The Alan Turing Institute

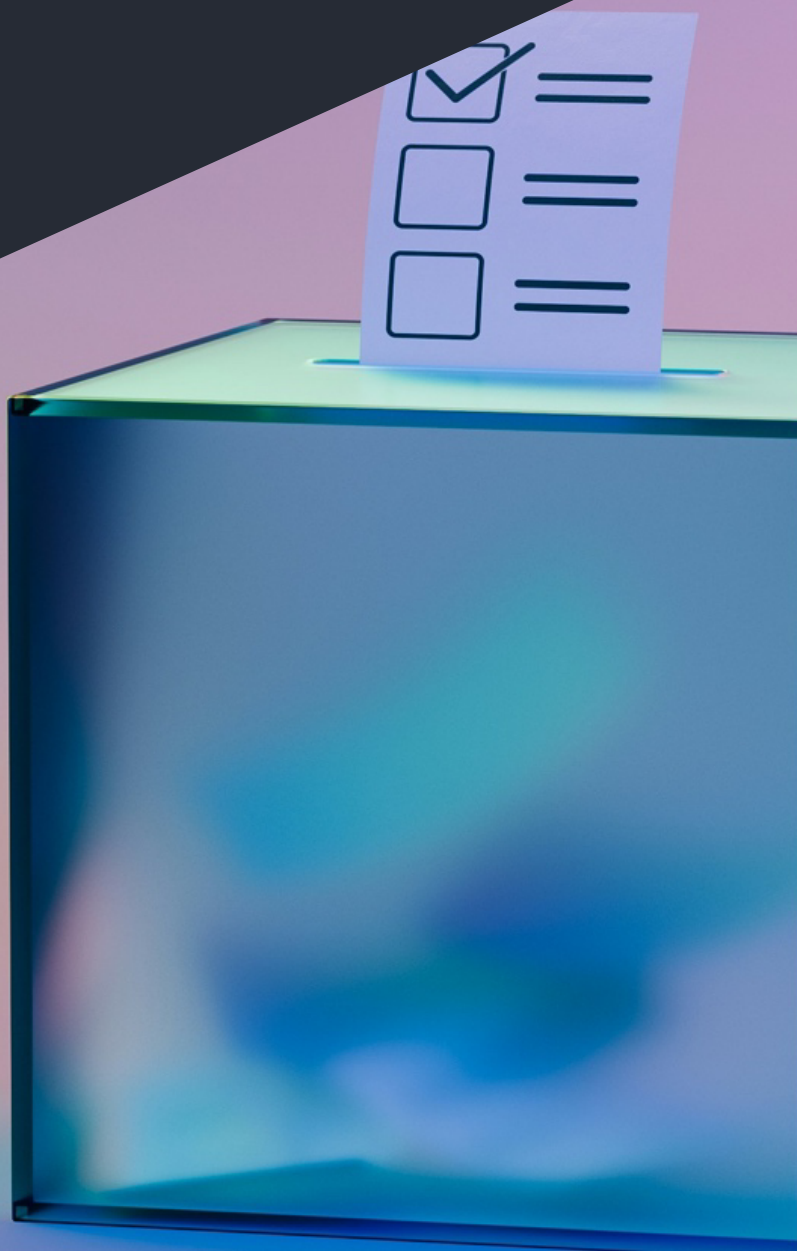
Centre for Emerging Technology and Security

RESEARCH REPORT

AI-Enabled Influence Operations: Safeguarding Future Elections

Sam Stockwell, Megan Hughes, Phil Swatton, Albert Zhang, Jonathan Hall and Kieran

November 2024



About CETaS

The Centre for Emerging Technology and Security (CETaS) is a research centre based at The Alan Turing Institute, the UK's national institute for data science and artificial intelligence. The Centre's mission is to inform UK security policy through evidence-based, interdisciplinary research on emerging technology issues. Connect with CETaS at cetas.turing.ac.uk.

This research was supported by The Alan Turing Institute's Defence and National Security Grand Challenge. All views expressed in this report are those of the authors, and do not necessarily represent the views of The Alan Turing Institute or any other organisation.

Acknowledgements

The authors are grateful to all those who took part in a workshop for this project, without whom the research would not have been possible. The authors are also grateful to: Tony A at the UK's National Cyber Security Centre; Anne-Louise Brown at the Australian Cyber Security Cooperative Research Centre; Dr Jonathan Bright at the Turing's Public Policy Programme; researchers at the Australian Strategic Policy Institute; and Daniel Jordan, Kevin Xu, Alice Crilly and Sam Abbott at the Department for Science, Innovation and Technology for reviewing an earlier version of the report. The figures in this Briefing Paper were designed by Emma Rowlands and Chris Raggett.

Cite this work as: Sam Stockwell, Megan Hughes, Phil Swatton, Albert Zhang, Jonathan Hall and Kieran, "AI-Enabled Influence Operations: Safeguarding Future Elections," *CETaS Research Reports* (November 2024).

Executive Summary

This CETaS Research Report examines hostile influence operations enabled or enhanced by artificial intelligence (AI), and methods to evaluate and counteract such activities during election cycles and beyond. It also includes evidence-based analysis of AI-enabled threats that emerged in the November 2024 US presidential election.

As 2024 draws to a close, more than 2 billion people in at least 50 countries will have voted in the biggest election year in history. At the start of the year, there were significant concerns over the proliferation of new generative AI models, which allow users to create increasingly realistic synthetic content. There has been persistent speculation about how these tools could disrupt key elections this year, many of which will have major consequences for international security.

There was a risk that a lack of empirical work on the impact of the threat would amplify public anxiety about it – which, in turn, could have undermined trust in electoral processes. Therefore, CETaS closely monitored key elections throughout the year, to understand if and how AI misuse affected these processes. As reflected in two Briefing Papers published in May and September 2024, CETaS consistently found no evidence that AI-enabled disinformation had measurably altered an election result in jurisdictions ranging from the UK and the European Union to Taiwan and India.

This final Research Report extends this global analysis to the US election and provides recommendations for protecting the integrity of future democratic processes from AI-enabled threats, with a focus on how UK institutions can counter such activities.

Key findings from the US election specifically are as follows:

- **There is a lack of evidence that AI-enabled disinformation has had a measurable impact on the 2024 US presidential election results.** However, this is primarily due to insufficient data on the impact of such disinformation on real-world voter behaviour. While social media metrics can provide insights into how users engage with this content, more empirical research is needed to understand how it influences large-scale voting intentions.
- **Despite this, deceptive AI-generated content *did* shape US election discourse by amplifying other forms of disinformation and inflaming political debates.** From fabricated celebrity endorsements to allegations against immigrants, viral AI-enabled content was even referenced by some political candidates and received widespread media coverage. Nevertheless, non-AI falsehoods continued to have a significant

impact and could not be ignored. They included: misleading claims by political candidates; conspiracy theories promoted by fringe online groups; and other tools of content manipulation, such as traditional video- and image-editing software.

- **AI-enabled disinformation in the US election was primarily endorsed or amplified by those with pre-existing beliefs aligned with its messages.** Given the extreme political polarisation of US society, the content predominantly helped reinforce prior ideological affiliations among the electorate. This echoes previous CETaS findings that alignment between disinformation and an individual's established political opinions is crucial in their decision to share the content.

Key findings for counteracting AI-enabled influence operations are as follows:

- **Digital literacy, a strong public broadcasting ecosystem and low levels of political polarisation are all factors that can increase public resistance to engagement with disinformation.** Such factors point to the importance of initiatives to foster a healthy information space at both the individual and societal levels.
- **There is no one-size-fits-all framework to evaluate hostile influence operations targeting future election cycles or wider society.** Instead, researchers should weigh the trade-offs between the different tools that are available and use the one most suited to the operation in question. In some cases, combining different frameworks will provide additional insights into these activities.
- **Given the signs that AI-enabled threats began to damage the health of democratic systems globally this year, complacency must not creep into government decision-making.** Ahead of upcoming local, regional and national elections – from Australia and Canada in 2025 to Scotland in 2026 – there is now a valuable opportunity to reflect on the evidence base and identify measures to protect voters.
- **Therefore, this report recommends the following actions to protect elections and wider society from AI-enabled influence operations and other disinformation activities.** These solutions have been informed by an extensive literature review and workshops with 47 cross-sector experts. They centre on the following four strategic objectives designed to help UK institutions target different aspects of the online disinformation process:
    - **Curtail generation** – measures that increase barriers to, or deter actors from, *creating* online disinformation in the first place.
    - **Constrain dissemination** – measures that reduce the effectiveness and virality of disinformation *circulating* on digital platforms.

- **Counteract engagement** – measures that target the ways that users *consume* disinformation on digital platforms, to reduce malicious influence.
- **Empower society** – measures that strengthen societal *capabilities* for exposing and undermining online disinformation.

CETaS UK election security recommendations

Curtail generation

1) **Digital provenance strategy for UK organisations:** The UK Department for Science, Innovation and Technology (DSIT) should establish an implementation strategy for automatically embedding provenance records in digital content produced by the UK Government and other sectors at its origin. This would strengthen the authenticity of credible information sources, and could draw on the US Office of Management and Budget's requirement to issue similar guidance by June 2025.

2) **Authenticity-by-design:** The Internet Engineering Task Force's Security Area should develop and implement authenticity-by-design principles across the internet ecosystem to protect information integrity, using structures such as the Starling Lab framework. The scheme should aim to embed tools into different parts of the internet infrastructure that automatically capture, store and verify digital provenance records securely.

3) **Clarifying existing laws:** The UK Ministry of Justice should conduct a review to understand weaknesses in existing legislation that may be exploited with malicious AI-generated content targeting political candidates or designed to undermine election integrity (including those related to defamation, privacy and electoral laws). This will help the Ministry understand whether existing laws are adequate to deter such activities or whether legislative reforms are required.

Constrain dissemination

4) **Deepfake detection benchmarking and guidance:** The UK AI Safety Institute and the Home Office should coordinate to develop standardised benchmarks and guidance for deepfake detection tools, providing minimum quality assurances for

those using them. The benchmarks should be continuously updated against new deepfake examples to maintain relevance, while the guidance should encourage developers to publish a list of key details before release, including: the purpose and scope of the tool; how it should be used and interpreted; the explainability of its outputs; and its limitations.

5) **Code of conduct on disinformation:** As part of its Phase Three roadmap for the Online Safety Act 2023, Ofcom should create a new Code of Conduct aimed at systematically targeting online disinformation. Drawing inspiration from the EU's Code of Practice on Disinformation, the new code should set out self-regulatory standards for different sectors on demonetising disinformation content creators; define unpermitted manipulative behaviours associated with bot accounts; provide tools for empowering users against disinformation; and require transparent incident reporting.

6) **Political party conduct:** The Electoral Commission should expand existing guidance for UK political parties on both the appropriate use of AI tools and clear redlines on misuse. In turn, political parties should update their internal codes of conduct with this guidance to create accountability for candidates and campaigners.

Counteract engagement

7) **Media validation app tools:** Ofcom should convene major UK communications app providers and the International Fact-Checking Network to design accessible and transparent fact-checking apps for UK users. These could replicate other initiatives, such as Taiwan's LINE app, which helps users verify content by providing trusted alternative news sources for cross-referencing.

8) **Election Incident Protocol:** The Cabinet Office should establish a UK Critical Election Incident Public Protocol based on the Canadian model. Involving a range of senior government experts, the protocol would inform the public of threats considered severe enough to undermine the integrity of elections. Any announcements made through the protocol should be based on a consensus and restricted to informing the public about the incident and how they can protect themselves.

9) **Election advert imprints:** The UK Government should table an amendment to Section 54 of the Elections Act 2022, which deals with imprints on digital campaign material during elections. This should introduce a new transparency provision legally requiring advert content that has been digitally edited to be embedded with secure provenance records detailing how it was edited and by whom.

10) **Decentralised fact-checking:** Social media platforms should invest greater resources in support of decentralised fact-checking initiatives, to help address the volume of disinformation circulated online. These initiatives should incorporate reputation and voting systems to provide quality control of, and a democratic consensus on, user-made notices.

11) **Media reporting guidance:** The Independent Press Standards Organisation should revise its existing guidance on 'reporting major incidents' to include key considerations for coverage of known hostile influence operations and viral disinformation content – drawing on insights from journalists and fact-checkers. Such information could include advice to refrain from linking to the original source content in online articles – thereby discouraging users from sharing it with others – and to frame the impact of the content in a way that does not exaggerate the threat of these activities to the wider public.

Empower society

12) **Regulator review:** DSIT's AI Central Risk Function should coordinate with both the Electoral Commission and Ofcom to analyse potential gaps in their respective

regulatory powers and remits. The review should focus on the effectiveness of both regulators in tackling all forms of online disinformation during elections, in accordance with the Online Safety Act 2023, the Elections Act 2022 and the Representation of the People Act 1983.

13) **Trusted researcher access:** The UK Government should ensure that the Digital Information and Smart Data Bill and other relevant future legislation include a provision for establishing a trusted research group on disinformation. This would require social media platforms to provide trusted members of the UK academic, research and civil society communities with access to data on identified hostile influence operations – akin to X's former data access model. To maintain impartiality, organisations and individuals should be selected by UK Research and Innovation's trusted research and innovation programme.

14) **Convening experts:** Ofcom should prioritise establishing the Advisory Committee on Disinformation and Misinformation, as set out by section 152 of the Online Safety Act 2023, to maintain a long-term focus on tackling disinformation. The committee should have a clear mandate for informing Ofcom's counter-disinformation activities, an independent chair not affiliated with any political party or tech platform, and diverse sectoral representation.

15) **Digital literacy programmes:** The Department for Education and DSIT should coordinate on establishing nationwide digital-literacy and critical-thinking programmes. Any schemes of this kind should be made mandatory in primary and secondary schools, while also being promoted to adults. Such initiatives would seek to improve societal resilience against disinformation and could include topics on: AI and algorithmic bias; deepfakes; evaluating information sources; understanding social media manipulation; and building a culture of content verification.

Introduction

Since CETaS published its Briefing Paper on the UK, EU and French elections in September 2024, most voting processes have concluded without being fundamentally reshaped or disrupted by AI. However, at the time of writing, the pivotal 2024 US presidential election had not taken place. Given the long time frame of the campaign, its narrow poll margins and the differences between the two main candidates on Russia and China policy, many observers believed the election would be the ultimate test of AI-generated disinformation.[1]

Yet as previous CETaS research concluded, there is a need to inform such judgements with evidence-based research and find a balance between assessing the severity of the threat and avoiding fearmongering.[2] AI threat reporting in the contest has focused on unpicking individual viral cases instead of systematic analysis of strategic themes and trends across the election cycle. Only well-grounded research can accurately inform the public and avoid unnecessary speculation.

The 'super year of elections' may be drawing to a close but AI misuse could still emerge in federal elections in Australia and Canada in 2025, as well as in regional elections such as those in Scotland in 2026. There is a risk that efforts to tackle these threats will be deprioritised on the incorrect assumption that, with many national elections now finished, malicious actors will have little incentive for further political interference. But maintaining a healthy information environment is also crucial outside election periods, as evidenced in the UK context by the recent use of disinformation to intensify far-right riots and political extremism.[3]

Therefore, policy responses and other protective measures should not be narrowly focused on securing election cycles only as official campaigning takes place.[4] Instead, they should identify long-term interventions that embed resilience, draw on the capabilities of different

[1] William Turton, "The US Election Threats Are Clear. What to Do About Them Is Anything But," *WIRED*, 15 May 2024, https://www.wired.com/story/election-threats-senate-hearing-ai-disinformation-deepfakes/.
[2] Sam Stockwell et al., "AI-Enabled Influence Operations: The Threat to the UK General Election," CETaS Briefing Papers (May 2024), 39, https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-threat-uk-general-election; Sam Stockwell, "AI-Enabled Influence Operations: Threat Analysis of UK and European Elections," *CETaS Briefing Papers* (September 2024), 6.
[3] Institute for Strategic Dialogue, "From rumours to riots: How online misinformation fuelled violence in the aftermath of the Southport attack," 31 July 2024, https://www.isdglobal.org/digital_dispatches/from-rumours-to-riots-how-online-misinformation-fuelled-violence-in-the-aftermath-of-the-southport-attack/.
[4] Helen Margetts, "The AI election that wasn't – yet," *UK Election Analysis*, https://www.electionanalysis.uk/uk-election-analysis-2024/section-6-the-digital-campaign/the-ai-election-that-wasnt-yet/.

sectors and empower citizens against disinformation. All such steps will help protect future elections – and wider society – against these threats.

Research methodology

This project seeks to answer the following research questions:

- **RQ1**: What factors make individual citizens and society either more vulnerable or resilient to engagement with disinformation, including AI-enabled content?
- **RQ2:** How has AI been maliciously deployed in the lead-up to the 2024 US presidential election?
- **RQ3:** Which existing evaluation frameworks gauge the impact of influence operations, and what are the barriers to effective measurement?
- **RQ4:** What initiatives can the UK implement to enhance election security and broader societal resilience against influence operations that incorporate novel AI tools?

Data collection for this study was conducted between June and November 2024, involving three core research activities:

1. **Literature review** covering journal articles, public reports and news articles on: AI misuse in the 2024 US election; public engagement with AI-enabled disinformation; challenges in evaluating influence operations; and countermeasures for improving election resilience.
2. **Social media analysis** of three different US deepfakes, to understand nodes of influence amplifying disinformation (see Section 2.2 for more details on the methodology used).
3. **Two workshops** designed to prioritise policy and technical recommendations identified by the project team. These sessions invited attendees to determine which solutions were most impactful and feasible in enhancing election resilience against AI threats; they involved 47 experts:
    - 20 from industry.
    - 12 from government and regulatory bodies.

- 10 from civil society.
- 5 from academia.

Report structure

The remainder of this report is structured as follows. Section 1 describes the factors that affect individual and societal engagement with disinformation content. Section 2 provides analysis of specific AI threats in the 2024 US presidential election cycle, as well as social media analysis of a selection of high-profile US deepfakes. Section 3 explores challenges and ways forward in evaluating the impact of these activities. Section 4 describes policy solutions that can help increase democracies' resilience against malicious AI-enabled influence operations. Finally, Section 5 details corresponding technical solutions.

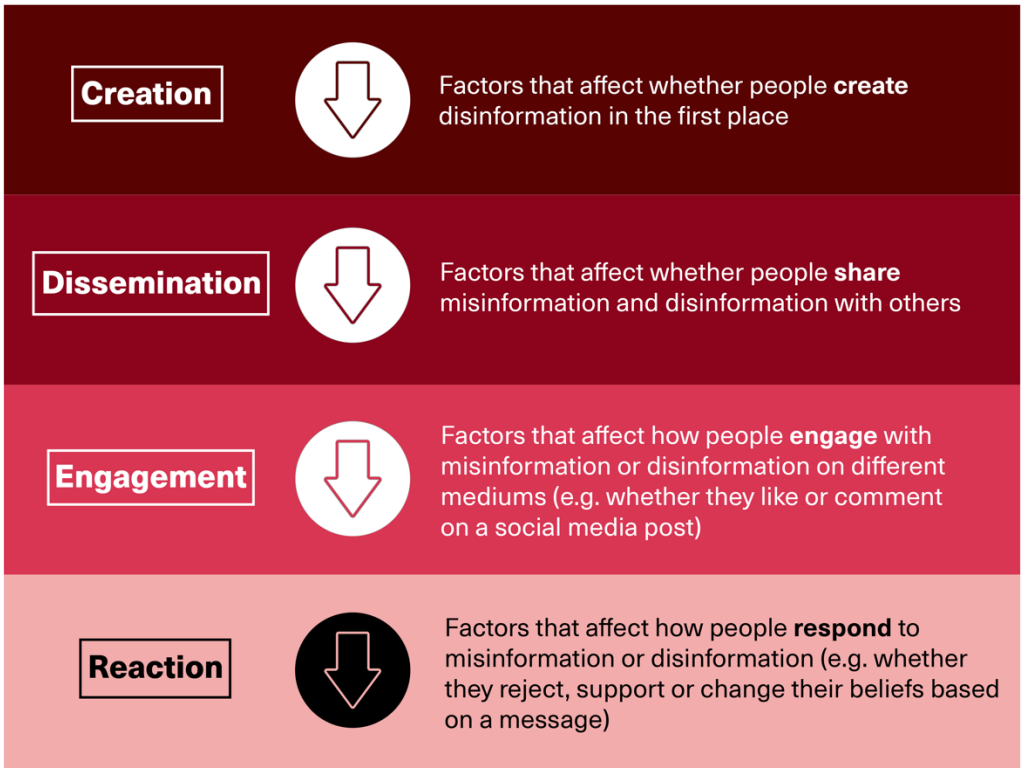
1. Public Vulnerability and Resilience against Deceptive Content

The ubiquity of social media has shifted responsibility for detecting falsehoods from professional journalists to everyday internet users. It is, therefore, important to understand how individuals interact with misinformation and disinformation to reduce public susceptibility to it – both during elections and beyond. CETaS defines misinformation as *unintentionally* misleading claims. In contrast, disinformation is *deliberate* falsehoods, including those shared as part of online influence operations that are intended to shape public opinion or behaviour. The analysis in this section focuses on the risk factors and protective factors that affect public vulnerability and resilience against both misinformation and disinformation.

