillsets to catalyse the work of prodemocracy voices; they save costs, pool resources, and avoid the duplication of efforts".[30]

Illustrative example:

Recognising that diverse organisations could each make specialised contributions to resilience, Taiwanese organisations built a domestic coalition of purpose.[31] Some partners also began collaborating internationally, allocating resources to sharing insight on local information operations through what were then untested frameworks, judging that doing so would help build resilience at a greater scale than by operating alone. They used data-sharing architectures such as DISARM, STIX, and OpenCTI to share insight on AI-facilitated electoral disinformation with partners in democracies including the Philippines and India.[32] Insight was shared both ways, cementing resilience across what is now a permanent transnational coalition.

3. International standards: Emerging technologies, enduring rights

International standards aim to guarantee a consistent level of minimum rights for all people. They are rooted in provisions which states have consented to, none of which are reneged just because technology advances. Particularly in contexts where citizens expect their leaders to respect international law and norms, they may offer a basis for incentivising collective action. Many clauses in existing standards and principles subscribed to by a majority of the world's states are directly relevant to the challenges posed by the manipulative use of AI. The following is a non-exhaustive, illustrative list of some of the key longstanding standards that can be drawn upon.[33]

Manipulative use of AI	Provision	Standard	Standard type
AI-enhanced information manipulation	"Voters should be able to form opinions independently, free of... manipulative interference of any kind"	UN Human Rights Committee, General Comment 25 on ICCPR Article 25	Legally binding international convention
Non-consensual use of personal data to train LLMs	"the same rights that people have offline must also be protected online, including the right to privacy"	UNGA Resolution 68/167, 18 December 2013	UN resolution adopted by general consensus of member states
AI-facilitated censorship	States must ensure that "the public has effective access to information"	UN Convention Against Corruption 2003	Legally binding international convention
Political campaigns' use of manipulative technologies	States must "enhance transparency in the funding of candidatures for elected public office and, where applicable, the funding of political parties"	UN Convention Against Corruption 2003	Legally binding international convention
State-backed information operations	States must "foster an enabling environment for freedom of expression"	UN/OSCE/OAS Joint Declaration on Freedom of Expression and the Internet 2017	Non-binding international guideline for all UN member states

Recent initiatives have also sought to define international standards specific to the challenge posed by AI and facilitatory technologies,[34] for example through the Global Partnership on Artificial Intelligence,[35] the OECD's AI principles,[36] the UK's AI Safety Summit,[37] the G7 Hiroshima Process,[38] and a range of UN initiatives.[39] These efforts have been helpful in fostering shared understanding between like-minded states.

Progress has also been made in defining standards for the secure development of AI systems[40] and for frontier AI.[41] There are several important standards defined by the International Organization for Standardization.[42] In addition, there is a growing set of legal obligations around the development and deployment of AI-facilitated technologies including the EU's AI Act. In addition, many countries' existing legal codes already prohibit related conduct, such as foreign interference and data misuse. Such efforts are significant because voluntary standards have not always incentivised compliance.[43]

The wide range of standards and principles can make them challenging to navigate, particularly as the frameworks continue to evolve. Efforts to build consensus between democratic actors around strong, harmonised standards remain vital. However, to translate these into action, international standards must be connected to incentives. The recommendations which follow seek to build upon the underlying principles of many of the above standards by promoting risk-assessed and incentive-driven collaboration as a means to bolster resilience.

4. Recommendations

The progressive realist approach to AI security involves three main principles. These correspond to a theory of change that identifies motive (incentives), means (coalitions of action), and opportunity (as identified through risk-based approaches) as necessary preconditions for action. The recommendations are structured accordingly.

1. Identify and act on the incentives around the manipulative use of AI

To reduce the risks of AI being used for manipulation, action on multiple fronts must focus on working with, and where necessary altering, the incentives that currently exist. These recommendations are targeted at policymakers seeking to build resilience against the use and facilitation of AI for manipulative purposes.

