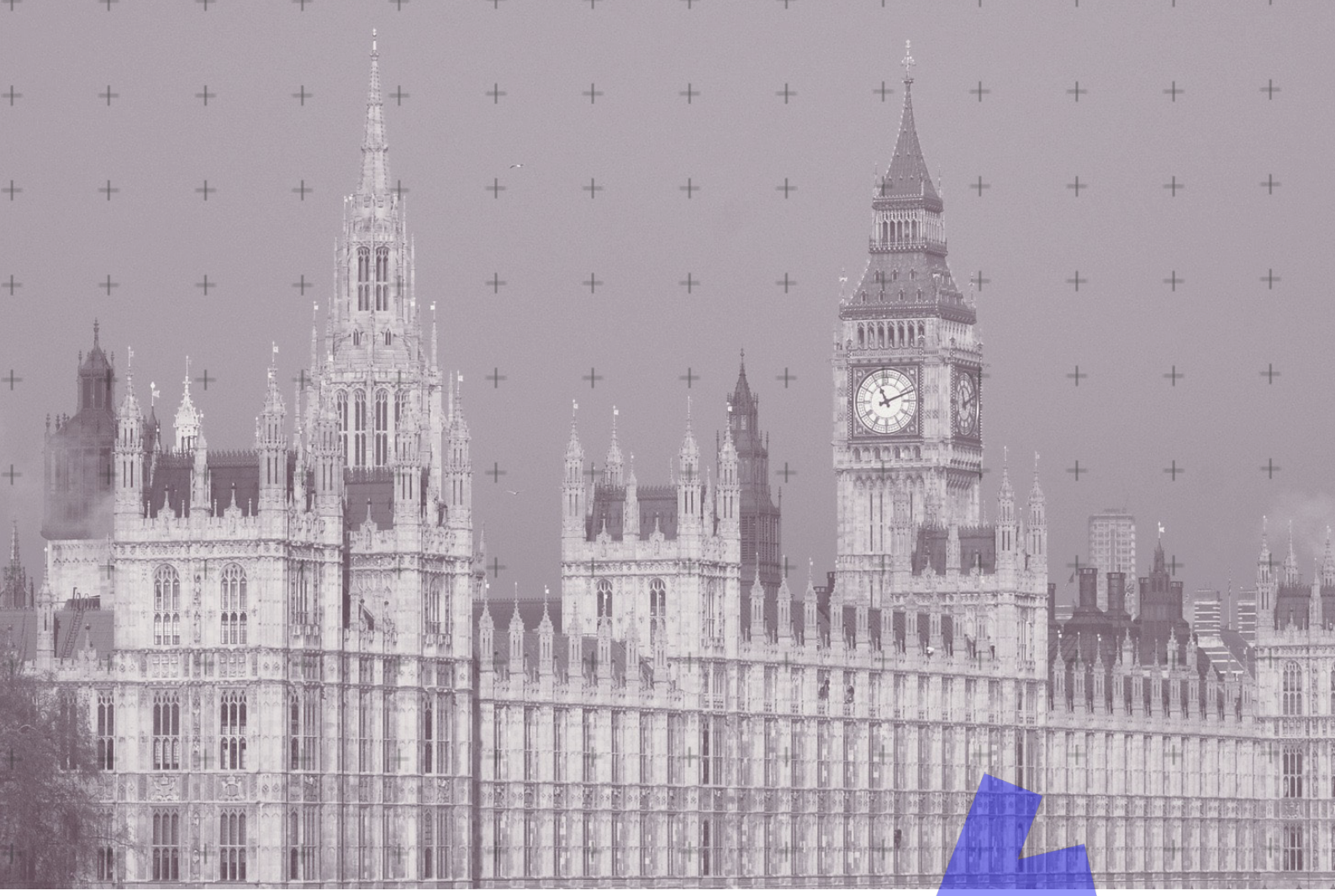

Written Evidence

Submitted to the Joint Committee on the National Security Strategy: DEFENDING DEMOCRACY

March 2024

Dr Ann Kristin Glenster
Stefanie Felsberger
Professor Gina Neff

MINDEROO
CENTRE FOR
TECHNOLOGY
& DEMOCRACY

Written evidence for the Joint Committee on the National Security Strategy (JCNSS) for its inquiry into Defending Democracy

Evidence submitted by the Minderoo Centre for Technology and Democracy at the University of Cambridge.

