



**FRIEDRICH NAUMANN  
STIFTUNG** Für die Freiheit.

**ISD** | Institute  
for Strategic  
Dialogue

**POLICY PAPER**

# **VILIFY, RIDICULE, DISINFORM**

**Political communication and media trust  
in the age of generative AI**

Christian Schwieter, Milan Gandhi

**ANALYSIS**

# Imprint

## Publisher

Friedrich-Naumann-Stiftung für die Freiheit  
Truman-Haus  
Karl-Marx-Straße 2  
14482 Potsdam-Babelsberg

/freiheit.org

/FriedrichNaumannStiftungFreiheit

/FNFreiheit

## Authors

Christian Schwieter, Milan Gandhi

## Editor

Zoe van Doren  
Global Themes Division, Berlin

## Contact

Phone: +49 30 22 01 26 34  
Fax: +49 30 69 08 81 02  
email: [service@freiheit.org](mailto:service@freiheit.org)

## Date

February 2024

## Notes on using this publication

This publication is an information offer of the Friedrich Naumann Foundation for Freedom. It is available free of charge and not intended for sale. It may not be used by parties or election workers for the purpose of election advertising during election campaigns (federal, state or local government elections, or European Parliament elections).

## Licence

Creative Commons (CC BY-NC-ND 4.0)

## About this Paper

This paper is a timely examination of political communication and media trust in the age of generative AI systems. After analysing real-world examples of generative AI use in a political context, it evaluates emerging technical and policy solutions and presents recommendations and reflections for policy practitioners. This paper is the outcome of a collaboration between the Friedrich Naumann Foundation for Freedom and the Institute for Strategic Dialogue.

## Acknowledgements

Any errors are the authors' alone. The authors are grateful to reviewers from Friedrich Naumann Foundation for Freedom and ISD.

# Table of contents

- EXECUTIVE SUMMARY** \_\_\_\_\_ **4**
  - About the Institute for Strategic Dialogue (ISD) \_\_\_\_\_ 4
  - Glossary \_\_\_\_\_ 4
  
- WHAT IS GENERATIVE AI?** \_\_\_\_\_ **6**
  - Artificial Intelligence \_\_\_\_\_ 6
  - Generative AI Systems \_\_\_\_\_ 6
  - Systems that generate text \_\_\_\_\_ 6
  - Systems that generate images, videos and audio \_\_\_\_\_ 7
  
- GENERATIVE AI AND POLITICAL COMMUNICATION** \_\_\_\_\_ **7**
  - Non-AI content used to target political opponents \_\_\_\_\_ 7
  - AI-generated content used for political campaigning \_\_\_\_\_ 8
  - AI-generated content used for political entertainment \_\_\_\_\_ 10
  - AI-generated content used in disinformation campaigns and influence operations \_\_\_\_\_ 12
  - Manipulated audio as a notable subset of AI-generated disinformation \_\_\_\_\_ 14
  
- SIX INSIGHTS FOR POLICYMAKERS** \_\_\_\_\_ **15**
  - Risks of mislabelling manipulated content as AI-generated \_\_\_\_\_ 15
  - Acknowledging the multi-modality of threats posed by generative AI \_\_\_\_\_ 15
  - Delimiting fair-use cases of AI in political campaigning and entertainment \_\_\_\_\_ 15
  - Non-political uses of AI affecting politics \_\_\_\_\_ 16
  - Continuities in disinformation strategies deployed \_\_\_\_\_ 16
  - Discrediting media evidence by alleging AI use \_\_\_\_\_ 16
  
- EMERGING POLICY AND TECHNICAL SOLUTIONS** \_\_\_\_\_ **17**
  - Detecting deepfakes \_\_\_\_\_ 17
  - Labelling deepfakes \_\_\_\_\_ 17
  - Authenticity and provenance \_\_\_\_\_ 17
  - Emerging legal rules and EU AI Act \_\_\_\_\_ 18
  
- CONCLUDING REMARKS: IMPACTS ON INDIVIDUALS AND SOCIETY** \_\_\_\_\_ **18**
  - Distribution and quality \_\_\_\_\_ 18
  - Demand-side analysis and belief formation \_\_\_\_\_ 19
  - The role of trust \_\_\_\_\_ 19
  
- APPENDIX** \_\_\_\_\_ **20**
  
- ABOUT THE AUTHORS** \_\_\_\_\_ **23**

# Executive Summary

This report examines political communication and media trust in the age of generative artificial intelligence systems (AI). Firstly, it provides a brief explainer of generative AI tools and techniques, looking separately at systems that generate text and those that generate or manipulate images, videos and audio.

By reference to real-world examples, the paper then surveys the ways in which generative AI systems have recently been used by political actors, distinguishing between three different use-cases: political campaigning, entertainment and disinformation campaigns. Building on this empirical analysis, the paper distils important insights for policymakers, which highlight the need to:

- Refrain from falsely labelling content as AI-generated to avoid overstating the technical capabilities and persuasive power of those spreading disinformation;
- Acknowledge the multimodality of threats posed by generative AI, in particular voice-generation;
- Delimit fair-use cases of generative AI for political campaigning, given these technologies are already widely used for legitimate political communication purposes;
- Raise awareness of how seemingly non-political uses of generative AI can be exploited for politics, in particular the creation of non-consensual intimate content.

This is followed by an evaluation of emerging technical and policy solutions, namely the detection and labelling of deep-fakes as well as the development of systems to certify content authenticity and provenance. The section concludes with a discussion of the emerging legal landscape, including the European Union's AI Act.

Finally, the authors provide concluding reflections, emphasising that regulating technologies, labelling deepfakes, and reducing the supply of disinformation are only partial solutions to a complex problem – restoring citizens' trust in democratic institutions, and in particular the news media, must be the overarching mission for those concerned about the spread of AI-generated disinformation.

## About the Institute for Strategic Dialogue (ISD)

The Institute for Strategic Dialogue (ISD) is an independent, non-profit organisation dedicated to safeguarding human rights and reversing the rising tide of polarisation, extremism and disinformation worldwide. Since 2006, ISD has been at the forefront of analysing and responding to extremism in all its forms. A global team of researchers, digital analysts, policy experts, frontline practitioners, technologists and activists have kept ISD's work systematically ahead of the curve on this fast-evolving set of threats. ISD has innovated and scaled sector-leading policy and operational programmes – on- and offline – to push back the forces threatening democracy and cohesion around the world today. ISD partners with governments, cities, businesses and communities, working to deliver solutions at all levels of society, to empower those that can really impact change. ISD is headquartered in London with a global footprint that includes teams in Washington DC, Berlin, Amman, Nairobi and Paris.

## Glossary

**Artificial Intelligence (AI)** is defined in the subsection titled 'Artificial Intelligence' below.

**Convolutional Neural Networks (CNNs)** are a type of deep learning algorithm optimised for processing grid-like data, such as images. A typical CNN consists of convolutional layers, paired with pooling layers, fully connected layers, and normalisation layers. CNNs are good at learning spatial hierarchies of features due to their structure, making them ideal for image recognition and object detection. Their design allows them to process visual data efficiently, making them a cornerstone in the AI sub-field of computer vision. Read more about CNNs here: ['What are convolutional neural networks?'](#), IBM (date unknown).

**Deepfake** is defined in the subsection titled 'Systems that generate images, videos and audio' below.

**Deep Learning** is a subset of machine learning (see 'Machine Learning' below). It employs artificial neural networks (ANNs), a methodology inspired by the functioning of a human or animal brain. ANNs are computational models consisting of node layers, which each contain "an input layer, one or more hidden layers, and an output layer".<sup>1</sup> They are particularly useful for clustering and classifying information. If a neural network has three or more layers of nodes through which data must pass, it is a deep-learning neural network – the intuition is that a greater number of layers makes the network literally deeper. In general, although not always true, the more node layers, the more capable the neural network at handling very large and complicated datasets and discovering patterns within unlabelled and unstructured data. As IBM explains, "[n]eural networks rely on training data to learn and improve their accuracy over time. However, once these learning algorithms are fine-tuned for accuracy, they are powerful tools in computer science and artificial intelligence, allowing us to classify and cluster data at a high velocity. Tasks in speech recognition or image recognition can take minutes versus hours when compared to the manual identification by human experts. One of the most well-known neural networks is Google's search algorithm."<sup>2</sup> A specific kind of ANN, a Transformer Model, is utilised in LLMs (see 'Transformer Models' below).

**Disinformation** is defined as false, misleading or manipulated content presented as fact that is intended to deceive or harm.

**Foreign Information Manipulation and Interference (FIMI)** is defined by the European Union Agency for Cybersecurity (ENISA) as "a mostly non-illegal pattern of behaviour that threatens or has the potential to negatively impact values, procedures and political processes. Such activity is manipulative in character, conducted in an intentional and coordinated manner. Actors of such activity can be state or non-state actors, including their proxies inside and outside of their own territory." ENISA explains that the term FIMI aims to refine the concept of disinformation by emphasising "manipulative behaviour, as opposed to the truth of content being delivered."<sup>3</sup>

**Generative AI (GenAI)** is defined within the subsection 'Generative AI Systems' below.

**Generative Adversarial Networks** are a type of machine learning model that involve two neural networks, a generator and a discriminator, which compete against each other. Utilising deep learning techniques, these networks operate in an unsupervised manner within a zero-sum game frame-

work. The generator's role is to create data that mimics real data, while the discriminator works to differentiate between genuine and artificially generated data. Through continuous interaction, both networks improve their functions, with the generator producing increasingly realistic data and the discriminator enhancing its ability to detect artificial data. This dynamic results in high-quality, believable outputs, such as lifelike images of human faces that do not correspond to real individuals. Read more about general adversarial networks here: 'generative adversarial network (GAN)', Kinza Yasar, TechTarget (2023); and 'Generative adversarial networks explained', Capex Hansen, IBM (2022).

