All-Party Parliamentary Group on Artificial Intelligence

# Citizen Participation in AI!

# Navigating Disinformation and Deep Fakes: Safeguarding Democratic Processes and Responsible AI Innovation

BIG INNOVATION CENTRE
Secretariat

26 March 2024
Policy Forum

# Table of Contents



Title: Citizen Participation in AI! Navigating Disinformation and Deep Fakes: Safeguarding Democratic Processes and Responsible AI Innovation

Publication Details:
Authoring Organisation: All-Party Parliamentary Group on Artificial Intelligence (APPG AI)
Date of Publication: May 2024
Publication Type: Parliamentary Brief | Policy Brief
Publisher: Big Innovation Centre

*From left to right: Simon Horswell (Onfido), Markus Anderljung (Centre for the Governance of AI), Aled Lloyd Owen (Onfido), Professor Gina Neff (University of Cambridge), Professor Birgitte Andersen (APPG AI Secretariat, Big Innovation Centre), Lord Clement-Jones CBE (APPG AI Chair), Stephen Metcalfe MP (APPG AI Chair), Sophie Murphy Byrne (Logically), Carl Miller (Demos), Dean Russell MP (APPG AI Vice Chair), The Lord Bishop Of Oxford, Bishops (APPG AI Parliamentary Member), and Viscount Stansgate (APPG AI Parliamentary Member).*

## INTRODUCTION

This document is a creative transcript with summary of an APPG AI meeting that took place on 26 March in the House of Lords Committee Room G, UK Parliament. The transcript exclusively contains crucial discussion elements; not all points are addressed.

This session aimed to demonstrate how we generate synthetic deepfakes and detect fraud using AI. It coincided with a broader discussion on democratic processes, policy and regulatory challenges.

## DETAILS

- Evidence Session: Navigating Disinformation and Deep Fakes - Safeguarding Democratic Processes and Responsible AI Innovation
- Time 5:30 pm – 7:00 pm (GMT)
- Date: Tuesday 26th of March 2024
- Venue: Committee Room G in the House of Lords.

## CONTACT

appg@biginnovationcentre.com

## SPEAKERS

1.**Carl Mille**r, Research Director, Centre for the Analysis of Social Media (CASM) at **Demos**
2. **Aled Lloyd Owen**, Global Policy Director, **Onfido,** joined by
Live showcase by **Simon Horswell**, Fraud Specialist Manager at **Onfido**
3. **Professor Gina Neff**, Executive Director, **Minderoo Centre for Technology and Democracy, University of Cambridge**
4. **Markus Anderljung**, Head of Policy, **Centre for the Governance of AI**
5.**Sophie Murphy Byrne**, Senior Manager, Government Affairs (EU&UK), **Logically**

## CHAIRS AND SECRETARIAT

The Meeting was chaired by **Stephen Metcalfe MP and Lord Clement-Jones CBE;** Co-Chairs of the All-Party Parliamentary Group on Artificial Intelligence.

Secretariat and Rapporteur: **Professor Birgitte Andersen**, CEO Big Innovation Centre

## Aim of Session: Navigating Disinformation and Deep Fakes - Safeguarding Democratic Processes and Responsible AI Innovation

The APPG AI is on high alert regarding the threat posed by disinformation and deepfakes in anticipation of this year's UK General Election and local elections. Similar challenges underpinning our democratic processes are being encountered on the international stage.

As technological advancements continue to shape our world, the prevalence of disinformation and deep fakes poses significant challenges to democratic processes and societal well-being. Through insightful discussions and collaborative efforts, we aim to identify actionable solutions to address these pressing issues and ensure a resilient and trustworthy digital environment for all.

## Questions was raised to inspire the discussion:

Navigating Deep Fake Advancements: Differentiating Genuine Information from Disinformation

- How can the public differentiate between genuine information and disinformation, especially considering the rapid advancement of deep fake technology?

Safeguarding Elections: Ensuring Responsible Use Amidst Deepfake Threats

- Given the significant threats posed by the misuse of generative AI and deepfake technology in disseminating disinformation, how can we ensure responsible national elections?

Protecting Democracy: Balancing Innovation with Defense Against Disinformation

- What practical measures and policies can or should be implemented to strengthen democratic processes and protect societal well-being against the harmful effects of disinformation and deep fakes while still fostering innovation and technological advancement?

Combating Disinformation: Enhancing Collaboration Across Governments, Tech Firms, Civil Society Organisations and Citizens

- What roles do governments, tech firms, civil society organisations and citizens play in combating disinformation, and how can collaboration be improved?

Evidence Giver:
Carl Miller

Evidence Giver:
Sophie Murphy Byrne

Evidence Giver:
Aled Lloyd Owen

Evidence Giver:
Professor Gina Neff

Evidence Giver:
Markus Anderljung

APPG AI Chair:
Stephen Metcalfe MP

APPG AI Chair: Lord
Clement-Jones CBE

Secretariat & Rapporteur:
Professor Birgitte Andersen

Showcase:
Simon Horswell

# FINDINGS

## ACTION FIELDS FOR POLICY AND STAKEHOLDER GROUPS

## ACTION FIELDS FOR POLICY AND STAKEHOLDER GROUPS

By adopting a comprehensive approach that combines **technological innovation**, **regulatory action**, **public awareness**, and **international cooperation**, we can hope to effectively **mitigate the harmful effects** of misinformation and deep fakes, **safeguarding democratic processe**s and promoting a **trustworthy online** environment.

The evidence can be summarised in the following focus areas for action points (see below). They are described in further detail on the next page.

### FOCUS AREAS:

Understanding the Threat

Detection and Prevention

Regulation and Legislation

Industry Collaboration and Responsibility

Public Awareness and Participation

International Cooperation

## ACTION FIELDS FOR POLICY AND STAKEHOLDER GROUPS

### Understanding the Threat:

- Recognise the distinction between disinformation and influence operations, understanding that the latter poses an organised threat to democratic processes by exploiting cognitive biases.
- Acknowledge that AI, particularly generative models, enables personalised manipulation and the creation of convincing deep fakes at scale.

### Detection and Prevention:

- Invest in advanced detection technologies to identify and mitigate the spread of misinformation and deep fakes, especially in critical sectors like finance and identity verification.
- Encourage collaboration between technology companies, government agencies, and independent researchers to develop effective detection tools.
- Implement stricter platform policies and tools to label AI-generated content and remove fake content that influences elections.

### Regulation and Legislation:

- Enact legislation that addresses the creation and dissemination of deep fakes, considering the limitations of prohibiting creation and watermarking.
- Strengthen existing regulatory frameworks like the Online Safety Act to manage foreign state-backed disinformation and address domestic disinformation in elections.
- Close loopholes in election laws to ensure regulations cover all forms of election content, including organic content, to prevent exploitation by malicious actors.

### Industry Collaboration and Responsibility:

- Hold AI developers accountable for the content their models generate, encouraging the implementation of safeguards like watermarks and provenance tags.
- Foster collaboration between AI companies, social media platforms, and government agencies to develop and deploy detection tools for AI-generated content.

### Public Awareness and Participation:

- Educate the public about the threat of misinformation and deep fakes, emphasising the importance of critical thinking and fact-checking.
- Promote public participation in identifying and reporting misinformation, empowering individuals to contribute to a healthier online environment.