Invest in researching the motivations of actors involved in the manipulative use of AI.

Where disinformation campaigns are concerned, there are multiple steps at which disruptive action may be taken, an idea that has been expressed as the "kill chain".[44] At earlier stages of disruption, intervention can avoid the costs associated with the manipulative activity. Likewise, supply chains and dependencies of AI technologies typically implicate many actors, whose diverse motivations may differ, from profit to ideology to perception of national interest. A more granular understanding of actor motivations may help to disrupt manipulation at earlier stages.

Disincentivise the use of platforms for manipulation.

Well-intentioned open information spaces also generate incentives for manipulation by actors opposed to the free exchange of ideas. Measures such as effective social media monitoring and technically competent investigative journalism may help rebalance incentives. In addition, platforms based in autocratic states have used AI to limit freedom of expression, for example by censoring criticism of their host state. Where such platforms are used in this way, consider creating common blacklists of manipulative platforms and actors to prevent such entities from influencing democratic processes.

Strengthen data protection safeguards.

This is particularly important when data can be subject to access requests in jurisdictions whose governments are implicated in manipulative practices. For example, companies subject to the PRC's legal framework are required to comply with requests by the government to divulge information about users.[45] In 2018 Apple was accused of moving iCloud user data to China in return for operating in the country.[46] It is important to ensure that citizens' rights over their data are upheld in law and practice wherever that data may be stored.

Expand blacklists, sanctions registers, and inclusion criteria to disincentivise manipulation.

The size and scale of blacklists and sanctions registers currently vary even between like-minded democracies. Foreign individuals and companies who deploy or develop AI technologies to undermine democratic institutions should be subject to specific sanctions including travel bans and asset freezes to disincentivise this behaviour. Greater sharing of known compromised companies and individuals may help better secure public tender processes and disincentivise manipulation.

2. Mitigate risk at every stage.

When informed by defined incentives, risk assessment processes can help establish evidence-based paths towards threat assessment and mitigation. There are plenty of resources available to facilitate risk management around AI. The US' National Institute of Standards and Technology has produced an AI Risk Management Framework[47] that can be used across public and private sectors, specific guidance on generative AI,[48] and a playbook that organisations can draw on to inspire their own approaches to risk mitigation.[49]

Put evidence at the heart of any approach

Risk-based approaches require a clear understanding of what does and does not work. The OECD notes "As AI adoption continues to grow, successful risk mitigation will require a solid evidence base".[50]

The evidence basis for countering the manipulative use of AI is growing. For example, in the domain of countering disinformation, a series of important evidential reviews, including those by USAID[51] and the Carnegie Endowment, have generated important insights.[52] There are many country-specific and intervention-specific studies from the flurry of elections in 2024, and it is an important time to integrate these insights into collective understandings of what works. This does not mean relying solely only on proven approaches – innovation remains critical – but it does mean that the evidential bar is higher than it was several years ago.

Equally, filling ongoing evidence gaps remains vital. Many 'known unknowns' remain, particularly relating to determining what works in specific geographies and improving the evidence basis from the Global South. One action taken by WFD in a context where democratic actors had been regularly targeted with AI-enhanced disinformation was to commission research on which institutions were trusted in a population and share this insight directly with key stakeholders, filling an important evidence gap in trust building.

Bolster transparency of algorithms and social platform content.

Transparency makes it easier to calculate and mitigate risk. Platform algorithms, including LLMs, are powerful gatekeepers that connect individuals to information. Algorithms incentivise user engagement, but the social costs may be unclear if its workings are opaque. For example, training data may contain generative AI-induced inaccurate or biased content at scale, a risk compounded by the growing use of synthetic data and the growing recognition that biases in LLM training data are extremely difficult to mitigate once incorporated.