**Large Language Models (LLM)** are statistical models that generate "plausible next words" to a user's prompt. LLMs employ deep learning and are trained on vast datasets, enabling them to produce coherent and contextually relevant responses. As they excel at language-related tasks, they are an applied example of the natural language processing AI subfield.

**Machine Learning** is a subfield of AI concerned with systems that automatically learn and improve from experience. For example, recommender systems utilised by digital platforms such as Facebook, YouTube, Netflix or Amazon analyse users' previous activity and preferences to recommend online content, movies, products and advertising etc.

**Misinformation** is defined as false, misleading or manipulated content presented as fact, irrespective of an intent to deceive.

**Shallowfake** (sometimes referred to as '*Cheapfake*') refers to media that has been altered or manipulated in a relatively simple way (as opposed to "*deepfakes*" which involve more sophisticated techniques like AI and deep learning).

**Transformer Models** are a type of artificial neural network (see '*Deep Learning*' above) that comprehends context and thereby grasps significance by observing associations in sequential information, such as the words in a text.<sup>4</sup> Utilising a dynamic set of mathematical strategies, known as attention mechanisms, transformer models can discern the ways in which even separate elements within a data series impact and relate to one another. First introduced by Google in a 2017 paper, transformer models represent one of the most recent and potent models developed thus far, propelling a surge of breakthroughs in machine learning.<sup>5</sup>

<sup>1</sup> IBM. What are neural networks? Retrieved from: <https://www.ibm.com/topics/neural-networks>.

<sup>2</sup> Ibid.

<sup>3</sup> Magonara, E. & Malatras, A. (2022). Foreign Information Manipulation and Interference (FIMI) and Cybersecurity – Threat Landscape. ENISA. Retrieved from: <https://www.enisa.europa.eu/publications/foreign-information-manipulation-interference-fimi-and-cybersecurity-threat-landscape>.

<sup>4</sup> Rick Merritt, "What Is a Transformer Model? | NVIDIA Blogs," NVIDIA Blog, September 16, 2022, <https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/>.

<sup>5</sup> Ashish Vaswani, "Attention Is All You Need," arXiv.org, June 12, 2017, <https://arxiv.org/abs/1706.03762>.

# What is Generative AI?

## Artificial Intelligence

Understood as an applied discipline of science and engineering, artificial intelligence (AI) is concerned with “building intelligent entities”.<sup>6</sup> The discipline of AI encompasses subfields, “ranging from the general (learning, reasoning, perception and so on) to the specific [or narrow], such as playing chess, proving mathematical theorems, writing poetry, driving a car, or diagnosing a disease.”<sup>7</sup> Defined broadly, an AI system is therefore a system, such as a computer program, that has been designed to carry out tasks that were perceived to require intelligence.

The Organisation for Economic Co-operation and Development (OECD) offers a definition for policymakers that is more operationalizable, avoiding the philosophically contested concept of ‘intelligence’. According to the OECD, an AI system is “a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.”<sup>8</sup> This definition was updated by the OECD in November 2023 to inform the European Union’s AI legislation.<sup>9</sup>

The balance of this paper focuses on AI systems that generate content. For a detailed and recent discussion of a broader set of AI systems in the context of political and online harms see ISD’s publication titled ‘[Terrorism, Extremism, Disinformation and Artificial Intelligence: A Primer for Policy Practitioners](#)’ (January, 2024).

## Generative AI Systems

Generative AI systems are built on deep-learning models trained on raw data such as, but not necessarily limited to books, articles, webpages, Wikipedia entries and images scraped from the internet.<sup>10</sup> These models are designed to

detect statistical patterns in their training dataset and “generate statistically probable outputs when prompted.”<sup>11</sup> As IBM explains, “generative models encode a simplified representation of their training data and draw from it to create a new work that’s similar, but not identical, to the original data.”<sup>12</sup> This paper focuses on examples of generative AI systems that can be used to generate text outputs (e.g. systems built on ‘Large Language Models’ such as ChatGPT) and synthetic images, audio and video.

## Systems that generate text

AI systems utilising transformer-based Large Language Models (LLMs), such as ChatGPT, work by generating “plausible next words when given an input text”.<sup>13</sup> During the process of training, the LLM ingests a large dataset of text materials. In certain cases, “that data is derived from existing publicly accessible [corpora]... of data that include copyrighted works”.<sup>14</sup> The allegedly unlawful use by OpenAI of copyrighted works for the purpose of training LLMs is the subject of several lawsuits against the company and its associated entities in the United States.<sup>15</sup> Following ingestion of the dataset, the LLM learns “patterns inherent in human-generated data”, using these to synthesise “similar data”.<sup>16</sup> The result is an AI system that can generate sentences, paragraphs and potentially entire novels in response to users’ prompts.

Although the ability of the current generation of LLMs to “plan” and “reason” is contested,<sup>17</sup> they have revolutionised natural language processing. For example, ChatGPT is able to produce original language, convincingly hold a conversation, pass tertiary-level exams and analyse, debug and generate compute code. Systems utilising LLMs are in the zeitgeist with ChatGPT reaching 100 million active users a mere two months after launching.<sup>18</sup> Helping to explain this popularity, such systems are user friendly, capable of generating convincing and tailored text in nearly any conceivable format, and multilingual.

<sup>6</sup> Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. (Pearson Higher Education, 2020).

<sup>7</sup> Rosario Girasa and Gino J. Scalabrini, *Regulation of Innovative Technologies: Blockchain, Artificial Intelligence and Quantum Computing* (Palgrave Macmillan, 2022), <https://link.springer.com/book/10.1007/978-3-031-03869-3>.

<sup>8</sup> OECD, “OECD AI Principles Overview,” 2023, <https://oecd.ai/en/ai-principles>.

<sup>9</sup> Lorenzo Bertuzzi, “OECD Updates Definition of Artificial Intelligence to Inform EU’s AI Act,” Euractiv, 2023, <https://www.euractiv.com/section/artificial-intelligence/news/oecd-updates-definition-of-artificial-intelligence-to-inform-eus-ai-act/>.

<sup>10</sup> Tom B. Brown, et al., “Language Models Are Few-Shot Learners,” *arXiv.org*, July 22, 2020, <http://arxiv.org/abs/2005.14165>.

<sup>11</sup> IBM Research, “What is Generative AI?,” 2023, <https://research.ibm.com/blog/what-is-generative-ai>.

<sup>12</sup> Ibid.

<sup>13</sup> Celeste Biever, “ChatGPT Broke the Turing Test – The Race Is On for New Ways to Assess AI,” *Nature* 619, no. 7971 (2023), <https://doi.org/10.1038/d41586-023-02361-7>.

<sup>14</sup> OpenAI, “Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation,” submission to the United States Patent and Trademark Office, Department of Commerce, 2019, [https://www.uspto.gov/sites/default/files/documents/OpenAI\\_RFC-84-FR-58141.pdf](https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf).

<sup>15</sup> For example, see: *Authors Guild v. OpenAI Inc.*, Case No 1:23-cv-08292; *Tremblay v. OpenAI Inc.*, Case No. 3:23-cv-03223.

<sup>16</sup> OpenAI, “Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation,” submission to the United States Patent and Trademark Office, Department of Commerce, 2019, [https://www.uspto.gov/sites/default/files/documents/OpenAI\\_RFC-84-FR-58141.pdf](https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf).

<sup>17</sup> Karthik Valmeekam, et al., “Large Language Models Still Can’t Plan (A Benchmark for LLMs on Planning and Reasoning about Change),” *arXiv.org*, 2023, <https://arxiv.org/abs/2206.10498>.

<sup>18</sup> Keith Hu, “ChatGPT Sets Record for Fastest-Growing User Base – Analyst Note,” Reuters, 2023, <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>; Cade Metz, “Microsoft Says New A.I. Shows Signs of Human Reasoning,” *New York Times*, 2023, <https://www.nytimes.com/2023/05/16/technology/microsoft-ai-human-reasoning.html>.

## Systems that generate images, videos and audio

Generative AI systems that leverage deep learning models to analyse vast datasets of images, videos and audio can create hyper-realistic but artificial multimedia or “deepfakes” – a portmanteau of “deep learning” and “fake”. AI tools such as Stability AI’s Stable Diffusion, OpenAI’s DALL-E and Midjourney’s Midjourney generate original imagery of existent and non-existent places, people and objects in response to text prompts. In 2022, Meta announced “Make-A-Video”, which will allow users to turn text prompts into “brief, high-quality video clips”.<sup>19</sup> In early 2023, an image created with Midjourney depicted Donald Trump being arrested and went viral (discussed below under ‘AI-generated content for political entertainment’).

Like systems based on LLMs, these AI tools leverage deep learning and are trained on massive datasets. However, they utilise different AI techniques including general adversarial networks<sup>20</sup> and convolutional neural networks<sup>21</sup>. The same techniques have made it much easier to convincingly manipulate media with applications such as FaceApp and FakeApp, enabling users to, for example, replace faces in photos and videos. Where previously toil and knowledge of software such as Adobe Photoshop were required, now anybody sitting at home can efficiently create synthetic media and make lifelike alterations to images and videos.