### International Cooperation:

- Collaborate with international partners to address the global nature of misinformation and deep fakes, share best practices, and coordinate efforts to combat these threats.
- Support initiatives like the EU Code of Practice on Disinformation and establish advisory committees to provide guidance and oversight.

# EVIDENCE

## Carl Miller, Research Director, Centre for the Analysis of Social Media (CASM) at Demos

**Disinformation vs. Influence Operations: Understanding the Digital Battlefield**

I've been very long in the Demons think tank, and I've spent about the last 13 or 14 years trying to make sense of the wilds and swirling mists of social media. And I wanted to bring back from that digital frontline one thought.

I think disinformation is a horrible way of trying to understand the problem we're facing. Because what it does is to set up the idea that you've got truth and lies wrapped in a struggle online, and that if only we can make the truth win, we'll all be okay.

The problem, at least when it comes to things happening online which genuinely undercut or threaten our democratic processes, is that it's not disinformation. It's concerted influence operations. These are campaigns. They are organised, evaluated, and often professionally conducted by a whole array of different kinds of actors: autocrats, extremist political mobilisations, and disaffected individuals. But they're united in seeing information spaces as a theatre of conflict or even a theatre of war.

**Democracy Under Threat? Unravelling the Anatomy of Influence Campaigns**

To understand how AI is likely to be used as a new tool or vector for threats like electronic interference, we need to see what these campaigns are doing.

I do not think that the main problem is people being spammed with deep fakes, somehow believing them and changing their view about the world. No, I think there are two things that we can already see these campaigns doing, which I think AI is going to be used to leverage and extenuate.

## SUMMARY

- *Carl Millar contrasts disinformation with influence operations, highlighting the latter as organised campaigns threatening democratic processes.*
- *He emphasises that influence operations exploit cognitive biases to confirm existing beliefs rather than simply spreading falsehoods.*
- *AI, particularly generative models, is noted as a tool to personalise manipulation, allowing bad actors to engage in one-on-one conversations at scale.*
- *The strategy of weaponising friendship involves using AI-generated models to mimic real individuals and foster trust with targeted audiences.*
- *Concerns are raised about the difficulty in identifying and defending against such information attacks and how trust in online information sources will diminish.*

**The Psychology of Persuasion: Leveraging Cognitive Biases in Influence Campaigns**

(1)
Number one, campaigns using deep fake, and even misinformation, is used to exploit people's cognitive biases, especially to confirm their existing preconceptions about the world.

We're far more likely to be influenced by having our own worldviews flattered and established. For the vast majority of the campaigns that we pull apart, that's how the influence vector works. It's confirming people's grievances, worldviews, ideas of what's right and wrong, true and false, and leading them in a certain direction, - it is not about lying to them in order to transform their views. Cognitive biases have long been woven into the way influence campaigns work.

(2)
But then secondly, it has something to do with friendship and direct one-on-one engagement. I think the game-changing kind of thing in AI, especially generative AI text and image models allow bad actors to target an audience in the thousands and still create one-on-one conversations. We've been able to manipulate images for a long time, but we haven't been able to bring people into these one-to-one conversations. If you ask any behavioural scientist why people get influenced or change their behaviour or view about the world, it's through the people they know, the friendships they have, the relationships they cherish.

**The Rise of Personalised Manipulation: AI's Role in Targeted Influence**

And if I put myself in the shoes of a bad actor thinking about how to intervene in an election, I would be thinking about how I can bring myself using automated models into those conversations, always ready there, willing to lend an ear, willing to ask how your day is, and over time, turning that into a vector of influence. Have you seen that news story? What do you think about this? What do you think about that? And also confirming people's conceptions of the world and exploiting that.

And to this end, what keeps me up at night is that there's obviously a big community of US researchers that have grown up trying to research and defend (ed. free) information spaces. However, I have absolutely no idea how I would spot an information attack.

**Weaponisation of Friendship: AI as the Trust Builder**

So, essentially, the end goal of information manipulation is to achieve a certain effect, either influencing attitudes or behaviours. The vector I see for this is the weaponisation of friendship rather than just amplification. This involves using a model, likely semi-automated AI-generated, that mimics a person, tailored to resemble the target audience. Over time, through chat interactions and human interventions, this model aims to foster a sense of closeness or familiarity with the target.

We understand that influence typically flows through trusted networks, which is how people are swayed. With the rise of deep fakes, trust in information may diminish. People may become less trusting of images from unfamiliar sources. As a result, our online world may become narrower, relying more on a select few individuals we trust. Those seeking influence online are likely to capitalise on this dynamic.

## Aled Lloyd Owen, Global Policy Director, Onfido

**Introduction**

I'm Aled Lloyd Owen from Onfido. We are an AI-powered ID verification provider and what we deal with is confirming people's identities, but also the detection of fraud. And we use AI to detect where identity fraud is taking place.

**The Rise of Deep Fakes in the Financial Sector and for Political Influence**

What we have seen in terms of the research that we conduct in our space is that AI and particularly ubiquitous generative AI models and readily available AI models are facilitating significant increases in the ability of hostile actors to be able to generate convincing and effective deep fakes at volume.

In the year to November 2023, we saw a 3000% increase in convincing deep fake delivering attacks on financial institutions seeking to effectively circumvent customer checks - and since last year we've seen an additional 40% increase on top of that in terms of the number of defects that we're encountering. So, it is a problem that continues to proliferate.

What we see in that financial services space is something that is replicated in terms of political influence and in terms of the ability to create types of generative, convincing deep fake images.

**Challenges in Detection**

We see that there is an increasing opportunity for hostile actors to stay a step ahead of the curve in terms of the sophistication of the deep fakes that they are able to produce.

Images, voices, etcetera, that can create convincing replication that can be delivered to a massive audience have a significant impact on influence and on undermining confidence.

## SUMMARY


*Aled Lloyd Owen from Onfido, an AI-powered ID verification provider, discusses the use of AI in confirming identities and detecting fraud.*

*The Rise of Deep Fakes:*

- *AI, particularly generative models, is enabling hostile actors to create convincing deep fakes at a significant volume.*
- *From November 2023 to the present, deep fake attacks on financial institutions increased by 3000%, with a subsequent 40% increase since.*

*Challenges in Detection:*

- *Hostile actors are staying ahead in sophistication, creating convincing deep fakes that undermine confidence.*
- *Detection in the identity verification and financial services industry is facilitated by regulation and funding availability.*

We are now able to detect deep fakes in our field (the identity verification industry, financial services and fraud detection) for two main reasons: Regulation and funding.

- Firstly, financial services are highly regulated, necessitating thorough checks.
- Secondly, sufficient funding is available for the detection of deep fakes, driven by both revenue generation from the services and the obligatory nature of these checks in the sector.

**Arms Race and Technology**

There is effectively an arms race to develop the right and effective machine learning-based tools to detect those deep fakes and create ever more convincing examples on the side of fakers.

That's really challenging because, first of all, it's a game of catching up. So those who are creating deep fakes in the first instance are at an advantage insofar as they are evolving and driving forward the available technology in order to create more effective, less detectable fraudulent images.

**Ethical Considerations and Legislative Responses**

The difficulty that is placed on those who are trying to keep up with that development is that not only then do they need access to the right levels of data and information in order to train their full detection technology within that AI space, but they also need to do that in a compliant way.

