DEMOS

GENERATIVE AI AND DEMOCRACY
IMPACTS AND INTERVENTIONS

BRIEFING PAPER

ELIZABETH SEGER

APRIL 2024

Published by Demos April 2024

15 Whitehall, London, SW1A 2DD
T: 020 3878 3955
hello@demos.co.uk
www.demos.co.uk

ABOUT THIS PROJECT

Demos is Britain's leading cross-party think tank. We put people at the heart of policy-making to create bold ideas and a more collaborative democracy.

This project is spans Demos' focus on Trustworthy Technology and on building a more Collaborative Democracy. With emerging technologies transforming our world at an ever faster pace, we work to build bridges between politicians, technical experts, and citizens to explore solutions, build trust, and create policy to ensure our technologies benefit the public. We have been digging into the near and long term impacts of generative AI on the health of our democratic institutions - institutions which are already under considerable stress from record low levels of trust in political leaders, rampant disinformation, and rising polarisation. We recommend immediate, targeted actions to mitigate risks of AI to impending elections. But more importantly, we emphasise the need for prolonged, multi-stakeholder processes that build partnerships between tech developers, politicians, and citizens to build a healthy and thriving future for democracy with technology.

This report was partially funded by Google DeepMind.

INTRODUCTION

Over the next year it is estimated that about half of the world's population will be going to the polls. At the same time, it is also an unprecedented time for AI with massive leaps in AI capability, functionality, and public accessibility. With very little time, expertise, or expense, individual actors can now produce artificial video and audio content of real people that is nearly indistinguishable from live recording. These "deepfakes", many worry, could have profound impacts on global 2024 elections and on the health and stability of democracy more broadly.

Concerns about the impacts of new and emerging technologies on democratic processes and institutions are not new.[1] The Cambridge Analytica scandal of 2016 brought global attention to how precision content targeting on social media can affect voter decisions.[2] Numerous scholars attend to the damaging effects of rampant online disinformation and social media "filter bubbles" that have emerged alongside increasing polarisation and distrust in political leaders.[3,4,5] AI experts even forewarn about the possibility of using generative AI to not only fake damaging video and audio of politicians, but also to create full interactive conversations.[6] Many have warned of the potential for irresponsible use by domestic actors[7] and by malicious foreign adversaries.[8]

The rash of 2024 elections around the world now presents a pressure point: this year's election cycles, saddled with the newly released generative AI capabilities, could result in significant damage to democratic institutions already stressed by rising distrust and dissatisfaction.

In this briefing paper:

Section 1 gives a brief primer on the current state of Generative AI capabilities.

Section 2 outlines what we posit are four pressing mechanisms by which generative AI challenges the stability of democracy.

- **Disinformation:** Producing convincing disinformation to influence what people believe and to sow distrust
- **Online abuse:** Producing online abuse material to force political disengagement of target individuals or demographics
- **Cyber attacks:** Scaling cyber attacks on election infrastructure and political campaigns
- **Concentration of power:** Strengthening concentration of social, economic, and political power in big tech companies

For each, we emphasise the additional 'marginal risk' generative AI poses to democracy. Marginal risk describes the additional risk the technology poses beyond that posed by existing technologies. For example, how much more of a challenge do AI image generators pose above and beyond that already posed by Photoshop? Attending to marginal risk is important to prevent undue hype and fear mongering and to ensure recommended interventions are proportional to the threat posed.

In addition to their direct impacts, the above mechanisms have the cumulative effect of further degrading trust in our information ecosystems,

1 Seger, E. et al. (2020). *Tackling threats to informed decision making in democratic societies Promoting epistemic security in a technologically-advanced world.* The Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/2020-10/epistemic-security-report_final.pdf
2 Melaugh, J., & Hern, A. (2018, May 7). *Cambridge Analytica: how did it turn clicks into votes?* The Guardian. Retrieved April 12, 2024, from https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie
3 Coeckelbergh, M. (2023). Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI Ethics* (3) 1341–1350 https://doi.org/10.1007/s43681-022-00239-4
4 Goldsworthy, A., Osborne, L., & Chesterfield, A. (2021). *Poles Apart: Why People Turn Against Each Other, and How to Bring Them Together.* United Kingdom: Random House.
5 O'Connor, C., & Weatherall, J. O. (2019). *The Misinformation Age: How False Beliefs Spread.* Yale University Press.
6 Horvitz, E. (2022). *On the Horizon: Interactive and Compositional Deepfakes.* ACM International Conference on Multimodal Interaction (ICMI '22). https://arxiv.org/abs/2209.01714
7 Chowdhury, R. (2024, April 9). *AI-fuelled election campaigns are here — where are the rules?* Nature. Retrieved April 12, 2024, from https://www.nature.com/articles/d41586-024-00995-9
8 Risk in Focus: Generative A.I. and the 2024 Election Cycle. (2024, January 18). Cybersecurity & Infrastructure Security Agency. Accessed April 11, 2024, from https://www.cisa.gov/sites/default/files/2024-01/Consolidated_Risk_in_Focus_Gen_AI_ElectionsV2_508c.pdf

in our own capacities for well-informed political engagement, and in the overall resilience of our democratic institutions to the modern technological age.

Section 3 attends to solutions.

