LESSONS LEARNED

Sam Stockwell *examines what one of the busiest years for elections teaches us about future AI threats*

As the 'year of elections' comes to a close, researchers are urging policy makers, regulators and social media platforms to reflect on the experiences of polls across the world and put measures in place to safeguard against future AI-enabled threats.

In a new report, researchers from the Alan Turing Institute's Centre for Emerging Technology and Security (CETaS) have called for new measures including a requirement for social media platforms to provide specific access to data on harmful disinformation campaigns, alongside a package of actions to both deter the creation and sharing of disinformation and help authorities and the public reduce its impact.

CETaS researchers analysed all major elections over the past year and in common with other key contests, the US election saw a number of examples of viral AI disinformation. These included AI-generated content used to undermine candidates, AI bot farms mimicking US voters or spreading conspiracy theories, through to fabricated celebrity endorsements.

And while researchers found a lack of evidence showing any measurable impact on the US election result, fears remain that AI-generated threats, and the hype that surrounds them, are eroding trust in the information environment and allowing harmful narratives to thrive.

The new report calls for action in four key areas.

Curtailing generation: Measures to increase barriers or deter the creation of online disinformation. This includes: reviewing defamation, privacy and electoral laws; and strengthening the authenticity of credible information through automatically embedding provenance records in digital content produced by the government and other sectors

Constraining dissemination: Measures to reduce the effectiveness and virality of disinformation. This includes: creating government co-ordinated benchmarks for deepfake detection tools; a new Ofcom Code of Conduct on online disinformation; and expanded Electoral Commission guidance for UK political parties on their use of AI in election campaigning

Counteracting engagement: Measures which target the ways that people consume disinformation on digital platforms to reduce malicious influence. This includes: urging the Independent Press Standards Organisation to revise guidance on 'reporting major incidents' to include key considerations for coverage on viral disinformation content – drawing on insights from journalists and fact-checkers.

Empowering society: Measures which strengthen societal capabilities for exposing and undermining online disinformation. This includes: analysing gaps in the powers of both the Electoral Commission and Ofcom; requiring social media platforms to provide data access on identified harmful disinformation campaigns for trusted members of the UK academic, research, and civil society community; and establishing nationwide digital literacy and critical thinking programmes.

More than 2 billion people went to the polls this year, providing us with unprecedented evidence of the types of AI-enabled threats we face and a golden window of opportunity to protect future elections. We should be reassured that there's a lack of evidence that AI has changed the course of an election result, but there can be no complacency. Researchers and others monitoring these issues must urgently be given better access to social media platform data, in order to effectively assess and counter the most serious malicious voter-targeting activities moving forward.

The report also looks at how the public engage with mis and disinformation, in a new era where AI is enabling dissemination of larger volumes of content as well as generating more personalised, realistic fake content. Digital literacy and critical thinking initiatives show promise, but surveys have shown that very few people have used resources that could build their resilience against disinformation. Report authors urge the government to introduce mandatory programmes in primary and secondary schools, along with providing materials for adults – covering issues like deepfakes, how to verify content, and how AI algorithms work.

Doctor Alexander Babuta, Director of CETaS at the Alan Turing Institute, notes: "Both malicious content itself and the fear of election interference can be equally damaging to democracy and sow confusion among voters about what's real and what's fake. Taking action now will mitigate the threats posed by these rapidly developing technologies but also reassure voters that future elections can be held securely in the age of AI."

AI-generated content has been used to undermine candidates while AI bot farms have mimicked voters and spread conspiracy theories

SAM STOCKWELL is lead author of the report and research associate at the Alan Turing Institute.