We're looking at deep fakes of images. We're dealing with biometrics, and therefore, we're dealing with the extra and additional burdens of data protection around biometrics. That's all very important, and it's right that that is adhered to. But it places additional burdens on those who are trying to detect deep fakes in order to stay one step ahead of the curve.

*Arms Race and Technology:*

- *There's an ongoing arms race between detection technology developers and those creating deep fakes.*

*Ethical Considerations and Legislative Responses:*

- *Compliance with data protection regulations poses challenges for detecting deep fakes.*
- *Legislative responses include prohibitions on deep fake creation and watermarking, but both have limitations.*

*Legislative responses aimed at combating deep fakes include prohibiting their creation and implementing watermarking. However, both approaches have* ***significant limitations****:*

*(i) Prohibition on Deep Fake Creation:*

- *Difficulty in enforcement due to challenges in distinguishing between harmful and benign uses.*
- *Hindrance to research and development of detection methods as creating deep fakes is often necessary for training detection systems.*
- *Limited effectiveness as malicious actors may disregard laws and regulations.*

When we look at some of the solutions proposed in this space, and particularly at some of the US responses to President Biden's deep fakes a couple of months ago, as well as some of those deep fakes around Taylor Swift, we see that they led to a legislative and conversational reaction.

**The responses have focused on effectively two key means of dealing with deep-faked images.**

***(i) Prohibition on the creation of deep fakes, but it is a "catch-22"!***

The first was a prohibition on the creation of deep fakes. Now, that's a difficult issue for a number of reasons, not least because if you want to detect and stay ahead of the curve of deep fakes, you need to be able to produce them.

You need either access to data that contains deep fakes to train your detection systems or the ability to develop your own synthetic data to reduce bias and ensure that you have the correct data volumes and the correct testing and development of potential.

Increasingly sophisticated deep fakes require conducting training on machines and staying as close as possible to being one step ahead of the curve. So, there are difficulties in considering prohibitions on deep fake production as a whole.

But there are obviously some limited use cases where potentially sensible steps could be taken to prohibit and limit the creation of those images. These have their limitations, of course, because ultimately, those who will adhere to the law are not those who will be doing these things in a harmful way. It is a catch-22!

*(ii) Watermarking:*

- *While useful for identifying genuine content, it is less effective for detecting deep fakes distributed through unofficial channels.*
- *Limited impact on deterring malicious actors who create deep fakes to deceive or influence.*
- *It relies on cooperation from individuals and organisations to implement watermarking effectively, which may not always be forthcoming.*

*Funding:*

- *Funding is crucial for developing effective deep fake detection, especially in well-established industries.*

*The discussion transitions to showcasing Onfido's fraud lab and deep fake detection technology by Simon.*

*Editor explanations: The catch-22 here is that you can't detect deep fakes until you have enough data to train your detection systems, but obtaining that data is difficult without being confronted by the regulatory challenges and without having established research for the development for deep fake systems in the first place.*

*To combat deep fakes effectively, you often need to mimic or simulate their creation process to develop detection methods. However, those engaging in mimicking must adhere to legal regulations aiming to combat deep fakes, while malicious actors creating deep fakes often disregard laws and regulations. This creates a significant challenge in maintaining parity with those creating harmful content.*

***(ii) Watermarking, but it has limitations***

The other option is of course, watermarking, and that was part of the US presidential statement on AI that came out in November of last year (2023) as part of the US response. Official videos of politicians, public figures, and government departments in the United States would have their images watermarked.

Now that's useful in a disinformation space to some extent because you can see what is genuine. But the fact is that as a politician, the video that is likely to catch you and likely to inflame tensions is not going to be the official video. Therefore, there is a there is a real limitation in terms of watermarking in terms of its benefits and effects.

Again, if you're being sensible and you're a good actor and you're conforming to the sort of social norms and the good use of AI and good use of technology, etc. all it fine. But it doesn't help us when it comes to those bad-actors using deep fake technologies to influence or deceive.

**Funding**

Detecting deep fakes is becoming increasingly challenging, especially in industries where there is a well-established infrastructure, regulatory framework, and financial resources. When considering this challenge within democratic contexts, funding becomes crucial for developing effective detection.

**Handover**

Now, I'll pass it over to Simon, who will showcase our fraud lab and our deep fake detection technology.

## Simon Horswell

### SHOWCASE

At the evidence session Simon Horswell from Onfido showcased (via PC and screens) Onfido's fraud lab, which serves to address the challenge of training machine learning models with sufficient data. He demonstrated how the lab bridges the gap between genuine and synthetic data, enabling the training of models at the necessary volumes without waiting for large amounts of real fraud cases to accrue.

During the presentation, Simon showcased real-time demonstrations illustrating how easily photos and ID documents (as passports) can be manipulated to create convincing simulations. He emphasised the significance of focusing on central facial features for recognition technology, as these features remain consistent over time, unlike hair or weight, which can change.

Simon also discussed the importance of replicating the signals that fraudsters use, noting that their technology utilises readily available software tools rather than proprietary ones. By using similar tools to the ones of the fraudsters, Onfido can more easily ensure effective detection even as fraudsters become more sophisticated.

Furthermore, he mentioned the seamless integration of their technology into various applications, particularly in video calls, where such video technology can fake-authenticate individuals despite mismatching their face with a provided document (e.g., passports and ID Cards) in real-time.

Overall, Simon's presentation highlighted the innovative approach of Onfido's fraud lab in combating fraud by leveraging both genuine and synthetic data and replicating the techniques used by fraudsters.

Mismatching face with a provided document (e.g., passports and ID Cards) in real-time (Example of Lord Clement Jones CBE)

Simon Horswell, Fraud Specialist Manager at Onfido

## SHOWCASE SUMMARY

- *Simon Horswell from Onfido presented how their fraud lab trains machine learning models with sufficient data.*
- *The lab bridges the gap between genuine and synthetic data, enabling training at necessary volumes without waiting for large amounts of real fraud cases.*
- *Demonstrations illustrated how easily photos and documents (such as passports and ID cards) can be manipulated to create convincing (but fake) simulations in real time.*
- *Simon emphasised the importance of focusing on central facial features for recognition technology, as they remain consistent over time (unlike hair or weight, which can change).*
- *He stressed the importance of replicating fraudsters' signals using readily available software tools, as fraudsters use readily available tools instead of proprietary ones.*

*Onfido aims to ensure effective detection even as fraudsters become more sophisticated.*

*They actively work on seamlessly integrating technology into various applications, especially in video calls for real-time authentication using facial matching with provided documents, where they determine deep fakes that appear convincingly real.*

*Overall, the presentation showcased Onfido's innovative approach in combating fraud by leveraging genuine and synthetic data and replicating fraudsters' techniques.*

Mismatching face with a provided document (e.g., passports and ID Cards) in real-time (Example of Keir Starmer MP)

## Professor Gina Neff, Executive Director, Minderoo Centre for Technology and Democracy, University of Cambridge

I'm Gina Neff. I run the Minderoo Centre for Technology and Democracy at the University of Cambridge.

**Responsible AI and Safeguarding the Democratic Process**
**Introduction**

The Minderoo Centre for Technology and Democracy is an academic research centre at the University of Cambridge, with world-leading expertise in the regulation and governance of emerging technologies. We are also part of the leadership of research and implementation projects in this area including the EU Horizon 2020-funded project AI4TRUST: AI-based-technologies for trustworthy solutions against disinformation; the UKRI and Tech Missions Fund backed, £33M project Responsible AI UK (RAI UK) and the Economic and Social Research Council Digital Good Network.