We outline recommendations for immediate action by government and AI companies to mitigate the threat generative AI poses to the pending 2024 elections. **Our recommendations fall in the following categories:**

- AI and social media company commitments
- Political party campaign commitments
- Public awareness campaign
- Real time research on AI impacts
- Starting work on fresh legislation

We conclude in Section 4 with a call for prolonged engagement in a longer term strategy for systematically and rigorously addressing these challenges to democracy through collaborative and multipronged efforts bringing together government, industry, and civil society.

1. GENERATIVE AI PRIMER

Generative AI describes a class of artificial intelligence that produces new media content. Current generative AI systems can be used to create text (e.g. GPT-4)[9], images (e.g. Midjourney)[10], audio (e.g. AudioCraft)[11], or video (e.g. Sora)[12] that is nearly indistinguishable from the real thing. The models function by learning patterns from training data (e.g. images of people) to generate unique outputs with the same properties.

Many companies host generative AI models behind user friendly API's (Application Programming Interfaces). These interfaces allow users to easily query the models to produce images, video, audio, or text without accessing the model directly. By mediating model access via API, companies can implement safety filters to limit the kind of content a model is able to produce, and watermarks to identify the images as artificially produced. The challenge, however, is that safety filters can often be bypassed by clever prompt injection, and watermarks are often easily removed. More so, there are many open-source model options (models that can be downloaded and run locally) that have no safety filters or that users can directly modify by removing safety filters or fine-tuning with specialised data sets.

Currently there are no foolproof technical solutions for identifying artificially generated content. Research is underway to increase the difficulty of watermark removal and to establish content provenance procedures to allow viewers to see how an image was produced and if and how it has been modified over the course of its life.[13] These solutions are, however, a work in progress.

9 https://openai.com/research/gpt-4
10 https://midjourney.co/
11 https://about.fb.com/news/2023/08/audiocraft-generative-ai-for-music-and-audio/
12 https://openai.com/sora
13 *AI Elections accord - A Tech accord to Combat Deceptive Use of AI in 2024 Elections.* (2024). Retrieved April 12, 2024, from https://www.aielectionsaccord.com/

2. MECHANISMS OF AI IMPACT ON DEMOCRACY

In this section we briefly outline four mechanisms by which generative AI may have an impact on the 2024 elections and the stability of democracy more generally. This is not an exhaustive list, but we posit they are some of the more pressing mechanisms at play.

I. Producing convincing disinformation to influence what people believe and to sow distrust

One straightforward mechanism by which generative AI can impact democracy is by producing high-quality and convincing disinformation to influence what people believe - about politicians, political issues and voting procedures - and to sow distrust in our information ecosystems and in ourselves as discerning information consumers.

(i) Influencing what people believe

Deepfakes often target political leaders, depicting them saying or doing things that never happened. For example, during the UK's Labour Party Conference a deepfake video was circulated on social media of party leader Keir Starmer verbally abusing a staffers.[14] In Slovakia, artificially generated audio seemingly evidenced Liberal Progressive leader, Michal Simecka, and journalist Monika Todova discussing how to rig the national election.[15] This situation was complicated by a moratorium on political campaigning and media coverage in the 48 hours prior to the election which made public debunking of the deepfake difficult. In the longer term, such media moratoriums may need to be reviewed for their efficacy in the modern technological age.

Deepfakes can also be produced with the aim of disenfranchising target voter groups with disinformation about voting procedures, times, and locations. For example, in January 2024 New Hampshire voters received deepfaked robocalls of US President Joe Biden telling voters not to vote in the presidential primaries. The voice told voters, "your vote makes a difference in November, not this Tuesday," and to save their votes for the general election.[16] For voters lacking a clear understanding of US presidential election procedures, the fake Biden audio could come across as a convincing

14 Bristow, T. (2023, October 9). *Keir Starmer suffers UK politics' first deepfake moment. It won't be the last.* POLITICO.eu. Retrieved April 12, 2024, from https://www.politico.eu/article/uk-keir-starmer-labour-party-deepfake-ai-politics-elections/
15 Meaker, M. (2023, October 3). *Slovakia's Election Deepfakes Show AI Is a Danger to Democracy.* WIRED. Retrieved April 12, 2024, from https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/
16 Elliott, V., & Kelly, M. (2024, January 23). *The Biden Deepfake Robocall Is Only the Beginning.* WIRED. Retrieved April 12, 2024, from https://www.wired.com/story/biden-robocall-deepfake-danger/

argument for staying home for the primaries conveyed by a trusted source.

These challenges are compounded in that generative AI can be used to scale 'microtargeting' campaigns whereby political content is tailored to individuals based on their unique personality traits and vulnerabilities as inferred from past content engagement.[17] Given a single template of a desired political message, generative AI can derive numerous personalised versions. Furthermore, experiments have demonstrated the potential for AI-driven influence operations (definition: an organised attempt to influence audience beliefs or outcomes); the Large Language Model (LLM) based system, CounterCloud, was deployed to autonomously identify political articles, to generate and publish counter-narratives, and then to direct internet traffic by writing tweets and building fake journalist profiles to create a veneer of authenticity.[18]