Inadequate transparency may open opportunities for manipulation. For example, LLMs can be trained to produce misleading outputs. Conversations on balancing transparency measures with protection must be inclusive, while transparency measures, such as access to training data and platform insight, may require capacity building and due consideration of other imperatives such as intellectual property protection. CSOs and researchers may require training to conduct input data analysis and model output analysis.[53]

Enforce access to social platform APIs for democratic researchers and election observers

Researcher insight is also vital to assess the status and risks of information operations, including AI-enhanced operations. The research community, including election observers, have been handicapped by platform closures of application programming interfaces (APIs), a means to access platform data. As long as safeguards on data protection are in place, reversal of these undue restrictions is appropriate, and legislation is an appropriate means to achieve this.

Legal processes are ongoing to determine whether this access can be mandated through existing measures such as the EU's Digital Services Act.

Empower overseas partners to implement risk-based approaches, particularly ahead of crucial moments such as elections.

Although information manipulation exists throughout the election cycle, evidence suggests that information operations intensify a month before and then during the period commencing 72 hours prior to election day. This short timeframe risks catching local actors off-guard who may face new AI-enhanced tactics and approaches that may have been deployed in other elections across the globe, but which may be unfamiliar to an election management body facing its first election in four or five years.

External risk assessments can be helpful in obtaining insight into emerging threats from other contexts, and in overcoming groupthink. Depending on the context, these could include specialised measures such as penetration testing and narrative forecasting. Risk assessments can provide local actors with clear indications of vulnerability gaps, prioritise finite in-country resources, and mobilise actors around shared objectives, while leaving legacies of increased domestic capacity.

Take seriously the potential advent of AGI and ASI in long-term strategic planning.

Many experts are concerned about the potential for future AGI to compromise human autonomy. AGI would surpass human cognitive abilities across many domains.[54] A longer-term threat, ASI is a hypothetical software-based system with fundamental cognitive capabilities far greater than any human.[55]

Fears include that such systems may by human intention or accident develop destructive capabilities that are impossible to mitigate or even identify, or that such capabilities could concentrate power sufficient to render democratic processes irrelevant. Analysts disagree on how long it may take for AGI to develop, but some argue that progress is occurring more swiftly than anticipated.[56] Given the speed with which current, nascent forms of AI have now surpassed human competence in many respects, it is prudent to treat the risks of AGI with the utmost seriousness.[57]

3. Take action through coalitions of purpose

The threats posed by the manipulative use of AI are shared across different actor types within societies, and also between like-minded international partners. By working together in risk-assessed and incentive-driven ways, actors may be able to do more with less.

Empower CSOs to contribute to global information-sharing coalitions.

AI makes it easier for malicious actors to spread information disorder across multiple platforms, languages, and geographies.[58] Autocratic approaches, such as the PRC's through its United Front strategy of influence, can span many geographies, organisations, and individuals. Many different actors across the globe may encounter different parts of the same core operation.

Effective data-sharing architectures have become critical in countering AI-enhanced cyber and information operations at scale.[59] These include common classification frameworks such as DISARM, threat intelligence operationalisation such as STIX, and repositories such as OpenCTI. Harmonisation holds the potential for securing global real-time insight on threat evolution. Consider supporting peer-to-peer capacity building between CSOs. It is also important to connect other institution types, such as election management bodies and political

parties, with the technical expertise to understand the evolving threat landscape.

Engage the voices of those facing a disproportionate use of AI for manipulation.

Stakeholders in geographies facing elevated levels of AI for manipulation, such as Ukraine and Taiwan, have unique expertise in mitigating the most well-resourced attacks. Bilateral training, peer-to-peer mentorship, and cybersecurity support from these experienced actors may help counterparts build resilience within their own geographies, particularly in places without local expertise around AI.

Strengthen collaboration on commercial standards for AI technologies.

Shared approaches can help avoid a race to the bottom on technology risk. For example, companies subject to the PRC's control have embedded their technologies in infrastructure across the globe as part of the Belt and Road initiative. They have often delivered technologies at lower cost than democratic actors are able to but have also raised fears around security vulnerabilities. Rights-respecting actors can take practical steps such as upholding agreed commercial and procurement standards, deepening AI supply chain audits, enhancing collaborative due diligence mechanisms, and broadening shared export controls to mitigate risks of AI-enhanced systems being used for manipulative practices.[60]

Build parliamentary resilience through peer-to-peer engagement.