So, I'm a social scientist, and I'm here in that capacity.

**Elections are social, cultural, and political**

Elections are social, cultural, and political in nature. Safeguarding democracy is a socio-technical issue. Disinformation campaigns are designed to derail democratic transparency and accountability. Such campaigns must be countered by protecting civil liberties and human rights, ensuring access to free and independent media, and strengthening the rule of law, accountability and transparency, and increased public awareness and participation. Technological solutions will never suffice in safeguarding the democratic process.

We are concerned that falsely attributing or the manufacturing of inaccurate political content and speech will have a significant impact on the safeguarding of the democratic process. As political parties and campaigners adopt and use generative AI, there are risks that regardless of intention, audio, visual, and textual content will mislead the public. (Reference 1) Applications built on Large Language Models (LLMs), such as ChatGPT, often provide inaccurate information or make up evidence for specific claims ('hallucinations'). Evidence from the US shows that public-facing ChatGPT is not fit for purpose for election information. (Reference 2)

## SUMMARY


*Nature of Elections and Safeguarding Democracy:*

- *Elections are complex socio-political events.*
- *Protecting democracy involves a combination of social and technological measures.*
- *Disinformation campaigns aim to undermine democratic processes and must be countered through various means.*

*Concerns Regarding Disinformation and AI:*

- *Misinformation, especially through generative AI, poses a significant threat to democratic processes.*
- *Large Language Models (LLMs) like ChatGPT can propagate inaccurate information and fabricate evidence.*
- *Low-tech disinformation methods, including 'cheap fakes,' also pose risks to public discourse integrity.*
- *Fact-checking alone is insufficient to combat the spread of disinformation, particularly when AI-generated content is involved*

We have seen examples of repeated prompts leading to false information concerning content from journal articles to speeches by politicians. (Reference 3) Furthermore, low-tech approaches to and the use of 'cheap fakes' in disinformation continue to pose a significant threat to the integrity of the public sphere. (Reference 4)

**Fact-checking is not an Adequate Solution**

The use of LLMs to spread disinformation poses considerable challenges for fact-checkers, journalists, civil society, and other stakeholders. These risks are augmented by the lack of available tools and methodologies to counter the spread of disinformation through AI for fact-checkers and journalists. Available tools to detect AI-generated photos and videos generated are often not sufficiently reliable or accurate. Deepfake audio content presents even larger challenges and is often shared on messaging platforms like Telegram or WhatsApp which are less researched and unmoderated. (Reference 5) Much of virality is about emotions and humour (as Carl Miller mentioned earlier), not facts and evidence, which is incredibly difficult to 'check' by an automated system. Our work with Royal Society on misinformation and science communication shows how misinformation is challenging many types of work, including science. (Reference 6) The solution to safeguarding democracy cannot be fact-checking alone.

**Protecting People's Right to Participate in the Public Sphere is Critical**

Protecting people's rights to participate in the public sphere is critical for maintaining a shared social reality. Disinformation campaigns often target marginalised groups in society by leveraging gendered or racialised stereotypes in campaigns, (Reference 7) such as was seen in efforts in the US. (Reference 8)

We are concerned about the damage to the public sphere through coordinated, sustained and unequal attacks on some members of society, such as women and members of marginal and vulnerable communities online, who are disproportionately the victims of online abuse. Deepfakes, which are overwhelmingly targeted at and harass women and minorized people, threaten to shut these voices out of the public sphere and conversation. This poses a significant threat to democracy.

*Challenges for Fact-Checkers and Journalists:*

- *Detection tools for AI-generated content are often unreliable or insufficient.*
- *Emotional and humorous content, prevalent in disinformation, is difficult to fact-check automatically.*
- *Misinformation affects various fields, including science communication.*

*Importance of Protecting Public Participation:*

- *Upholding people's rights to engage in public discourse is crucial for maintaining a cohesive society.*
- *Disinformation campaigns often target marginalized groups, exacerbating societal divisions.*
- *Deepfakes disproportionately target women and marginalized communities, threatening democratic inclusivity.*

*Proposed Solutions:*

1. *Voluntary Code of Conduct: Establish guidelines for using generative AI in political campaigns and support independent journalism.*
2. *Better Monitoring: Enable independent researchers to monitor online platforms' health to identify threats and inform solutions.*
3. *Improved Tools and Platform Policies: Develop better tools and policies to prevent abuse, including stronger guardrails against disinformation campaigns and enhanced safety-by-design approaches.*

**Solutions**

As first steps to remedy these problems, we propose the following three solutions:

1. A voluntary **Code of Conduct:** for the use of generative AI tools in political campaigns and an ongoing commitment to strengthening independent journalism.

2. **Better monitoring**: the EU AI Act will allow independent researchers to monitor the health of online platforms, the Online Safety Act may eventually do the same. (Reference 9) The AI4Trust project has learned that building tools outside the platforms is challenging because of data access and scale. Mechanisms for mandatory researcher access to data is therefore crucial to inform society of threats and solutions that work to safeguard democracy.

3. **Better tools**: online platforms must design better tools for users to prevent and stop chronic abuse. There should be stronger guardrails to ensure generative AI are not used by malign individuals or organisations for the design and spread of disinformation campaigns. Protecting members of marginalised and minority groups from disinformation campaigns is vital to protect against domestic and foreign attacks on democratic processes. To this end, we argue platforms need to be encouraged to enhance safety-by-design approaches and upstream solutions, such as improved *human* content moderation, effective handling of user complaints and improved reporting mechanisms.

**References**


- 1) This has been evident in recent elections in Indonesia, Pakistan, and India where AI-generated avatars have been utilised to improve politicians' public image or to create videos of deceased politicians expressing support for election campaigns: James Purtill, "AI is Changing How Elections Are Fought, From Deepfake Endorsements to Chatbot Campaigners," ABC News Australia, 29 February 2024, https://www.abc.net.au/news/science/2024-02-21/ai-elections-deepfakes-generative-campaign-endorsement-democracy/103483710.

- 2) See Julia Angwin, Alondra Nelson, and Rina Palta, Seeking Reliable Election Information? Don't Trust AI (AI Democracy Project Report, 27 February 2024), https://www.proofnews.org/content/files/2024/02/SeekingReliableElectionInformationDontTrustAI.FullReport-Methodology.pdf

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

- 4) So-called cheap fakes are created through cropping images, sharing images out of context, or altering the speed of videos. These require less technical know-how, have been shared significantly more online, and are documented to have caused more harm over the past years. For example, in 2019, 96 percent of deep fake videos online were pornographic content, see Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen, The State of Deepfakes: Landscape, Threats, and Impact (Deeptrace Report, September 2019), https://regmedia.co.uk/2019/10/08/deepfake_report.pdf. Overall, see Britt Paris and Joan Donovan, Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence (Data & Society Report, September 2019), p. 11, https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1.pdf.

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

- 6) https://royalsociety.org/news-resources/projects/online-information-environment/

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

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

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

## Markus Anderljung, Head of Policy, Centre for the Governance of AI

I'm Markus Anderljung, Head of Policy at the Centre for the Governance of AI. We're a non-profit research organization focused on advanced AI systems such as GPT-5, -6, and -7. We aim to understand the impact of increasingly capable AI systems and how policymakers should respond.