(ii) Sowing distrust

Even if not deployed to change minds or to convince people of specific untruths, generative AI can still have the troubling effect of disintegrating citizen trust in their information environment and in their own ability to distinguish fact from fiction. Citizens also need not knowingly encounter deepfaked content for trust to dwindle. The very existence of the technology is enough to arouse suspicion. This dynamic results in a 'liars dividend' – as the public becomes more educated about the threats posed by deepfakes, people are more likely to believe false claims about content being artificially generative. In this way lairs can avoid accountability merely by raising suspicion that a real recording of them doing or saying something untoward is fake.[19] More so, when information consumers don't know what to believe or where to turn, they tend to either disengage or turn inward. Familiar communities of like minded individuals provide a sense of stability and affirmation.[20] These bubbles are fertile ground for increasingly polarised discourse and extremist beliefs that are so detrimental to well-functioning democracy.[21]

Marginal risk:

With respect to producing convincing disinformation, we posit that the capability uplift for well-organised and well-resourced actors (e.g. foreign state actors) is minor. Such actors have had access to, and the financial backing and expertise to operate, tools such as photoshop and state of the art video production equipment to produce convincing and misleading fake content. They have also benefited from well-organised methods of distribution.

On the other hand, the capability uplift conveyed by generative AI for poorly-resourced individual actors is high. Generative AI errases the financial and knowledge barriers to producing high quality and targeted fake content that can be easily distributed via social media platforms.

With respect to trust, it is worth noting that trust in politicians and information environments is already low, with disinformation and 'fake news' running rampant and often disseminated by the politicians we are meant to trust. Disintegrating trust in our political and epistemic environments should not be fully or even mostly credited to generative AI, but generative AI is adding another layer of uncertainty and mistrust to already stressed systems of democratic discourse and communication.

II. Producing online abuse material to force political disengagement of target individuals or demographics

Another mechanism of impact on democracy is through the use of generative AI to produce online abuse material to intimidate individuals or demographic groups away from political engagement.

One of the leading uses of image generation systems to date is to produce pornographic content. In 2022 after the open-source launch of the image generator Stable Diffusions, US Congresswoman Anna Eshoo issued a letter calling out the model's use for generating pornographic images of real people, and of violently beaten asian women in particular.[22] The OECD[23] and the Carnegie

17 Simchon et al. (2024). The persuasive effect of political microtargeting in the age of generative artificial intelligence. *PNAS Nexus*, 3(1). https://doi.org/10.1093/pnasnexus/pgae035
18 Knight, W. (2023, August 29). *It Costs Just $400 to Build an AI Disinformation Machine*. WIRED. Retrieved April 12, 2024, from https://www.wired.com/story/400-dollars-to-build-an-ai-disinformation-machine/
19 Goldstein, J., Lohn, A. (2024). *Deepfakes, Elections and Shrinking the Liar's Dividend.* Brennan Center for Justice. https://www.brennancenter.org/our-work/research-reports/deepfakes-elections-and-shrinking-liars-dividend
20 O'Connor, C., & Weatherall, J. O. (2019). *The Misinformation Age: How False Beliefs Spread.* Yale University Press.
21 Goldsworthy, A., Osborne, L., & Chesterfield, A. (2021). *Poles Apart: Why People Turn Against Each Other, and How to Bring Them Together.* United Kingdom: Random House.
22 *Eshoo Urges NSA & OSTP to Address Unsafe AI Practices.* (2022, September 22). Congresswoman Anna Eshoo. Retrieved April 12, 2024, from https://eshoo.house.gov/media/press-releases/eshoo-urges-nsa-ostp-address-unsafe-ai-practices
23 Caira, C., Russo, L., & Aranda, L. (2023, March 8). *Artificially Inequitable? AI and closing the gender gap.* OECD AI Policy Observatory. Retrieved April 12, 2024, from https://oecd.ai/en/wonk/closing-the-gender-gap

Endowment for International Peace[24] have also noted the disproportionate use of generative AI to engage in gendered disinformation[25] campaigns primarily against female politicians from minority groups. The result is a silencing effect as women disengage and avoid visible, political roles to avoid being targeted.[26]

Open and free political engagement is a core pillar of a well-functioning democracy, but women and minorities can not participate fully in democracy - as politicians, as journalists, as activists - if they are under threat of extremely damaging and dangerous online abuse.

Marginal risk:

Efforts to intimidate individuals and groups out of political engagement is not new, however generative AI significantly lowers the bar to playing the game. As in the previous section, the capability uplift for previously well-resourced and well-organised actors is likely minor. The uplift provided by generative AI to individual actors is significant. Accordingly, there has been a massive uptick in online abuse through hyper realistic fake imagery - primarily deepfake porn targeted at women, including elected politicians.[27]

III. Scaling cyber attacks on election infrastructure and political campaigns

Cybercriminals can also use generative AI tools to reduce costs and increase scale of cyberattacks on election infrastructure.[28,29]

(i) Spear phishing:

Malicious actors can use generative AI to pose as trusted individuals for the purpose of theft, extracting sensitive information.[30] Large language models like ChatPGT and Bard are proficient at generating convincing spear phishing emails that lure recipients to click on links or download files that contain malware.[31] AI voice cloning tools have also been successfully used to pose as trusted individuals over phone calls. In one case $35million was stolen from a Japanese firm by scammers posing as the company's CEO approving a bank transfer.[32]

In the context of elections, election officials are required to open email attachments as part of election administration. These attachments could contain malware to breach system security and gain access to election records and voter registration information. Voice cloning tools could also be used to impersonate election officials to gain access to sensitive administration or security information.The information gained could then, in turn, be used to inform more sophisticated influence operations.

These same tools could also be used to mount spear phishing attacks against politicians and campaign workers to gain access to secrets or to disseminate damaging content from internal accounts.