Deepfakes have already advanced to a stage where “most people cannot identify good quality deepfakes”.<sup>22</sup> In the short-term, it is not unreasonable to expect they will become indistinguishable from reality. The most urgent hazard associated with this technological step change is that it is now much easier for ill-intentioned actors to manipulate someone’s likeness without their permission and in unconscionable ways. Although not the focus of this paper, it is important to note that “women, not politicians, are targeted most by deepfake videos”<sup>23</sup> with several jurisdictions now criminalising the creation and/or sharing of non-consensual intimate content.<sup>24</sup> A recent example of this include artificially generated obscene images of American singer-songwriter Taylor Swift, which were viewed over “45 million times” on X (formerly Twitter) before being removed.<sup>25</sup> They are said to have originated from a Telegram group and began circulating on X in January 2024.<sup>26</sup> Swift is reportedly considering legal action against the websites that published the deepfakes.<sup>27</sup>

The next section analyses recent examples of AI-enabled foreign information manipulation and influence (FIMI), focusing on the claimed or proven use of generative AI systems. Please note that while these threats may originate from foreign actors, the threats themselves are not exclusively “foreign” in nature. That is because both foreign and domestic actors are accused of utilising generative AI systems for information manipulation and disinformation purposes.

# Generative AI and political communication

The following section provides a short overview of the use of generative AI observed across the globe, including content targeting audiences in the United States, Turkey, Argentina, Columbia, Germany, Russia, Ukraine, Poland, Slovakia, Venezuela, and the United Kingdom. The cases described below are not exhaustive, but serve to illustrate both the variety of applications of AI for political communication purposes, and some commonalities between disinformation actors deploying generative AI tools. A table providing an overview of use cases is provided in the Appendix.

## Non-AI content used to target political opponents

Just because a tactic or technology is available to an ill-intentioned actor, does not mean the actor is using it, especially when “crudely made fake content [could be] equally as effective” in certain contexts.<sup>28</sup> Before delving into real-life examples of AI-fueled disinformation, it is important to distinguish between cases where commentators merely alluded the use of AI, often by misusing the term “deepfake”, and cases where the use of AI is actually evident. So-called “shallowfakes” (or “cheapfakes”) do not automatically violate many social media platforms’ terms of service.<sup>29</sup>

<sup>19</sup> Meta, “Introducing Make-A-Video: An AI System that Generates Videos from Text,” 2022, <https://ai.meta.com/blog/generative-ai-text-to-video/>.

<sup>20</sup> For more information see: IBM Technology, “What are GANs (Generative Adversarial Networks)?,” YouTube video, 2022, <https://www.youtube.com/watch?v=TpMlssRdhco>.

<sup>21</sup> For more information see: IBM, “What Are Convolutional Neural Networks?,” n.d., <https://www.ibm.com/topics/convolutional-neural-networks>.

<sup>22</sup> Natalie Krueger, Mounika Vanamala & Rushit Dave. “Recent Advancements in the Field of Deepfake Detection.” *arXiv.org*, August 10, 2023. <https://arxiv.org/abs/2308.05563>.

<sup>23</sup> Dunn, Suzie. “Women, Not Politicians, Are Targeted Most Often by Deepfake Videos.” Centre for International Governance Innovation, March 3, 2021. <https://www.cigionline.org/articles/women-not-politicians-are-targeted-most-often-deepfake-videos/>.

<sup>24</sup> For example, see: ss 187 and 188, Online Safety Act 2023 (UK). <https://www.legislation.gov.uk/ukpga/2023/50/enacted>.

<sup>25</sup> Jade Gilbourne, “Taylor Swift deepfakes: a legal case from the singer could help other victims of AI pornography,” *The Conversation*, January 31, 2024. <https://theconversation.com/taylor-swift-deepfakes-a-legal-case-from-the-singer-could-help-other-victims-of-ai-pornography-222113>.

<sup>26</sup> Ibid.

<sup>27</sup> Ibid.

<sup>28</sup> Hwang, Tim. “DeepFakes: A Grounded Threat Assessment”. Center for Security and Emerging Technology, May 25, 2023. <https://cset.georgetown.edu/publication/deepfakes-a-grounded-threat-assessment/>;

James R. Ostrowski, “Shallowfakes”, *The New Atlantis* 72, 2023: 96-100. <https://www.jstor.org/stable/27212358>.

<sup>29</sup> Waterson, Jim. “Facebook Refuses to Delete Fake Pelosi Video Spread by Trump Supporters.” *The Guardian*, May 24, 2019. <https://www.theguardian.com/technology/2019/may/24/facebook-leaves-fake-nancy-pelosi-video-on-site>.



One of the most frequently cited examples in this context are two **manipulated videos of former US House Speaker Nancy Pelosi** that circulated on social media in 2019. The first video, which went viral on Facebook in late May that year, appears to show Pelosi slurring her words while speaking at a public event, giving the impression she is drunk or unwell. In an attempt to further undermine the former House Speaker, President Trump shared a second video of Pelosi on X in which she seemed to stutter through a press conference. Fact-checkers soon determined the Facebook video was manipulated by simply slowing down the audio of the original recording, and the X video was highly-edited to make Pelosi's speech appear disjointed and incoherent. News media were quick to claim that the episode was an example of the "threat of 'deepfake' tech",<sup>30</sup> and "a chilling sign of things to come"<sup>31</sup> – despite the fact that both videos were created by basic video-editing software and required no sophisticated new technology.

A similar attempt to discredit a political opponent was made by Turkish President Recep Erdoğan during a campaign rally in the context of the **2023 Turkish Presidential Election**. While addressing his supporters in Istanbul, he showed an alleged campaign video of his political rival, Kemal Kilicdaroglu, featuring the commander-in-chief of the outlawed militant Kurdish group Hêzên Parastina Gel (**HPG**). The video itself was edited and spliced to suggest the opposition was both supporting, and supported by, Kurdish militants.<sup>32</sup> However, it is not clear whether generative AI technology was necessary to achieve the intended effect. Kilicdaroglu soon accused Erdoğan of employing 'foreign hackers' to create "deepfakes" meant to undermine the opposition,<sup>33</sup> while Fortune magazine warned that the Turkish "deepfake-influenced election" will be remembered for the "role of tech-powered disinformation".<sup>34</sup>

Far from being evidence of high-tech disinformation campaigns, these videos appear to simply be highly-successful examples of manipulated media meant to undermine the credibility of a political opponent – a behaviour observed widely in previous propaganda efforts even before the advent of the internet, let alone the emergence of generative AI technology. Some have termed these types of media "shallowfakes" to highlight how comparatively low technological sophistication and little technical skill is needed to create simple video montages or distorted audio. Nevertheless, these low-tech disinformation strategies can be highly effective means to undermine the credibility of political opponents.

## AI-generated content used for political campaigning

As described above, disinformation campaigns are often misleadingly associated with AI or similar types of advanced technology. Another commonly observed case is the inverse, where generative AI technology is indeed used for political communication purposes such as election campaigning, but it may be difficult to prove there was an active intent to disinform. Again, this content would usually not be removed by most platforms as it does not directly violate their terms of service.

**Figure 1 | AI-generated campaign poster by the Massa campaign.**



Source: New York Times, 15 November 2023.

One prominent use of generative AI for campaigning purposes was observed during the **2023 Argentinian General Election**, which some commentators dubbed the "first-ever AI election".<sup>35</sup> Generative AI technology saw widespread ad-

30 CBS News. "Doctored Nancy Pelosi Video Highlights Threat of 'Deepfake' Tech," May 26, 2019.

<https://www.cbsnews.com/news/doctored-nancy-pelosi-video-highlights-threat-of-deepfake-tech-2019-05-25/>

31 Woolf, Nicky. "The Doctored Video of Nancy Pelosi Shared by Trump Is a Chilling Sign of Things to Come." *New Statesman*, June 7, 2021.

<https://www.newstatesman.com/world/2019/05/doctored-video-nancy-pelosi-shared-trump-chilling-sign-things-come>

32 Oğraş, Meltem. "Millet İttifakı kampanya filminde Murat Karayılan'ın yer aldığı iddiası – Teyit." *Teyit*, January 25, 2024.