I'll focus on the question: To what extent will the use of AI undermine elections in 2024, and what can be done about it?

**Impact of AI on Content Quality:**

AI systems' ability to produce high-quality content has increased tremendously in recent years. Text, images, and audio are now nearly indistinguishable from the real thing to those who aren't experts or paying close attention. Video is not quite there yet, but significant progress is being made.

However, it's not clear how much these technologies will impact elections.

**Factors to Consider:**

Firstly, we should expect them to primarily affect close elections. So if current polling in the UK holds, the UK is more likely to be used as a testing ground for attempts at election interference rather than a concerted effort.

Secondly, I don't expect individual pieces of content to have a huge impact. Though there's been reporting that the Slovak election was swayed by two deepfakes of the President, it's not clear whether this did indeed change the election. While Smer and Hlas, the parties now in government, saw a bump compared to pre-election polls, so did the Progressives.

## SUMMARY


- *Markus Anderljung highlights the increasing capability of AI systems like GPT-5, -6, -7 to produce convincing content.*
- *He emphasises the need to understand AI's potential election influence and how policymakers should respond.*

*Impact of AI on Elections:*

- *AI's ability to create high-quality content raises concerns about its influence on elections.*
- *While AI-generated content can be indistinguishable from real content, its impact on elections remains uncertain.*
- *Deepfakes' potential influence is mediated by trust, making it difficult to change behavior solely through fake content.*

*Potential Effects of Deepfakes:*

- *Deepfakes could exploit trust dynamics, such as combining them with existing narratives about candidates or spreading fake official messages.*
- *Better deepfakes, combined with AI-generated trusted relationships, could lead to belief change.*
- *Deepfakes might be effective in populations unaware of the problem, leading to a "Liar's Dividend" effect.*

**Challenges of Trust:**

The ability to change someone's behavior with AI-generated content is partly difficult because it is hugely mediated by trust. Assuming people know that it's possible to create fake content, it will only be believed if it is seen as trustworthy: if it comes from a trusted source or accords with a voter's existing worldview. For now, deepfakes are still distinguishable from the real thing by experts, so they get debunked before mainstream media reports on them as if they are real. See the deepfake audio of Keir Starmer verbally abusing staff that came out in October last year: no major outlet treated it as real, resulting in no damage.

**Potential Risks and Exploitation:**

However, the effects from deepfakes could exploit this centrality of trust in different ways:

- Deepfakes combined with longstanding attempts to build a particular narrative about a specific candidate.
- Fake official-seeming messages that are more targeted, such as robocalls on election day saying the polling station location has changed.
- Better deepfakes: Content that is wholly indistinguishable from the real thing, including by authoritative sources.
- AI systems being used to create trusted relationships: influencers and chatbots could achieve belief change.
- Deepfakes can be used in populations where there hasn't yet been an "immune response" to them. Where it is not widely known just how good deepfakes are.
- Lastly, as awareness of deepfakes increases, it might be possible to dismiss real content as fake: something often called the Liar's Dividend.

*Proposed Solutions:*

*Responsibility of AI Developers:*

- *Developers should assess their models' ability to create persuasive content and introduce safeguards like watermarks and provenance tags.*
- *Government intervention may be necessary to ensure developers take appropriate actions.*

*Adapting to Misinformation:*

- *Social media platforms should label AI-generated content and remove fake content influencing elections.*
- *Collaboration with AI companies to develop detectors for AI-generated content and adjust recommender systems.*
- *Governments should introduce penalties for election interference attempts and support initiatives like a Bellingcat for election interference tracking.*

***Analysis of AI's Impact on UK Elections and Global Implications***

*Markus Anderljung offers a nuanced perspective on the impact of AI on UK elections in 2024. His analysis acknowledges the UK's relatively robust institutional framework and public trust in electoral processes. However, he argues that even within this context, there are potential vulnerabilities. While established trust structures may help mitigate some risks associated with AI-generated misinformation, Markus Anderljung highlights the possibility of localised impacts, especially in smaller electoral races (e.g. local elections) or regions with lower levels of institutional trust. In these settings, where scrutiny and resources may be limited compared to national elections, the influence of AI-generated content could be more pronounced.*

*Globally, Markus Anderljung suggests that elections with narrower margins may be more susceptible to AI's influence. His global perspective extends beyond the UK to consider elections in regions where political landscapes are more diverse and institutional capacities vary widely.*

*He posits that elections with narrower margins—where the difference between winning and losing candidates is small—may be particularly vulnerable to the influence of AI-generated content. In such contexts, even minor shifts in voter sentiment or turnout could have significant electoral consequences. Markus Anderljung also emphasises the importance of considering awareness levels regarding AI capabilities. In regions where understanding of AI technology is limited among both policymakers and the general public, the potential impact of AI-generated misinformation may be more profound.*

*Markus Anderljung mentions the potential interest of Russian election interference operators in exploiting AI technologies to influence elections, particularly in the UK. He suggests that the UK might serve as a testing ground for such interference efforts due to its political significance and relatively stable institutional framework. Markus Anderljung implies that while the UK may not be as susceptible to election interference as other jurisdictions, such as those with narrower margins or lower institutional trust, it could still be a target for experimentation by foreign entities like Russia. This highlights the broader geopolitical implications of AI-related electoral vulnerabilities and the need for vigilance in defending against potential interference from foreign actors.*

Summary of Analysis

- AI Impact in UK Elections: Anderljung notes UK's robustness but highlights vulnerabilities, especially in smaller races.
- Global AI Influence: He suggests narrower-margin elections globally are more susceptible to AI's impact, stressing the importance of awareness.
- Foreign Interference: Anderljung warns of Russian interest in exploiting AI in UK elections, emphasizing vigilance against interference.

Proposed Solutions: So, what can be done?

**Ensure AI developers take responsibility:**

- Evaluate their models' ability to produce persuasive text and authentic-seeming audio, video, and images.
- Map those capabilities to specific safeguards, including making the production of election-relevant content difficult and introducing watermarks and content provenance tags.

**Ensure society adapts to AI-generated misinformation:**

- Social media platforms have an incredibly important role to play:
    i. They should commit to tagging AI-generated content as such and remove fake content aimed at influencing elections.
    ii. To tag AI-generated content, social media platforms need to build high-quality detectors of AI-generated content, likely in collaboration with AI companies.
    iii. Furthermore, they can adjust recommender systems, reducing the saliency of divisive political content around election time.
    iv. They must also enable third-party research on their platforms so we can study the prevalence and effectiveness of content aimed at influencing elections.
    v. Accompanying content with provenance data is essential: Authentic content, such as pictures taken by phones or official government communication, should have content provenance data added to it.

- Furthermore, the public may need to be educated.
- Media institutions should also work hard to maintain and deserve the trust of the public.
- Governments should consider introducing new criminal penalties for attempts at election interference.
- There should be a Bellingcat for election interference, tracking down and debunking AI-generated content being used to influence elections covertly.

*Markus Anderljung provides an acknowledgement to Dr Valerie Belu for her contribution to the evidence preparations.*

## Sophie Murphy Byrne, Senior Manager, Government Affairs (EU&UK), Logically

I'm Sophie Murphy Byrne. I primarily manage government relations in the EU for Logically, but I also provide support in the UK. For those who aren't familiar with us, Logically operates in the counter-disinformation sector, focusing on three key areas. Firstly, we're a fact-checking organization through our independent subsidiary, Logically Fact. Secondly, we operate as an OSINT (Open-source intelligence) organization, employing open-source intelligence investigators to analyze disinformation trends and narratives online. Thirdly, we're an AI company developing and deploying our own models to support and accelerate these processes.