By Dr Ann Kristin Glenster, Stefanie Felsberger and Professor Gina Neff.

22 March 2024

The Minderoo Centre for Technology and Democracy is an academic research centre at the University of Cambridge, with world-leading expertise in the regulation and governance of emerging technologies. We are also part of the EU Horizon 2020 funded project AI4TRUST: AI-based-technologies for trustworthy solutions against disinformation.

www.mctd.ac.uk

Summary of Submission:

In this submission we make the following points:

- Disinformation online poses a significant threat to the UK's democracy
- AI-generated content, especially deepfakes, poses a particularly potent threat
- Disinformation must be countered with the pursuit of a socio-technical approach
- Public awareness, digital literacy, and media regulation should be strengthened
- Researchers' access to data should be made a priority to ensure that the National Security Strategy is adequately informed
- Article 12 of the National Security Act 2023 can be a tool to combat disinformation
- The free speech and civic participation of women and marginal groups in democratic processes should be strengthened
- Political speech must be protected and balanced with the need for maintaining a functioning public sphere; voluntary codes of conduct across parties should be welcomed

What are the actual and perceived threats to the UK's democracy, and from where do those threats originate?

1. The threats to the UK's democracy are multifaceted and complex. A comprehensive answer to this question is beyond the scope of this submission. Rather, we will focus on the threat from a continued assault on societal trust from the fragmented and often hostile online environment.

2. Overall strengthening the public square and the climate for an independent and reliable media sector is key to fostering trust within society. This requires strengthening media regulation, increasing digital literacy, and enforcing the Online Safety Act.

3. We are concerned about the damage to the public sphere through coordinated, sustained and unequal attacks on some members of society, such as women and members of marginal and vulnerable communities online, who are disproportionately the victims of online abuse. Protecting the freedom of speech of all British citizens is crucial to countering hate speech and propaganda that undermines trust in society and trust in our democratic process and institutions.

What role are emerging technologies, such as generative AI, expected to play in upcoming elections? How will threats to the UK democracy evolve in the medium and long term? How prepared is the UK for addressing these threats?

4. With the use of generative AI becoming cheaper and more accessible, political parties and campaigners are more likely to use generative AI to share and create audio, visual, and textual content: not necessarily with malign intentions but still with the capacity to mislead. This has been evident in recent elections in Indonesia, Pakistan, and India where AI-generated avatars have been utilised to improve politicians' public image or to create videos of deceased politicians expressing support for election campaigns.[1] Voluntary codes of conduct on the use of AI in campaigns across political parties should be welcomed.

5. Nonetheless, generative AI can disrupt democratic processes in several specific ways. Applications built on Large Language Models (LLMs), such as ChatGPT, often provide inaccurate information or make up evidence for specific claims ('hallucinations'). Evidence from the US shows that public-facing ChatGPT is not fit for purpose for election

[1] James Purtill, "AI is Changing How Elections Are Fought, From Deepfake Endorsements to Chatbot Campaigners," ABC News Australia, 29 February 2024, https://www.abc.net.au/news/science/2024-02-21/ai-elections-deepfakes-generative-campaign-endorsement-democracy/103483710.

information.[2] We have seen examples of repeated prompts leading to false information concerning content from journal articles to speeches by politicians.[3] We are concerned that falsely attributing or the manufacturing of inaccurate political content and speech will have a significant impact on upcoming elections.

6. We are also concerned with the potential use of LLMs or generative AI by a range of new and different actors. For example, in scaling up disinformation campaigns, which will target more individuals in fine-tuned ways than has previously been possible. The use of LLMs to spread disinformation poses considerable challenges for fact-checkers, journalists, civil society, and other stakeholders. We advocate for stronger guardrails to ensure that LLMs are not used by malign individuals or organisations for these purposes.