1.1 Risk factors associated with vulnerability

To understand the impact of misinformation and disinformation on individuals, it is helpful to break down the different stages of the content lifecycle:

Figure 1. Online misinformation and disinformation life cycle



Source: Authors' analysis.

There is only sparse data on the known motivations of people who *generate* misinformation and disinformation.[5] However, some factors have been suggested based on historical cases. These include various foreign and domestic actors' desire to influence election results, sow political division or undermine media integrity in a country, as well as hyper-partisan media outlets' aim to distort facts to suit organisational agendas.[6]

An increasing body of evidence helps explain the reasons why individuals *disseminate*, *engage* and positively *react* to this content. For example, individuals who consume misinformation and disinformation are more likely to have conspiratorial outlooks, distrust public institutions, experience stress and frustration, or lack critical-thinking and information-verification habits.[7] Individuals who exaggerate their knowledge of topics and score lower on tests of analytical thinking are also more likely to believe fake news stories.[8] When it comes to demographics, some studies show that older people are more likely to *share* misinformation or disinformation online when they view it, but younger people – particularly those under the age of eighteen – are more likely to *believe* misleading narratives.[9] Other studies have found that men are more likely than women to disseminate political disinformation.

Users who rely on social media (rather than traditional media) for news and political engagement will also be more likely to encounter misinformation and disinformation – and will, therefore, be more at risk of consuming it. Disinformation may be spread by groups with ideological agendas, such as climate-change deniers, or by those seeking to benefit

[5] Sophie Lecheler and Jana Laura Egelhofer, "Disinformation, Misinformation, and Fake News Understanding the Supply Side" in *Knowledge Resistance in High-Choice Information Environments*, ed. Jesper Strömbäck et al. (Routledge: 2022), 73-80, https://library.oapen.org/bitstream/handle/20.500.12657/54482/1/9781000599121.pdf.
[6] Ibid.
[7] Valentin Stoian-Iordache and Irena Chiru, "2. Aggravating Factors for the Dissemination of Disinformation: 2.1. Individual and group factors" in *Handbook on Identifying and Countering Disinformation*, ed. Christina Ivan et al. (DOMINOES Project: 2023), https://dominoes.ciberimaginario.es/21-individual-factors.html; Jonáš Syrovátka, Nikola Hořejš and Sarah Komasová, "Towards a model that measures the impact of disinformation on elections," *European View* 22, no. 1 (2023), https://doi.org/10.1177/17816858231162677.
[8] Gordon Pennycook and David G. Rand, "Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking," *Journal of Personality* 88, No. 2 (March 2019: 185-200), https://doi.org/10.1111/jopy.12476.
[9] Andrew Guess, Jonathan Nagler and Joshua Tucker, "Less than you think: Prevalence and predictors of fake news dissemination on Facebook," *Science Advances* 5, no. 1 (January 2019); Center for Countering Digital Hate, "Belief in conspiracy theories higher among teenagers than adults, as majority of Americans support social media reform, new polling finds," 16 August 2023, https://www.science.org/doi/10.1126/sciadv.aau4586; https://counterhate.com/blog/belief-in-conspiracy-theories-higher-among-teenagers-than-adults-as-majority-of-americans-support-social-media-reform-new-polling-finds/; Manjul Gupta et al., "Fake news believability: The effects of political beliefs and espoused cultural values," *Information & Management* 60, no. 2 (March 2023), https://www.sciencedirect.com/science/article/pii/S0378720622001537.

financially from social media advertising revenue on high-engagement posts.[10] Interestingly, users are likely to perceive themselves as less vulnerable to the impact of deepfakes than others.[11] Characterised as third-person perception bias, this phenomenon could have at least two implications: individuals overestimate their ability to identify deepfakes, and are unlikely to cross-reference sources; and individuals underestimate others' ability to detect deepfakes, and perceive such content as having more influence on voters than it actually does.[12]

Crucially, any individual is vulnerable to engaging with (and believing) misinformation and disinformation when it supports their established opinions or worldview.[13] Accordingly, misinformation and disinformation are more likely to enhance existing perspectives rather than change them,[14] as evidenced by CETaS analysis of AI-generated falsehoods in the US election campaign (see Section 2). Individuals knowingly share fake news when it conforms to their prior viewpoints for a variety of other reasons, including a desire to maintain social relations and a sense of belonging, seek entertainment or engage in debates.[15]

1.2 The effects of AI on risk factors

Generative AI can lead to greater exposure to misinformation and disinformation, given its capacity to disseminate both larger volumes of content and more personalised and realistic fake content. However, online disinformation has historically been concentrated in small proportions of platform users. During the 2016 US presidential election, 1% of roughly 2

[10] Sebastian Valenzuela et al., "The Paradox of Participation Versus Misinformation: Social Media, Political Engagement, and the Spread of Misinformation," *Digital Journalism* 7, No. 6 (June 2019: 802-823), https://doi.org/10.1080/21670811.2019.1623701; Eva Surawy Stepney and Clare Lally, *Disinformation: sources, spread and impact* (UK Parliament POST: 25 April 2024), https://researchbriefings.files.parliament.uk/documents/POST-PN-0719/POST-PN-0719.pdf.

[11] Saifuddin Ahmed, "Examining public perception and cognitive biases in the presumed influence of deepfakes threat: empirical evidence of third person perception from three studies," *Asian Journal of Communication* 33, No. 3 (November 2021: 308-331), https://doi.org/10.1080/01292986.2023.2194886.

[12] Ibid.

[13] Cornelia Sindermann, Andrew Cooper and Christian Montag, "A short review on susceptibility to falling for fake political news," *Current Opinion in Psychology* 36 (December 2020: 44-48), 46-47 https://doi.org/10.1016/j.copsyc.2020.03.014; Gillian Murphy et al., "False Memories for Fake News During Ireland's Abortion Referendum," *Psychological science* 30, No. 10 (August 2019: 1449–1459) https://doi.org/10.1177/0956797619864887; Gordon Pennycook and David G. Rand, "The Psychology of Fake News," *Trends in Cognitive Science* 25, No. 5 (May 2021: 388-402), 399, https://www.cell.com/trends/cognitive-sciences/fulltext/S1364-6613(21)00051-6?rss=yes&fbclid=IwAR2-SBHPbK-NV-ShyhlJOeermbdp4njIMOhqT59XuQNn1f58qG-GpZtwpso.

[14] Ruben Arcos et al., "Responses to digital disinformation as part of hybrid threats: a systematic review on the effects of disinformation and the effectiveness of fact-checking/debunking," *Open Research Europe* 2, No. 8 (January 2022), https://open-research-europe.ec.europa.eu/articles/2-8.

[15] Miriam J. Metzger et al., "From Dark to Light: The Many Shades of Sharing Misinformation Online," *Media and Communication* 9, No. 1 (February 2021: 134–143), 136, https://www.cogitatiopress.com/mediaandcommunication/article/view/3409/1955.

million Twitter (now X) users were exposed to 80% of all fake news posts analysed by researchers, while only 0.1% of those users were responsible for sharing 80% of fake news posts.[16]

Nevertheless, social media companies employ recommender systems based on users' engagement with and time spent on their platforms, aiming to maximise advertising revenue. These dynamics may lead to increased exposure to misinformation and disinformation in two ways. Firstly, algorithms built around popularity-based metrics are vulnerable to manipulation by social bots.[17] Bots that repeatedly engage with content can drive misinformation and disinformation to users' feeds. Secondly, the rise of platforms such as TikTok – which uses interest graphs rather than social graphs, and surfaces personalised content based on individuals' online behaviours rather than content produced and shared within a network – could lead users to view content they would not usually see on other platforms.[18] While some platforms restrict political content from unknown accounts to prevent the spread of disinformation, social media is quickly becoming one of the most popular channels for users to receive their news.[19] Given this, political misinformation and disinformation may no longer be confined to a small proportion of users.

AI not only increases the likelihood of exposure to misinformation and disinformation online, but also allows for more personalised political messaging. Microtargeting involves the analysis of personal data on groups or individuals (such as browsing habits or social interactions) to personalise content and influence users' actions.[20] Like microtargeting, so-called 'message-tailoring' can be effective in various circumstances, such as efforts to influence voter turnout (depending on the alignment between message and audience)[21] and to prevent voter defection.[22] Some off-the-shelf generative AI models can be used to

[16] Nir Grinberg et al., "Fake news on Twitter during the 2016 U.S. presidential election," *Science* 363, No. 6425 (January 2019: 374-378), https://www.science.org/doi/10.1126/science.aau2706.
[17] Emilio Ferrara et al., "The Rise of Social Bots," *Communications of the ACM* 59, No. 7 (July 2016: 96-104), 98, https://dl.acm.org/doi/pdf/10.1145/2818717.
[18] Kazuki Nakayashiki, "Letting the Interest Graph Guide You," *Medium*, 7 January 2022, https://medium.com/@kazuki_sf_/letting-the-interest-graph-guide-you-faf5e30c178a; Zeve Sanderson, Solomon Messing and Joshua A. Tucker, "Misunderstood mechanics: How AI, TikTok, and the liar's dividend might affect the 2024 elections," *Brookings Institution*, 22 January 2024, https://www.brookings.edu/articles/misunderstood-mechanics-how-ai-tiktok-and-the-liars-dividend-might-affect-the-2024-elections/.
[19] Alex Heath, "Meta's new opt-out setting limits visibility of politics on Instagram and Threads," *The Verge*, 25 March 2024, https://www.theverge.com/2024/3/25/24111604/meta-setting-downranks-politics-instagram-threads; Ofcom, "TV loses its crown as main source for news," 25 September 2024, https://www.ofcom.org.uk/media-use-and-attitudes/attitudes-to-news/tv-loses-its-crown-as-main-source-for-news/?language=en.
[20] Information Commissioner's Office, "Microtargeting," https://ico.org.uk/for-the-public/microtargeting/.
[21] Katherine Haenschen, "The Conditional Effects of Microtargeted Facebook Advertisements on Voter Turnout," *Political Behavior* 45 (2023: 1,661-1,681), 1,675-1,679, https://doi.org/10.1007/s11109-022-09781-7.
[22] Mathieu Lavigne, "Strengthening ties: The influence of microtargeting on partisan attitudes and the vote," *Party Politics* 27, No. 5 (2021: 965-976), https://journals.sagepub.com/doi/abs/10.1177/1354068820918387.

automate microtargeting, making it easier for creators to produce targeted disinformation at scale.[23]

However, there is a lack of conclusive research into the *impact* of targeted AI-generated disinformation. In many cases, the persuasive effect of microtargeted messaging generated by a large language model is not statistically different to non-microtargeted messaging generated by the same tools.[24] Regardless, AI-generated messages can have similarly persuasive effects as human-generated messages, and may already be catching up to the "capacity of everyday people."[25] Anthropomorphic cues, in particular, can help make AI-generated messaging more persuasive, such as through ongoing message interactivity and the use of human characteristics or images.[26]

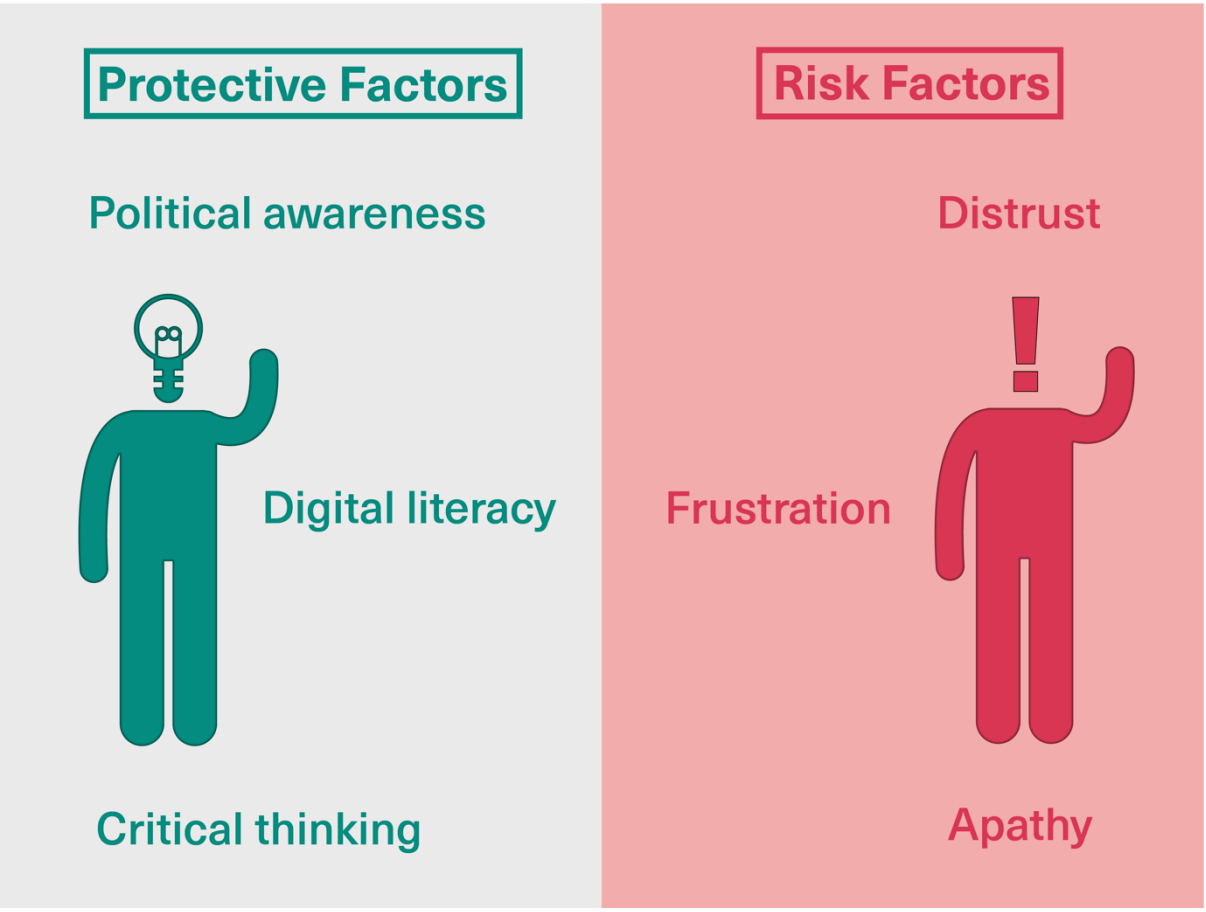
1.3 Protective factors associated with resilience

While no individual is immune to engagement with misinformation and disinformation, those who critically assess both the credibility and quality of information are less likely to believe falsehoods than those who do not.[27] People who seek to inform or persuade others are also less likely to intentionally share misleading claims – though the extent to which this applies in an election context is unclear.[28] More specifically, domain knowledge (i.e. expertise in a subject) and proficiency in digital hygiene can significantly improve an individual's ability to detect fake news.[29] Similarly, those demonstrating a combination of general knowledge about politics, digital literacy and a propensity for cognitive reflection performed best in a deepfake detection experiment.[30] Such protective factors point to the importance of cross-referencing news sources, and of education initiatives that teach users to question and critically assess online information (see Figure 2 below).

---

[23] Almog Simchon, Matthew Edwards, Stephan Lewandowsky, "The persuasive effects of political microtargeting in the age of generative artificial intelligence," *PNAS Nexus* 3, No. 2 (February 2024: 1-5), 3, https://doi.org/10.1093/pnasnexus/pgae035.
[24] Kobi Hackenburg and Helen Margetts, "Evaluating the persuasive influence of political microtargeting with large language models," *PNAS* 121, No. 24 (May 2024), https://www.pnas.org/doi/10.1073/pnas.2403116121.
[25] Hui Bai et al., "Artificial Intelligence Can Persuade Humans on Political Issues," *OSF Preprints*, October 2023, 7, https://osf.io/preprints/osf/stakv.
[26] Yunju Kim & Heejun Lee, "The Rise of Chatbots in Political Campaigns: The Effects of Conversational Agents on Voting Intention," *International Journal of Human–Computer Interaction* 39, No. 20 (2023: 3984-3995), https://www.tandfonline.com/doi/full/10.1080/10447318.2022.2108669.
[27] Aljaž Zrnec, Marko Poženel and Dejan Lavbič, "Users' ability to perceive misinformation: An information quality assessment approach," *Information Processing and Management 59*, No. 1 (January 2022), 10-13, https://www.sciencedirect.com/science/article/pii/S0306457321002211#b41.
[28] Metzger et al. (2021).
[29] Zrnec et al. (2022).
[30] Sara Savat, "Political deepfake videos no more deceptive than other fake news, research finds," *Phys.org*, 19 August 2024, https://phys.org/news/2024-08-political-deepfake-videos-deceptive-fake.html.

Figure 2. Resilience and vulnerability to misinformation and disinformation



Source: Authors' analysis.

Beyond specific individual factors, there are also societal dynamics that affect citizens' resilience against deceptive content. For example, a strong and trusted information environment can insulate the public from misinformation and disinformation. In countries where there is a fragmented media landscape consisting of a weak public broadcasting service and a number of prominent populist news outlets, citizens are more vulnerable to exposure to harmful falsehoods.[31] Indeed, during the 2019 UK general election, the more that users received campaign news via professional news organisations rather than social media sources, the more they could distinguish true information from false content.[32] Therefore, it is crucial to build trust in professional news and public broadcasting outlets.

[31] Edda Humprecht, Frank Esser, and Peter Van Aelst, "Resilience to Online Disinformation: A Framework for Cross-National Comparative Research," *The International Journal of Press/Politics* 25, No. 3 (January 2020: 493-516), https://doi.org/10.1177/1940161219900126.

[32] Cristian Vaccari, Andrew Chadwick and Johannes Kaiser, "The Campaign Disinformation Divide: Believing and Sharing News in the 2019 UK General Election," *Political Communication* 40, No. 1 (September 2022: 4-23), 4, https://www.tandfonline.com/doi/full/10.1080/10584609.2022.2128948.

Initiatives such as the Reuters Institute's Trust in News project hold lessons in how to do so.[33]

The extent of political partisanship within a society can also affect people's resilience against misinformation and disinformation. Although citizens from different parts of the political spectrum may agree that the proliferation of fake news poses challenges, entrenched divisions lead to fundamental disagreements over who is responsible for spreading falsehoods.[34] These tendencies can not only foster unhealthy scepticism and automatic dismissal of news sources that go against the prior worldviews of respective political groups, but can also reinforce echo chambers where trusted voices are marginalised.[35] There are several promising responses to this issue (such as deliberative polling), but there is also a need for more research into understanding the effectiveness of these interventions at scale, since many of them have only been tested in small-scale trials.[36]

[33] Sayan Banerjee et al., "Strategies for building trust in news: What the public say they want across four countries," *Reuters Institute for the Study of Journalism*, 21 September 2023, https://reutersinstitute.politics.ox.ac.uk/strategies-building-trust-news-what-public-say-they-want-across-four-countries.