(ii) Malware production:

Large language model coding abilities can also be used to write the malware that is delivered to election officials. Ransomware is of particular concern. Ransomeware encrypts systems or data locking out the system owners. The ransomware actors demand payment to decrypt the system, though there is no guarantee access will be regained or that data will not be permanently lost or leaked. In the context of an election, a ransomware attack could cause significant disruption to electoral processes and risk leaking the personal identification information of millions of registered voters.

There is also concern that generative AI will not only increase the pace of customised malware production, but the variability of malware design. Greater variability in malware design would place more pressure on detection systems built to identify

24 di Meco, L., & Brechnenmacher, S. (2020, November 30). *Tackling Online Abuse and Disinformation Targeting Women in Politics.* Carnegie Endowment for International Peace. Retrieved April 12, 2024, from https://carnegieendowment.org/2020/11/30/tackling-online-abuse-and-disinformation-targeting-women-in-politics-pub-83331

25 Judson, E., Atay, A., Krasodomski-Jones, A., Lasko-Skinner, R., & Smith, J. (2020). *Engendering Hate: The Contours of State-Aligned Gendered Disinformation Online.* DEMOS. Retrieved April 3, 2024, from https://demos.co.uk/wp-content/uploads/2020/10/Engendering-Hate-Report-FINAL.pdf

26 Shames, S. L. (2014). *The Rational Non-Candidate: A Theory of Candidate Deterrence*. Doctoral dissertation, Harvard University. https://dash.harvard.edu/handle/1/12271801

27 Donegan, M. (2023, March 13). *Demand for deepfake pornography is exploding. We aren't ready | Moira Donegan*. The Guardian. Retrieved April 12, 2024, from https://www.theguardian.com/commentisfree/2023/mar/13/deepfake-pornography-explosion

28 *Risk in Focus: Generative A.I. and the 2024 Election Cycle.* (2024, January 18). Cybersecurity & Infrastructure Security Agency. Accessed April 11, 2024, from https://www.cisa.gov/sites/default/files/2024-01/Consolidated_Risk_in_Focus_Gen_AI_ElectionsV2_508c.pdf

29 *Cybersecurity Toolkit and Resources to Protect Elections.* (n.d.). CISA. Retrieved April 12, 2024, from https://www.cisa.gov/cybersecurity-toolkit-and-resources-protect-elections

30 Gupta, AI.. (2018). *The evolution of fraud: Ethical implications in the age of large-scale data breaches and widespread artificial intelligence solutions deployment.* International Telecommunication Union Journal, 1. http://handle.itu.int/11.1002/pub/812a022b-en.

31 Hazell, J. (2023). *Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns.* 10.48550/arXiv.2305.06972.

32 Brewster, T. (2021, October 14). *Fraudsters Cloned Company Director's Voice In $35 Million Heist*, Police Find. Forbes. Retrieved April 12, 2024, from https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/

less-variable human designed malware.[33,34] While it is prudent to keep this risk in mind, in the near term at least, the greatest risk remains with the increased quantity of attacks as opposed to novel kinds of attacks.

Marginal risk:

This campaign season we are likely to increase the quality of spear phishing attacks. Where grammatical errors and oddities in idiomatic expression due to poor translation used to be good indicators of spear phishing attempts, large language models trained on idiomatic speech in multiple languages are doing away with these tells.

With respect to artificially generated malware, a noticeable jump in malware quality or sophistication in the nearterm are unlikely due to limitations in training data. On the other hand, we should expect an uptick in the quantity of attacks as generative AI lowers the barrier of entry to those actors who have previously lacked the necessary expertise.[35]

IV. Further concentrating power in big tech companies

Recent advances in generative AI capabilities are further concentrating social, economic, and political power around the world's largest tech companies.

AI promises to be an extremely financially lucrative technology integrated into everyday systems, in our homes, at work, in education, finance, and healthcare. Leading AI developers are attracting massive financial investments to the tune of billions. Some speculate that the wealth accruing to AI tech giants may one day measure in full percentages of global GDP.[36] Before the generative AI boom, the profits of individual big tech companies already eclipsed the GDP of some countries,[37] and in 2022 the European Union opened a tech 'embassy' in San Francisco to deal with California's tech giants directly as independent global economic and political forces.[38]

Extreme wealth puts these companies in a position to influence policy through lobbying and campaign funding. Aside from the political influence that comes with wealth, private AI companies are also in a position to make decisions about AI design and development that will have profound impacts on citizens in their everyday lives, in how they work, socialise, consume information, seek help, and recreate. For example, with little oversight companies make decisions about whether it is safe to release a generative AI model for public consumption, and decisions about how content ranking algorithms prioritise content in online searches and news feeds. These factors raise questions about the extent to which democratic governments are capable of protecting citizen interests and to effectively preserve citizen rights when control of such a pervasive and influential technology is in the hands of few.

Regulatory efforts such as the EU AI Act and the Biden Admininstartions Executive Order on Safe and Trustworthy AI order are taking steps to regulate, to hold private companies accountable in their decision-making to democratic governments. Though some propose that, in the long run, society may benefit from the implementation democratic mechanisms more directly to influence corporate AI governance,[39] or to democratise AI via publicly owned and operated AI models and compute infrastructure.[40]

Marginal risk:

Wealthy corporations have always wielded significant political influence through their lobbying might. Increasing wealth concentration and control over a pervasive technology will solidify this positioning, though perhaps the more concerning prospect is that the generative AI technologies the tech giants control are pervasive and becoming inextricably intertwined with daily life. This puts tech giants in an incredibly powerful position politically with the tools and access to citizens to raise the potentiality, if they chose to do so, for mass information manipulation, swaying public opinion, and undermining democratic processes.