7. These risks are augmented by the lack of available tools and methodologies to counter the spread of disinformation through AI for fact checkers and journalists. While tools to detect photos and videos generated by AI are available, these are often not sufficiently reliable or accurate for fact-checkers to assess 100% veracity and accuracy. Deepfake audio content presents even larger challenges and is often shared on messaging platforms like Telegram or Whatsapp which are less researched and unmoderated.[4]

8. While generative AI will play an important role in upcoming elections, we also draw attention to the difference between actual and perceived threats from generative AI. While LLMs have garnered the public and media's attention, low-tech approaches to and the use of 'cheap fakes' in disinformation continue to pose a significant threat to the integrity of elections. So-called cheap fakes are created through cropping images, sharing images out of context, or altering the speed of videos. These require less technical know-how, have been shared significantly more online, and are documented to have caused more harm over the past years.[5]

---

[2] See Julia Angwin, Alondra Nelson, and Rina Palta, *Seeking Reliable Election Information? Don't Trust AI* (AI Democracy Project Report, 27 February 2024), https://www.ias.edu/sites/default/files/Angwin-Nelson-Palta_SeekingReliableElectionInformationDontTrustAI_2-27-24.pdf.

[3] "Post-Graduate Science Students Break Large Language Model Guardrails at Royal Society AI Safety Event," The Royal Society, 07 November 2023, https://royalsociety.org/news/2023/11/ai-safety-red-teaming/.

[4] Yasmine Houri, Emmanuel Lazega, Camille Roth, Paola Tubaro, Camille Roth, Elena Pavan, Gina Neff, Hugo Leal, and Stefanie Felsberger, *D4.1 Social Dynamics of Mis/Disinformation* (AI4TRUST Report, November 2023).

[5] In 2019, 96 percent of deep fake videos online were pornographic content, see Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen, The State of Deepfakes: Landscape, Threats, and Impact (Deeptrace Report, September 2019), https://regmedia.co.uk/2019/10/08/deepfake_report.pdf. Overall, see Britt Paris and Joan Donovan, *Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence* (Data & Society Report, September 2019), p. 11, https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1.pdf.

What can be done to improve public awareness of disinformation, fraud, and technological interference such as that through AI or deep fakes?

9. Disinformation campaigns attempt to derail democratic accountability and transparency by spreading false and misleading information about political processes.

10. Elections are just one of many democratic processes at risk of being disrupted by disinformation campaigns. It is crucial that the negative impact of disruption campaigns directed at all democratic processes are given adequate attention in the National Security Strategy. These processes include the protection of civil liberties and human rights, accountability and transparency, access to free and independent media, the rule of law, civic education and awareness, and public participation.

11. Second, the use of disinformation to interfere in democratic processes is both a social and a technical problem, which means that it needs to be addressed with socio-technical solutions and approaches. Due to the nature of AI-generated content, social approaches and solutions, such as media literacy training, alone do not suffice. Simultaneously, technical solutions – e.g., detection tools, labelling, or watermarking – alone do not suffice in countering disinformation.

12. It is crucial that independent researchers have access to social media platform data in safe, secure and privacy-protecting ways to increase understanding of the threats and impact of these campaigns, and to monitor for emerging threats. Data access enables a better understanding of scale, tactics, and targets of mis and disinformation campaigns. It also enables an understanding of the effectiveness of countermeasures, such as content labelling, debunking, and other socio-technical tools. Enforceable regulation mandating researcher access to data is therefore key to ensure that academia, media, and civil society can inform policymakers and legislators in shaping the National Security Strategy. Researcher access to data should be a significant concern to the Committee when considering how robust, forward-looking, and accurate the information it relies on is when shaping its understanding of threats to UK democracy.

13. Public awareness must also be raised. This includes the promotion of independent and diverse media, media literacy, sustainable funding for fact-checking, and the development and adoption of better tools to detect and verify AI-generated content (audio, visual, and textual).

How effective is the UK's legislative framework for defending democracy, including the new powers under the National Security Act 2023?