[34] Kaitlin Peach et al., "Seeing lies and laying blame: Partisanship and U.S. public perceptions about disinformation," *HKS Misinformation Review*, 14 February 2024, https://misinforeview.hks.harvard.edu/article/seeing-lies-and-laying-blame-partisanship-and-u-s-public-perceptions-about-disinformation.

[35] Morgan Kelly, "Political polarization and its echo chambers: Surprising new, cross-disciplinary perspectives from Princeton," *Princeton University*, 9 December 2021, https://www.princeton.edu/news/2021/12/09/political-polarization-and-its-echo-chambers-surprising-new-cross-disciplinary.

[36] Saima May Sidik, "How to tackle political polarization – the researchers trying to bridge divides," *Nature*, 1 March 2023, https://www.nature.com/articles/d41586-023-00573-5.

2. AI-Enabled US Election Threat Analysis

This section presents two case studies. The first is a qualitative analysis of AI-enabled election threats throughout the US election campaign. The second is a network analysis of three US election deepfakes, designed to identify the key nodes of influence on social media platforms that were amplifying disinformation.

2.1 Qualitative analysis of AI-enabled US election threats

The first case study builds on CETaS research into the AI threat landscape in the 2024 UK and European elections.[37] It presents a breakdown of cases by threat category and the actors responsible, their likely intentions and methods, and the impact of these cases.

2.1.1 Smear campaigns

Table 1. AI-enabled smear campaigns identified in the US election

Summary	Instances reported[38]	Impact
AI-generated videos, images and audio of political candidates	24	High user engagement with fake content amplified the disinformation.[41]

[37] Stockwell (2024).

[38] Based on cited examples in news articles and public reports between 22 May and 8 November 2024.

[41] McCarthy (2024a); Laws (2024); Flynn Nicholls, "Donald Trump Paints DNC As Communist Rally in New Kamala Harris Attack Ads," *Newsweek*, 18 August 2024, https://www.newsweek.com/donald-trump-paints-dnc-communist-rally-new-kamala-harris-attack-ads-1940811.

making false or controversial statements.[39] AI-generated videos and images of political candidates depicted in contentious activities and fake medical conditions.[40]		Users' uncertainty about the authenticity of content damaged trust in the integrity of online sources.[42]

[39] Bill McCarthy (a), "Biden deepfake spreads online after withdrawal from 2024 race," *AFP Fact Check*, 22 July 2024, https://factcheck.afp.com/doc.afp.com.364R2N9; Jasmine Laws, "Kamala Harris Deepfake Removed By TikTok After Going Viral," *Newsweek*, 24 July 2024, https://www.newsweek.com/kamala-harris-deepfake-removed-tiktok-viral-1929003; Monir Ghaedi, "Fact check: Viral video claims Biden-Harris call made by AI," *DW*, 24 July 2024, https://www.dw.com/en/fact-check-viral-video-claims-biden-harris-call-made-by-ai/a-69753771; NewsGuard (a), "Russian Deep Fake: "Obama" Admits Dems Planned Trump Shooting," 12 August 2024, https://www.newsguardrealitycheck.com/p/russian-deep-fake-obama-admits-dems; Ara Eugenio, "Video of Trump slamming supporters of the Philippines' Duterte is AI-generated," *AFP Fact Check*, 21 August 2024, https://www.msn.com/en-us/news/politics/video-of-trump-slamming-supporters-of-the-philippines-duterte-is-ai-generated/ar-AA1pi5ga; Bill McCarthy, "Prof. Hany Farid Reveals Video of Biden Botching Ukraine History Is a Deepfake," *UC Berkeley School of Information*, 22 March 2024, https://www.ischool.berkeley.edu/news/2024/prof-hany-farid-reveals-video-biden-botching-ukraine-history-deepfake; Wenhao Ma, "AI videos of US leaders singing Chinese go viral in China," *VOA News*, 16 September 2024, https://www.voanews.com/a/ai-videos-of-us-leaders-singing-chinese-go-viral-in-china/7787160.html; Clint Watts (b), "As the U.S. election nears, Russia, Iran and China step up influence efforts," *Microsoft Threat Analysis Center*, 23 October 2024, https://blogs.microsoft.com/on-the-issues/2024/10/23/as-the-u-s-election-nears-russia-iran-and-china-step-up-influence-efforts/.

[40] Euronews, "Deepfake claiming Kamala Harris was a sex worker circulating less than a day after her first rally," 24 July 2024, https://www.euronews.com/next/2024/07/24/deepfake-claiming-kamala-harris-was-a-sex-worker-circulating-less-than-a-day-after-her-fir; Jordan Liles, "Fact Check: Fake Photo of Trump, Gunman Thomas Crooks Planning Assassination Attempt Generated by AI," *Snopes*, 1 August 2024, https://www.snopes.com/fact-check/trump-crooks-fbi-photo/; Nicholls (2024); Ghaedi (2024); Alex Oliveira, "Elon Musk's Grok AI is flooding social media with absolutely wild images of Trump and Harris," *New York Post*, 22 August 2024, https://nypost.com/2024/08/22/us-news/elon-musks-grok-ai-is-flooding-social-media-fake-images-of-trump-and-harris/; Bill McCarthy (b), "AI-generated Donald Trump image spreads after guilty verdict," *AFP Fact Check*, 4 June 2024, https://factcheck.afp.com/doc.afp.com.34UR4UB; Clare Duffy, "Elon Musk's AI photo tool is generating realistic, fake images of Trump, Harris and Biden," *CNN*, 16 August 2024, https://edition.cnn.com/2024/08/15/tech/elon-musk-x-grok-ai-images/index.html; Niamh Ancell, "Elon Musk incites AI image battle after posting fake picture of Kamala Harris on X," *Cyber Security News Today*, 3 September 2024, https://cybernews.com/ai-news/elon-musk-kamala-harris-fake-image-artificial-intelligence/; Maggie Harrison Dupré, "Trump Posts AI-Generated Image of Kamala Harris as Joseph Stalin, But Instead It Just Looks Like Mario," *Futurism*, 3 September 2024, https://uk.news.yahoo.com/trump-posts-ai-generated-image-203616303.html; True Media, "Secret Service Smiling After Assassination Attempt," https://detect.truemedia.org/media/analysis?id=pta1lc10gmjD2TEcisPkcazVDHc.jpeg; Chris Mueller, "Kamala Harris with Jeffrey Epstein? No, images are altered | Fact check," *USA Today*, 25 July 2024, https://www.msn.com/en-us/news/politics/kamala-harris-with-jeffrey-epstein-no-images-are-altered-fact-check/ar-BB1qQlnX; Philip Marcelo, "Image claiming to show Trump dancing with underage girl is fake," *AP News*, 23 June 2023, https://apnews.com/ap-fact-check/image-claiming-to-show-trump-dancing-with-underage-girl-is-fake-00000188e8ebdc10ad9bf8eb72850000; David Gilbert (a), "Russian Propaganda Unit Appears to Be Behind Spread of False Tim Walz Sexual Abuse Claims," *WIRED*, 21 October 2024, https://www.wired.com/story/russian-propaganda-unit-storm-1516-false-tim-walz-sexual-abuse-claims/; Taija PerryCook, "Image of Trump and Epstein on Private Plane Isn't Real," *Snopes*, 16 August 2024, https://www.snopes.com/fact-check/fake-trump-epstein-on-private-plane/.

[42] Laws (2024); Eugenio (2024).

Actors, motives and tradecraft

CETaS identified far higher – albeit still relatively moderate – volumes of AI-enabled viral disinformation in the US campaign cycle than in elections earlier in the year. There are likely several factors that explain this difference, including:

1) The high levels of political polarisation and division in the US, leading to a greater risk of domestic AI misuse.

2) Marginal poll differences between the major political candidates, encouraging those seeking to influence election results to disseminate greater volumes of disinformation.

3) The major political parties' contrasting policy positions on key campaign issues – such as Russia's war in Ukraine and US relations with NATO – attracting foreign interference.

4) A far longer campaign cycle compared to snap elections in the UK and France, allowing more time for malicious actors to plan and implement hostile influence operations.

As in the UK general election, smear campaigns – or deepfake content implicating political candidates in controversial but fabricated activities and statements – were the most prevalent category of AI disinformation that emerged in the US election. Involving a mixture of video, image and audio formats, these deepfakes presented candidates as making expletive-filled remarks, using firearms, consuming illegal drugs or suffering from critical medical conditions.[43] The July 2024 assassination attempt on then Republican candidate (and now president-elect) Donald Trump led to the emergence of a handful of deepfakes that reinforced conspiracy theories surrounding the incident.[44]

In several cases, much of the misleading content was created by or went viral thanks to domestic political commentators and social media influencers with similar ideological leanings, who sometimes combined it with conspiracy theories or inflammatory language.[45] US officials have accused Russia of relying on "witting and unwitting Americans to promote

[43] McCarthy (2024a); McCarthy (2024b); Oliveira (2024); Duffy (2024).
[44] Liles (2024); NewsGuard (2024a).
[45] Olivia Little, "Fake Harris audio spreads like wildfire on TikTok after Biden's announcement," *Media Matters for America*, 22 July 2024, https://www.mediamatters.org/tiktok/fake-harris-audio-spreads-wildfire-tiktok-after-bidens-announcement; Ghaedi (2024); Euronews (2024).

and add credibility to narratives that serve [Kremlin] interests."[46] Indeed, some political influencers have already claimed that Russian state broadcasters tricked them into promoting such content.[47] Corresponding to trends CETaS previously identified, this highlights the vital role that influential human users, rather than the AI-generated content itself, play in increasing others' exposure to disinformation.[48]

Smear campaigns in the contest also involved a tactic CETaS previously noted in which AI content is integrated with credible media branding to increase its perceived authenticity – thereby encouraging users to reshare the sources without questioning their origin.[49] This included two efforts targeting President Joe Biden: a news outlet's logo was used in a deepfake video, and AI-generated medical images were embedded in the format of a professional news article.[50]

Impact

Deepfakes targeting US politicians saw high levels of user engagement on social media platforms, with some fake videos of Biden and Democratic Party candidate Kamala Harris receiving millions of views.[51] This viral content amplified harmful falsehoods across digital communities and polluted the online information environment, undermining voters' capacity to be informed with facts.[52] Indeed, one recent survey showed that 48% of US respondents felt influenced by deepfakes targeting political candidates in relation to who they voted for in the election.[53]

Despite this, one must exercise caution about the extent to which those who were engaging with or being persuaded by this content were already aligned with its underlying political narratives.[54] In many viral cases, the most influential accounts that increase the reach of

[46] Christopher Bing, Katie Paul and Raphael Satter, "Russia focusing on American social media stars to covertly influence voters," *Reuters*, 9 September 2024, https://www.reuters.com/world/russia-focusing-american-social-media-stars-covertly-influence-voters-2024-09-09/; David Klepper, "Russia is relying on unwitting Americans to spread election disinformation, US officials say," *AP News*, 30 July 2024, https://apnews.com/article/russia-trump-biden-harris-china-election-disinformation-54d7e44de370f016e87ab7df33fd11c8.

[47] Phil McCausland, "Right-wing US influencers say they were victims of alleged Russian plot," *BBC News*, 5 September 2024, https://www.bbc.co.uk/news/articles/crrlv7jdnq8o.

[48] Stockwell (2024), 21.

[49] Stockwell et al. (2024), 30; Klepper (2024).

[50] McCarthy (2024a); Ghaedi (2024).

[51] McCarthy (2024a); Laws (2024).

[52] Krasodomski (2024).

[53] Jasdev Dhaliwal, "How To Survive the Deepfake Election with McAfee's 2024 Election AI Toolkit," *McAfee Blog*, 28 October 2024, https://www.mcafee.com/blogs/internet-security/how-to-survive-the-deepfake-election-with-mcafees-2024-election-ai-toolkit/.

[54] Stockwell (2024),14-17.

these misleading claims were outspoken supporters of the political candidate who benefited from the content, often reposting deepfakes from like-minded users.[55] Therefore, rather than swaying large numbers of undecided voters, such disinformation more likely consolidated pre-existing beliefs – including discriminatory views of women.[56] As highlighted in Section 1, this is not only problematic in itself but also exacerbates wider societal challenges, such as deepening political polarisation, and plays into the hands of malicious actors by weakening societal cohesion.

Some users reposted AI-generated smear campaigns in confusion over whether they were genuine, further amplifying disinformation.[57] In line with observations CETaS made in the UK and European elections, this has detrimental effects beyond the election context.[58] There is a risk that, as users' confidence in their capacity to identify synthetic media declines, so does their trust in the integrity of online sources.

2.1.2 Voter targeting

Table 2. AI-enabled voter targeting efforts identified in the US election

Summary	Instances reported[59]	Impact
Kremlin- and Beijing-affiliated AI bot farms mimicking US voters, spreading disinformation on campaign issues and analysing	14	High user engagement with fake content amplified the disinformation.[63] Beijing-affiliated microtargeting of specific political candidates failed to garner meaningful voter engagement.[64]

[55] Little (2024); Ghaedi (2024); Euronews (2024).
[56] Little (2024); Ghaedi (2024); Euronews (2024).
[57] Laws (2024); Eugenio (2024).
[58] Stockwell et al. (2024), 3.
[59] Based on cited examples in news articles and public reports between 22 May and 8 November 2024.
[63] James Titcomb, "Bots push conspiracy theory that Trump shooting was staged," *Telegraph*, 15 July 1024, https://www.telegraph.co.uk/business/2024/07/15/bots-push-conspiracy-theory-trump-shooting-was-staged/; Clint Watts (a), "China tests US voter fault lines and ramps AI content to boost its geopolitical interests," *Microsoft Threat Analysis Center*, 4 April 2024, https://blogs.microsoft.com/on-the-issues/2024/04/04/china-ai-influence-elections-mtac-cybersecurity/; Sherman (2024).
[64] Microsoft Threat Analysis Center (a), "Russia, Iran, and China engaging in influence activity in final weeks before Election Day 2024," 23 October 2024, https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/MTAC-Election-Report-5-on-Russian-Influence.pdf.

divisive positions for amplification.[60] Social bots spreading conspiracy theories and disinformation on campaign issues, security incidents and baseless claims of election fraud.[61] Social bots conducting smear campaigns against specific political candidates and bolstering support for others.[62]		

Actors, motives and tradecraft

Mirroring recent elections, the US presidential campaign was subject to several high-profile voter-targeting efforts that used automated social media bots.[65] Although there was at least one operation linked to domestic users, the vast majority had the hallmarks of foreign hostile interference – including links to Russia and China.[66] One investigation found that pre-

[60] Wending (2024); Margarita Franklin et al., "Adversarial Threat Report: Second Quarter," *Meta*, August 2024, https://transparency.meta.com/en-gb/integrity-reports-q2-2024/; The Graphika Team, "The #Americans: Chinese State-Linked Influence Operation Spamouflage Masquerades as U.S. Voters to Push Divisive Online Narratives Ahead of 2024 Election," August 2024, https://public-assets.graphika.com/reports/graphika-report-the-americans.pdf; Val Dockrell, "Thousands of fake accounts targeting U.S. voters exposed," *National Security News*, 26 August 2024, https://nationalsecuritynews.com/2024/08/thousands-of-fake-accounts-targeting-u-s-voters-exposed/.

[61] Titcomb (2024); Global Witness, "No ifs, many bots? Partisan bot-like accounts continue to amplify divisive content on X, generating over 4 billion views since the UK general election was called," 31 July 2024, https://www.globalwitness.org/en/campaigns/digital-threats/no-ifs-many-bots-partisan-bot-accounts-continue-amplify-divisive-content-x-generating-over-4-billion-views-uk-general-election-was-called/; DFRLab, "Inauthentic Chinese X accounts amplifying Trump shooting and Biden withdrawal conspiracy theories," 30 July 2024, https://dfrlab.org/2024/07/30/china-x-trump-biden-harris/; Katie Polglase et al., "'My identity is stolen': Photos of European influencers used to push pro-Trump propaganda on fake X accounts," *CNN*, 28 August 2024, https://edition.cnn.com/2024/08/28/europe/fake-maga-accounts-x-european-influencers-intl-cmd/index.html; Mike Wending, "FBI issues warning over two fake election videos," *BBC News*, 2 November 2024, https://www.bbc.co.uk/news/articles/cly2qjel083o; Amy Sherman, "Video shows the FBI reported that "three linked groups have been apprehended for rigging early voting by mail in ballot," *PolitiFact*, 3 November 2024, https://www.politifact.com/factchecks/2024/nov/03/tweets/this-video-about-ballot-fraud-is-not-from-the-fbi/.

[62] Darren Linvill and Patrick Warren, "Digital Yard Signs: Analysis of an AI Bot Political Influence Campaign on X" in *Media Forensics Hub Reports* 7, 30 September 2024, https://open.clemson.edu/mfh_reports/7/; Darren Linvill and Patrick Warren, "Hub Brief: Spamouflage Targeting of U.S. Senator Marco Rubio" in *Media Forensics Hub Reports* 8, 14 October 2024, https://regmedia.co.uk/2024/10/21/clemson_university_spamouflage_targeting_senator_marco_rubio.pdf/; Sarah Staffen, "AI-driven bot network trying to help Trump win US election," *DW*, 5 November 2024, https://www.dw.com/en/2024-us-election-ai-driven-social-media-bots-donald-trump-supporters/a-70695526.

[65] Stockwell et al. (2024), 24-27; Stockwell (2024), 15-17.