**Unlocking Persuasion: Generative AI's Versatile Influence Techniques**

(i) Breaking Barriers: Generative AI Lowers Entry Barriers to Disinformation

Disinformation is not new. The emergence of generative AI has only compounded the challenges it poses. While social media can facilitate the visibility and rapid spread of disinformation, generative AI dramatically lowers the barrier for creating and disseminating realistic fake content, and magnifies existing societal risks by transforming the scale, scope and likely effectiveness of online influence operations.

With generative AI, the costs of running disinformation operations are markedly reduced. Mass campaigns, such as that conducted by Russia during the 2016 US election, have historically been well-funded and organised undertakings. The Internet Research Agency, which ran this operation, had an operational budget of $12.2 million in 2017 and a staff of about 400 people.

However, if that operation were to be replicated today, this would cost under $1000. With the creation of AI systems like CounterCloud, the total cost of delivering these campaigns could be lowered to $400. CounterCloud's AI identifies specific articles and uses a Large Language Model to create targeted content, including fake comments, images, and sound clips. It can also engage in social media activities like self-promotion, trolling, and shaping narratives (Reference 1)

## SUMMARY


*Introduction:*

- *Sophie Murphy Byrne manages government relations in the EU and provides support in the UK for Logically.*
- *Logically operates in three key areas: fact-checking through Logically Fact, open-source intelligence (OSINT) analysis, and AI development.*

*Generative AI and Disinformation:*

- *Breaking Barriers: Generative AI significantly reduces the cost of running disinformation campaigns, making it more accessible.*
- *Flooding the Zone: Generative AI tools like Midjourney contribute to overwhelming the public with disinformation by easily generating content to manipulate narratives.*
- *Microtargeting Revolution: Generative AI enables personalised disinformation tailored to individuals based on their digital footprint, potentially leading to fully automated trolling.*

(ii) Flooding the Zone: How Generative AI Amplifies Disinformation Overload

Generative AI can facilitate the tactic of “flooding the zone”, whereby with vast amounts of disinformation are pumped out to overwhelm and confuse the public. In 2023, a Logically study found that, in response to 85% of prompts, generative AI tools such as Midjourney quickly and easily generated images related to common narratives used to manipulate elections in the US, India and the UK (Reference 2). Eight months later in March 2024, Logically reran the study on Midjourney and found that the platform's prompt safety guardrails had deteriorated — the acceptance rate of prompts has increased from 80% to 95% — despite significant discussion around potential misuse.

(iii) Microtargeting Revolution: Generative AI's Role in Personalized Disinformation

Generative AI is also revolutionising microtargeting. It enables the creation of content that resonates with individuals based on their digital footprint and supports intelligent chatbots that can manipulate opinions and behaviours through seemingly genuine interaction. Researchers have already demonstrated that a user’s Facebook profile offers a strong indication of a variety of different personal characteristics: a person's gender can be predicted with 93% accuracy, their politics with 85% accuracy, their ethnicity with 95% accuracy and their sexual orientation with 88% accuracy (Reference 3). Generative AI facilitates the use of this kind of psychographic data to essentially create disinformation that is personalised to the targeted audience.

In the longer term, rapid advancements in conversational AI and its use on social media platforms could plausibly lead to fully automated trolling, removing the human operator from the loop. An automated trolling system could leverage platforms’ chatbots to engage in conversations via comments and use natural language-generation systems to produce compelling arguments, while employing machine-learning-powered sentiment analysis to optimise its messaging.

*Evaluation of Current Approaches:*

- *The focus on technological solutions like watermarking AI-generated content faces challenges and does not address the issue of people trusting fake images and videos.*
- *Comprehensive measures beyond watermarking are needed, involving governments, regulatory bodies, and social media platforms to detect and curb AI-generated disinformation.*

*Harnessing AI Innovation:*

- *AI should be used to detect the tactics behind disinformation campaigns, identify attempts by foreign states to manipulate public opinion and scale up fact-checking efforts.*
- *Pre-emptive measures like AI-powered information threat feeds can help curb the virality of disinformation.*

*Policy Recommendations:*

*The UK Government and Ofcom should take legislative and regulatory actions to address disinformation:*

- *The Online Safety Act should require social media platforms to manage foreign state-backed disinformation and address domestic disinformation.*
- *The Elections Act needs to close loopholes to cover all forms of election-related content, including unpaid 'organic' content.*
- *Ofcom should set clear guidelines for identifying behaviours associated with disinformation and establish the Advisory Committee on Misinformation and Disinformation swiftly.*

**Evaluation of the current approach**

The debate about managing the disinformation threat posed by generative AI has thus far focused on technological approaches to verify whether content is indeed AI-generated or not. This needs to move on.

The idea of 'watermarking' content is just one among a range of necessary measures, and still faces challenges in ensuring accuracy. As it currently stands, there is as no reliable way to watermark AI-generated text and deepfake audio. OpenAI's cancellation of a synthetic text detection project due to an accuracy rate of just 26% highlights this issue. These technological solutions also do not address the fact that people generally give credence to claims in fake images and videos.

The Government, Ofcom and social media platforms need to look beyond watermarking and deploy a comprehensive portfolio of capabilities which aim to detect and curb the dissemination of AI-generated disinformation-related content.

**Harnessing AI innovation to tackle AI risks**

AI can create risks, but it also offers the opportunity to promote a safer and more trusted online information environment. The best way to tackle AI-generated disinformation risks lies in leveraging AI models not just to detect ‘fake’ content, but to also identify the types of tactics, techniques and procedures (TTPs) that disinformation actors use to coordinate and disseminate such content to audiences via social media platforms.

AI should be deployed as a safety tool to detect the TTPs (tactics, techniques, and procedures) behind disinformation campaigns. AI tools can be used to identify attempts by foreign states to actively manipulate public opinion. Moreover, AI-powered tools can scale up fact-checking efforts by triaging which content fact-checkers should spend their time verifying, thereby accelerating their work. Preemptive measures like AI-powered information threat feeds can also alert platforms to emerging disinformation trends before they become widespread, enabling the deployment of timely countermeasures to curb their virality. Thus, it is essential to equip those who are responsible for risk management with the tools they need to deal with the scale and nature of the problem.