14. In our view, Article 12 on sabotage of the National Security Act 2023 may offer a potent tool to combat disinformation online when these are orchestrated by *foreign powers*. According to Article 12(1)(a), the offence of sabotage is defined as conduct that causes damages to any asset. Article 12(3) specifies that 'asset' may include "electronic systems and information." We argue that "electronic information …" includes information that is intended to influence the outcome of elections, and thus satisfies the test set out in Article 12(1)(c) of being "prejudicial to the safety and interests of the United Kingdom." Thus, it is conceivable that the National Security Act 2023 can be a significant tool in holding individuals to account as part of the National Security Strategy to defend democracy.

How does the Foreign Influence Registration scheme strengthen the resilience of the UK political system against covert foreign influence?

15. Civil liberties and human rights form the cornerstones of free participation of all citizens in democratic processes. Thus, the Foreign Influence Registration scheme is not enough to counter foreign influence. Approaches to countering disinformation or Foreign Information Manipulation and Interference (FIMI), must also be designed to uphold civil liberties and human rights, especially data protection and privacy rights.

16. Furthermore, disinformation campaigns often target marginalised groups in society by leveraging gendered or racialised stereotypes in campaigns,[6] such as was seen in efforts in the US.[7] Protecting members of marginalised and minority groups from disinformation campaigns is vital to protect against domestic and foreign attacks on democratic processes. To this end, we argue platforms need to be encouraged to enhance safety-by-design approaches and upstream solutions, such as improved *human* content moderation, effective handling of user complaints and improved reporting mechanisms.

---

[6] Ellen Judson, Asli Atay, Alex Krasodomski-Jones, Rose Lasko-Skinner, Josh Smith, *Engendering Hate: The Contours of State-Aligned Gendered Disinformation Online* (Demos Report, October 2020), https://demos.co.uk/wp-content/uploads/2023/02/Engendering-Hate-Oct.pdf. and Rita Jonusaite, Maria Giovanna Sessa, Kristina Wilfore, and Lucina Di Meco, *Gender-Based Disinformation 101: Theory, Examples, and Need for Regulation* (EU Disinfo Lab, 12 October 2022),https://www.disinfo.eu/wp-content/uploads/2022/10/20221012_TechnicalDocumentGBD-2.pdf.

[7] See John Kelly, *Statement of Dr. John W. Kelly, Chief Executive Officer Washington*, DC (Briefing for the United States Senate Select Committee on Intelligence, 1 August 2018), https://www.intelligence.senate.gov/sites/default/files/documents/os-jkelly-080118.pdf., and John Kelly, *Responses to Questions for the Record by Dr. John W. Kelly, Chief Executive Officer Washington*, DC (For Senate Select Committee on Intelligence: Foreign Influence Operations and Their Use of Social Media Platforms, 30 August 2018). https://www.intelligence.senate.gov/sites/default/files/documents/Completed Questions for the Record_Kelly.pdf.

In addition, policymakers play a crucial role in holding platforms accountable by leveraging the Online Safety Bill.[8]

Concluding remarks

17. In this submission, we have drawn on our academic research to argue that disinformation, especially when presented in the form of AI-generated content and deepfakes, poses significant challenges to the defence of the UK's democracy. The National Security Strategy must therefore take a broad approach and support the development and adoption of socio-technical solutions that will strengthen civic participation, civil rights, human rights, and the rule of law. Coordinated online harassment campaigns are part of the problem of mis- and disinformation and they target vulnerable or marginalised groups differentially. Protecting democratic participation means ensuring that all voices have the opportunity to participate in public debates. Measures should be adopted to counter foreign influence and manipulation of the media to ensure the integrity of upcoming elections.

[8] Gina Neff and Rumman Chowdhury, "Platforms Are Fighting Online Abuse—but Not the Right Kind," Wired, 28 February 2023, https://www.wired.com/story/platforms-combat-harassment-but-theyre-focusing-on-the-wrong-kind/.

About the Minderoo Centre for Technology and Demoracy

The Minderoo Centre for Technology and Democracy is an independent team of academic researchers at the University of Cambridge, who are radically rethinking the power relationships between digital technologies, society and our planet.

For more information visit www.mctd.ac.uk

Get in touch: email us at minderoo@crassh.cam.ac.uk