[66] DFRLab (2024); Wending (2024).

existing bot accounts connected with disinformation content during the UK election campaign pivoted to spreading falsehoods about US politics after the former ended.[67]

Echoing previous CETaS findings, these bot activities primarily focused on circulating misleading content and narratives on key campaign issues – such as Russia's war in Ukraine.[68] Chinese-affiliated bots directed their efforts more towards amplifying conspiracy theories around the July 2024 assassination attempt, to confuse voters about the facts.[69] Similar bots linked to China also sought to use polling features on different campaign issues to better understand the demographics of US voters' intentions, which could then feed into future influence operations.[70]

Outside these operations, other bot activities targeted specific political candidates with either smear campaigns or favourable coverage, while amplifying falsehoods on disaster relief efforts for Hurricane Helene and Hurricane Milton.[71] Finally, in the buildup to the vote, bot accounts helped scale the dissemination of baseless claims of widespread voter fraud, in an attempt to undermine the integrity of the election.[72]

Social bot activities targeting the US public were more sophisticated than those in other elections. In one case, an AI-enhanced software package was used to create multiple fake user profiles on X, which could then generate posts and even repost, like and comment on the posts of other bots in the network.[73] In another operation, AI-manipulated images of female models were scraped and used on fake accounts advocating support for Trump's campaign.[74]

Impact

In several of these voter-targeting efforts, conspiracy theories and specific policy positions on politically divisive issues were amplified to online audiences, helping inflame conversations between users.[75] However, it is important to note that much of this

[67] Global Witness (2024).
[68] Wending (2024); Stockwell (2024), 15-17.
[69] Titcomb (2024).
[70] Watts (2024a).
[71] Microsoft Threat Analysis Center (2024a); Associated Press, "Russia spread hurricane disinformation after Helene, Milton in an effort to undermine American leadership," *Fast Company*, 24 October 2024, https://www.fastcompany.com/91215986/russia-hurricane-helene-milton-disinformation-disaster-relief-fema.
[72] Sherman (2024); Steffen (2024).
[73] Steven Lee Myers and Julian E. Barnes, "U.S. and Allies Take Aim at Covert Russian Information Campaign," *New York Times*, 10 July 2024, https://www.nytimes.com/2024/07/09/business/russian-bots-artificial-intelligence-propaganda.html.
[74] Polglase et al. (2024).
[75] Titcomb (2024); Global Witness (2024).

disinformation originally went viral through fringe political commentators and social media influencers, reflecting how bots acted as a force multiplier for deceptive content that originated from human users.[76]

Worryingly, bots also targeted smear operations against US political candidates outside the presidential race, focusing on those running in parallel congressional or state ballots.[77] Given that voters may be more open-minded about candidates in smaller-scale election contests – and that fewer fact-checking organisations and election security officials monitor AI-generated disinformation in these races – there is a higher risk that viral rumours amplified by fake accounts will affect the results.[78]

2.1.3 AI misattribution

Table 3. AI-generated misattribution cases identified in the US election

Summary	Instances reported[79]	Impact
Candidate endorsement by major political party incorrectly accused of being AI-generated.[80] Election withdrawal of political candidate wrongly alleged as being a deepfake.[81]	6	High user engagement with fake content amplified the disinformation.[84] Users' uncertainty about the authenticity of content damages their trust in the integrity of online sources.[85]

[76] Titcomb (2024); Bing et al. (2024); Klepper (2024).
[77] Sasha Issenberg, "Why Kamala Harris and Donald Trump Don't Need to Worry About Deepfakes," *Politico*, 27 October 2024, https://www.politico.com/news/magazine/2024/10/27/2024-elections-deepfakes-00184863.
[78] Ibid.
[79] Based on cited examples in news articles and public reports between 22 May and 8 November 2024.
[80] Ghaedi (2024).
[81] FP Explainers, "'Orange President', 'AI Biden': The conspiracy theories surrounding Joe Biden's address at Oval Office," *Firstpost*, 25 July 2024, https://www.firstpost.com/explainers/ai-biden-ornage-president-the-conspiracy-theories-surrounding-joe-biden-oval-address-13796712.html.
[84] Ghaedi (2024); Billie Schwab Dunn, "Joe Rogan Tests Biden's Kamala Harris Call, 'Likely' AI," *Newsweek*, 25 July 2024, https://www.msn.com/en-us/news/other/joe-rogan-tests-biden-s-kamala-harris-call-likely-ai/ar-BB1qBLlg?ocid=BingNewsSerp.
[85] Ibid.

Candidate rally size falsely accused of being AI-generated.[82] Candidates wrongly accused of faking involvement in campaign events through AI-generated content.[83]		

Actors, motives and tradecraft

One threat that CETaS identified in the UK general election emerged more prominently during the US campaign: the misattribution of political candidates and election activities as AI-generated. This included misleading claims that footage of the address in which Biden withdrew from the race must be synthetic due to his supposedly unnatural skin tone, as well as similar claims about a phone call in which Biden endorsed Harris as the Democratic Party's nominee.[86] Trump himself even dismissed genuine photos of a Democratic Party rally as fabricated, leading his supporters to claim that other photos from Democrat rallies were AI-generated fakes.[87]

Impact

As was clear in the UK election, misattributing content or individuals as AI-generated erodes trust in the information environment and allows conspiracy theories to thrive.[88] This is due to the increasingly realistic aesthetics of AI content, which make it difficult for users to distinguish fact from fiction. When prominent political candidates such as Trump employ these methods, it incentivises the party's core supporter base to amplify that disinformation, which entrenches their pre-existing beliefs.[89] Dismissing the activities of the political

[82] Shane Goldmacher, "Trump Falsely Claims That the Crowds Seen at Harris Rallies Are Fake," *New York Times*, 12 August 2024, https://www.nytimes.com/2024/08/11/us/politics/trump-harris-crowds-ai.html; Bill McCarthy (c), "Photo of Harris rally falsely claimed to show fabricated crowd," *AFP Fact Check*, 19 August 2024, https://factcheck.afp.com/doc.afp.com.36ED9CW.

[83] Sarah Thompson, "Fact Check: Image After Assassination Attempt Does NOT Show Six Fingers On Trump's Raised Fist – Trolling Edit," *Lead Stories*, 26 August 2024, https://leadstories.com/hoax-alert/2024/08/fact-check-image-after-assassination-attempt-does-not-show-six-fingers-on-trumps-raised-fist-trolling-edit.html; Uliana Malashenko, "Fact Check: People's Hands In 'Walz's For Trump' Image Do NOT Prove Image Was Generated By AI," *Lead Stories*, 5 September 2024, https://leadstories.com/hoax-alert/2024/09/fact-check-peoples-hands-in-walzs-for-trump-image-do-not-prove-image-was-generated-by-ai.html.

[86] Ghaedi (2024); FP Explainers (2024).

[87] Goldmacher (2024); McCarthy (2024c).

[88] Stockwell (2024), 23-24.

[89] Oremus (2024); McCarthy (2024c); Mohar Chatterjee, "Trump's crafty new use of AI," *Politico*, 22 August 2024, https://www.politico.com/newsletters/digital-future-daily/2024/08/22/trump-crafty-new-use-ai-00175822.

opposition as AI forgeries – despite a lack of evidence – helps divert attention away from any controversies faced by the candidate in question, placing the burden of proof on the targeted individual to debunk them.[90]

2.1.4 Parody and satire content

Table 4. AI-developed parody and satire disinformation identified in the US election

Summary	Instances reported[91]	Impact
AI-generated parody image of a political candidate's rally size.[92] AI-generated parody video including discriminatory remarks and conspiracy theories about a political candidate.[93] AI-generated 'parody' music video of a political candidate falsely depicted as being in poor health.[94] AI-generated memes reinforced baseless claims about immigrants' behaviour.[95]	5	High user engagement with fake content amplified the disinformation.[96] AI-generated parody content reinforced viral disinformation on social media, and was referenced by political candidate during a live TV debate.[97] Prominent social media owner retweeted AI-generated parody video containing disinformation without disclosure, thereby misleading users.[98]

[90] Dan Merica and Ali Swenson, "Trump's post of fake Taylor Swift endorsement is his latest embrace of AI-generated images," *AP News*, 20 August 2024, https://apnews.com/article/trump-taylor-swift-fake-endorsement-ai-fec99c412d960932839e3eab8d49fd5f.

[91] Based on cited examples in news articles and public reports between 22 May and 8 November 2024.

[92] Reuters Fact Check, "Fact Check: AI image of crowd at Arizona Harris-Walz rally is from parody account," *Reuters*, 12 August 2024, https://www.reuters.com/fact-check/fact-check-ai-image-crowd-arizona-harris-walz-rally-is-parody-account-2024-08-12/.

[93] Ali Swenson, "A parody ad shared by Elon Musk clones Kamala Harris' voice, raising concerns about AI in politics," *AP News*, 29 July 2024, https://apnews.com/article/parody-ad-ai-harris-musk-x-misleading-3a5df582f911a808d34f68b766aa3b8e.

[94] David Gilbert (b), "A Russian Propaganda Network Is Promoting an AI-Manipulated Biden Video," *WIRED*, 26 June 2024, https://www.wired.com/story/russia-disinformation-network-ai-generated-biden-video/.

[95] David Ingram, "How AI images of cats and ducks powered the pet-eating rumor mill in Springfield, Ohio," *NBC* News, 15 September 2024, https://www.nbcnews.com/tech/misinformation/ai-images-cats-ducks-powered-pet-eating-rumor-mill-rcna171065.

[96] McCarthy (2024a); Laws (2024); Ingram (2024).

[97] Reuters Fact Check (2024); Ingram (2024).

[98] Swenson (2024).

Actors, motives and tradecraft

Deepfakes labelled as parody continue to create challenges for countering disinformation while protecting free speech. Although some blur the lines between satire and harmful falsehoods, others can be made with no malicious intent but change in meaning when shared by different users.[99] This includes deepfake videos that are labelled as satirical by their creators but that contain discriminatory remarks about political candidates.[100] Reflecting a continuing theme across this election report series, several of these parody clips were created by users located in the US, opening up new sources of domestic risk.[101] Indeed, social media platforms such as X have even been accused of financially incentivising users to post provocative claims through AI-generated satire on both sides of the US political campaign.[102]

In one case, the Kremlin-affiliated Doppelganger disinformation network utilised bot accounts to amplify and distort a viral AI-manipulated parody music video. Referencing conspiracy theories about the previous US presidential election in 2020 being 'stolen', the bots disseminated clips from the video in 13 languages and trimmed the video arbitrarily to evade content moderation tools.[103] Equally concerning, Elon Musk – the owner of X – reshared an AI-generated parody video involving harmful tropes about Harris without initially labelling the content as either synthetic or satire.[104]

Although memes can often appear non-offensive or lacking any deceptive motive, they can also be used as a vehicle for spreading discriminatory narratives.[105] For example, baseless claims that Haitian immigrants were eating pets in the US state of Ohio started out as rumours on fringe social media platforms, before users began spreading AI-generated memes embedded with these false narratives.[106] While some were focused on uncontroversial depictions of animals, others were more openly prejudiced and included

[99] Swenson (2024); Scott Rosenberg, "Deepfakes' parody loophole," *Axios*, 30 July 2024, https://www.axios.com/2024/07/30/ai-deepfake-parody-musk-first-amendment; Andrew R. Chow, "AI's Underwhelming Impact On the 2024 Elections," *Time*, 30 October 2024, https://time.com/7131271/ai-2024-elections/.
[100] Swenson (2024).
[101] Swenson (2024); Stockwell (2024), 18-19.
[102] Marianna Spring, "How X users can earn thousands from US election misinformation and AI images," *BBC News*, 29 October 2024, https://www.bbc.co.uk/news/articles/cx2dpj485nno.
[103] Gilbert (2024b).
[104] Swenson (2024).
[105] Dan Merica, Garance Burke and Ali Swenson, "AI is helping shape the 2024 presidential race. But not in the way experts feared," *AP News*, 21 September 2024, https://apnews.com/article/artificial-intellgence-memes-trump-harris-deepfakes-256282c31fa9316c4059f09036c70fa9.
[106] Ingram (2024).

racist assertions about immigrants.[107] Following several viral cases, Trump posted this content on his own social media platform, Truth Social.[108]

Impact

Despite receiving criticism for resharing a so-called 'parody' deepfake containing discriminatory remarks about Harris and conspiracy theories about the 'deep state', Musk did not delete the original clip or his own repost. This was due to the classification of the video as a political meme.[109] Accordingly, it was considered exempt from X's content moderation polices. The case could set a dangerous precedent whereby malicious actors seeking to undermine election security can post disinformation under the guise of a parody and avoid its removal under platform protections.[110] This is evidenced by, for instance, the Doppelganger bot activities connected to the 'parody' music video of Biden.[111]

As well as posting misleading allegations about Haitian immigrants through AI-generated memes, Trump referenced these rumours during a live TV election debate – reinforcing their perceived legitimacy.[112] However, it remains unclear to what extent the AI content itself – as opposed to other content formats through which it was amplified – caused this baseless claim to influence Trump's decision.[113]

Once again, it appears that domestic political bloggers who were ideologically aligned with the political narratives and candidate benefiting from the memes were responsible for amplifying them.[114] This indicates how such content may be acting as a "visual parallel" to evidence-free claims that political candidates make in real life, reinforcing the pre-existing beliefs of their core supporters through humour and the truth they want to believe.[115] Indeed, in the case of AI-generated content referencing the debunked pet-eating story,

---

[107] Ibid.
[108] Ibid.
[109] Swenson (2024).
[110] Rosenberg (2024).
[111] Gilbert (2024b).
[112] Henry J. Gomez et al., "How a fringe online claim about immigrants eating pets made its way to the debate stage," *NBC News*, 13 September 2024, https://www.nbcnews.com/politics/donald-trump/trump-fringe-online-claim-immigrants-eating-pets-debate-trump-rcna170759.
[113] Ibid.
[114] Ingram (2024).
[115] Will Oremus, "Trump's AI fakes of Harris and Swift aren't meant to fool you," *Washington Post*, 19 August 2024, https://www.washingtonpost.com/technology/2024/08/19/trump-taylor-swift-ai-fakes-dnc-kamala-harris/; Sam Stockwell, "Propaganda by Meme: The impact of generative AI on extremist memes," *CETaS Expert Analysis* (May 2024); Perry Carpenter, "Opinion: Deepfakes didn't disrupt the election, but they're changing our relationship with reality," *The Hill*, 6 November 2024, https://www.msn.com/en-us/news/politics/opinion-deepfakes-didn-t-disrupt-the-election-but-they-re-changing-our-relationship-with-reality/ar-AA1tC02z?ocid=BingNewsSerp.

positive engagement with its messages requires the user to have both insider knowledge of the meme and certain views of immigrants.[116]

2.1.5 AI-generated knowledge sources

Table 5: AI-generated knowledge sources identified in the US election

Summary	Instances reported[117]	Impact
Social media networks affiliated with Russia and Iran spread AI-based disinformation on campaign issues using fake US news sources.[118]	4	High user engagement with Kremlin-affiliated fake content amplified the disinformation.[119] Iranian-affiliated activities failed to achieve meaningful audience engagement.[120]

Actors, motives and tradecraft

A smaller number of viral cases in the US election involved the use of AI tools for fabricating news stories promoting specific narratives or policy positions on campaign issues. The persistent CopyCop operation deployed by the Kremlin-affiliated Doppelganger network shifted its focus towards the US election following those in the UK and Europe.

Using the same tactics as before, new fake news sites were set up with a focus on political topics such as Russia's war in Ukraine.[121] One prominent network scraped and altered

[116] Ingram (2024).
[117] Based on cited examples in news articles and public reports between 22 May and 8 November 2024.
[118] Recorded Future, "Russia-Linked CopyCop Expands to Cover US Elections, Target Political Leaders," Insikt Group, June 2024, 21, https://www.recordedfuture.com/research/copycop-expands-to-cover-us-elections-target-political-leaders; Microsoft Threat Analysis Center (b), "Iran steps into US election 2024 with cyber-enabled influence operations," 9 August 2024, https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/5bc57431-a7a9-49ad-944d-b93b7d35d0fc.pdf; OpenAI, "Disrupting a covert Iranian influence operation," 16 August 2024, https://openai.com/index/disrupting-a-covert-iranian-influence-operation/; NewsGuard (b), "USNewsper.com is not American or News; It's Lithuanian AI!," 5 August 2024, https://www.newsguardrealitycheck.com/p/usnewspercom-is-not-american-or-news.
[119] Paul Myers et al., "A Bugatti car, a first lady and the fake stories aimed at Americans," *BBC News*, 3 July 2024, https://www.bbc.co.uk/news/articles/c72ver6172do.
[120] OpenAI (2024).
[121] Recorded Future (2024).

articles from politically mainstream and conservative-leaning US outlets, as well as Russian state-affiliated outlets, before spreading them to US election-themed websites.[122]

To further bolster the credibility of the fake stories, Doppelganger operatives created YouTube videos, often featuring people who falsely claimed to be whistleblowers or independent journalists. Some of the videos were narrated by actors, while others used AI-generated voices. Several of them appeared to be shot against similar-looking backgrounds – another suggestion of a centralised effort to spread fake news stories.[123]

The US election was also subject to fake news operations with ties to users in Iran and Lithuania. Accounts linked to an Iranian influence operation sought to use chatbots to generate election disinformation, though they were shut down shortly after being discovered.[124] The same Iranian network also created fake website domains. Masquerading as domestic US news outlets, the sites catered to different US voter demographics but focused on circulating divisive content in an attempt to inflame online discourse.[125] Finally, one case involved a Lithuanian company that used AI to produce fake articles on US politics, which were then amplified by AI-generated bots.[126]

These cases reveal how generative AI tools are *enhancing* and *accelerating* the fake news aspects of foreign influence operations, as opposed to *revolutionising* them.[127] Despite their advantages, these methods are still costly. For example, foreign malicious actors need to overcome model restrictions that can prevent them from creating the desired content while remaining undetected.[128]

Impact

Although generative AI has benefits in its rapid creation and dissemination of realistic fake news stories or websites, most efforts of this kind in the US election failed to garner meaningful audience engagement.[129] For example, Iranian-affiliated social media posts

[122] Emma Woollacott, "AI-Powered Russian Influence Network Targets U.S. Elections," *Forbes*, 24 June 2024, https://www.forbes.com/sites/emmawoollacott/2024/06/24/ai-powered-russian-influence-network-targets-us-elections/.
[123] Myers et al. (2024).
[124] OpenAI (2024).
[125] Microsoft Threat Analysis Center (2024b).
[126] NewsGuard (2024b).
[127] Office of the Director of National Intelligence, "Election Security Update as of Mid-September 2024," 23 September 2024, https://www.dni.gov/index.php/fmic-news/3998-election-security-update-20240923.
[128] Ibid.
[129] NewsGuard (2024b); Microsoft Threat Analysis Center (2024b); *Reuters*, "OpenAI blocks Iranian group from ChatGPT, says it targeted US election," *VoA News*, 16 August 2024, https://www.voanews.com/a/openai-blocks-iranian-group-from-chatgpt-says-it-targeted-us-election/7745899.html.

containing links to the fabricated articles received few or no likes, shares or comments, and were not shared across social media networks.[130]

As leaked files from the Doppelganger network show, the biggest advantage that disinformation operators gained from these synthetic articles was Western media outlets' own anxious coverage of the project, as opposed to any widespread exposure among voters.[131] Once more, this highlights the importance of ensuring that researchers and journalists do not exaggerate the impact of these cases – and thereby avoid playing into the hands of malicious actors.[132]

2.1.6 Deceptive political advertising

Table 6. AI-generated deceptive political adverts identified in the US election

Summary	Instances reported[133]	Impact
Fabricated endorsement of a political candidate by a celebrity and a deceased civil rights activist.[134]	2	Content of false endorsement from the celebrity backfired and may have led to them publicly supporting the opposition candidate.[135]

Actors, motives and tradecraft

Echoing findings from the Indian and Indonesian elections earlier in the year,[136] there were two instances during the US election where AI tools were used for a fake high-profile candidate endorsement. In one case, Trump falsely asserted that Taylor Swift endorsed his

[130] *Reuters*, "OpenAI blocks Iranian group from ChatGPT, says it targeted US election," *VoA News*, 16 August 2024, https://www.voanews.com/a/openai-blocks-iranian-group-from-chatgpt-says-it-targeted-us-election/7745899.html.
[131] Thomas Rid, "The Lies Russia Tells Itself," *Foreign Affairs*, 30 September 2024, https://www.foreignaffairs.com/russia/lies-russia-tells-itself.
[132] Stockwell (2024), 28.
[133] Based on cited examples in news articles and public reports between 22 May and 8 November 2024.
[134] Siladitya Ray, "Trump Reposts AI-Generated Images Claiming Taylor Swift Fans Support Him," *Forbes*, 19 August 2024, https://www.forbes.com/sites/siladityaray/2024/08/19/trump-reposts-ai-generated-images-claiming-he-has-support-from-taylor-swift-fans/; Nur Ibrahim, "Audio of MLK Endorsing Trump Is Deepfake," *Snopes*, 5 November 2024, https://www.snopes.com/fact-check/mlk-endorsing-trump-deepfake/.
[135] Steven Musil and Gael Cooper, "Taylor Swift Endorses Kamala Harris, Calling Out Donald Trump's AI Deepfake Post," *CNET*, 12 September 2024, https://www.cnet.com/tech/services-and-software/taylor-swift-endorses-kamala-harris-calling-out-donald-trumps-ai-deepfake-post/#google_vignette.
[136] Stockwell et al. (2024), 21-23.

campaign, using what turned out to be a series of AI-manipulated images on X.[137] However, it was his supporter base who initially posted their own AI-generated content of Swift appearing to support the Republican candidate, before Trump himself posted the videos and images, believing them to be a legitimate endorsement.[138] This reflects how grassroots political activists can now shape discourse at the official party level through the circulation of realistic but fabricated election content at scale.