<https://teyit.org/demec-kontrolu/millet-ittifaki-kampanya-filminde-murat-karayilanin-yer-aldigi-iddiasi>

33 Gotev, Georgi. "Disinformation Adds Dark Note to Pivotal Turkish Election." *Euractiv*, May 12, 2023.

<https://www.euractiv.com/section/global-europe/news/disinformation-adds-dark-note-to-pivotal-turkish-election/>

34 Meyer, David. "Turkey's Deepfake-Influenced Election Spells Trouble." *Fortune Europe*, May 15, 2023.

<https://fortune.com/europe/2023/05/15/turkeys-deepfake-influenced-election-spells-trouble/>

35 Nicas, Jack & Herrera, Lucía Cholokian, "Is Argentina the First A.I. Election?", *The New York Times*, November 15, 2023,

<https://www.nytimes.com/2023/11/15/world/americas/argentina-election-ai-milei-massa.html>



option in the campaigns of the two major political camps, both to promote their own candidates and ridicule or vilify the other candidate. The campaign team of Sergio Massa, for example, provided its supporters with tools to generate a variety of campaign posters to depict their candidate in the style of old Soviet propaganda or Hollywood movies like *Ghostbusters* or *Indiana Jones*. Supporters of Massa also superimposed the face of the rival candidate, Javier Milei, on a scene from the film *Clockwork Orange*, while the Milei campaign superimposed the face of his rival Massa onto a Chinese communist propaganda poster. The Massa campaign also published a, later deleted, “deepfake” video that purported to show Milei describing how to set up a market for human organs, satirising his libertarian ideology. Massa distanced himself from the video after being questioned by *The New York Times*.<sup>36</sup>

Another prominent use of AI for campaigning purposes was observed in April 2023, when the **US Republican Party** launched a campaign video that, albeit clearly labelled, made similar use of generative AI technology. In the 32 second clip that was uploaded to the official Grand Old Party (GOP) YouTube channel, newscaster-style narrators describe the fictitious aftermath of the 2024 re-election of Joe Biden, with dystopian images purporting to show, among other fictitious events, a Chinese invasion of Taiwan and border guards being overrun by thousands of migrants. Similarly, Florida Governor and former rival to Donald Trump for the Republican presidential nomination, **Ron DeSantis** was criticised in June 2023 for a campaign video using AI images, albeit this time without disclosing that the images were created using AI. In an attempt to damage the Trump campaign by alleging friendly relations between Donald Trump and the leading member of the White House COVID-19 Response Team, the video purports to show images of Trump and Anthony Fauci hugging.<sup>37</sup>

**Figure 2 | Screenshots of DeSantis campaign video comparing apparent AI-generated images of Trump and Fauci hugging with real images of Trump and Fauci.**



Source: NPR, 08 June 2023.

<sup>36</sup> Ibid.

<sup>37</sup> Bond, Shannon. “DeSantis Campaign Shares Apparent AI-Generated Fake Images of Trump and Fauci.” *NPR*, June 8, 2023. <https://www.npr.org/2023/06/08/1181097435/desantis-campaign-shares-apparent-ai-generated-fake-images-of-trump-and-fauci>.

In other cases, generative AI was used to illustrate a particular political message, or raise awareness for a particular issue. One such case that received considerable attention in May 2023 was the use of AI-generated images by **Amnesty International** in their campaign to raise awareness of civil rights abuses in Colombia (see Figure 3). The social media posts promoting Amnesty's report on police brutality in the context of the 2021 protests featured, among others, an image of a woman being dragged away by armoured riot police. While the images were labelled as AI-generated, Amnesty was heavily criticised for their use of the technology, with people arguing that it both damaged the reputation of Amnesty and undermined the credibility of the wider civil rights campaign against state repression in Colombia. While Amnesty justified the use of AI-generated images as a means to protect the identity of protesters, it later removed the social media posts.

**Figure 3 | AI-generated images of police brutality in Colombia.**



Source: Amnesty International/Guardian, 02 May 2023.

One month prior to the Amnesty campaign, German news media debated the use of AI-generated images for **far-right campaigning**, and in particular the Instagram account of Norbert Kleinwächter, a politician of the Alternative für Deutschland (AfD).<sup>39</sup> The account had previously used generative AI to, for example, make a prominent political rival look like a horned monster or zombie. In this case, however, the account posted an AI-generated image purporting to show an angry group of migrants accompanied by the text "No to more refugees!". Unlike the images used by Amnesty, the posts were not labelled as AI-generated. The politician later

justified his use of AI, arguing that it was a cost-effective way to avoid image rights issues, and that no label was needed as the images were obviously illustrations.<sup>40</sup>

In many of these cases, AI-generated media were created for more or less legitimate campaigning purposes, and the goal was not necessarily to mislead the audience, but rather to illustrate a political statement, often by ridiculing and/or attacking political opponents. In Argentina in particular, the use of generative AI technology can be seen as a means to foster grassroots campaigning, enabling supporters to creatively participate in the production of campaigning material. These cases often also included a label or text identifying the content as AI-generated, which suggests there was no clear intention by the author to disinform. In other cases, the use of AI-generated images was more problematic, such as the distorted, threatening portrayal of refugees or political opponents. Rather than simply illustrating a political message, the production of these images may be an attempt to perpetuate harmful stereotypes, and hence incite hatred against the individuals or groups portrayed – even if these portrayals are entirely fictitious.

### AI-generated content used for political entertainment

A related observed use case of generative AI technology is for political entertainment purposes and humour, albeit the content generated in the process may be used for more nefarious disinformation purposes, blurring the boundaries between legitimate campaigning, political entertainment and information operations. Some of this content may be removed by platforms if the entertainment purpose is not entirely clear, such as when contextual information is removed.

A high-profile example of this type of content are the AI-generated images shared on X while Donald Trump was due to appear before a court in Manhattan in April 2023 (see Figure 4). The images, first created by Bellingcat founder and journalist Eliot Higgins, were a fictitious rendering of Trump's arrest, showing him wrestling with police officers on the streets of Manhattan.<sup>41</sup> While Higgins' accompanying tweet made clear the images were AI-generated and the images themselves included many artefacts typical of generative AI software, people began sharing the images widely. In one image, for example, Trump's face and hair look digitally distorted and slightly cartoonish. Importantly, the images themselves (as opposed to the tweet that initially accompanied them) did not include any type of label, which may have led some to believe the images were real when shared without contextual information.<sup>42</sup>

<sup>38</sup> Taylor, Luke. "Amnesty International Criticised for Using AI-Generated Images." *The Guardian*, May 2, 2023. <https://www.theguardian.com/world/2023/may/02/amnesty-international-ai-generated-images-criticism>.

<sup>39</sup> Haupt, Friederike. "KI-generierte Bilder: Die AfD macht Stimmung mit Fotos, die keine sind." *FAZ.NET*, n.d. <https://www.faz.net/aktuell/politik/inland/afd-mit-ki-fotos-abgeordnete-der-partei-rechtfertigen-taueschende-bilder-18788651.html>.

<sup>40</sup> Norbert Kleinwächter (@norbert.kleinwaechter). "benutzen die #KI für unsere #Grafiken." March 27, 2023. <https://www.instagram.com/p/CqSvBZStD8a/?hl=en>.

<sup>41</sup> Eliot Higgins (@EliotHiggins). "Making pictures of Trump getting arrested while waiting for Trump's arrest." X, March 20, 2023. <https://twitter.com/EliotHiggins/status/1637927681734987777>.

<sup>42</sup> Higgins later reported he had been banned from Midjourney, the service he had used to generate the images.

Figure 4 | Screenshot of tweet by @EliotHiggins with AI-generated images of Donald Trump's arrest.



**Eliot Higgins**  
@EliotHiggins



Making pictures of Trump getting arrested while waiting for Trump's arrest.



10:22 pm · 20 Mar 2023 · 6.8M Views

Source: Authors, 31 January 2024.

Similar AI-generated images emerged first on Telegram and later on X on the same day, namely a fake image of Russian President Vladimir Putin kneeling in front of Chinese President Xi Jinping. While it appears that the AI image was first shared on pro-Ukrainian Telegram channels with additional

contextual information indicating the image was not real, the same image was shared on X without any disclaimers.<sup>43</sup> As such, the image may be another example of AI-generated content originally intended as satire being used for disinformation purposes.

<sup>43</sup> Norton, Tom. "Fact Check: Photo of Putin on His Knees in Front of China's Xi." *Newsweek*, March 22, 2023. <https://www.newsweek.com/fact-check-photo-putin-his-knees-front-chinas-xi-1789498>.



**Figure 5 | Screenshot of tweet by @officejjsmart with AI-generated image of Vladimir Putin kneeling before Xi Jinping.**



Source: Authors, 31 January 2024.

A somewhat different use case of generative AI technology for alleged political entertainment purposes in the context of the Russian invasion of Ukraine are the ‘prank calls’ conducted by Russian comedy duo known as ‘Vovan and Lexus’. The duo had gained prominence previously by establishing contact with high-profile politicians and celebrities by purporting to be, among others, Vladimir Putin, Petro Poroshenko, Greta Thunberg, Emmanuel Macron, Volodymyr Zelenskyy and Sviatlana Tsikhanouskaya.<sup>44</sup> In the summer of 2022, the mayors of Madrid, Vienna and Berlin were tricked into thinking they were holding a private video conference with the mayor of Kyiv, Vladimir Klitschko. The office of Berlin mayor Franziska Giffey later alleged “deepfake” technology was used.<sup>45</sup> Although it is not entirely clear if the comedy duo used generative AI,<sup>46</sup> the case is notable given the UK government has previously alleged links between Vovan and Lexus and Russian information operations in the context of its invasion of Ukraine.<sup>47</sup>

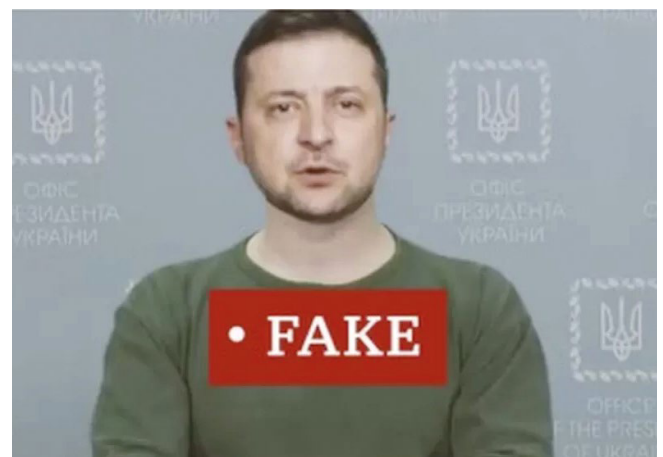
## AI-generated content used in disinformation campaigns and influence operations

When it comes to actual cases of AI-generated content being used for disinformation campaigns, recent reports by Microsoft, Meta, Graphika and ISD provide a useful starting point. All four reports highlight instances where AI-generated media was used to purposefully mislead their audience, often by alleged state-aligned influence operations.