33 Fritsch, L., Jaber, A., Yazidi, A. (2022). *An Overview of Artificial Intelligence Used in Malware.* In: Zouganeli, E., Yazidi, A., Mello, G., Lind, P. (eds) Nordic Artificial Intelligence Research and Development. NAIS 2022. Communications in Computer and Information Science, vol 1650. Springer, Cham. https://doi.org/10.1007/978-3-031-17030-0_4

34 Stoecklin, M. P., Jang, J., & D. Kirat, D. (August 8, 2018) DeepLocker: *How AI Can Power a Stealthy New Breed of Malware.* Security Intelligence. Accessed April 4, 2024, from https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/

35 *Global ransomware threat expected to rise with AI, NCSC warns.* (2024, January 24). National Cyber Security Centre. Retrieved April 12, 2024, from https://www.ncsc.gov.uk/news/global-ransomware-threat-expected-to-rise-with-ai

36 O'Keefe, C., Cihon, P., Flynn, C., Garfinkel, B., Leung, J., and Dafoe, A. (2020). The Windfall Clause: Distributing the Benefits of AI. Centre for the Governance of AI Research Report. Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/windfallclause/

37 Daly, K. (2020, July 27). *Facebook, Google, Apple and Amazon's combined market cap vs. GDP.* Axios. Retrieved April 12, 2024, from https://www.axios.com/2020/07/27/big-techs-power-in-4-numbers

38 Bertuzzi, L. (2022, July 29). *New EU office in the Silicon Valley mulls Big Tech diplomacy.* Euractiv. Retrieved April 12, 2024, from https://www.euractiv.com/section/digital/news/new-eu-office-in-the-silicon-valley-mulls-big-tech-diplomacy/

39 Seger, E., Ovadya, A., Siddarth, D., Garfinkel, A., and Dafoe, A. (2023). *Democratising AI: Multiple Meanings, Goals, and Methods.* In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. P. 715– 722. DOI: 10.1145/3600211.3604693.

40 Kerry, C. F., Meltzer, J. P., Renda, A., & Wyckoff, A. W. (2024, March 4). *How public AI can strengthen democracy | Brookings.* Brookings Institution. Retrieved April 12, 2024, from https://www.brookings.edu/articles/how-public-ai-can-strengthen-democracy/

3. POTENTIAL SOLUTIONS

The previous section outlines four mechanisms by which generative AI might pose threats to democracy. This section attends to solutions.

In talking about the threats AI poses to democracy there is risk of implying that democracy has, until now, been a beautifully well-functioning institution, now under threat from a dangerous new technology. This is not the case. Democratic institutions have been under significant pressure, suffering from record low levels of trust in politicians, climbing cost of living, rampant disinformation, buckling public services. AI adds another layer. Therefore, In the short term run-up to the remaining 2024 we should aim to ensure that generative AI is not the straw that breaks the back of already stressed democratic institutions. In the long run, the challenges AI poses to democracy will need to be addressed through systemic multistakeholder engagement and simultaneous consideration of the deeper societal issue that AI exacerbates. There will be no silver bullet.

3.1 IMMEDIATE ACTIONS TO MITIGATE ELECTION HARMS FROM AI

This section presents recommendations for immediate actions. We have identified interventions that we believe are straightforward and achievable on a 2-3 month time scale to help mitigate detrimental AI impacts on impending elections.

I. Company commitments

First, as controllers of the technologies in question, AI and social media companies are best positioned to take actions that will have immediate, direct impacts to mitigate negative impacts of generative AI on the 2024 elections.

At the 2024 Munich Security Conference, major companies pledged to work collaboratively to advance goals for limiting risks of deliberately Deceptive AI Election Content. However, the specific commitments are loosely worded to accommodate limited technical feasibility.[41] For example:

- "Working toward attaching machine-readable information….to content that is generated by users with models"
- "Seeking to detect the distribution of deceptive AI election content"
- "Seeking to appropriately address deceptive AI election content we detect"

Despite limitations of technical feasibility, there are still specific, concrete steps companies can take in the near term to protect democratic processes.

(i) Review and update recommendation algorithms.

In the run up to elections, social media platforms

41 *AI Elections accord - A Tech accord to Combat Deceptive Use of AI in 2024 Elections.* (2024). Retrieved April 12, 2024, from https://www.aielectionsaccord.com/

should update their recommendation algorithms to reduce extremist and violent content in newsfeeds, even if doing so comes at the expense of optimising user engagement.

During the January 6 riots in the United States, Meta modified its content recommendation algorithms to stem the spread of extremist, divisive, and violent content on the platform in order to not unduly fan the flames or insurrection.[42] Many of these changes have since been rolled back, but the instance demonstrated that such modifications are possible and effective. More so, researchers have shown that recommendation algorithms can be constructed to promote mutual understanding and to lessen partisan animosity.[43]

(ii) Redirect election queries

Controllers of popular large language models should commit to redirect queries for election information to trusted external sources.