In another case, an audio recording of a speech by Martin Luther King Jr. was manipulated with AI to insert endorsements of Trump.[139] Similar to the original posters of the fake Taylor Swift content, the user behind this audio deepfake was part of Trump's core voting bloc, referring to themselves as "Trump's Online War Machine."[140]

Impact

Although the AI-generated campaign endorsement of Swift went viral and received a high level of user views, it did not have the desired effect. Indeed, Swift not only publicly announced her backing for Trump's political opponent, but also explicitly cited his decision to post the fake AI-generated endorsement as a motivating factor in doing so.[141]

Nevertheless, this may not always be the intention of those circulating such fabricated endorsements. Political candidates can seek to consolidate the support of their core voter base by promoting alternate realities, provoking reactions and creating "illusions of support" for their own campaigns.[142] Indeed, one investigation found at least 70 social media posts promoting fake VIP endorsements and snubs by celebrities in the US election for different candidates, including content manipulated with basic editing software.[143]

---

[137] Rachel Looker, "Trump falsely implies Taylor Swift endorses him," *BBC News*, 19 August 2024, https://www.bbc.co.uk/news/articles/c5y87l6rx5wo.

[138] Merica and Swenson (2024).

[139] Ibrahim (2024).

[140] Marco Margaritoff, "MLK Jr.'s Daughter Slams 'Vile' Deepfake Video Of Civil Rights Leader Endorsing Trump," *Huffington Post*, 5 November 2024, https://www.huffingtonpost.co.uk/entry/mlk-jr-daughter-vile-deepfake-video-civil-rights-leader-endorsing-trump_n_6729f76fe4b0be8c956a7768.

[141] Musil and Cooper (2024).

[142] Merica and Swenson (2024).

[143] AFP Fact Check, "Fake celebrity endorsements, snubs plague US presidential race," 22 August 2024, https://www.msn.com/en-us/entertainment/news/fake-celebrity-endorsements-snubs-plague-us-presidential-race/ar-AA1qS8Kf?ocid=BingNewsVerp.

2.2 Network analysis of US election deepfakes

As highlighted throughout the previous subsection, a select few individuals often have a great deal of influence in shaping how electoral disinformation cascades across social media networks, and in exposing it to wider digital communities. By identifying the political affiliations of these influential individuals, one can better understand both the likely intentions of those who create disinformation and the types of voters who consume their posts.

In August 2024, the Australian Strategic Policy Institute identified a network of inauthentic accounts across X and YouTube likely connected to a covert social media operation linked to the Chinese government called Spamouflage.[144] These accounts were responsible for disseminating viral deepfakes to undermine support for the president of the Philippines. However, to date, CETaS has not identified similar network analysis of viral deepfakes in the 2024 US presidential election – analysis that would help inform those monitoring and countering these harmful activities.

2.2.1 Methods and limitations

This subsection employs network analysis in three case studies of AI-generated images and videos that targeted US political candidates during the 2024 presidential election:

- A deepfake 'parody' video of Harris featuring discriminatory remarks and conspiracy theories about her nomination as a candidate.
- A deepfake image of a US Secret Service agent smiling during the attempted assassination of Trump.
- A deepfake video depicting US politicians and other notable figures robbing a store.

The case studies were chosen to represent different political targets and to yield insights into the dynamics of influence across the political spectrum. In this analysis, influence is defined by the ability to promote both exposure and engagement with content that the individual has posted. Exposure is defined by the number of views an interaction has achieved, and engagement by summing individual interaction mechanisms – such as likes,

[144] Albert Zhang, "China's high stakes and deepfakes in the Philippines," *Australian Strategic Policy Institute*, 2 August 2024, https://www.aspistrategist.org.au/chinas-high-stakes-and-deepfakes-in-the-philippines/.

reposts and replies. Users were sampled based on those who had interacted with the original deepfake content between the date of the original posts and September 2024.

Owing to restrictions on access to this type of data, CETaS commissioned analytics company Meltwater to provide relevant data across the three case studies. This included information on: the date and time of the interaction; interaction type (quote, repost, reply, etc.); engagement measure counts; and user profiles. Following this, CETaS determined the top ten users who had the longest reach with any reshares of the original deepfakes – and then qualitatively analysed these accounts to assess previous sharing behaviour and political leanings, and to gain information on who operated them.

Despite overcoming the barriers to data access through Meltwater, some critical variables were missing from the data – such as that on the users behind further amplification of the deepfakes beyond the initial nodes of influence. This made it harder to understand the broader networks that increased the virality of the content. Other limitations included: the opaque nature of the pipelines that provide this data, which made it difficult to understand whether the data had been manipulated; and the selection of accounts based on their active engagement with these deepfake cases, which did not provide insights into the extent of the content's influence on wider online communities.

2.2.2 Findings

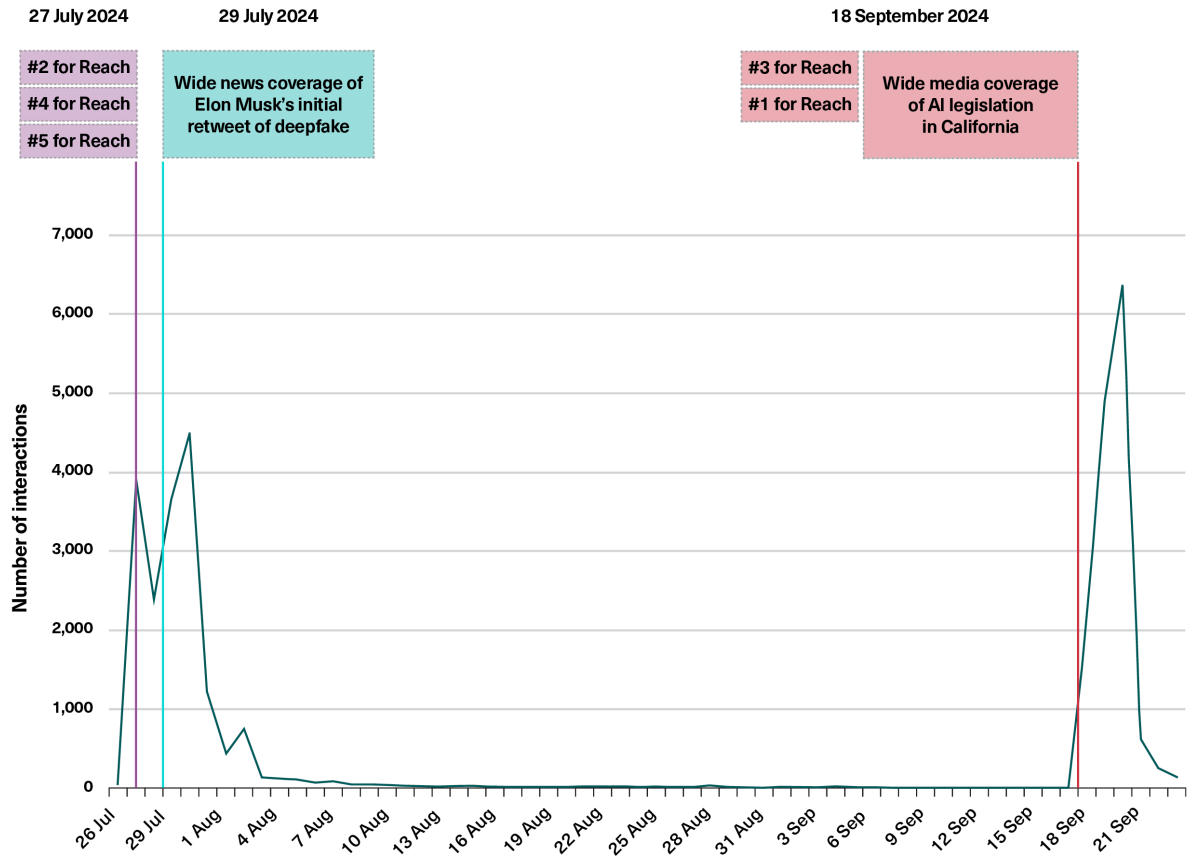
The first case study of the so-called 'parody' deepfake of Harris had by far the most engagement of the three cases. Analysis of the top ten influential users sharing the video revealed they were all prominent figures in US society who displayed right-leaning political views. Notably, nine out of the ten individuals had previously shared content identified as deepfakes or disinformation on their X accounts. None of these users exhibited the characteristics typical of bot accounts.

While the parody Harris video reached an initial viral peak in late July, Musk's resharing of the video a few days later led to another sharp spike in user engagement (see Figure 3 below), demonstrating the significant impact that individuals with a large follower base can have on virality.[145] The video gained additional attention after it received coverage from media outlets and came up a few months later in discussions of legislation to combat deepfakes (see Figure 3 below). The coverage likely drew more viewers to the video, resulting in increased engagement. This sequence of events illustrates the unintended

[145] Swenson (2024).

consequences of citing original content in articles, as it often amplifies engagement rather than suppresses it.

Figure 3. Time-series of interactions with Kamala Harris 'parody' deepfake video



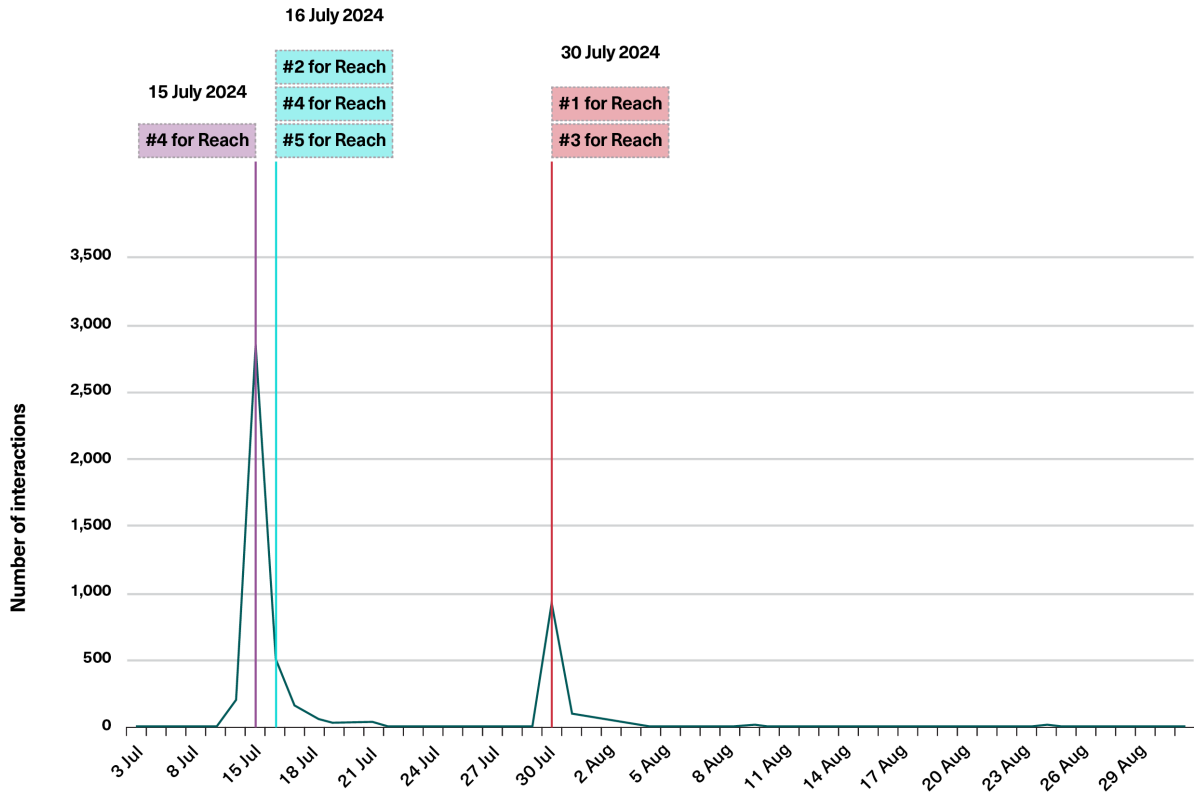
Source: Authors' analysis.

In the case of a deepfake image depicting a US Secret Service agent smiling during the July 2024 assassination attempt, engagement was more sustained over time – influenced by public debates and fact-checking efforts (see Figure 4 below). Of the top ten influential users amplifying the content, five were corporations. They included three US news providers and one US fact-checking organisation. Nine out of ten of these users clarified that the image was a fake. Additionally, three out of ten users commented on how Meta allegedly blocked the unaltered version of the deepfake, with some implying that the company did so intentionally. An interesting aspect of this case study is that some highly influential actors were not attempting to convince users that the deepfake was real but were instead accusing Meta of censoring the original image. Overall, the influential actors displayed minimal intention to deceive, with half being overtly unbiased news outlets.

Compared to the first case study, the distribution of engagement in the second case study is much less concentrated in the highest 1% of individuals with the most engagement. With

the Harris deepfake, ~1% of users sampled were responsible for ~96% of interactions; whereas ~1% of users sampled were responsible for ~26.4% of interactions in the second case study. One plausible explanation for this is that the network sharing the latter content lacked prominent figures who increase virality and engagement, such as Musk.

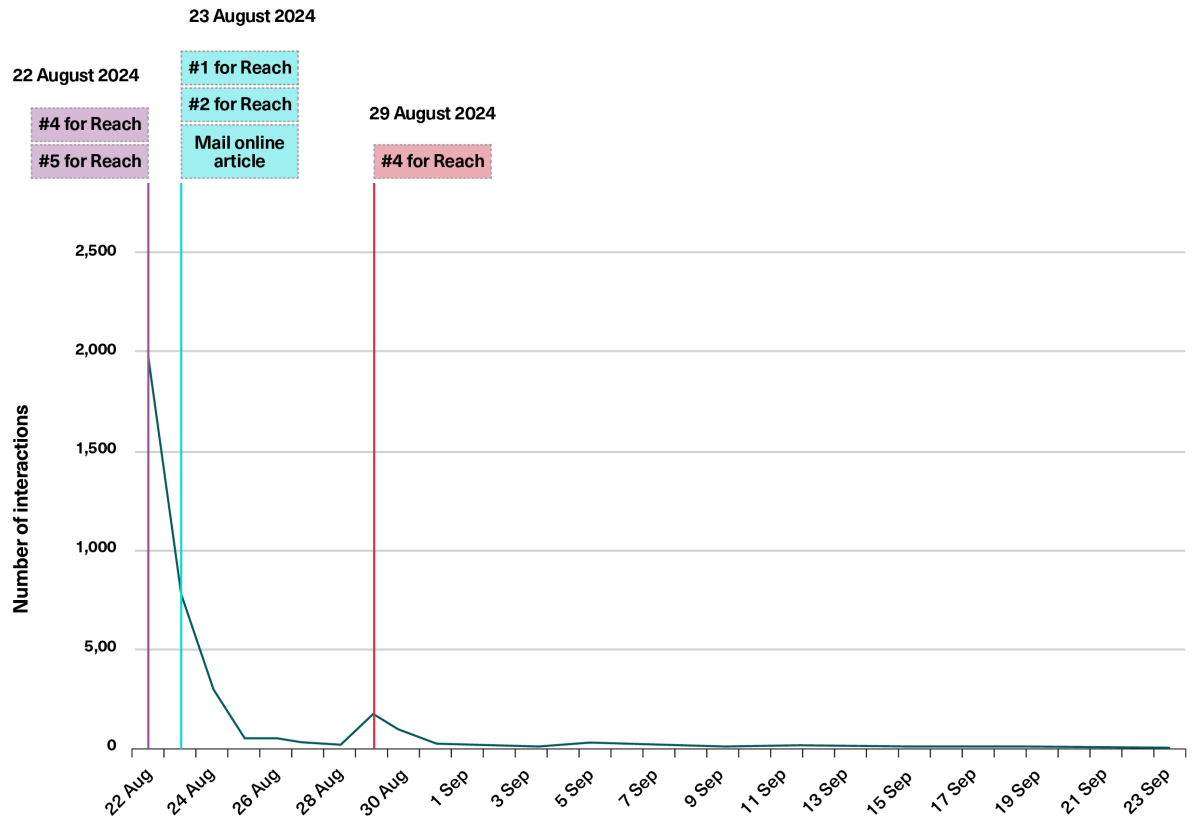
Figure 4. Time series of interactions with smiling US Secret Service agent deepfake image



Source: Authors' analysis.

Analysis of the Grok-made robbery deepfake video was limited due to insufficient data, but the video appeared to receive moderate engagement without significant spikes in viewership. The most influential user sharing this video was a co-founder of a French news outlet, who did not provide additional commentary on the content (see Figure 5 below). The second-most influential user was a videographer known for producing content that promotes conspiracy theories about topics such as COVID-19 vaccinations and 9/11. The remaining influential accounts represented a mix of nationalities, and those with identifiable political alignments tended to be right-wing (see Figure 5 below).

Figure 5. Time series of interactions with deepfake video of a store robbery



Source: Authors' analysis.

Overall, the analysis reveals that a diverse range of users were responsible for amplifying these deepfakes to other users and increasing their virality. Echoing findings elsewhere in this report, many of these accounts showed clear signs of political affiliation with the beneficiaries of the misleading claims, and of sharing the content because it aligned with their pre-existing views. Yet interestingly, several news outlets and fact-checking organisations were also inadvertently responsible for boosting the exposure of such content to online audiences, reflecting the importance of understanding when and how they should report on these cases.

3. Evaluating Influence Operations in the Age of AI

There is a risk that malicious actors will use AI tools to target future political processes. This section explores the challenges researchers face when trying to evaluate hostile influence operations and disinformation campaigns, before analysing existing frameworks designed to assist in overcoming such challenges.

3.1 Challenges in evaluating influence operations

3.1.1 Identifying intentions and attribution

One of the key challenges in evaluating hostile influence operations is that different actors may seek to achieve different objectives.[146] While some campaigns will try to shape election outcomes, others could have wider ambitions to undermine public confidence in the integrity of an election, or to erode trust in democratic institutions or processes. Some simply intend to sow public confusion about the truth, increase users' susceptibility to disinformation in future activities or stir hatred between individuals on different sides of the political spectrum.

Technical measures to obfuscate internet activity – such as virtual private networks (VPNs), proxy users and social bot accounts – also create barriers to attribution. Accordingly, those behind hostile influence operations can maintain a degree of plausible deniability in relation to any operation, owing to the difficulties in finding evidence that reveals a direct link between the two.[147] This provides greater incentives to, and lowers the risks for, malicious actors conducting these activities at scale – and increases the costs for those trying to counter them.

3.1.2 Defining impact

Another challenge is in how to measure the impact of influence operations. Many assessments focus on social media metrics such as likes or reshares, on the premise that

[146] Jon Bateman and Dean Jackson, "Countering Disinformation Effectively: An Evidence-Based Policy Guide," Carnegie Endowment for International Peace, 2024, 13-14, https://carnegie-production-assets.s3.amazonaws.com/static/files/Carnegie_Countering_Disinformation_Effectively.pdf.

[147] James Pamment and Victoria Smith, "Attributing Information Influence Operations: Identifying Those Responsible for Malicious Behavior Online," NATO Strategic Communications Centre of Excellence and European Centre of Excellence for Countering Hybrid Threats, July 2022, 16, https://stratcomcoe.org/pdfjs/?file=/publications/download/Nato-Attributing-Information-Influence-Operations-DIGITAL-v4.pdf?zoom=page-fit.

this type of engagement could indicate that a user has been successfully influenced by the content.[148] However, these metrics only measure *outputs* tied to the content in question rather than *outcomes* related to the users involved.[149] In other words, while like or share counts give an indication of the *virality* of a piece of disinformation on a platform, they do not necessarily equate to changes in the *behaviour* or *attitudes* of those engaging with it. As such, there is a danger that relying on these output metrics could exaggerate the actual effects of a hostile influence operation. In turn, this can create a false impression that the integrity of an election has been compromised, which could damage public confidence.