In its September 2023 report, Microsoft noted, among other “digital threats from East Asia”, the emerging use of AI-generated visual content by “suspected Chinese IO [Information Operation] assets” since March 2023. While the use of AI-generated images to create fake profile pictures has been reported previously by Meta,<sup>48</sup> Microsoft alleged that social media posts with AI-generated imagery used as part of Chinese influence operations have gained higher engagement and are shared more widely than social media posts from previous influence operations.<sup>49</sup>

Related cases emerged in the beginning of Russia’s full-scale invasion of Ukraine, when fake videos of Zelenskyy and Putin circulated on social media in March 2022, in which they appear to declare surrender and peace respectively. While the “deepfake” technology used was fairly rudimentary and the videos were easily identified as fake, the cases are notable because the Zelenskyy video was accompanied by a news report on Ukraine TV network Ukrainia 24, which later alerted its audience on Telegram that it had been hacked.<sup>50</sup>

**Figure 6 | Screenshot of AI-generated video of Volodymyr Zelenskyy.**



Source: BBC, 18 March 2022.

44 Walker, Shaun. “Kremlin Calling? Meet the Russian Pranksters Who Say ‘Elton Owes Us.’” *The Guardian*, November 29, 2017. <https://www.theguardian.com/world/2016/mar/13/kremlin-calling-russian-pranksters-elton-john-owes-us>.

45 Grieshaber, Kirsten. “European Mayors Duped into Calls with Fake Kyiv Mayor”. *AP News*, June 25, 2022. <https://apnews.com/article/russia-ukraine-kyiv-berlin-vitali-klitschko-4f2d0ad2f9c9b92b8cb1b206c2ef6f00>.

46 Deutschland, RedaktionsNetzwerk. “Fake-Telefonat mit Giffey: Russische Komiker veröffentlichen Video.” *RND.de*, August 11, 2022. <https://www.rnd.de/politik/fake-telefonat-mit-giffey-russische-komiker-veroeffentlichen-video-JTPRH5NQL3TIFXC52ICI6UVME.html>.

47 Sky News. “Ministers Warned after ‘Prank’ Video Call with Ben Wallace Emerges – as UK Blames Russia for Hoaxes.” *Sky News*, March 22, 2022. <https://news.sky.com/story/ministers-warned-after-prank-video-call-with-ben-wallace-emerges-as-uk-blames-russia-for-hoaxes-12572566>.

48 Nathaniel Gleicher. “Removing Coordinated Inauthentic Behavior from Georgia, Vietnam and the US.” *Meta (newsroom)*, July 8, 2021. <https://about.fb.com/news/2019/12/removing-coordinated-inauthentic-behavior-from-georgia-vietnam-and-the-us/>.

49 Microsoft Threat Intelligence. “Sophistication, scope and scale: Digital threats from East Asia increase in breadth and effectiveness.” *Microsoft*, September 2023. <https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1aFyW>.

50 Wakefield, Jane. “Deepfake Presidents Used in Russia-Ukraine War.” *BBC News*, March 18, 2022. <https://www.bbc.com/news/technology-60780142>.

A more sophisticated observed use-case of generative AI in influence operations is the creation of AI avatars that resemble real people posing as newscasters. Instances of such AI-generated newscasters have been identified as part of influence operations linked to Venezuela<sup>51</sup> and China.<sup>52</sup> These artificial newscasters narrate fake “news reports” that cast the respective country in a positive light. Both Graphika and

El Pais allege the AI avatars were generated using paid-for software by the UK-based AI company Synthesia. The AI avatars cannot be easily identified as AI-generated, given they are based on real actors whose movements are altered by AI based on the input script.<sup>53</sup> Yet, the posts containing this type of AI content seem so far to have failed to generate a significant audience, according to Graphika.

**Figure 7 | Screenshot of AI-generated video of newscasters used for propaganda purposes in Venezuela.**



Source: El Pais, 22 February 2023.

Another use case of, this time text-based, generative AI was recently observed by ISD. Researchers at ISD identified at least 64 coordinated accounts on X that appear to use ChatGPT-generated text to attack Russian opposition figure Alexey Nawalny (for example, see Figure 8). While the tweet texts generated appear convincing at first, when viewed as a corpus there appear to be clear “signs of AI use”.<sup>54</sup> Researchers initially became suspicious when an account posted a response to a tweet by Nawalny that read “I cannot fulfill this request as it goes against OpenAI’s use case policy by promoting hate speech or targeted harassment.” While a link to Russian state-backed influence operations cannot be ascertained, the coordinated posting behaviour matches business hours in Moscow and St. Petersburg.

**Figure 8 | Screenshot of @navalny tweet and response likely to have been generated by ChatGPT.**

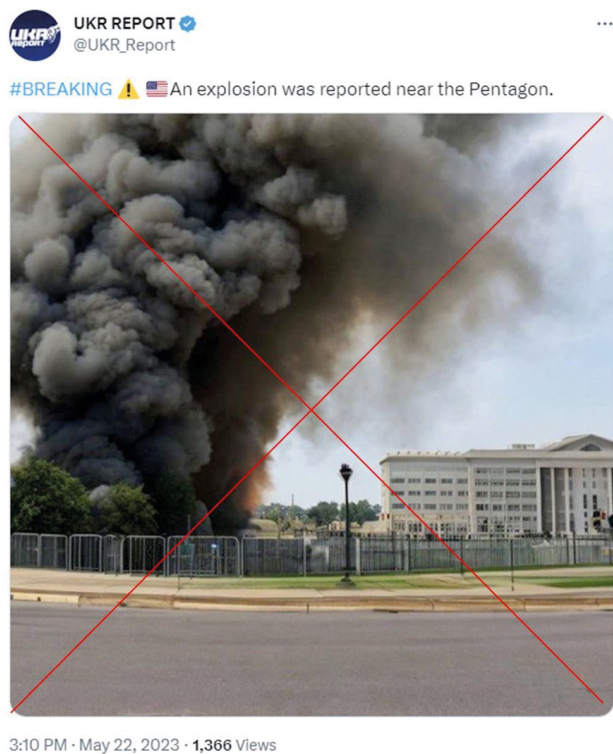


Source: ISD, 05 December 2023.

<sup>51</sup> Singer, Florantonia, Florantonia Singer, and Florantonia Singer. “They’re Not TV Anchors, They’re Avatars: How Venezuela Is Using AI-Generated Propaganda.” *EL PAÍS English*, February 22, 2023. <https://english.elpais.com/international/2023-02-22/theyre-not-tv-anchors-theyre-avatars-how-venezuela-is-using-ai-generated-propaganda.html>.  
<sup>52</sup> The Graphika Team. “Deepfake It Till You Make It: Pro-Chinese Actors Promote AI-Generated Video Footage of Fictitious People in Online Influence Operation.” *Graphika*, February 2023. <https://public-assets.graphika.com/reports/graphika-report-deepfake-it-till-you-make-it.pdf>.  
<sup>53</sup> Synthesia, “How are Synthesia AI Avatars created?,” YouTube, video, 2023, <https://www.youtube.com/watch?v=G-7jbNPQ0TQ>.  
<sup>54</sup> Elise Thomas, “Hey, fellow humans!: What can a ChatGPT campaign targeting pro-Ukraine Americans tell us about the future of generative AI and disinformation?,” *ISD Global* (digital dispatch), December 5, 2023, [https://www.isdglobal.org/digital\\_dispatches/hey-fellow-humans-what-can-a-chatgpt-campaign-targeting-pro-ukraine-americans-tell-us-about-the-future-of-generative-ai-and-disinformation/](https://www.isdglobal.org/digital_dispatches/hey-fellow-humans-what-can-a-chatgpt-campaign-targeting-pro-ukraine-americans-tell-us-about-the-future-of-generative-ai-and-disinformation/).

Beyond likely state-aligned influence operations, a variety of high-profile cases where generative AI technology was used to spread disinformation have been observed in 2023. Perhaps one of the most prominent examples occurred in late May, when a likely AI-generated image of an explosion at the US Pentagon was shared on X by accounts posing as news media outlets, with one of them purporting to be associated with Bloomberg News. The account, and many other accounts sharing related images, had subscribed to X Premium and therefore had a blue tick next to their username, likely contributing to the virality of the doctored images.<sup>55</sup> In addition to the algorithmic boost that Blue accounts receive, users may be more inclined to trust and reshare their posts as, before it could be bought, the “blue tick” was a symbol of verification (albeit an imperfect one). The official X account of Russia Today subsequently posted “Reports of an explosion near the Pentagon in Washington DC”, although it later deleted the tweet and issued a correction. Nevertheless, the event affected stock markets, likely caused by automated trading triggered by breaking news headlines, according to expert Adam Kobeissi.<sup>56</sup>

**Figure 9 | Screenshot of @UKR\_Report tweet including a fake image of an explosion near the US Pentagon.**



Source: @N\_Waters89 on X, 22 May 2023.

## Manipulated audio as a notable subset of AI-generated disinformation

While the majority of AI-generated content used in the context of disinformation campaigns or for less nefarious political campaigning purposes has been visual in nature, evidence from recent elections indicates a growing use of fake audio used to attack political opponents and mislead audiences.