There is ample evidence that text generators provide false information about elections in response to queries due to the nature of how the system functions.[44] Large language models are not search engines, though they are often used as such. Rather, they are prediction systems, presenting strings of words that statistically represent what a good answer to the given query looks like based on the system's training data. That training data could include media coverage of past elections, and current election information may not be represented in the model's training data at all. Accordingly, inaccurate answers are common.

Anthropic has committed to redirecting election queries made to its large language model, Claude, to trusted external sources.[45] We ask that other providers of popular large language model providers (e.g. Alphabet Inc.'s Gemini, and OpenAI's ChatGPT and GPT4) commit to doing the same.

(iii) Commit to labelling artificially generated media in so far as is possible

To the extent that that is technically possible, AI companies should watermark content produced with their generative AI tools, and social media platforms should label artificially generated content. The purpose is to give citizens more context on the source of the information they are consuming so that they can better appraise for themselves the reliability of the material. This recommendation is in line with recommendations by Meta's independent Oversight Board.[46]

We acknowledge that watermarks can often be easily removed, are not included in some open-source models, and that current methods of automatically identifying and labelling synthetic media are imperfect. In the near term, social media platforms should therefore also consider implementing mechanisms by which users can flag content they suspect of being artificially generated for review.

II. Political party commitments

In the lead up to election, political parties should mutually commit to a code of conduct for the use of generative AI in the production and dissemination of campaign materials. These commitments should include agreements to not use generative AI to produce materially misleading content, to not amplify misleading artificially generated media produced by others, and to communicate these guidelines throughout the political party.

Political party commitments around the use of generative AI will do little to address threats from foreign adversaries employing generative AI to sow disinformation and discord, but it is an important step towards underpinning trust in a country's own political actors – trust which, in the UK, is at a 40 year low.[47]

At Demos we have worked in partnership with the respected fact-checking organisation Full Fact to devise a workable text for an agreement that the political parties could all commit to. We have roadtested this text with the party campaign teams and other civil society partners, and we are now encouraging the political parties' leaderships to consider cross party talks toward agreement on the text. *The full text can be seen in annex A.*

42 Haugen, F. (2021, October 23). *Facebook missed weeks of warning signs over Capitol attack, documents suggest.* The Guardian. Retrieved April 12, 2024, from https://www.theguardian.com/technology/2021/oct/23/facebook-whistleblower-january-6-capitol-attack

43 Thorburn, L. (2023, October 27). *How to redesign social media algorithms to bridge divides.* The Conversation. Retrieved April 12, 2024, from https://theconversation.com/how-to-redesign-social-media-algorithms-to-bridge-divides-216321

44 Mufarech, A. (2024, February 28). *AI Chatbots Not Ready for Election Prime Time, Study Shows.* Bloomberg.com. Retrieved April 12, 2024, from https://www.bloomberg.com/news/articles/2024-02-27/ai-chatbots-not-ready-for-election-prime-time-study-shows

45 *Preparing for global elections in 2024.* (2024, February 16). Anthropic. Retrieved April 12, 2024, from https://www.anthropic.com/news/preparing-for-global-elections-in-2024

46 *Oversight Board Upholds Meta's Decision in Altered Video of President Biden Case.* (2024, February). Oversight Board. Retrieved April 12, 2024, from https://oversightboard.com/news/1068824731034762-oversight-board-upholds-meta-s-decision-in-altered-video-of-president-biden-case/

47 Clemence, M., & King, L. (2023, December 14). *Trust in politicians reaches its lowest score in 40 years.* Ipsos. Retrieved April 12, 2024, from https://www.ipsos.com/en-uk/ipsos-trust-in-professions-veracity-index-2023

III. Public awareness campaign

Governments should support and execute public education campaigns to inform citizens of the kind of misleading deepfake material they are likely to encounter in the course of the political campaign season.

There is a risk that this kind of education campaign could backfire, sowing further mistrust in information environments and in individuals' ability to distinguish fact from fiction. It is therefore important that such a campaign not only warn of the risk, but also inform citizens where to find information that is more likely to be reliable. For example, citizens should be cautious about information gleaned from social media platforms and encouraged to consume news and press releases directly from well-known and professional journalistic sources that maintain high standards for fact checking and editorial independence.

IV. Real time AI impacts research

These 2024 elections provide an opportunity to study in real time the impacts of generative AI on democracy. Governments must provide ample support for academic and civil society organisations to investigate uses and impacts of AI in the 2024 elections. Research should closely attend to the most prevalent (mis)uses of generative AI during campaign seasons around the world, identify what interventions are proving most effective, and engage with citizenry to understand how artificially generated content is being experienced by and influencing consumers.

It may be the case (it hopefully will be the case) that 2024 gives us a golden opportunity to learn from numerous global election cycles about how emerging AI capabilities will interact with political proceedings, but while the technology is still fledgling enough for those impacts to be relatively limited. It will be a small window of opportunity; let's not squander it.

V. Start working on legislation

Legislative processes can be slow. The dynamics influencing election integrity and the resilience of our democratic institutions are complex, and while legal intervention is urgently needed, legislation needs to be carefully crafted in collaboration with citizens and technical expertise to ensure the laws are well suited and avoid unintentional adverse consequences.

For this reason, necessary debate and legislative processes can and should start immediately to help protect and improve democratic institutions.