Alongside defining impact, there is the related issue of distinguishing causation from correlation in specific content. A deepfake targeting a political candidate may go viral just before polling day and be seen as convincing by many users. Yet the extent to which that isolated incident leads to any substantial change in voter behaviour – or can be pinned down as the main reason for causing such a shift – is inherently hard to determine.[150] This is further complicated by the fact that election cycles have relatively long durations, in which the volume of information consumed by users means they have only a limited ability to recall exposure to specific deepfakes.[151]

3.1.3 Assessment scope

Finally, there are trade-offs in the structure of evaluations. Researchers need to make decisions based on resources and objectives about whether they examine the short term (i.e. days and weeks) or the long term (i.e. months and years) in measuring exposure to disinformation.[152] Although the former tends to be less resource-intensive and can quickly inform counter-narratives, it cannot account for the potential duration or long-term effects of the influence attempt.[153] In contrast, long-term evaluations are much more expensive. Often, researchers will only discover an influence operation that has already started and must work out its intentions, methods and impact as it goes on.[154]

---

[148] Joint Committee on the National Security Strategy, "Oral evidence: Defending democracy," 18 March 2024, https://committees.parliament.uk/oralevidence/14514/pdf/.
[149] Ibid; Ben Nimmo, "The Breakout Scale: Measuring the Impact of Influence Operations," Brookings Institution, September 2020, 6-8, https://www.brookings.edu/wp-content/uploads/2020/09/Nimmo_influence_operations_PDF.pdf/
[150] Bateman and Jackson (2024), 13-14.
[151] Tvesha Sippy et al., "Behind the Deepfake: 8% Create; 90% Concerned," *The Alan Turing Institute*, 5, https://www.turing.ac.uk/sites/default/files/2024-07/behind_the_deepfake_full_publication.pdf.
[152] Jon Bateman et al., "Measuring the Effects of Influence Operations: Key Findings and Gaps from Empirical Research," Carnegie Endowment for International Peace, 2021, https://carnegieendowment.org/research/2021/06/measuring-the-effects-of-influence-operations-key-findings-and-gaps-from-empirical-research?lang=en&center=global.
[153] Ibid.
[154] Nimmo (2020), 2-3.

3.2 Measuring hostile influence operations

Despite the challenges described above, several organisations have sought to develop frameworks that equip researchers with the skills to mitigate against, and facilitate evaluations of, influence operations. The table below provides a non-exhaustive list of these evaluation frameworks, including analysis of their strengths and weaknesses. There is no one-size-fits-all framework in this space. As such, researchers should weigh the trade-offs between the tools listed and use the one most suited to the type of operation in question. In some cases, combining different frameworks will help provide valuable insights into these activities.

Table 7. Non-exhaustive overview of evaluation frameworks for influence operations

Framework	Summary	Benefits	Limitations
ABCD	Uses four criteria (actors, behaviours, content, and distribution) to assess influence operations.[155]	Helpful in identifying specific areas in which to target interventions across the four categories. Focus on behaviours and distribution mechanisms can also aid early detection.	Focuses heavily on digital platforms, potentially overlooking other important channels, such as traditional media or offline networks.
DISARM	Focuses on informing countermeasures to influence operations based on historical and hypothetical scenarios.[156]	Has been successfully deployed in real-world contexts, such as countering disinformation during the COVID-19 pandemic.[157]	Difficult for beginners to navigate given the number of features and layers, requiring time and resources for training researchers before it can be used effectively.
MITRE SP!CE	Divides influence operations into four components (plan,	Divides evaluations into a series of indicators that capture the different	Complexity means it can be time-consuming to use effectively in disinformation

[155] Alexandre Alaphilippe, "Adding a 'D' to the ABC disinformation framework," *Brookings Institution*, 27 April 2020, https://www.brookings.edu/articles/adding-a-d-to-the-abc-disinformation-framework/.
[156] DISARM Foundation, "DISARM Framework," https://www.disarm.foundation/framework.
[157] Ibid.

	enable, engage and assess) and provides scores based on six key performance indicators.[158]	objectives of influence operations to a high degree of granularity.	scenarios requiring rapid responses.
SCOTCH	Divides influence operations into six categories (source, channel, objective, target, composition and hook).[159]	Allows for faster identification and analysis of influence operations compared to more detailed frameworks.	Relies heavily on digital tools and datasets that can be expensive or restricted to researchers.
The Breakout Scale	Divides influence operations into six categories based on whether they remain on one platform or travel across multiple channels (including in real life) and whether they remain in one community or spread through many.[160]	Captures not only the impact of hostile influence operations on online audiences but also whether such operations translate into real-world threats. The scale is actor-agnostic and can be used to analyse other threats, such as conspiracy theories.	Less useful in evaluating the tactics or narratives used by malicious actors due to focus on the spread of narratives rather than the content itself. The value of the framework depends on the availability of social media data, which is often restricted to researchers.
NATO Capability Assessment	Divides assessments of influence operations into four categories (objectives, indicators, scenarios and process maturity).[161]	Recognises that there is no one-size-fits-all solution to this problem and provides tailored guidance to different types of influence operations.	Complex and resource-intensive, meaning that those with constrained resources may struggle to implement it. NATO member states may have contrasting approaches to countering disinformation, creating challenges in

[158] Matt Venhaus et al., "Structured Process for Information Campaign Evaluation (SP!CE): An Analytic Framework, Knowledge Base, and Scoring Rubric for Operations in the Information Environment," MITRE, November 2021, 5-1, https://users.cs.fiu.edu/~markaf/doc/o13.venhaus.2021.mitre.210039_archival.pdf.

[159] Sam Blazek, "SCOTCH: A framework for rapidly assessing influence operations," *Atlantic Council*, 24 May 2021, https://www.atlanticcouncil.org/blogs/geotech-cues/scotch-a-framework-for-rapidly-assessing-influence-operations/.

[160] Nimmo (2020), 6-8.

[161] James Pamment, "A Capability Definition and Assessment Framework for Countering Disinformation, Information Influence, and Foreign Interference," NATO Strategic Communications Centre of Excellence, November 2022, 4-5, https://stratcomcoe.org/pdfjs/?file=/publications/download/Defining-Capabilities-DIGITAL.pdf?zoom=page-fit.

			applying this framework to different national contexts.
AI modelling	Use of generative AI models to simulate multiple election scenarios with disinformation factored into different hypothetical outcomes.[162]	Helps provide an idea of when disinformation of a given magnitude and frequency affects simulated election outcomes.	Only provides statistical data to estimate the impact of disinformation in a hypothetical scenario, as opposed to guidance on countermeasures and other responses.

Source: Authors' analysis.

[162] Dorje C. Brody, "Generative AI model shows fake news has a greater influence on elections when released at a steady pace without interruption," *The Conversation*, 16 April 2024, https://theconversation.com/generative-ai-model-shows-fake-news-has-a-greater-influence-on-elections-when-released-at-a-steady-pace-without-interruption-227332.

4. Policy Responses to AI-Enabled Election Threats

As the nascent evidence base on AI misuse in elections has grown over the last 12 months, there is a valuable opportunity to reflect on best practices and lessons for countering these activities. This section analyses legal and policy measures to mitigate against AI-enabled election threats. The measures below are non-exhaustive, having been selected as the most promising based on a combination of literature review and workshop insights. (Technical countermeasures are discussed in Section 5.)

4.1 Legal and regulatory measures

4.1.1 National security and online safety legislation

Introduced in 2023, the National Security Act (NSA) is intended to protect the UK from national security threats. In particular, the NSA makes it a criminal offence to interfere in UK elections and other democratic processes, applying to individuals who know or ought reasonably to know that they are acting on behalf of foreign powers, or who intend to benefit them.[163] The Act also criminalises attempts outside the electoral cycle to interfere with political decision-making on behalf of foreign powers, and to intimidate individuals entering politics.[164]

Yet while the NSA provides a much-needed refresh of the law on counteracting novel methods of hostile election interference, its deterrence effect depends substantially on the prospect of its enforcement – which remains unclear. Moreover, it has historically been challenging to obtain evidential proof of state attribution in any election interference, including in the 2020 US presidential election.[165] Since the NSA is concerned solely with state threats, it also provides no legal enforcement powers against domestic actors responsible for creating and amplifying AI-enabled disinformation.[166]

[163] Such conduct may amount to the general offence of foreign interference (sections 13-15) or the specific offence of foreign interference in elections (section 16).
[164] Section 14(1)(d) provides that interfering with whether or how any person participates in relevant political processes or makes political decisions amounts to an interference effect.
[165] Robyn Dixon and Catherine Belton, "Russia dismisses charges of election meddling; Putin claims he backs Harris," *Washington Post*, 5 September 2024, https://www.washingtonpost.com/world/2024/09/05/putin-russia-america-meddling/.
[166] Ardi Janjeva et al., "The Rapid Rise of Generative AI: Assessing risks to safety and security," *CETaS Research Reports* (December 2023), 36-38; Stockwell (2024), 13.

Section 13 of the NSA (the foreign Interference offence) has also been made a priority offence in the UK's Online Safety Act 2023 (OSA). The OSA seeks to improve users' online safety by establishing new legal duties on digital platforms, requiring them to take down illegal content and content that is harmful to children.[167] Section 72 of the OSA requires the largest platforms to ensure that they have adhered to their own terms and conditions – including by removing AI-generated and other forms of disinformation that meet certain thresholds. Ofcom has powers to enforce these requirements, including by issuing substantial fines and other business-disruption measures.[168]

Despite placing greater pressure on platforms to tackle disinformation, it is unclear whether the regulator can enforce its OSA-imposed obligations. This is particularly due to the challenges platforms face in identifying creators of malicious, deceptive content, who can often mask their location or true identity through VPNs and bot accounts.[169] Given that DSIT's recently created AI Central Function is tasked with helping regulators address new AI threats, it should coordinate with Ofcom to analyse potential gaps in their regulatory powers or remit.[170] This would help Ofcom effectively tackle online disinformation during elections and beyond, in accordance with OSA requirements.[171]

Finally, under section 152 of the OSA, Ofcom is required to establish an advisory committee on misinformation and disinformation.[172] The committee will provide Ofcom with advice on how providers of regulated services should deal with disinformation on such services, among other commitments. Although Ofcom has stated that the committee may not begin work until the end of the year, it is vital to prioritise the establishment of this body.[173] The regulator should ensure that the committee has: a clear mandate informing Ofcom's activities in this space; an independent chair with no current or prior affiliation with any political party or tech platform; and diverse sectoral representation.[174]

---

[167] Department for Science, Innovation and Technology, "Online Safety Act: explainer," 8 May 2024, https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer.
[168] Joe Tyler-Todd and John Woodhouse, "Preventing misinformation and disinformation in online filter bubbles," *House of Commons Library*, 15 January 2024, https://commonslibrary.parliament.uk/research-briefings/cdp-2024-0003/.
[169] Catherine Kim, "How deepfakes could upend the 2024 elections," *Politico*, 7 February 2024, https://www.politico.com/newsletters/politico-nightly/2024/07/02/how-deepfakes-could-upend-2024s-elections-00166347.
[170] Department for Science, Innovation and Technology, "Implementing the UK's AI regulatory principles: initial guidance for regulators," 6 February 2024, https://www.gov.uk/government/publications/implementing-the-uks-ai-regulatory-principles-initial-guidance-for-regulators/102aa401-60f6-46e8-95dc-b48150eba7dd.
[171] CETaS policy workshop, 17 September 2024.
[172] Robert Cann, "Ofcom should move now to set up its Advisory Committee on Disinformation and Misinformation," Full Fact, 5 April 2024, https://fullfact.org/blog/2024/apr/ofcom-should-move-now-to-set-up-its-advisory-committee-on-disinformation-and-misinformation-and-make-progress-on-media-literacy/.
[173] Ibid.
[174] Ibid; CETaS policy workshop, 17 September 2024.

4.1.2 Domestic electoral law and election news reporting

While the NSA and the OSA seek to enhance protections against all forms of influence operations, electoral law is specifically designed to protect the integrity of UK elections. In limited cases, this applies to harmful falsehoods. Under section 106 of the Representation of the People Act 1983, it is illegal to publish a "false statement of fact" about a candidate's "personal character or conduct."[175] Although this could apply to some AI-enabled threats such as deepfakes, the fragmented and sometimes dated nature of electoral law means that many of these provisions were not established with new technologies in mind.[176] Nor does the law prohibit wider falsehoods in elections, such as those concerning an opposing party's policies.

Prior to this Act, the Elections Act 2022 was the most recent update to the rules governing elections. The earlier Act stripped the UK's election regulator, the Electoral Commission, of its enforcement powers to bring forward prosecutions against those who broke electoral law.[177] This means that the police and Crown Prosecution Service now need to work with the Electoral Commission to pursue prosecutions.[178] However, given the large remit and current capacity constraints on these organisations, there are widespread concerns about the effectiveness of this regime.

Therefore, DSIT's AI Central Function should also integrate the Electoral Commission into a review similar to that of Ofcom. This should focus on whether the Electoral Commission has the necessary remit and enforcement powers to address AI misuse cases during elections, in line with relevant electoral legislation such as the Representation of the People Act 1983 and the Elections Act 2022.[179]

During recent elections, CETaS has observed the challenges media organisations face in helping expose or debunk disinformation cases without inadvertently exposing more individuals to this content.[180] Although there are established standards for the conduct of media organisations during elections and major national security incidents, there is a lack of

[175] Will Hazell, "Stop AI deepfakes undermining elections by updating law, says watchdog," *Telegraph*, 14 May 2023, https://www.telegraph.co.uk/news/2023/05/14/stop-ai-deepfakes-undermining-elections-by-updating-law/.
[176] Ibid.
[177] Electoral Commission, "The Electoral Commission's ability to bring prosecutions," https://www.electoralcommission.org.uk/news-and-views/elections-act/electoral-commissions-ability-bring-prosecutions.
[178] Ibid.
[179] CETaS policy workshop, 17 September 2024.
[180] Stockwell et al. (2024), 20.

guidance on how the media should navigate malicious activities designed to deceive voters.[181]

In light of this, the Independent Press Standards Organisation should revise its guidance on "reporting major incidents" to include key considerations for coverage of known hostile influence operations and viral disinformation content, drawing on insights from journalists and fact-checkers.[182] The new guidance could recommend against linking to the original source content in online articles – to avoid encouraging users to share misleading claims – and recommend against exaggerating the threat of these activities to the wider public.[183]

As some of these incidents could undermine public trust in the integrity of elections, it is also vital for election security officials to help restore confidence in the process when they arise. Canada's Critical Election Incident Public Protocol is an example of a mechanism designed to facilitate this type of strategic communication.[184] Involving a diverse range of senior civil servants, the Protocol sets a significantly high threshold for a public announcement – based on the extent to which the incident could damage the electoral process.[185] Moreover, announcements require a consensus and are restricted to what is known about the incident and what voters can do to protect themselves, helping maintain political impartiality.[186]

Accordingly, the Cabinet Office should liaise with the Canadian government on these issues and adopt a similar protocol for UK elections.[187] The initiative should receive buy-in from political parties to enhance its legitimacy, while adapting the Canadian model of selecting several senior experts from diverse backgrounds to determine the most appropriate course of action. Any public announcement should require a consensus between these officials.

---

[181] Vikki Julian, "IPSO Blog: Election reporting," IPSO Blog, 8 November 2019, https://www.ipso.co.uk/news-press-releases/blog/ipso-blog-election-reporting/; IPSO, "Guidance on reporting major incidents," 2 October 2024, https://www.ipso.co.uk/resources/guidance-on-reporting-major-incidents/.

[182] IPSO (2024).

[183] Melinda McClure Haughey, Rachel Moran-Prestridge and Emma S. Spiro, "Recommendations for journalists covering election rumors in 2024," Center for an Informed Public, 3 October 2024, https://www.cip.uw.edu/2024/10/03/recommendations-journalists-covering-election-rumors-2024/.

[184] Morris Rosenberg, "Report on the assessment of the 2021 Critical Election Incident Public Protocol," Government of Canada: May 2020, https://www.canada.ca/en/democratic-institutions/services/reports/report-assessment-2021-critical-election-incident-public-protocol.html; Government of Canada, "Cabinet Directive on the Critical Election Incident Public Protocol," https://www.canada.ca/en/democratic-institutions/services/protecting-democracy/critical-election-incident-public-protocol/cabinet.html.

[185] Rosenberg (2020), 26.

[186] Ibid, 27-32.

[187] Full Fact, "Full Fact Report 2024: Trust and truth in the age of AI," April 2024, 63-65, https://fullfact.org/media/uploads/ff2024/18042024-full_fact_report_corrected.pdf.

4.1.3 International deepfake legislation

Outside the UK, several nations and blocs have proposed or introduced new legislation that applies to AI-enabled election threats such as deepfakes. This ranges from legislation banning the creation of such content to requirements for disclosure of the use of AI tools, to time-sensitive bans during election periods (see table below).[188]

Table 8. Overview of international deepfake regulation

Regulation type	Summary	Country examples
Deepfake creation bans	Prohibitions on creating malicious AI-generated content such as robocalls or deepfakes, with fines and other penalties for perpetrators	US; EU Brazil
Deepfake disclosure requirements	Mandatory requirements for those circulating AI-generated election content to disclose that they have used AI tools	EU; Brazil
Deepfake election restrictions	Prohibitions on circulating AI-generated disinformation in a specific time frame before and/or after an election cycle	Singapore; South Korea

Source: Authors' analysis.

[188] Cristina Vanberghen, "The AI Act vs. deepfakes: A step forward, but is it enough?," *Euractiv*, 26 February 2024, https://www.euractiv.com/section/artificial-intelligence/opinion/the-ai-act-vs-deepfakes-a-step-forward-but-is-it-enough/; Nadine Yousif, "US FCC makes AI-generated robocalls illegal," *BBC News*, 8 February 2024, https://www.bbc.co.uk/news/world-us-canada-68240887; Beatriz Farrugia, "Regulating the use of AI for Brazilian elections: what's at stake," DFRLab, 29 May 2024, https://dfrlab.org/2024/05/29/regulating-the-use-of-ai-for-brazilian-elections-whats-at-stake/; Yonhap News Agency, "90-day ban on deepfake political ads passes parliamentary special committee," 5 December 2023, https://en.yna.co.kr/view/AEN20231205006400315; Nurdianah Md Nur, "Singapore's new law bans digitally manipulated content during elections," *The Edge Singapore*, 15 October 2024, https://sg.news.yahoo.com/singapore-law-bans-digitally-manipulated-150000386.html.