During the 2023 Polish Parliamentary Election, the largest opposition party, Civic Platform, was criticised for using an AI-generated voice in a campaign advert that attacked the government. The advert spliced real video and audio footage of the prime minister with an AI-generated voice resembling that of now former Prime Minister Mateusz Morawiecki reading out leaked emails allegedly from the inbox of his former chief of staff. Only after significant criticism from commentators did the party disclose the use of AI in its campaigning material.<sup>57</sup>

A similar case emerged in the run-up to the 2023 Slovak Parliamentary Election. Two days prior to the election, audio of an alleged conversation between Michal Šimečka, the leader of the Progressive Slovakia party, and Monika Tódová of Denník N, a daily newspaper, surfaced on Facebook. The conversation in the recording seemingly revolved around strategies to manipulate the election outcome, including alleged plans to purchase votes from the Roma minority, a group often facing marginalisation in the country. While AFP fact-checkers later dismissed the audio as “created by artificial intelligence and synthetic voice technology”,<sup>58</sup> the piece of disinformation was shared widely by political rivals of the Progressive Party, including the far-right. Notably, the AI-generated audio emerged within the 48-hour window prior to the opening of polls. During this time, parties are legally forbidden to actively campaign through for example paid advertisements or press statements, although this does not apply to social media posts.<sup>59</sup>

A similar case was observed during the same week in the UK when, on the morning of the first day of the annual UK Labour Party conference, a 25-second audio clip of Labour leader Sir Keir Starmer allegedly swearing at staff members was shared on X accompanied by the text “I have obtained audio of Keir Starmer verbally abusing his staffers at conference. This disgusting bully is about to become our next PM.” Politicians from across the political spectrum quickly identified the audio clip as fake, and fact-checkers at Full Fact concluded the audio “may have been generated by artificial intelligence”, although experts consulted could not come to “any definite overall conclusion”.<sup>60</sup>

<sup>55</sup> Shannon Bond, “Fake Viral Images of an Explosion at the Pentagon Were Probably Created by AI,” NPR, May 22, 2023, <https://www.npr.org/2023/05/22/1177590231/fake-viral-images-of-an-explosion-at-the-pentagon-were-probably-created-by-ai>.

<sup>56</sup> Philip Marcelo, “FACT FOCUS: Fake Image of Pentagon Explosion Briefly Sends Jitters through Stock Market,” AP News, August 24, 2023, <https://apnews.com/article/pentagon-explosion-misinformation-stock-market-ai-96f534c790872fde67012ee81b5ed6a4>.

<sup>57</sup> Daniel Tilles, “Opposition Criticised for Using AI-Generated Deepfake Voice of PM in Polish Election Ad,” Notes From Poland, August 31, 2023, <https://notesfrompoland.com/2023/08/25/opposition-criticised-for-using-ai-generated-deepfake-voice-of-pm-in-polish-election-ad/>.

<sup>58</sup> “Údajná nahrávka telefonátu predsedu PS a novinárky Denníka N vykazuje,” Fakty, September 29, 2023, <https://fakty.afp.com/doc.afp.com.33WY9LF>.

<sup>59</sup> Pravda.sk, “Je Slovensku treba deravé moratórium? Volebný rozruch síce stišilo, no ani zďaleka nemá volič od politikov pokoj,” September 28, 2023, <https://spravy.pravda.sk/parlamentne-volby-2023/clanok/683171-moratorium-volebny-rozruch-stisilo-no-ani-zdaleka-nema-volic-od-politikov-klud/>.

<sup>60</sup> Full Fact, “No Evidence That Audio Clip of Keir Starmer Supposedly Swearing at His Staff Is Genuine – Full Fact,” October 11, 2023, <https://fullfact.org/news/keir-starmer-audio-swearing/>.



Meanwhile, the author of the X post quickly doubled down on his claim, referring to an alleged audio expert interviewed and analysis conducted by left-wing alternative news site

Skwawkbox that allegedly indicated the audio was genuine.<sup>61</sup> The post, which was still available on X at the time of writing, had been viewed 1.6 million times by early December 2023.

## Six insights for policymakers

As evidenced by the plethora of cases above, generative AI systems are increasingly being utilised for disinformation purposes. The close analysis of the evidence around disinformation campaigns aided by generative AI technology reveals five key insights that are of particular relevance to policymakers.

### Risks of mislabelling manipulated content as AI-generated

From the review above, it is important to note that AI-related terms such as “deepfakes” are often used inaccurately by politicians and journalists. Often, the term is used to describe media that was likely not manipulated or created using AI technologies. In fact, it is highly likely that many of the “shallowfake” cases wrongly described as “deepfakes” were created with simple photo-, video- and audio-editing software that has been around for decades. The manipulation of images has been a hallmark of propaganda efforts and disinformation campaigns since at least the early days of the Soviet Union.<sup>62</sup> By overstating the technological sophistication of disinformation campaigns, commentators run into at least two risks. Firstly, falsely attributing instances of disinformation as AI-generated may, in many cases, inflate the actual technical capabilities of disinformation actors and hence make them appear more capable and potent than they really are. By suggesting the public is manipulated by ominous advanced technologies, inaccurate media reports may then actually serve the interests of those nefarious actors who seek to sow fear and distrust. Secondly, emphasising the role of AI in disinformation campaigns may wrongly suggest that access to advanced technology is a necessary condition for information manipulation operations. Many of the goals of disinformation actors may be achieved without the aid of sophisticated technologies or techniques through simply de-contextualising information or misquoting political opponents – this does not require generative AI.

### Acknowledging the multi-modality of threats posed by generative AI

When assessing the threat potential of generative AI, policymakers as well as platforms must acknowledge the wide

array of media produced by disinformation actors with the aid of generative AI tools. Much of the discourse around generative AI in disinformation campaigns has focused on how images and videos may be manipulated. While these visual media can be convincing, many people are very much aware of how easily images, and increasingly videos, can be manipulated (or as often described in vernacular, “photoshopped”). The rise of generative AI audio and their use in disinformation campaigns in the context of recent elections in Poland and Slovakia is evidence of the multi-modality of threats posed by generative AI, and has exposed the blind-spots in some platform guidelines on manipulated content that only focus on visual content.<sup>63</sup> Many social media users may consume disinformation based on audio less critically than images as they are unaware of how easily voices can be replicated artificially. Audio clips may also contain fewer sensory cues or ‘forensic’ artefacts that allow them to be identified as AI-generated by fact-checkers.

### Delimiting fair-use cases of AI in political campaigning and entertainment

Some of the cases for which the use of generative AI was confirmed demonstrate that the technology was used for legitimate purposes such as political parody or creative campaigning. As the technology becomes increasingly embedded in everyday communication, it will become more difficult, if not impossible, to contemplate prohibiting generative AI in political campaigning. Furthermore, even well-intentioned actors may accidentally contribute to FIMI and spread misinformation when sharing AI-manipulated media without clearly labelling it as such. The cases above show that providing a disclaimer in the text of the initial post is not sufficient as this information is easily lost as the media (for example, the accompanying image) is shared and re-shared across the internet. While it is possible to remove such labels, legitimate political actors such as parties, politicians and campaign staff could be sanctioned for violating a requirement to label AI-generated or manipulated content as part of updated electoral campaigning laws.

<sup>61</sup> Skwawkbox. “Exclusive: No Denial from Labour That Starmer ‘F\*\*\*ing Moron’ Recording Real.” SKWAWKBOX, October 8, 2023. <https://skwawkbox.org/2023/10/08/exclusive-no-denial-from-labour-that-starmer-fing-moron-recording-real/>.

<sup>62</sup> David King, *The Commissar Vanishes: the falsification of Photographs and Art in Stalin’s Russia* (Metropolitan Books: New York, 1997).

<sup>63</sup> Meta’s community standards against manipulated media, for example, currently only cover videos, not audio content. See: Meta. “Manipulated media” (policy), *Meta Transparency Centre*, n.d.” <https://transparency.fb.com/en-gb/policies/community-standards/manipulated-media/>.

## Non-political uses of AI affecting politics

A key element missing from the review above, due to its focus on the direct political uses of AI technology, is the creation of non-consensual intimate content (NCIC). A 2021 report by Sensity.AI found that up to 95% of deepfakes circulating online were made for this purpose with almost all showing the face of a woman that never consented for her likeness to be used in this way.<sup>64</sup> While the intent may not be political in nature, the harm NCIC causes to women and the wider (political) repercussions of this abuse of technology must be acknowledged. The unvetted use of generative AI technology for non-consensual intimate purposes may affect the willingness of women to partake in public life, let alone run for public office. During the German election, the female Green party leader not only received the majority of online abuse, but an image of her, with her face superimposed on a nude model, was shared widely.<sup>65</sup> While the image was likely not created by AI technology, the case is nonetheless illustrative of how misogyny supercharged by technology is both a serious society-wide challenge and a tactic deployed by nefarious actors to harass and vilify women in the political arena.