**Policy Recommendations**

In the UK, the National Cyber Security Centre has recently warned that the next election "will be the first to take place against the backdrop of significant advances in AI" (Reference 4). The Government and Ofcom have, respectively, a legislative and a regulatory role that they should not shy away from:

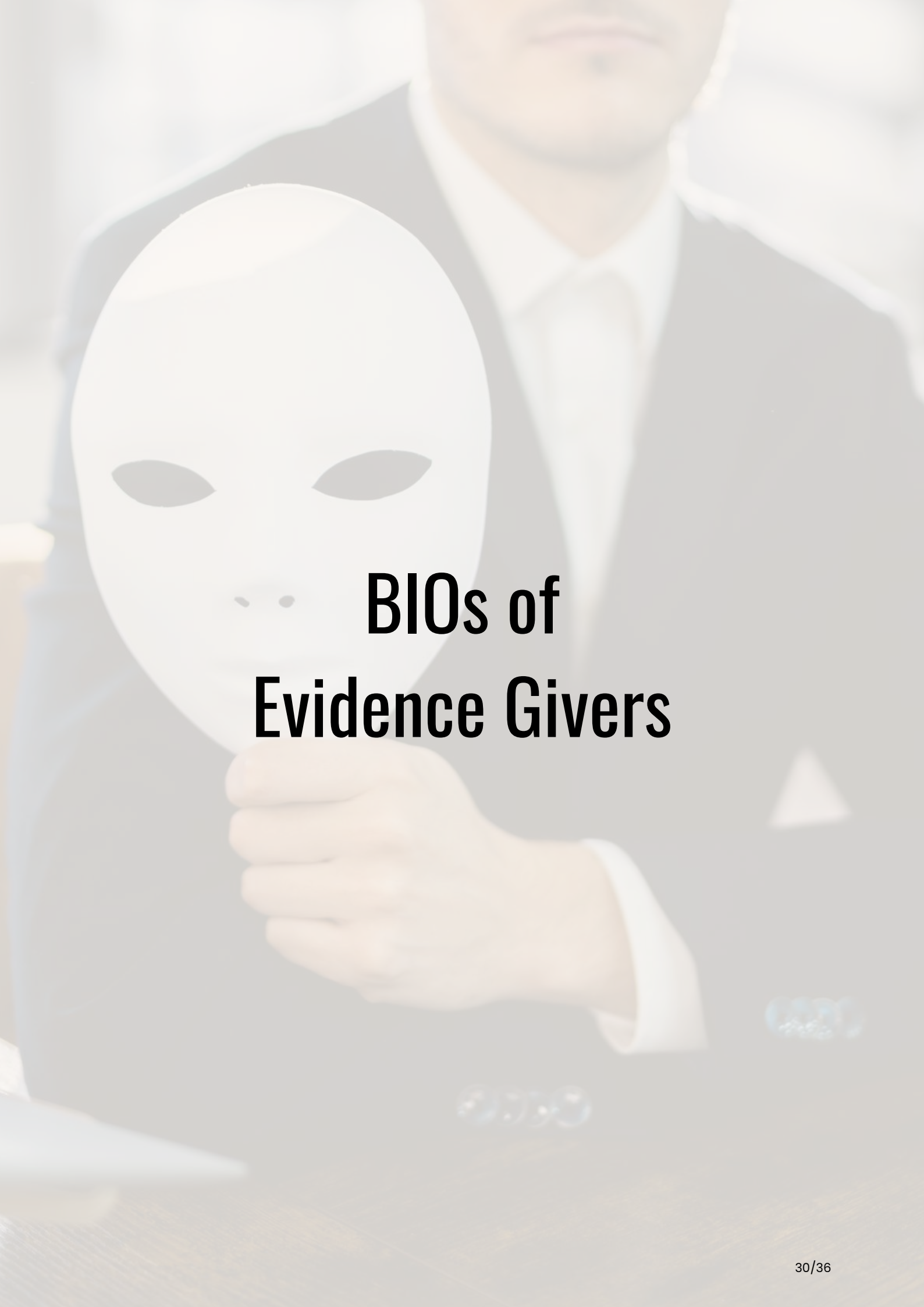
- The Online Safety Act requires social media platforms to manage foreign state-backed disinformation proactively. However, domestic disinformation falls into a regulatory grey area. The use of generative AI by an ordinary UK citizen to try to interfere in an election would not necessarily be illegal.

- The Elections Act (2022) only partially addresses this. It contains loopholes which mean that unpaid 'organic' election content may not be covered by the Act's digital imprint regime. There is a risk that malicious actors could exploit this.

- Ofcom needs to use its newly acquired powers to set out a very clear set of generic behaviours associated with foreign state-backed disinformation and social media platforms should be required to demonstrate that they can identify these behaviours to meet their new regulatory requirements. This model has already been adopted in the context of the EU Code of Practice on Disinformation. Ofcom also needs to ensure the swift establishment of the Advisory Committee on Misinformation and Disinformation.

**References**

1) MJ Banias, Inside Countercloud: A fully Autonomous AI Disinformation System (The Debrief, August 2023)
2) Logically, Generating Election Mis and Disinformation Evidence: Spot the difference (2023)
3) Kosinksi et al., Private traits and attributes are predictable from digital records of human behaviour in PNAS (March 2013)
4) UK National Cyber Security Centre, Annual Review, (November 2023



# BIOs of Evidence Givers

C**arl Miller, Research Director, CASM Centre for the Analysis of Social Media (CASM) at Demos.**

Carl Millar is the Research Director of the Centre for the Analysis of Social Media (CASM) at Demos. He is interested in how social media is changing society, and how researching it can inform important decisions. This includes:

- Digital politics and digital democracy
- Cybercrime, and the hacking community
- Cyber-bullying, hate crime, misogyny and abuse online
- Information warfare and online disinformation
- 'Fake news', digital and citizen journalism
- Automated decision-making, Internet governance and digital addiction
- Building new methods and technology to study social media data

He researches and writes widely on these issues, including for Wired, New Scientist, the Sunday Times, the Telegraph, and the Guardian. He is a Visiting Research Fellow at King's College London. His first book is The Death of the Gods: The New Global Power Grab, an examination of the new centres of power and control in the twenty-first century, published by Penguin RandomHouse in August 2018.

*Summary:*

- *Research Director at CASM, Demos, focusing on the societal impacts of social media and its role in decision-making.*
- *Research interests cover digital politics, cybercrime, cyber-bullying, hate crime, misogyny, abuse online, information warfare, online disinformation, and 'fake news'.*
- *Explores topics such as digital and citizen journalism, automated decision-making, internet governance, digital addiction, and developing new methods to study social media data.*
- *Contributes to various publications including Wired, New Scientist, Sunday Times, Telegraph, and Guardian.*
- *Holds a Visiting Research Fellowship at King's College London.*
- *Author of "The Death of the Gods: The New Global Power Grab", Penguin RandomHouse 2018.*

**Aled Lloyd Owen, Global Policy Director, Onfido**

Aled is Global Policy Director at Onfido. He provides strategic policy leadership to ensure Onfido remains at the cutting edge of developments in digital identity verification, AI, regulation and compliance.

Aled has over a decade of experience engaging with complex emerging technology, policy, security, AI and data protection challenges as a UK government official, counsellor to the European Union and US-based academic.

*Summary:*

- *Aled serves as the Global Policy Director at Onfido, offering strategic policy guidance in digital identity verification, AI, regulation, and compliance.*
- *With over a decade of experience, he has engaged with diverse challenges in emerging technology, policy, security, AI, and data protection.*
- *Aled's background includes roles as a UK government official, counsellor to the European Union, and academic in the United States.*

**Professor Gina Neff, Executive Director, Minderoo Centre for Technology and Democracy, University of Cambridge**

As the Executive Director of the Minderoo Centre for Technology and Democracy at the University of Cambridge, Gina Neff oversees a multidisciplinary team of researchers and practitioners studying the social, political, and ethical implications of digital technologies. With over 15 years of experience in academia and public policy, she am passionate about advancing conversations about society's role in responsible technology and creating positive social change through evidence-based interventions and advocacy.