In all cases, however, legislators assume that existing laws are inadequate to tackle threats such as deepfakes, warranting entirely new legislation. Yet in creating such legislation, there is a risk that policymakers will rush into drafting new bills that duplicate pre-existing statutes or prompt legal challenges to their alleged infringement of freedom of speech.[189] Therefore, it would be beneficial to first subject existing laws to "expert examination as to their fitness of purpose [...] in the AI era," which would allow policymakers to determine the best ways to fill any gaps.[190]

For example, the Elections Act 2022 may be suited to tackling some of the challenges at the centre of the international efforts above. In particular, the law requires campaign materials published online to include imprints disclosing that they are digital content produced for elections.[191] However, there is no requirement for these materials to include imprints when they have been *digitally edited*, such as through generative AI tools or even basic content editing software. This threatens to confuse voters about the authenticity of online campaign adverts, as the publishers of this material can opt not to clarify such alterations. Given this challenge, the UK Government should table an amendment to Section 54 of the Elections Act, requiring the content of campaign adverts that have been digitally edited to be embedded with content provenance records detailing how it was edited and by whom.[192]

4.1.4 Defamation and privacy legislation

Generative AI's capacity to imitate individuals without their consent presents formidable challenges to existing privacy and defamation legislation. Many jurisdictions require an individual's consent to the use of their likeness or personal data as part of basic privacy protections, but social media platforms that allow content sharing complicate the situation. This is because personal images, videos and audio clips can still be obtained through indirect consent mechanisms, such as by agreeing to a platform's terms and conditions when creating an account.[193]

[189] AP News, "Judge blocks new California law cracking down on election deepfakes," 3 October 2024, https://apnews.com/article/california-deepfake-election-law-ee5a3d7cba3e9f5caddf91b127e4938a.
[190] Alice Dawson and James Ball, Generating Democracy: AI and the Coming Revolution in Political Communications, Demos, January 2024, 23, https://demos.co.uk/wp-content/uploads/2024/01/Generating-Democracy-Report-1.pdf.
[191] Electoral Commission, "Introducing digital imprints," https://www.electoralcommission.org.uk/news-and-views/elections-act/introducing-digital-imprints; Electoral Commission, "Statutory guidance on digital imprints," https://www.electoralcommission.org.uk/statutory-guidance-digital-imprints.
[192] CETaS policy workshop, 17 September 2024.
[193] Sara H. Jodka, "Manipulating reality: the intersection of deepfakes and the law," *Reuters*, 1 February 2024, https://www.reuters.com/legal/legalindustry/manipulating-reality-intersection-deepfakes-law-2024-02-01/,

While legislation can offer some recourse when individuals are falsely defamed, deepfakes create a high burden of proof on the victim to prove the inauthenticity of such content. [194] As AI-generated material becomes more realistic, these requirements become even harder to meet. Accordingly, the UK Ministry of Justice should conduct a review into the appropriateness of relevant defamation, privacy and electoral laws in the area, before determining whether new laws or amendments are required to prevent malicious actors from exploiting any loopholes. [195] This should cover not only the AI model output stage but also model training, because data used to train these tools could replicate an individual's likeness without their active or implicit consent.

4.2 Policy measures

4.2.1 Digital literacy and critical thinking initiatives

Digital literacy is a relatively new concept that builds on older media literacy initiatives, which are designed to teach individuals to critically engage with the media.[196] Correspondingly, digital literacy focuses on helping citizens engage with online media in "wise, safe and ethical ways" – thereby encompassing issues around AI-enabled disinformation.[197]

In 2020, there were 170 digital and media literacy initiatives in the UK designed to tackle misinformation and disinformation.[198] However, one representative survey of 2,000 adults living in the UK found that just 3% had taken a media literacy course and 7% had used self-help resources such as fact-checking tiplines.[199]

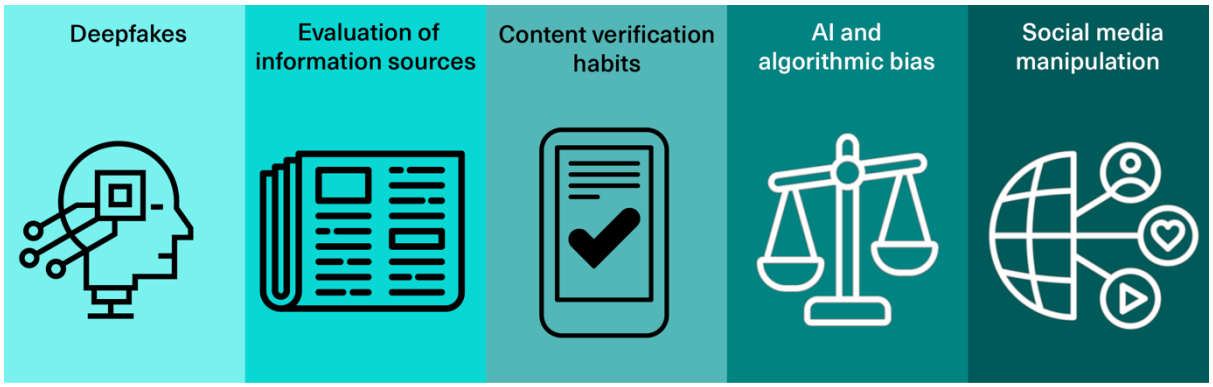
The UK's Online Media Literacy Strategy, launched in July 2021, aims to improve national media literacy skills through multi-year plans designed to inform and empower users.[200] However, it is unclear whether the latest plan, from late 2023, will receive adequate

[194] Ibid.
[195] CETaS policy workshop, 17 September 2024.
[196] Luci Pangrazio, Anna-Lena Godhe, and Alejo González López Ledesma, "What is digital literacy? A comparative review of publications across three language contexts" in *E-Learning and Digital Media* 17, No. 6 (2020: 442-459), 453-456, https://journals.sagepub.com/doi/epdf/10.1177/2042753020946291.
[197] Ibid.
[198] Department for Digital, Culture, Media and Sport, "Online Media Literacy Evidence Review – Executive Summary," October 2020, 2, https://assets.publishing.service.gov.uk/media/61129356d3bf7f0443acba68/2020-10-27_Executive_Summary_ACCESSIBLE.pdf.
[199] Florence E. Enock et al., "How do people protect themselves against online misinformation?," The Alan Turing Institute, May 2024, 4, https://www.turing.ac.uk/sites/default/files/2024-05/how_do_people_protect_themselves_from_misinformation.pdf.
[200] Department for Science, Innovation and Technology and Department for Digital, Culture, Media & Sport, "Online Media Literacy Strategy," 10 August 2021, https://www.gov.uk/government/publications/online-media-literacy-strategy.

investment to tackle existing and future digital literacy needs.[201] Early and higher education in the UK includes no standardised digital citizenship or critical media literacy courses that could help young people navigate online spaces safely.[202]

In light of these issues, the Department for Education and DSIT should coordinate to develop nationwide digital literacy and critical thinking programmes. Such schemes should be made mandatory in primary and secondary schools, while also being promoted for adults as optional courses. They should encourage a hands-on collaborative learning experience to help foster healthy habits, and should adopt an interdisciplinary approach involving critical thinking, psychology, media studies and ethics.[203] The programme's modules could also be based on various aspects of digital hygiene (see Figure 6 below).[204]

Figure 6. Overview of digital literacy and hygiene topics



Source: Adapted from Shalevska (2024).

4.2.2 Election security planning and political party conduct

With AI-enabled influence operations targeting voters during elections, it is essential to equip election officials with the necessary skills to respond to such threats. The Cabinet

[201] Full Fact (2024), 53-54; Department for Science, Innovation and Technology, "Year 3 Online Media Literacy Action Plan (2023/24)," 23 October 2024, https://www.gov.uk/government/publications/year-3-media-literacy-action-plan-202324/year-3-online-media-literacy-action-plan-202324.

[202] London School of Economics, "Dedicated digital citizen curriculum needed to help pupils navigate online dangers and tackle 'digital divide'," 11 June 2024, https://www.lse.ac.uk/News/Latest-news-from-LSE/2024/f-June-2024/Dedicated-digital-citizen-curriculum; Fiona Abades-Barclay and Shakuntala Banaji, "LSE – Common Sense Digital Citizenship Curriculum Evaluation," London School of Economics, 2024, 9-13, https://www.lse.ac.uk/media-and-communications/assets/documents/research/projects/LSE-%E2%80%94-Common-Sense-Digital-Citizenship-Curriculum-Evaluation-Report.pdf; Nadia Naffi, "Deepfakes: How to empower youth to fight the threat of misinformation and disinformation," *The Conversation*, 28 January 2024, https://theconversation.com/deepfakes-how-to-empower-youth-to-fight-the-threat-of-misinformation-and-disinformation-221171.

[203] CETaS policy workshop, 17 September 2024.

[204] Elena Shalevska, "The Future of Political Discourse: AI and Media Literacy Education" in *Journal of Legal and Political Education* 1, July 2024, 50-61, https://e-jlia.com/index.php/jlpe/article/view/1504/511.

Office and the National Cyber Security Centre (NCSC) have published mitigation plans for political candidates and election officials facing influence operations,[205] but it is unclear whether they have adopted other measures. The NCSC could explore a pilot scheme similar to one in the US state of Utah, which gave political candidates in recent congressional elections the ability to authenticate their digital identity for free, to protect against deepfakes.[206]

Alongside this, UK election security teams could draw on the tabletop exercises in states such as Arizona and Colorado. These activities tested how polling officials, the media and other stakeholders with a role in protecting election integrity would respond to AI threats in a series of emergency scenarios.[207] The organisers observed that, following the sessions, those involved adopted a greater array of cybersecurity practices when sharing confidential election information.[208]

Political parties also have an important role to play in establishing norms and accountability for the appropriate use of generative AI during election campaigns. Given that recent European and US elections saw political candidates post AI-generated election adverts without disclosure, there is a risk that such deceptive practices will become normalised and replicated at the grassroots level in future elections (also see Section 2).[209]

Although the Electoral Commission published its expectations for political parties' behaviour in the recent UK general election, this may not go far enough in its recommendations for the appropriate use of AI.[210] Consequently, the Electoral Commission should expand these expectations into more detailed guidance. This could draw on an open letter from Demos that outlined four key steps parties could take to protect voters from AI-enabled disinformation.[211] The Electoral Commission should encourage political parties to

[205] Cabinet Office, "Online disinformation and AI threat guidance for electoral candidates and officials," 17 June 2024, https://www.gov.uk/government/publications/security-guidance-for-may-2021-elections/online-disinformation-and-ai-threat-guidance-for-electoral-candidates-and-officials.
[206] Payton Davis, "Will Utah lead the way in combating election deepfakes?," *Deseret News*, 9 July 2024, https://www.yahoo.com/news/utah-lead-way-combating-election-225241670.html.
[207] David Evan Harris et al., "How Election Officials Can Identify, Prepare for, and Respond to AI Threats," Brennan Center for Justice, 8 May 2024, https://www.brennancenter.org/our-work/research-reports/how-election-officials-can-identify-prepare-and-respond-ai-threats.
[208] Ibid.
[209] Stockwell (2024), 11-14.
[210] Electoral Commission, "New advice for voters on disinformation, and for campaigners using generative AI," 17 June 2024, https://www.electoralcommission.org.uk/media-centre/new-advice-voters-disinformation-and-campaigners-using-generative-ai.
[211] Demos, "Open Letter to UK Political Parties," April 2024, https://demos.co.uk/wp-content/uploads/2024/04/AI-Pledge-Open-Letter_PDF_Final.pdf.

update their codes of conduct based on any new guidance, to ensure accountability for candidates and campaigners when using AI tools.

4.2.3 Fact-checking initiatives

Several organisations in the UK and globally play an important role in exposing election disinformation and helping voters obtain factual information about campaign developments and voting processes. Yet new AI tools make these activities increasingly challenging.

The increasingly realistic nature of generative AI outputs – along with the declining barriers to entry for those who use them to create disinformation – increases the resources, time and expertise required in fact-checking efforts. Television is still the most popular source of election news for UK voters (49%) but, *outside* an election context, more people (71%) get the news from social media platforms than from any other medium.[212] Given the rapid flow of information on social media, coupled with the way that recommender algorithms can amplify viral AI-generated disinformation for millions of users, fact-checking teams face a trade-off in minimising the public's exposure of falsehoods as quickly as possible while ensuring that their work is accurate and robust.

Community-based (or decentralised) fact-checking initiatives have emerged as a potential way to help professional organisations cope with the volume of online disinformation in circulation. These mechanisms outsource fact-checking to users, with the aggregated reviews of a large audience potentially resulting in accuracy "comparable to that of experts."[213] There are some disadvantages to relying on community-based measures as the sole source of content verification, particularly when it comes to politically divisive content.[214] Despite this, recent experiments have revealed that such initiatives can also increase trust in digital content, as well as reduce the spread of misleading posts by 62% in some contexts.[215] Following these promising results, social media platforms should invest greater resources in community-based measures to disseminate fact-checks quickly and

[212] Ofcom, "UK General Election news and opinion-formation survey 2024," September 2024, 5, https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/tv-radio-and-on-demand-research/tv-research/news/news-consumption-2024/uk-general-election-survey-2024-report.pdf?v=379617; Ofcom, "News consumption in the UK: 2024," September 2024, 3-4, https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/tv-radio-and-on-demand-research/tv-research/news/news-consumption-2024/news-consumption-in-the-uk-2024-report.pdf?v=379621.

[213] Chiara Patricia, Drolsbach, Kirill Solovev and Nicolas Pröllochs, "Community notes increase trust in fact-checking on social media" in *PNAS Nexus* 3, No. 7, July 2024, 3-14, https://doi.org/10.1093/pnasnexus/pgae217.

[214] Center for Countering Digital Hate, "Rated Not Helpful," October 2024, 9, https://counterhate.com/wp-content/uploads/2024/10/CCDH.CommunityNotes.FINAL-30.10.pdf.

[215] Drolsbach et al. (2024); Yuwei Chuai at al., "Community-based fact-checking reduces the spread of misleading posts on social media," *arXiv,* 13 September 2024, https://arxiv.org/abs/2409.08781.

make them more accessible.[216] Crucially, these schemes should integrate reputation scores and voting systems.[217] The former assign positive scores to users who consistently provide accurate fact-checks – as a form of quality control – while the latter require democratic consensus in the form of a majority vote to determine the veracity of a claim.

Solutions that embed fact-checking directly into communication apps have also shown promise. In Taiwan, the messaging app LINE includes an in-app chatbot that allows users to submit website links for analysis and verification against previously fact-checked content. If the chatbot cannot match any of the existing data, users can forward their message for manual fact-checking.[218] During the COVID-19 pandemic, the service supported more than 230,000 users and analysed more than 40,000 messages to help counter health-related disinformation.[219]

Given the success of these initiatives, Ofcom should engage with the International Fact-Checking Network and UK-based communication app providers to explore how such services can be replicated for users.[220] Any such services should place the user journey at the heart of app design to maximise engagement, while helping users verify content by promoting alternative trusted news sources for cross-referencing.[221]

4.2.4 Disinformation research access

During the early 2000s, disinformation researchers supported the work of online safety teams in monitoring hostile influence operations on social media platforms – which, in turn, informed countermeasures designed to protect users.[222] At the time, historical data was relatively easy to access and research organisations were respected for their role in exposing hostile actors' efforts to influence citizens.[223] However, social media platforms have since shifted towards restricting researcher activities and data access, in the aftermath of what has been described as the 'techlash' against these platforms during the 2010s,

[216] CETaS policy workshop, 17 September 2024.
[217] FP Team, "What is Decentralized Fact-checking?," *Fact Protocol*, 8 January 2023, https://fact.technology/learn/what-is-decentralized-fact-checking.
[218] Elizabeth Lange and Doowan Lee, "How One Social Media App Is Beating Disinformation," *Foreign Policy*, 23 November 2020, https://foreignpolicy.com/2020/11/23/line-taiwan-disinformation-social-media-public-private-united-states/.
[219] Ibid.
[220] CETaS policy workshop, 17 September 2024; CETaS technical workshop, 19 September 2024.
[221] CETaS technical workshop, 19 September 2024.
[222] David Karpf, "Back to Basics: Studying Digital Campaigning While Our Objects of Analysis Are in Flux," *PolComm*, https://politicalcommunication.org/article/back-to-basics/.
[223] Ibid.

along with stringent new privacy obligations imposed by legislation such as the General Data Protection Regulation 2018.[224]

Social media platforms such as Facebook have revoked researchers' access to their data and, after initially shifting to bespoke tools such as CrowdTangle, are now closing these systems down.[225] Others, such as X (formerly Twitter), have made large-scale data access prohibitively expensive.[226] In parallel, lawsuits have deterred disinformation researchers from carrying out their work in the area.[227]

As access to social media data has deteriorated, some jurisdictions have sought to overcome these challenges through legislation. For instance, the EU's Digital Services Act now requires tech firms to provide independent researchers with access to data on how their services affect politics, including elections.[228] The UK's OSA does not have the same requirements.[229] However, given that most social media platforms are based in the US, UK and EU regulators will face difficulties in enforcing any rules they introduce – as recent developments in Brazil and Australia have shown.[230]

Taking these challenges into account, legislation such as that proposed in the Digital Information and Smart Data Bill should include a provision for establishing a trusted research group on disinformation. This amendment would require social media platforms to provide access to data on identified hostile influence operations to trusted members of the UK academic, research and civil society communities.[231] Drawing on X's former data access model, platforms should give free application programming access to these communities as

[224] Ibid; Annabel Latham, "Cambridge Analytica scandal: legitimate researchers using Facebook data could be collateral damage," *The Conversation*, 20 March 2018, https://theconversation.com/cambridge-analytica-scandal-legitimate-researchers-using-facebook-data-could-be-collateral-damage-93600.
[225] Alex Krasodomski, "The US election will take place in a polluted information space," Chatham House, 11 September 2024, https://www.chathamhouse.org/2024/09/us-election-will-take-place-polluted-information-space; Carissa Goodwin and Dean Jackson, "Global Perspectives on Influence Operations Investigations: Shared Challenges, Unequal Resources," Carnegie Endowment for International Peace, February 2022, https://carnegieendowment.org/research/2022/02/global-perspectives-on-influence-operations-investigations-shared-challenges-unequal-resources?lang=en.
[226] Krasodomski (2024).
[227] Ibid.
[228] Christian Hetrick, "How to spot AI fake content—and what policymakers can do to help stop it," *Phys.org*, 4 July 2024, https://phys.org/news/2024-07-ai-fake-content-policymakers.html.
[229] Center for Countering Digital Hate, Social Media's Role in the UK Riots: Policy Responses and Solutions, August 2024, 5, https://counterhate.com/wp-content/uploads/2024/08/240819-Convening-Policy-Paper-FOR-DESIGN-WEBSITE.pdf.
[230] Karpf; Ione Wells and Andre Rhoden-Paul, "X suspends business in Brazil over censorship row," *BBC News*, 17 August 2024, https://www.bbc.co.uk/news/articles/cgjv857plevo; Jake Evans and Jordyn Butler, "eSafety drops case against Elon Musk's X over church stabbing videos," *ABC News*, 5 June 2024, https://www.abc.net.au/news/2024-06-05/esafety-elon-musk-x-church-stabbing-videos-court-case/103937152.
[231] CETaS policy workshop, 17 September 2024; CETaS technical workshop, 19 September 2024.

they seek to understand users' behaviour and trends in malicious online activities.[232] To maintain impartiality in the selection process, organisations and individuals should be chosen through UK Research and Innovation's trusted research and innovation programme.[233]

[232] Huo Jingnan, "Twitter's new data access rules will make social media research harder," *NPR*, 9 February 2023, https://www.npr.org/2023/02/09/1155543369/twitters-new-data-access-rules-will-make-social-media-research-harder.
[233] UK Research and Innovation, "Trusted research and innovation," 8 July 2024, https://www.ukri.org/manage-your-award/good-research-resource-hub/trusted-research-and-innovation/.

5. Technical Solutions to AI-Enabled Election Threats

Various technical measures can help protect elections against AI-enabled interference. The technical countermeasures described in this section are non-exhaustive, including only those the project team considers to be most promising based on a combination of literature review and workshop insights.

5.1 Prevention methods

The best way to tackle AI-enabled election disinformation is to prevent the generation of such content in the first place. Most generative AI models are trained on large, uncurated datasets scraped from the web. This is a major reason why these models often generate harmful content.[234] Careful dataset curation – including limits on training data on high-profile individuals likely to be targeted by deepfakes, such as politicians – would, *in principle*, help restrict malicious actors' exploitation of AI tools.[235] However, *in practice*, such dataset curation is likely unworkable due to the sheer scale of the datasets on which large models are trained.