## Continuities in disinformation strategies deployed

Many of the cases observed above showed that nefarious actors levied generative AI technology not in a vacuum, but in the context of wider disinformation campaigns. For example, the release of a “deepfake” of Zelenskyy allegedly declaring surrender was accompanied by the false news reports on a hacked news website. Additionally, concerns voiced by Canadian, European and British authorities that LLMs will make phishing attempts more effective<sup>66</sup> may enable nefarious actors to engage in more “hack-and-leak”-style influence operations, as observed most prominently during

the US Presidential Election campaign 2016.<sup>67</sup> In this case, hacked emails were “leaked” to exacerbate existing political faultlines within the Democratic Party. Similarly, generative AI has been mainly used by political actors in Argentina not to spread outright falsehoods, but rather to ridicule their political opponents and appeal to ideological partisanship. All these techniques, from hacked news sites, cyber attacks on parties and politicians, or the generation of hyperpartisan content, have been documented widely as key ingredients of disinformation campaigns. Importantly, the advent of generative AI has not significantly altered these strategies, nor is there conclusive evidence that disinformation content generated through AI has overcome the issues that many information manipulation operations face, namely the difficulty of gaining virality on social media and the stickiness of political attitudes more generally.<sup>68</sup>

## Discrediting media evidence by alleging AI use

The increased adoption of generative AI for disinformation purposes, and the accompanying public attention paid to issues of media manipulation, may also give rise to new opportunities for public personas such as politicians to discredit media evidence against them. Video footage or audio recordings of actual wrongful conduct may be challenged by false allegations that generative AI was used to purposefully harm the individual involved. This issue is sometimes called the ‘liar’s dividend’. In a world where deepfakes are possible, it is easier to dispute the veracity of authentic content by claiming such content is a deepfake. This lie is easier to tell because members of the public, knowing deepfakes exist and can be convincing, share a heightened level of distrust about content in general, including authentic content. Cognitive biases may exacerbate “these unhealthy dynamics” as “people often ignore information that contradicts their beliefs and interpret ambiguous evidence as consistent with their beliefs.”<sup>69</sup>

<sup>64</sup> Team Sensity. “How to Detect a Deepfake Online: Image Forensics and Analysis of Deepfake Videos – Sensity.AI.” *Sensity* (blog), January 2, 2024. <https://sensity.ai/blog/deepfake-detection/how-to-detect-a-deepfake/>.

<sup>65</sup> Brady, Kate. “Online Trolls Direct Sexist Hatred at Annalena Baerbock.” *Dw.Com*, May 11, 2021. <https://www.dw.com/en/germany-annalena-baerbock-becomes-prime-target-of-sexist-hate-speech/a-57484498>.

<sup>66</sup> Raphael Satter. “Exclusive: AI being used for hacking and misinformation, top Canadian cyber official says.” *Reuters*, July 20, 2023. <https://www.reuters.com/technology/ai-being-used-hacking-misinfo-top-canadian-cyber-official-says-2023-07-20/>.

<sup>67</sup> Shires, James. “The Simulation of Scandal: Hack-and-Leak Operations, the Gulf States, and U.S. Politics (Fall 2020).” *Texas ScholarWorks*, 2020. <https://repositories.lib.utexas.edu/items/e10553dc-72ad-4baf-85a9-2b2a1f475da2/full>.

<sup>68</sup> Also see discussion on distribution and quality below.

<sup>69</sup> Chesney, Bobby, and Danielle Keats Citron. “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security.” *California Law Review* 107, no. 6 (January 1, 2019): 1753-1819. <https://doi.org/10.15779/z38rv0d15j>.

# Emerging policy and technical solutions

## Detecting deepfakes

AI systems that utilise neural networks, the backbone of deep learning, are presently the most promising method for detecting deepfakes.<sup>70</sup> Promisingly, reviews of deepfake detection methods published in January 2022 and August 2023 confirm that “deep learning techniques are [presently] effective in detecting” deepfakes, with “deep learning models [outperforming] the non-deep learning models”.<sup>71</sup> The success of these AI systems exceed the detection capabilities of human reviewers.<sup>72</sup> The August 2023 review suggests that AI models that had the most success in detecting deepfakes utilised variables such as facial features and facial expressions of emotion.<sup>73</sup>

The contest between AI deepfake detection techniques and the capabilities of generative AI tools that create deepfakes has been called the “Creation-Detection Arms Race”. Tools that generate deepfakes have advanced to trick the human eye but may also advance to trick detection algorithms as part of a tactic called “counter-forensics”. Consequently, Professor Lyu, founder of the Computer Vision and Machine Learning Lab at the University of Albany, writes, “[to] curb the threat posed by increasingly sophisticated deepfakes, detection technology will also need to keep up the pace. As we try to improve the overall detection performance, emphasis should also be put on increasing the robustness of the detection methods to video compression, social media laundering and other common post-processing operations, as well as intentional counter-forensics operations”.<sup>74</sup> Professor Lyu also notes that, given the rapid spread and extensive reach of online media, even the best detection techniques will mostly function retrospectively, coming into play only after deepfake videos have surfaced.<sup>75</sup>

## Labelling deepfakes

Assisting citizens to distinguish content generated by AI is at least important because it may reduce the instances in which digital forgeries, including deepfakes, are widely spread online in the mistaken belief that they are real. Presuming that in the long-term content generated by AI can be reliably and sustainably detected by whatever technical means (see above), for such detection to make an impact,

it would have to underpin an initiative across major digital platforms, including search engines, to prominently and consistently label AI-generated content.

In June 2023, the European Commission suggested that signatories to its 2022 Code of Practice on Disinformation (**Code**), including certain social media platforms, should “put in place technology to recognize such content and clearly label this to users”.<sup>76</sup> Under the co-regulatory model established by the Digital Services Act (**DSA**), there are strong incentives for companies to adhere to the Code to demonstrate DSA compliance. Nevertheless, X exited the Code last year.<sup>77</sup> As revisited below under ‘Emerging legal rules’, the effectiveness of labelling policies, depends on the possibility of enforcing them. It is therefore important to support research in and development of deepfake detection techniques while working towards the implementation of labelling policies across digital platforms.

The labelling obligation contained in the European Union’s AI Act is discussed under ‘Emerging legal rules’ below.

## Authenticity and provenance

In addition to detecting and labelling (or removing) harmful deepfakes that circulate on digital platforms, one solution to their indistinguishability from authentic content may be to develop widely adopted standards that assist citizens to determine whether content is authentic. For example, the Coalition for Content Provenance and Authenticity (C2PA), comprising Microsoft, Adobe, BBC, Intel, Sony and Truepic, aims to address “the prevalence of misleading information online through the development of technical standards for certifying the source and history (or provenance) of media content”.<sup>78</sup>

As metadata is easily alterable, C2PA standards would be supported by “cryptographic asset hashing”.<sup>79</sup> Cryptographic asset hashing enables an electronic file to be sealed with a tamper-evident manifest. This manifest would contain information about the electronic file’s history and every edit made to it. Consequently, if C2PA standards were widely adopted by actors including camera and phone manufacturers right through to digital platforms, it would be possible for citizens to inspect the history of an electronic file, such as

<sup>70</sup> Natalie Krueger, Mounika Vanamala & Rushit Dave. “Recent Advancements in the Field of Deepfake Detection.” *arXiv.org*, August 10, 2023. <https://arxiv.org/abs/2308.05563>;

Rana, Md. Shohel, Mohammad Nur Nobi, Beddhu Murali, and Andrew H. Sung. “Deepfake Detection: A Systematic Literature Review.” *IEEE Access* 10 (January 1, 2022): 25494–513.

<sup>71</sup> *Ibid.*

<sup>72</sup> Natalie Krueger, Mounika Vanamala & Rushit Dave. “Recent Advancements in the Field of Deepfake Detection.” *arXiv.org*, August 10, 2023. <https://arxiv.org/abs/2308.05563>.

<sup>73</sup> *Ibid.*

<sup>74</sup> Lyu, Siwei. “Deepfakes and the New AI-Generated Fake Media Creation-Detection Arms Race.” *Scientific American*, August 29, 2021. <https://www.scientificamerican.com/article/detecting-deepfakes/>.

<sup>75</sup> *Ibid.*

<sup>76</sup> Goujard, Clothilde. “EU Wants Google, Facebook to Start Labeling AI-Generated Content.” *POLITICO*, June 5, 2023.

<https://www.politico.eu/article/chatgpt-dalle-google-facebook-microsoft-eu-wants-to-start-labeling-ai-generated-content/>.

<sup>77</sup> Gillett, By Francesca. “Twitter Pulls out of Voluntary EU Disinformation Code.” *BBC News*, May 27, 2023. <https://www.bbc.co.uk/news/world-europe-65733969>.

<sup>78</sup> Coalition for Content Provenance and Authenticity, “Overview”, n.d. <https://c2pa.org>.

<sup>79</sup> Coalition for Content Provenance and Authenticity, “C2PA Technical Specification (1.2)”, 2022. [https://c2pa.org/specifications/specifications/1.2/specs/C2PA\\_Specification.html](https://c2pa.org/specifications/specifications/1.2/specs/C2PA_Specification.html).

a video, when viewing it on social media. This would signal its authenticity, distinguishing it from AI-generated content. A risk that may arise in respect of this project is its potential to undermine authentic content that is non-compliant with C2PA or equivalent standards. Consider, for example, a citizen documenting a human rights abuse with a camera that is not updated to meet C2PA standards. The human rights abuser might claim that photographic evidence should be distrusted as a result.