Professor Neff's research focuses on how technological change affects work and workplaces and how data and AI transform our lives and futures. She has published several books and articles on these topics, drawing from her expertise in qualitative research, data analysis, and research design. She is also an experienced mentor and teacher, having held professorships and fellowships at Oxford, University of Washington, and UCLA. Professor Neff enjoys developing leaders, connecting and engaging with diverse audiences, and bringing evidence to bearon the challenges and opportunities of the digital age.

*Summary:*

- *Gina Neff leads research at the Minderoo Centre for Technology and Democracy, focusing on digital technology's societal implications.*
- *With 15+ years in academia and public policy, she advocates for responsible technology and social change through evidence-based methods.*
- *Neff's research delves into how technology affects work, workplaces, and society's transformation via data and AI.*
- *She's an accomplished author using qualitative research and data analysis in her work.*
- *Neff has taught and mentored at Oxford, University of Washington, and UCLA.*

**Markus Anderljung, Head of Policy, Centre for the Governance of AI**

Markus leads GovAI's policy team, aiming to produce rigorous recommendations for governments and AI companies. His research focuses on e.g. the frontier AI regulation, responsible cutting-edge development, national security implications of AI, and compute governance. He is an Adjunct Fellow at the Center for a New American Security, and a member of the OECD AI Policy Observatory's Expert Group on AI Futures. He was previously seconded to the UK Cabinet Office as a Senior Policy Specialist, GovAI's Deputy Director, and Senior Consultant at EY Sweden

*Summary:*

- *Markus heads the policy team at GovAI, striving to generate rigorous recommendations for governments and AI firms.*
- *His research interests encompass frontier AI regulation, responsible cutting-edge development, national security ramifications of AI, and compute governance.*
- *Markus holds the position of Adjunct Fellow at the Center for a New American Security and is a member of the OECD AI Policy Observatory's Expert Group on AI Futures.*
- *Previously, he served as a Senior Policy Specialist at the UK Cabinet Office, Deputy Director at GovAI, and Senior Consultant at EY Sweden.*

**Sophie Murphy Byrne, Senior Manager, Government Affairs (EU&UK), Logically**

Sophie is responsible for government affairs across the UK and EU at Logically, a British-based scale-up operating globally to fulfil its mission to tackle the harms associated with mis- and disinformation. Before Logically, Sophie spent four years working in EU Affairs in Brussels, including in the Permanent Representation of Ireland to the EU. Her primary focus is tech policy, specifically as regards the regulation of artificial intelligence and digital service providers.

*Summary:*

- *Sophie oversees government affairs for the UK and EU at Logically, a British scale-up addressing harm associated with mis- and disinformation.*
- *She brings four years of experience in EU Affairs, including work at the Permanent Representation of Ireland to the EU.*
- *Sophie's expertise lies in tech policy, particularly in regulating artificial intelligence and digital service providers.*

# ABOUT APPG AI

# ABOUT


**APPG AI Officers:**

Stephen Metcalfe MP, APPG AI Chair, Conservative
Lord Clement-Jones CBE, APPG AI Chair, Liberal Democrat
Dawn Butler MP, Vice Chair, Labour
Dean Russell MP, Vice Chair, Conservative
Sir Mark Hendrick MP, Honorary Officer, Labour
Justin Madders MP, Honorary Officer, Labour

**Parliamentary APPG AI Members – House of Commons**

Sir Peter Bottomley MP, Conservative
Anthony Browne MP, Conservative
Liam Byrne MP, Labour
Dr. Lisa Cameron MP
Ruth Cadbury MP, Labour
Jon Cruddas MP, Labour
Clive Efford MP, Labour
Simon Fell MP, Conservative
Patrick Grady MP, SNP
Chris Green MP, Conservative
Dame Eleanor Laing MP, Conservative
Scott Mann MP, Conservative
Anna McMorrin MP, Labour
Carol Monaghan MP, SNP
Damien Moore MP, Conservative
Layla Moran MP, Liberal Democrat
Lee Rowley MP, Conservative
Gary Sambrook MP, Conservative
Alex Sobel MP, Labour
Craig Tracey MP, Conservative
Matt Warman MP, Conservative

**Parliamentary APPG AI Members – House of Lords**

Lord Janvrin, Crossbench
Lord Knight Of Weymouth, Labour
Baroness Susan Kramer, Liberal Democrat
Baroness McGregor-Smith, Conservative
Lord Ian Strathcarron, Conservative
Lord Ravensdale, Crossbench
Lord Ranger of Northwood, Conservative
Baroness Rock, Conservative
Viscount Stansgate, Labour
Lord Taylor of Warwick, Conservative
Lord Wei, Conservative
Lord Willetts, Conservative
The Earl of Erroll, Crossbench
Lord Freyberg, Crossbench
Lord Fairfax of Cameron, Conservative
The Earl of Glasgow, Liberal Democrat
Lord Haskel, Labour
The Lord Bishop Of Oxford, Bishops
Baroness Uddin, Labour
Lord Richard Inglewood, Non-affiliated

**APPG AI Advisory Board:**

Lawrence Turner, Founder, AMI Limited
Dr Scott Steedman CBE, Director of Standards, BSI Group
Professor Ashley Braganza, Brunel University London
Zoe Webster, AI Director, BT Group
Paul Dixon, Head of Public Sector, Capgemini UK
Markus Anderljung, Head of Policy, Centre for the Governance of AI
Charles Kerrigan, Partner, Banking & Int. Finance, CMS Tax Law
Yatin Mahandru, Head of Public Sector & Health, Cognizant
Sulabh Soral, Chief AI Officer, Deloitte
Edward Fu, Head of Government Affairs, Duolingo
Sarah Reynolds, Partner, EY Law
Joel Roberts, Head of Corporate Affairs, Hewlett Packard Enterprise
Sara El-Hanfy, Head of AI & Machine Learning, Innovate UK
Aled Owen, Global Policy Director, Onfido
John Buyers, Partner, Osborne Clarke
Professor David Leslie, Queen Mary University of London
Richard Chiumento, Director, Rialto
Shaun O'callaghan, Chief Information Officer, Homes, Santander UK
David Elcombe, Managing Director, WindWorkX

**Secretariat:**

Big Innovation Centre is appointed as the Group's Secretariat.

The Secretariat is responsible for delivering the programme for the APPG AI, organising the outputs, advocacy and outreach, and managing stakeholder relationships and partnerships.

**Contact:**

Professor Birgitte Andersen, CEO, Big Innovation Centre
appg@biginnovationcentre.com

APPGs are informal cross-party groups in the UK Parliament. They are run by and for Members of the Commons and Lords. The All-Party Parliamentary Group on Artificial Intelligence (APPG AI) functions as the permanent, authoritative voice within the UK Parliament (House of Commons and House of Lords) on all AI-related matters, and it has also become a recognisable forum in the AI policy ecosystem both in the UK and internationally.

# ACCESS APPG AI RESOURCES, EVENTS AND FULL PROGRAMME

Pavilion proudly hosts the All-Party Parliamentary Group on Artificial Intelligence (APPG AI), providing a centralised hub for all its resources, including and event registrations.


Please use the same username and password across all web and mobile app devices, avoiding the hassle of multiple accounts.
Click below:

visit our website bicpavilion.com

Download on the App Store

GET IT ON Google Play


All-Party Parliamentary Group on
Artificial Intelligence
appg@biginnovationcentre.com

## SECRETARIAT

Big Innovation Centre is appointed by the UK Parliament as the Group's Secretariat.

BIG INNOVATION CENTRE