Parliaments which have already concluded processes to build resilience through legislation, such as the EU and UK, are uniquely placed to support peers overseas who are facing the same challenges. Numerous resources can support this engagement, including guidelines WFD has produced on how parliaments can seize the opportunities of AI while upholding democratic principles.[61] Specialist peer-to-peer networks can also help. For example, the Inter-Parliamentary Alliance on China is available to support politicians to share best practice in scrutinising legislation for risks related to the PRC's use of technology.

As AI's role in government expands rapidly, it poses significant threats to the balance of power in democratic institutions. Parliaments must modernise and enhance their oversight capabilities to ensure AI deployment is transparent, ethical, and accountable. By doing so, they can safeguard democratic principles, maintain necessary checks and balances, and avoid further executive dominance.[62] AI is helping transform legislatures from paper-based organisations into data-driven institutions.[63]

Endnotes

1 https://www.rand.org/content/dam/rand/pubs/research_reports/RRA3200/RRA3295-1/RAND_RRA3295-1.pdf

2 https://www.csis.org/analysis/russian-bot-farm-used-ai-lie-americans-what-now

3 https://www.wired.com/story/deepseek-ai-china-privacy-data/

4 https://www.justice.gov/opa/pr/seven-hackers-associated-chinese-government-charged-computer-intrusions-targeting-perceived

5 https://www.newsguardtech.com/special-reports/ai-tracking-center/

6 https://balkaninsight.com/2024/06/28/serbian-authorities-use-high-tech-surveillance-to-monitor-opponents-birn-report/

7 https://www.rferl.org/a/china-surveillance-cameras-europe-dahua-hikvision/32930737.html

8 https://freedomhouse.org/report/freedom-net/2018/rise-digital-authoritarianism

9 https://www.cna.org/our-media/newsletters/china-ai-and-autonomy-report

10 https://foreignpolicy.com/2024/01/23/taiwan-election-china-disinformation-influence-interference/

11 https://cdn.openai.com/threat-intelligence-reports/disrupting-malicious-uses-of-our-models-february- pdate.pdf

12 https://fimi-isac.org/wp-content/uploads/2024/10/FIMI-ISAC-Collective-Findings-I-Elections.pdf

13 https://openai.com/index/disrupting-deceptive-uses-of-AI-by-covert-influence-operations/

14 https://www.rand.org/nsrd/news/nsrd-upfront/2024/12/chinese-social-media-manipulation.html

15 https://cdn.openai.com/threat-intelligence-reports/disrupting-malicious-uses-of-our-models-february-2025-update.pdf

16 Inter-Parliamentary Union, 'Use cases for AI in parliaments: Narrowing of the attack threshold for parliamentary applications' (2024).

17 https://www.parlamericas.org/uploads/documents/PressRelease-DigitalCaucus-June15-en.pdf

18 https://time.com/7012799/kauna-malgwi/

19 https://www.wfd.org/sites/default/files/2024-09/wfd_2024_ai_in_action_-_final.pdf

20 https://home.treasury.gov/news/press-releases/jy2766

21 COMELEC Resolution 11064 of 17 September 2024.

22 https://www.ned.org/wp-content/uploads/2024/10/NED_Leveraging-AI-for-Democracy-Report.pdf

23 https://www.wfd.org/what-we-do/resources/democratic-approach-global-ai-safety

24 https://foundation.mozilla.org/en/insights/trustworthy-ai-whitepaper/path-forward/creating-regulations-and-incentives/

25 https://www.oecd.org/en/publications/assessing-potential-future-artificial-intelligence-risks-benefits-and-policy-imperatives_3f4e3dfb-en.html

26 https://www.csis.org/analysis/ai-safety-institute-international-network-next-steps-and-recommendations

27 https://www.csis.org/analysis/ai-safety-institute-international-network-next-steps-and-recommendations