Policy makers in collaboration with civil society, citizens and technical experts should consider legal mechanisms for reinforcing the voluntary commitment suggested above. **These measures might include:**

- Requirements of social media companies for content labelling and removing abusive content
- Requirements of AI developers and providers for researching and implementing content provenance measures
- Laws dictating acceptable use of generative AI in political campaigns

Additional legislation to consider includes:

- Robust antitrust legislation to limit accrual of power and resources to tech giants
- Criminalization of non consensual artificially generated porn and other abusive material (as in the UK)[48]
- Reviewing and updating election silence rules to ensure political parties and trusted news media can deny and debunk misleading content whenever it emerges
- Clarifying and granting relevant powers to regulatory bodies such as, in the UK, the Electoral Commission and Ofcom

This is not an exhaustive list. We encourage more thinking on the subject and expect more policy insights to emerge over the following year.

3.2 LONG TERM INTERVENTION TO STRENGTHEN DEMOCRACY

The 2024 elections will come and go, and the hype around AI and democracy will likely lessen as polls close in large Western democracies. However, we must not lose momentum in addressing the impacts of AI on democracy with urgency. Elections will continue to be held around the globe where AI impacts will remain very relevant. Furthermore, well-functioning democratic systems underpinned by trust in political leaders and well-informed civic participation is needed year round to navigate complex societal challenges.

48 Siddique, H. (2023, June 26). *Sharing deepfake intimate images to be criminalised in England and Wales.* The Guardian. Retrieved April 12, 2024, from https://www.theguardian.com/society/2023/jun/27/sharing-deepfake-intimate-images-to-be-criminalised-in-england-and-wales

In addition to the immediate actions outlined above, tech companies, government, and civil society must prepare for prolonged engagement to address the deeper societal issues facing democracy and to ensure emerging and evolving AI capabilities do not add further hindrance.

I. Collaboratively analyse and build on 2024 findings

The first step is to analyse the wealth of information we will have gathered from a year of global 2024 elections. Derive insights from 2024 to (i) forcast how AI impacts on democracy with change as technology becomes more capable and accessible, (ii) identify the most crucial intervention points, and (iii) convene multistakeholder conversations on key topics to inform next steps.

II. Legislation

We need to maintain urgency in pursuing legislative interventions to mitigate negative impacts of generative AI on democratic integrity. Consider to what extent voluntary commitments by companies and political campaigns have been successful, and where might society benefit from legislative reinforcement. Also push forward multistakeholder deliberation on other legislative measures. We recommend employing public participation methods such as citizens assemblies and town squares to accrue perspectives and insights from the diverse communities the technology will impact, and to help ground public trust in the decisions made.

III. Continue research and investment in technical solutions

As AI capabilities change and improve, AI developers companies will need to maintain investment in researching and developing technical solutions. Robust watermarking and content provenance are tricky challenges but extremely important for allowing confirmation of content origin and veracity. Furthermore, malicious actors bent on misusing AI to cause harm will continuously seek new vulnerabilities and workarounds to circumvent safety restriction and strip content identification properties. Companies might substantiate their commitments in the 2024 AI Elections Accord by committing a significant percentage of their profits back towards these efforts.

IV. Update digital media literacy education

In the short term, a flash media literacy campaign may help citizens to be cautious about how they consume political information around the elections, but in the long term a more substantial overhaul in how media literacy is taught in schools and in adult continuing education is needed. From the invention of the printing press through to the internet era we have had a slow shift in educational emphasis on information memorization to information location - information became readily available but you needed to know where to find it. Now we need a similar shift in emphasis from information location to information evaluation - information is actively pushed at you, but you need to be able to distinguish between fact and fiction, and trustworthy from untrustworthy sources.

CONCLUSION

The rise of generative AI technologies raises urgent concerns about safeguarding the integrity of imminent elections and shoring up public faith in the resilience of democratic institutions. However, AI represents just one front in a broader array of forces straining democracy's foundations in our rapidly evolving technological landscape. Ensuring democracy's long-term health will require a more comprehensive and sustained campaign - a multi-pronged, collaborative endeavour that unites key stakeholders across sectors and building partnership between government, tech industry, civil society, and citizens to tackle our most pressing societal challenges.

ANNEX A
CROSS-PARTY AGREEMENT TEXT

CONTEXT

Voters should have access to accurate information in order to make informed decisions at elections. The use of generative AI in campaigning brings new potential for political parties in communicating with voters. However, this is a complex and evolving issue which will require governments, political leaders, tech companies and civil society organisations to come together to devise systemic solutions to the emerging risks.

In the immediate run up to the UK election, being clear about the use of synthetic content, and considered about the amplification of it, will be critical in building electoral trust and transparency and in protecting election integrity. This cross-party agreement aims to bring parties together across political lines to help achieve this.

In an era of diminished trust in politics, this is an area where UK political parties can demonstrate collaborative political leadership and model best practice. This may go some way in building trust with the UK public and garnering respect on the international stage.

WE COMMIT:

1. To not use generative AI tools to produce materially misleading content; that is content that may confuse citizens into believing something is real when it is not.
2. To clearly label if generative AI is used in a non-trivial way*, for example to claim that individuals had said something they hadn't, to change the location of a real event, or depict images that didn't happen, including for the creation of satirical content, with the disclosure being located where it is likely to be noticed by the receiver.
3. To not amplify materially misleading synthetic content, including from third parties, and where appropriate and a significant risk, to be a responsible actor in calling this out in such a way that does not contribute toward further amplifying this content.
4. To ensure that party staff, members, campaigners and supporters are all given clear guidelines for the transparent use of generative AI and synthetic content in election campaigning. These guidelines will be made public.