One alternative approach to the challenge involves controlling the output of AI models through techniques such as reinforcement learning with human feedback. In this, a second 'reward' model is trained on human-annotated text to predict how 'good' or 'desirable' a particular response is. This can not only improve the general quality of outputs but also to reduce the production of harmful content.[236] Another approach involves model guardrails, which filter the *output* of models to detect features associated with harmful content.[237]

Many of these preventative techniques are only partially effective. This is because many AI models have already been open-sourced, meaning that it is too late to influence their

[234] Abeba Birhane et al., "On Hate Scaling Laws for Data-Swamps," *arXiv*, 28 June 2023, 2, https://arxiv.org/abs/2306.13141.
[235] Amandalynne Paullada et al., "Data and its (dis)contents: A survey of dataset development and use in machine learning research" in *Patterns: Cell Press* 2, No. 11, 12 November 2021, 9-10, https://www.cell.com/patterns/fulltext/S2666-3899(21)00184-7; Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (1 March 2021: 610-623), 615-616, https://dl.acm.org/doi/abs/10.1145/3442188.3445922; Ofcom, "Deepfake Defences: Mitigating the Harms of Deceptive Deepfakes," 23 July 2024, 18, https://www.ofcom.org.uk/siteassets/resources/documents/consultations/discussion-papers/deepfake-defences/deepfake-defences.pdf?v=370754.
[236] Yuntao Bai et al., "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback," *arXiv*, 12 April 2022, https://arxiv.org/abs/2204.05862.
[237] Yi Dong et al., "Building Guardrails for Large Language Models," *arXiv*, 2 February 2024, https://arxiv.org/abs/2402.01822v1.

development or training. Such models are also more likely to accept instructions to generate disinformation, while malicious actors can fine-tune their operations to circumvent output restrictions.[238] However, technical methods for preventing harmful content generation can still increase the resource costs for, and technical barriers facing, malicious actors. Accordingly, they have a place in the design of future AI models.

5.2 Content detection methods

Given that it is impossible to prevent all malicious or even irresponsible use of generative AI, it is inevitable that some deceptive content will enter the online information environment. Therefore, effective content moderation is essential. However, the sheer volume of posts on social media far outstrips the capacity of human moderators.[239] As such, automated detection of malicious content could become a useful aid, enabling the triage or flagging of AI-generated content.[240]

So far, the most promising content-based detection methods are based on AI. Approaches based on deep learning models have proven to be the most effective at detecting both AI-generated text and deepfakes.[241] However, these tools are only as good as the data on which they are trained – with their development dependent on the availability of good performance benchmarks.[242] In this context, 'good' means access to not only a *representative* sample but also real data. This has proven challenging in the deepfake domain, with the most common

[238] Angus R. Williams et al., "Large language models can consistently generate high-quality content for election disinformation operations," *arXiv*, 13 August 2024, 20, https://arxiv.org/abs/2408.06731; Xiangyu Qi et al., "Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!," *arXiv,* 5 October 2023, https://arxiv.org/abs/2310.03693; Nicolas Carlini et al., "Are aligned neural networks adversarially aligned?," *arXiv*, 26 June 2023, https://arxiv.org/abs/2306.15447.

[239] R. Michael Alvarez, Frederick Eberhardt and Mitchell Linegar, "Generative AI and the Future of Elections," Center for Science, Society and Public Policy, July 2023, 5, https://lindeinstitute.caltech.edu/documents/25475/CSSPP_white_paper.pdf.

[240] Ibid.

[241] Rowan Zellers et al., "Defending Against Neural Fake News" in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, no. 812 (8 December 2019: 9054 – 90), https://proceedings.neurips.cc/paper/2019/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html; Amal Naitali et al., "Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions" in *Computers* 12, no. 10, 23 October 2023, 10-11, https://www.mdpi.com/2073-431X/12/10/216; Gan Pei et al., "Deepfake Generation and Detection: A Benchmark and Survey," *arXiv*, 16 May 2024, https://arxiv.org/abs/2403.17881; Md Shohel Rana et al., "Deepfake Detection: A Systematic Literature Review" in *IEEE Access* 10, 2022, 25,494-25,513, https://ieeexplore.ieee.org/document/9721302.

[242] Amandalynne Paullada et al. (2021); David Donoho, "50 Years of Data Science" in *Journal of Computational and Graphical Statistics* 26, No. 4, December 2017, 745-766, https://www.tandfonline.com/doi/full/10.1080/10618600.2017.1384734.

benchmarks lacking even containing media generated by diffusion models – an increasingly common architecture in the creation of deepfakes.[243]

A more fundamental problem with the detection approach is that innovation in defensive methods will doubtlessly inspire innovation in malicious generation techniques.[244] For instance, an early deepfake detection method focused on the fact that individuals depicted in deepfake videos typically did not blink.[245] In response, the creators of deepfakes innovated to include blinking.[246] In light of this, the UK's AI Safety Institute and Home Office should coordinate with each other to create and continually update deepfake detection benchmarks, to help designers of such tools maintain their effectiveness against changes in adversaries' tradecraft.[247]

These detection models could be adopted by not only social media platforms and researchers but also members of the public, through browser plug-ins and apps. Yet caution is needed when presenting the outputs of these models to users. For instance, a user may not understand the practical implications of a 95% probability that an image is AI-generated. Therefore, deepfake detection developers need guidance on the technical information to provide to users of such tools.[248] This should include key details on: the purpose of the tool; how it should be used and interpreted; the explainability of a model's output; and its limitations. The AI Safety Institute's remit on AI model safety testing and the Home Office's recent Deepfake Detection Challenge initiative indicate why the two organisations should coordinate with each other on these activities.[249]

---

[243] Natali et al. (2023); Pei et al. (2024); Phil Swatton and Margaux Leblanc, "What are deepfakes and how can we detect them?," *The Alan Turing Institute*, 7 June 2024, https://www.turing.ac.uk/blog/what-are-deepfakes-and-how-can-we-detect-them.

[244] Noémi Bontridder and Yves Poullet, "The role of artificial intelligence in disinformation" in *Data & Policy* 3, No. e32, 25 November 2021: 1-21, 8-9, https://www.cambridge.org/core/journals/data-and-policy/article/role-of-artificial-intelligence-in-disinformation/7C4BF6CA35184F149143DE968FC4C3B6; Michael Brückner and Tobias Scheffer, "Stackelberg games for adversarial prediction problems" in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 21 August 2011, 547-555), https://dl.acm.org/doi/abs/10.1145/2020408.2020495.

[245] Bontridder & Poullet (2021), 8-9.

[246] Ibid.

[247] CETaS technical workshop, 19 September 2024.

[248] Ibid.

[249] Department for Science, Innovation and Technology, AI Safety Institute and Michelle Donelan, "AI Safety Institute releases new AI safety evaluations platform," 10 May 2024, https://www.gov.uk/government/news/ai-safety-institute-releases-new-ai-safety-evaluations-platform; Kate Shanks, "Innovative solutions unveiled at the Deepfake Detection Challenge Showcase," *Accelerated Capability Environment*, 30 July 2024, https://ace.blog.gov.uk/2024/07/30/innovative-solutions-unveiled-at-the-deepfake-detection-challenge-showcase/.

5.3 Social bot detection methods

Given the inherent problems in content detection, another approach is to focus on the detection of fake *accounts* (or social bots). While there are many types of AI-generated disinformation content, there are often similar patterns of behaviour in the dissemination and amplification of such content. Typically, those seeking to interfere in democratic processes will use a combination of bot accounts and troll farms to post malicious content and amplify both their own and other selected content.[250]

The Kremlin-supported influence operation in the 2016 US presidential election made extensive use of troll farms to shape voter discourse, while evidence from the 2024 London mayoral election showed how bots sought to spread false allegations of voter fraud.[251] In many cases, these accounts tended to focus on well-established disinformation methods of 'astroturfing' (i.e. flooding social media posts with political narratives to influence users) and 'information laundering' (i.e. creating a fake news source citing disinformation to improve its perceived credibility).[252]

One can detect bot accounts based on many different cues, including behavioural, linguistic and network patterns.[253] As in other domains, neural networks have become increasingly predominant for such detection.[254] And some researchers have worked to automate the detection of malicious troll accounts.[255] However, as with content detection, innovation in bot detection techniques can lead to innovation in inauthentic behaviour. Therefore, the most effective response is to identify and target the *owners* of bot accounts (rather than the accounts themselves) using techniques such as demonetisation.

---

[250] Mike Wendling, "US officials uncover alleged Russian 'bot farm'," *BBC News*, 10 July 2024, https://www.bbc.co.uk/news/articles/c4ng24pxkelo; Foreign, Commonwealth & Development Office, Elizabeth Truss and Nadine Dorries, "UK exposes sick Russian troll factory plaguing social media with Kremlin propaganda," 1 May 2022, https://www.gov.uk/government/news/uk-exposes-sick-russian-troll-factory-plaguing-social-media-with-kremlin-propaganda

[251] Adam Badawy et al., "Characterizing the 2016 Russian IRA influence campaign" in *Social Network Analysis and Mining* 9, No. 31, 8 July 2019, https://doi.org/10.1007/s13278-019-0578-6; Stockwell et al. (2024), 25-27.

[252] Stockwell (2024), 15-17; David Schoch et al., "Coordination patterns reveal online political astroturfing across the world" *in Scientific Reports* 12, No. 4572 (17 March 2022), https://www.nature.com/articles/s41598-022-08404-9; Joe Littell, "The Future of Cyber-Enabled Influence Operations: Emergent Technologies, Disinformation, and the Destruction of Democracy" in *The Great Power Competition Volume 3*, ed. Adib Farhadi et al., Springer, 2022, 197-227, https://link.springer.com/chapter/10.1007/978-3-031-04586-8_10.

[253] Emilio Ferrara, "Social bot detection in the age of ChatGPT: Challenges and opportunities" in *First Monday* 28, No. 6, 15 June 2023, https://firstmonday.org/ojs/index.php/fm/article/view/13185.

[254] Ibid.

[255] Fatima Ezzeddine et al., "Exposing influence campaigns in the age of LLMs: a behavioral-based AI approach to detecting state-sponsored trolls" in *EPJ Data Science* 12, No, 1 9 October 2023, https://epjds.epj.org/articles/epjdata/abs/2023/01/13688_2023_Article_423/13688_2023_Article_423.html.

Ofcom should issue a new draft Code of Conduct for social media platforms, drawing inspiration from the EU's Code of Practice on Disinformation, which seeks to establish a set of definable behaviours associated with disinformation operators based on the use of tools such as bot accounts.[256] The new guidance should provide similar self-regulatory standards to counter disinformation, including that on demonetising the creators of disinformation content; defining unpermitted manipulative behaviours associated with bot accounts; creating tools to empower users against disinformation; and establishing transparent incident reporting.[257]

5.4 Content provenance

If one is unable to prevent the generation or viral dissemination of AI-enabled disinformation, the next best step is to help users make informed judgements concerning such media. Content provenance encompasses a broad set of techniques for digitally signing media, to preserve information on how it was created. That is, one should know which device and software was used to create the media and who created or edited it.[258] It is important to stress that this involves an assessment only of the *source* of content, not of its *accuracy*.[259]

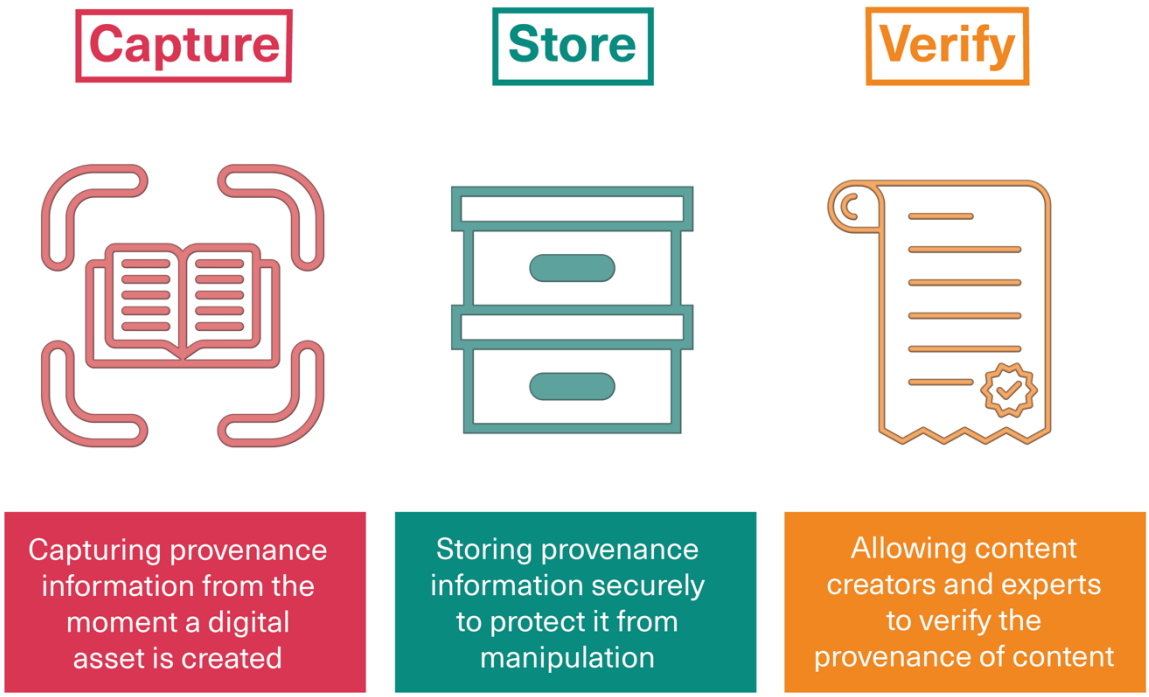
Watermarking is one process to ensure that AI-generated content carries provenance information. These watermarks can be clearly visible, but they are often invisible to the human eye.[260] They can be applied either during content generation or afterwards.[261] However, watermarks can often be easily removed by malicious actors, sometimes simply by screenshotting the image in question.[262]

A more effective approach to provenance may lie in embedding authenticity-by-design principles in different parts of the internet ecosystem – from the devices used to create

[256] The EU Commission, "The 2022 Code of Practice on Disinformation," https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation.
[257] Ibid.
[258] Charlie Halford, "Mark the good stuff: Content provenance and the fight against disinformation," *BBC Research and Development*, 5 March 2024, https://www.bbc.co.uk/rd/blog/2024-03-c2pa-verification-news-journalism-credentials.
[259] C2PA Specifications, "C2PA Explainer," https://c2pa.org/specifications/specifications/1.4/explainer/Explainer.html.
[260] Sasha Luccioni et al., "AI Watermarking 101: Tools and Techniques," *Hugging Face*, 26 February 2024, https://huggingface.co/blog/watermarking.
[261] Vivien Chappelier, "Robust image watermarking with Stable Signature + IMATAG's BZH," *Hugging Face*, 22 January 2024, https://huggingface.co/blog/imatag-vch/stable-signature-bzh.
[262] David Evan Harris and Lawrence Norden, "Meta's AI Watermarking Plan Is Flimsy, at Best: Watermarks are too easy to remove to offer any protection against disinformation," *IEEE Spectrum*, 4 March 2024, https://spectrum.ieee.org/meta-ai-watermarks.

content to their publication on online platforms (see Figure 7 below).[263] Based on frameworks such as that of the Starling Lab, this method involves creating and preserving provenance data "throughout the entire lifecycle of a digital media asset," establishing a verifiable digital chain of custody to help users understand the context of an information source.[264]

Figure 7. Authenticity-by-design framework



Source: Adapted from the Starling Lab framework.

Given the need to universalise and adapt these measures in various parts of the internet, international standards organisations will be vital to their implementation. The Internet Engineering Task Force's Security Area, which focuses on the creation of a trusted internet, should facilitate the uptake of authenticity-by-design principles among users, applications and devices.[265] The initiative should aim to embed tools to automatically capture, store and verify digital provenance records into every part of the internet's infrastructure – without undermining user privacy.

[263] Lindsay Walker and Adam Rose, "The Starling Lab Framework," *FFDWeb*, 6 March 2024, https://www.ffdweb.org/blog/the-starling-lab-framework.
[264] Ibid.
[265] IETF, "Security & privacy," https://www.ietf.org/technologies/security/.

There is a risk that users will misinterpret content provenance markings. While such credentials provide a useful audit trail of who created the content and when, they can be mistaken for an indication that the content is factual. [266] Malicious actors could exploit this by providing transparent information about the origins of their content but embedding it within misleading messages.

Yet in many ways, the main objective of content provenance is to increase user engagement with verifiable and credible information sources.[267] As a result, users could be more sceptical of content originating from sources that did not include such credentials or that originated from unverified accounts, helping undermine malicious actors who sought to pollute the information space with falsehoods.

Given the importance of rolling out these features ahead of future elections, DSIT should publish an implementation strategy for automatically embedding secure provenance records in digital content produced by the UK Government and other trusted sources, such as news outlets, at its origin. This could draw on the US Office of Management and Budget's requirement to issue similar guidance by June 2025.[268]

[266] K.J. Kevin Feng et al., "Examining the Impact of Provenance-Enabled Media on Trust and Accuracy Perceptions" in *Proceedings of the ACM on Human-Computer Interaction* 7, No. 270, 4 October 2023, 1-42, https://dl.acm.org/doi/abs/10.1145/3610061.
[267] Halford (2024).
[268] Microsoft, "Protecting the Public from Abusive AI-Generated Content," 2024, 33, https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1nuJx.

Conclusion

AI technology may not have fundamentally reshaped or disrupted the outcomes of key elections in 2024, but this is no cause for complacency.

Those involved in election security should learn from the emerging body of evidence on AI use as they work to protect future elections. Upcoming national elections in Canada and Australia could become targets of AI misuse, as could local and regional elections around the world. Therefore, it is crucial to develop initiatives to share knowledge between nations in the area.

Future research should avoid amplifying unnecessary speculation about AI-enabled threats to democracy. Public debates on the issue in early 2024 often lacked any evidence base. Frequent coverage can allow minor cases of disinformation to gain traction that they would have otherwise lacked. Accordingly, journalists and researchers need to inform the public without overplaying the impact of AI misuse, striking a careful balance in line with some of the recommendations above.

The gaps for further research in these areas include:

1) Studies of how engagement with hostile influence operations damages public trust in the information environment, the media and government institutions.

2) Controlled experiments to understand public engagement with printed versus online AI-generated political disinformation.

3) Surveys and focus groups to understand how exposure to deepfakes and other forms of AI-generated election disinformation affects voting behaviour and political beliefs.

4) Behavioural analysis of users who generate disinformation.

5) Tests of promising AI-based solutions to extreme political polarisation at the national level.

About the Authors

Sam Stockwell is a Research Associate at the Centre for Emerging Technology and Security (CETaS). His research interests focus on the intersection between national security and the online domain, particularly in relation to countering radicalisation and disinformation through both policy and technical solutions. Prior to joining the Turing, Sam worked on a wide portfolio of defence and security research at RAND – spanning military workforce issues, human security concerns, UK defence strategy and emerging technologies.

Megan Hughes is a Research Associate at CETaS. Her research explores the impact of AI on intelligence tradecraft and the information environment. Prior to joining the Turing, Megan worked as an Analyst within the Defence and Security research group at RAND Europe. Her research has informed strategy and policy at the UK Home Office, the UK Ministry of Defence, the European Commission and the United Nations Development Programme.

Dr Phil Swatton is a Data Scientist in the Turing's Applied Research Centre for Defence and Security. He works on a range of projects on and adjacent to deep learning, including that on the effect of dataset similarity on transfer attack success and low-cost measures for pre-trained model selection. Prior to joining the Turing, Phil obtained a PhD from the University of Essex. His thesis focused on the measurement of ideology from survey data, and the application of those measurements under the paradigm of measurement inference to empirical research in political science.

Albert Zhang is an analyst at the Australian Strategic Policy Institute specialising in cyber, technology and security matters. His research delves into the intersection of foreign interference, encompassing influence operations, disinformation and cyber-espionage, covering emerging technologies such as artificial intelligence, extended reality and persuasive technologies.

Jonathan Hall is a Senior Visiting Fellow at CETaS. He is a leading barrister and the UK's Independent Reviewer of Terrorism Legislation, first appointed in May 2019 by the Home Secretary and reappointed in March 2022 for a further three-year term. In February 2024, he was also appointed the first Independent Reviewer of State Threat Legislation.

Kieran is a Data Scientist within the UK Government.

Centre for Emerging Technology and Security

RESEARCH REPORT