28 https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

29 https://www.oecd.org/en/topics/ai-risks-and-incidents.html

30 https://www.ned.org/stronger-together-coalitions-as-a-catalyst-for-information-integrity/#coalitions

31 https://medium.com/doublethinklab/taiwan-power-a-model-for-resilience-to-foreign-information-manipulation-interference-70ea81f859b7

32 https://medium.com/doublethinklab/watch-our-neighbourhood-collaborative-approaches-to-address-fimi-during-the-india-election-727e3e5f0171

33 https://www.wfd.org/commentary/how-election-support-field-must-adapt-artificial-intelligence

34 https://carnegieendowment.org/posts/2024/03/why-we-need-a-global-ai-compact?lang=en

35 https://gpai.ai/2023-GPAI-Ministerial-Declaration.pdf

36 https://oecd.ai/en/ai-principles

37 https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

38 https://www.mofa.go.jp/files/100573471.pdf

39 https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf

40 https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf

41 https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024

42 https://www.iso.org/sectors/it-technologies/ai

43 Consider the 2018 Santa Clara Principles on Transparency and Accountability on Content Moderation, signed up to by companies including Facebook (Meta) and Twitter (now X).

44 https://carnegieendowment.org/research/2023/03/phase-based-tactical-analysis-of-online-operations?lang=en

45 Article 22 of the Counter-Espionage Law of the PRC (2014, revised 2023); Article 35 of the Data Security Law of the PRC (2021).

46 https://www.article19.org/resources/apple-cares-about-digital-rights-unless-youre-in-china/

47 https://www.nist.gov/itl/ai-risk-management-framework

48 https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

49 https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook

50 https://www.oecd.org/en/topics/ai-risks-and-incidents.html

51 https://pdf.usaid.gov/pdf_docs/PA0215JW.pdf

52 https://carnegieendowment.org/2024/01/31/countering-disinformation-effectively-evidence-based-policy-guide-pub-91476

53 https://knowledgehub.transparency.org/assets/uploads/kproducts/Algorithmic-Transparency_2021.pdf

54 https://arxiv.org/abs/2306.12001

55 https://www.ibm.com/think/topics/artificial-superintelligence

56 https://www.stateof.ai

57 https://www.pewresearch.org/internet/2023/06/21/as-ai-spreads-experts-predict-the-best-and-worst-changes-in-digital-life-by-2035/

58 https://www.disinfo.eu/outreach/our-webinars/building-fimi-resilience-through-models-of-practice-taiwan-2024-election/

59 https://fimi-isac.org/wp-content/uploads/2024/10/FIMI-ISAC-Collective-Findings-I-Elections.pdf

60 https://www.ncsc.gov.uk/collection/machine-learning-principles/secure-development/secure-supply-chain

61 Fitsilis, Fotis, Von Lucke, Jorn, and De Vrieze, Franklin, 'Guidelines for AI in parliaments' (WFD, 2024); Fitsilis, Fotis, Von Lucke, Jorn, and De Vrieze, Franklin, 'Inception, development and evolution of guidelines for AI in parliaments', in: The Theory and Practice of Legislation (10 March 2025).

62 De Vrieze, Franklin, 'Parliaments must modernise to avoid AI-induced executive dominance', Blog for University of Birmingham (2024).

63 De Vrieze, Franklin, 'Democratic innovation through AI in parliaments'

Westminster Foundation for Democracy (WFD) is the UK public body dedicated to supporting democracy around the world. Operating internationally, WFD works with parliaments, political parties, and civil society groups as well as on elections to help make political systems fairer, more inclusive and accountable.

www.wfd.org

@WFD_Democracy

@WFD_Democracy

Westminster Foundation for Democracy (WFD)

Scan here to sign up to WFD news

WFD

Westminster Foundation for Democracy is an executive Non-departmental Public Body sponsored by the Foreign, Commonwealth & Development Office.

Foreign, Commonwealth & Development Office