** Trivial altering of content is content that is altered or generated in such a way that is inconsequential to the viewer's perception of it. This is exempt from disclosure under this commitment. This may include edits that do not materially change the implied context or content of an event.*

Licence to publish

Demos – Licence to Publish

The work (as defined below) is provided under the terms of this licence ('licence'). The work is protected by copyright and/or other applicable law. Any use of the work other than as authorized under this licence is prohibited. By exercising any rights to the work provided here, you accept and agree to be bound by the terms of this licence. Demos grants you the rights contained here in consideration of your acceptance of such terms and conditions.

1 Definitions

a 'Collective Work' means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this Licence.

b 'Derivative Work' means a work based upon the Work or upon the Work and other pre-existing works, such as a musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work or a translation from English into another language will not be considered a Derivative Work for the purpose of this Licence.

c 'Licensor' means the individual or entity that offers the Work under the terms of this Licence.

d 'Original Author' means the individual or entity who created the Work.

e 'Work' means the copyrightable work of authorship offered under the terms of this Licence.

f 'You' means an individual or entity exercising rights under this Licence who has not previously violated the terms of this Licence with respect to the Work, or who has received express permission from Demos to exercise rights under this Licence despite a previous violation.

2 Fair Use Rights

Nothing in this licence is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3 Licence Grant

Subject to the terms and conditions of this Licence, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) licence to exercise the rights in the Work as stated below:

a to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works;

b to distribute copies or phono-records of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works; The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. All rights not expressly granted by Licensor are hereby reserved.

4 Restrictions

The licence granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a You may distribute, publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this Licence, and You must include a copy of, or the Uniform Resource Identifier for, this Licence with every copy or phono-record of the Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Work that alter or restrict the terms of this Licence or the recipients' exercise of the rights granted hereunder. You may not sublicence the Work. You must keep intact all notices that refer to this Licence and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this Licence Agreement. The above applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this Licence. If You create a Collective Work, upon notice from any Licensor You must, to the extent practicable, remove from the Collective Work any reference to such Licensor or the Original Author, as requested.

b You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended

for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works.

c If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or any Collective Works, you must keep intact all copyright notices for the Work and give the Original Author credit reasonable to the medium or means You are utilizing by conveying the name (or pseudonym if applicable) of the Original Author if supplied; the title of the Work if supplied. Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5 Representations, Warranties and Disclaimer

a By offering the Work for public release under this Licence, Licensor represents and warrants that, to the best of Licensor's knowledge after reasonable inquiry:

i Licensor has secured all rights in the Work necessary to grant the licence rights hereunder and to permit the lawful exercise of the rights granted hereunder without You having any obligation to pay any royalties, compulsory licence fees, residuals or any other payments;

ii The Work does not infringe the copyright, trademark, publicity rights, common law rights or any other right of any third party or constitute defamation, invasion of privacy or other tortious injury to any third party.

b Except as expressly stated in this licence or otherwise agreed in writing or required by applicable law, the work is licenced on an 'as is' basis, without warranties of any kind, either express or implied including, without limitation, any warranties regarding the contents or accuracy of the work.

6 Limitation on Liability

Except to the extent required by applicable law, and except for damages arising from liability to a third party resulting from breach of the warranties in section 5, in no event will licensor be liable to you on any legal theory for any special, incidental, consequential, punitive or exemplary damages arising out of this licence or the use of the work, even if licensor has been advised of the possibility of such damages.

7 Termination

a This Licence and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this Licence. Individuals or entities who have received Collective Works from You under this Licence, however, will not have their licences terminated provided such individuals or entities remain in full compliance with those licences. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this Licence.

b Subject to the above terms and conditions, the licence granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different licence terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this Licence (or any other licence that has been, or is required to be, granted under the terms of this Licence), and this Licence will continue in full force and effect unless terminated as stated above.

8 Miscellaneous

a Each time You distribute or publicly digitally perform the Work or a Collective Work, Demos offers to the recipient a licence to the Work on the same terms and conditions as the licence granted to You under this Licence.

b If any provision of this Licence is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this Licence, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

c No term or provision of this Licence shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

d This Licence constitutes the entire agreement between the parties with respect to the Work licenced here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This Licence may not be modified without the mutual written agreement of Demos and You.

DEMOS

Demos is a champion of people, ideas and democracy. We bring people together. We bridge divides. We listen and we understand. We are practical about the problems we face, but endlessly optimistic and ambitious about our capacity, together, to overcome them.

At a crossroads in Britain's history, we need ideas for renewal, reconnection and the restoration of hope. Challenges from populism to climate change remain unsolved, and a technological revolution dawns, but the centre of politics has been intellectually paralysed. Demos will change that. We can counter the impossible promises of the political extremes, and challenge despair – by bringing to life an aspirational narrative about the future of Britain that is rooted in the hopes and ambitions of people from across our country.

Demos is an independent, educational charity, registered in England and Wales. (Charity Registration no. 1042046)

Find out more at **www.demos.co.uk**

DEMOS

PUBLISHED BY DEMOS APRIL 2024

15 WHITEHALL, LONDON, SW1A 2DD
T: 020 3878 3955
HELLO@DEMOS.CO.UK
WWW.DEMOS.CO.UK