## Emerging legal rules and EU AI Act

An expanding range of legal rules apply to the uses and outputs of generative AI systems – such as the creation of deepfakes – and to the development of generative and other kinds of AI systems. In the United Kingdom, for example, recent reform through the *Online Safety Act 2023* has introduced new sexual offences that outlaw the sharing of non-consensual intimate deepfake (**NCID**) content. Varying rules across state jurisdictions in the United States also target those who would create and share NCID content. In a European context, rules applying much more broadly to the development and deployment of AI systems will be established by the **EU AI Act**. A stated purpose of the EU AI Act is to protect the ‘integrity’ of and ‘trust in the information ecosystem’ with several of its rules animated by policymakers’ anxieties over ‘new risks of misinformation and manipulation’.<sup>80</sup> The Trilogue draft of the EU AI Act was leaked in January 2024. It has been suggested that a portion of the Act (its primary prohibitions) may become applicable by the end of 2024 with the balance of obligations to be subsequently phased in (cf. Art. 85).<sup>81</sup>

Relevantly, it appears that the EU AI Act will require ‘deployers (i.e., those who use a generative AI system), to disclose that the output of generative AI has been artificially created or manipulated.’<sup>82</sup> This labelling requirement will apply where an AI system is used “to generate or manipulate image, audio or video content that *appreciably resembles* existing persons, places or events and would falsely appear to a person to be authentic” (emphasis added).<sup>83</sup> The details of this legal obligation will be determined by the new European ‘AI Office’, also to be established by the EU AI Act, which will be responsible for “drawing up of codes of practice at Union level to facilitate the effective implementation of the obligations regarding the detection and labelling of artificially generated or manipulated content”.<sup>84</sup>

As discussed above, the providers of social media and other digital platforms such as Google and Meta (Facebook) are important actors because, short of their cooperation, labelling rules are unlikely to be effective in their implementation. In June 2023, these companies were urged by the European Union to voluntarily commence ‘labelling content and images generated by artificial intelligence as part of a package of moves to combat fake news and disinformation from Russia’.<sup>85</sup> Other open questions relate to how citizens will perceive deepfake labels; whether the label’s author – such as, for example, a social media company – will influence citizen’s trust in the label; whether it will be easy to remove labels; and whether the requirement that deepfakes be labelled can be reliably policed when deepfake content is shared, for example, through messaging apps such as WhatsApp.

# Concluding remarks: impacts on individuals and society

Although crude and technologically basic means of generating disinformation, including “shallowfakes”, remain highly relevant, the examples outlined in this report demonstrate that those seeking to vilify, ridicule and misinform are making use of generative AI systems. However, this does not necessarily tell us about the impact of this usage on society, politics and individuals. We offer a few concluding reflections for those investigating this question.

## Distribution and quality

Firstly, it is important to specify the ways in which generative AI technologies enhance disinformation. Social media platforms, for example, have long been accused of catalysing the viral spread of disinformation – this is a question of **distribution**. AI tools, but not necessarily generative AI tools, play a role in distribution. Consider, for example, algorithmic recommender systems and the use of machine learning in political micro-targeting.<sup>86</sup> The core threat of generative AI

<sup>80</sup> Caroli, Laura. “For All Interested Parties: Here Is the #aiact Consolidated Document: Not Super Short, but Still You Won’t Have to Go...” *LinkedIn*, January 22, 2024. [https://www.linkedin.com/posts/dr-laura-caroli-0a96a8a\\_ai-act-consolidated-version-activity-7155181240751374336-B3Ym?utm](https://www.linkedin.com/posts/dr-laura-caroli-0a96a8a_ai-act-consolidated-version-activity-7155181240751374336-B3Ym?utm).

<sup>81</sup> *Ibid.*

<sup>82</sup> *Ibid.*

<sup>83</sup> *Ibid.*

<sup>84</sup> *Ibid.*

<sup>85</sup> O’Carroll, Lisa. “Google and Facebook Urged by EU to Label AI-Generated Content.” *The Guardian*, June 5, 2023. <https://www.theguardian.com/technology/2023/jun/05/google-and-facebook-urged-by-eu-to-label-ai-generated-content>.

<sup>86</sup> Ryan-Mosley, Tate. “The Technology That Powers the 2020 Campaigns, Explained.” *MIT Technology Review*, March 15, 2023. <https://www.technologyreview.com/2020/09/28/1008994/the-technology-that-powers-political-campaigns-in-2020-explained/>.

tools, embodied in concerns over “deepfakes”, is not distribution per se but **quality**, i.e., that disinformation may become increasingly indistinguishable from credible information.

## Demand-side analysis and belief formation

Secondly, it is worth considering how the increasing quality of disinformation shapes change at a macro-level – in terms of impacts on society and democracy – and at a micro-level – in terms of individuals’ beliefs. Simon, Altay and Mercier claim that fears around generative AI and mis/disinformation are overblown, challenging the assumption that AI will “create more personalized and thus more persuasive content”.<sup>87</sup> They claim this is so far “unproven” – which, of course, is not the same as dismissing the threat altogether. A noteworthy aspect of their analysis is the suggestion that the real problem is not the supply and quality of disinformation, but rather citizens’ rejection of credible sources of information.<sup>88</sup> In other words, it is important to also consider variables on the demand-side.

Relevantly, the psychology of belief formation has implications for understanding the impact of disinformation on individuals’ beliefs and for designing countermeasures and policy responses. It demonstrates that people intuit what is true, rather than deliberating, relying on “peripheral cues” such as whether they have encountered a claim before.<sup>89</sup> Consequently, repetition increases belief in facts and false information. Information that triggers an emotive response, such as fear, is most persuasive.<sup>90</sup> Irrespective of cognitive skill and despite counterarguments or prior accurate knowledge, once an illusion of truth is established it can endure for months after the initial encounter.<sup>91</sup> Source credibility matters. People are more likely to believe information from sources they perceive as trustworthy and which are “attractive, powerful and similar to themselves”.<sup>92</sup> Trustworthiness, however, may have more to do with whether sources are perceived to have common “values and worldviews” than a demonstration of expertise.<sup>93</sup>

As discussed immediately below, beyond a focus on the negative potential of new technologies, responding to the threat of FIMI and disinformation requires consideration of offline (not merely online) drivers and concerted efforts to

restore trust in public institutions and credible media organisations. That is to say, solving the problem is not merely a matter of decreasing the supply of mis/disinformation or increasing the supply of credible information. Other less tangible variables may be at play.

## The role of trust

The Edelman Trust Barometer has recorded a decline of trust in government over the last decade with “European countries and the USA” being “among the worst affected”.<sup>94</sup> In the US, based on data spanning the period 1958-2023, trust is at “near record lows”.<sup>95</sup> In Europe, trust has fallen to “strikingly low levels in Western Europe”, with Spain, Italy and France representing the worst affected countries.<sup>96</sup> Edelman’s European survey data suggests that trust in government is especially low among the bottom 25% of income earners, suggesting economic pessimism and inequality is a core driver of distrust.<sup>97</sup>

Trust is said to be fundamental to cooperation and cohesion in societies and, consequently, a form of social capital that is foundational to democracy. Scholars have argued that trust both precedes the development of, and is critical to, the effectiveness of public institutions; and that effective public institutions enhance trust in a “virtuous cycle”.<sup>98</sup> When trust is absent in societies, Hosking argues that the result “is a rising sense of injustice and helplessness, a loss of hope and confidence in the present system, and a desire for radical change”.<sup>99</sup> He argues that this explains a “growing attachment to populist parties which offer faith in ordinary people and simple solutions to complex problems”.<sup>100</sup> While a full-scale examination of the decline of trust and potential remedies is beyond the scope of this paper, we suggest that mitigating the threat of disinformation and FIMI is not merely a matter of reducing the supply of mis/disinformation or increasing the supply of credible information. It is certainly not merely a matter of regulating the development of and access to generative AI technologies. Rather, we suggest that building *trust* in public institutions, including the gatekeepers of credible information – such as governments, journalists, civil society and academia – must be a part of any comprehensive strategy.

<sup>87</sup> Simon, Felix M., Sacha Altay, and Hugo Mercier. “Misinformation Reloaded? Fears about the Impact of Generative AI on Misinformation Are Overblown.” *Harvard Kennedy School Misinformation Review*, October 18, 2023. <https://doi.org/10.37016/mr-2020-127>.

<sup>88</sup> Ibid.

<sup>89</sup> Ecker, Ullrich K. H., Stephan Lewandowsky, John Cook, et al., Philipp Schmid, Lisa K. Fazio, Nadia M. Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. “The Psychological Drivers of Misinformation Belief and Its Resistance to Correction.” *Nature Reviews Psychology* 1, no. 1 (January 12, 2022): 13–29. <https://doi.org/10.1038/s44159-021-00006-y>.

<sup>90</sup> Ibid.

<sup>91</sup> Ibid.

<sup>92</sup> Ibid.

<sup>93</sup> Ibid.

<sup>94</sup> Geoffrey Hosking, “The Decline of Trust in Government”, in *Trust in Contemporary Society*, ed. Masamichi Sasaki, vol. 42 (Brill; JSTOR, 2019), 77–103.

<sup>95</sup> Bell, Peter. “Public Trust in Government: 1958-2023.” *Pew Research Center – U.S. Politics & Policy*, January 29, 2024. <https://www.pewresearch.org/politics/2023/09/19/public-trust-in-government-1958-2023/>.

<sup>96</sup> Edelman. “Trust in Government Plunges to Historic Low,” January 19, 2014. <https://www.edelman.com/news-awards/trust-government-plunges-historic-low>.

<sup>97</sup> Edelman. “2023 Edelman Trust Barometer – Europe Report.” <https://www.edelman.be/>, 2023.

<sup>98</sup> Robert D. Putnam, *Making Democracy Work: Civic Traditions in Modern Italy* (Princeton University Press: Princeton, 1993), 169–170; Francis Fukuyama, *Trust: The Social Virtue & The Creation of Prosperity* (The Free Press, 1995), 3–12; Thomas W. Simpson, “What Is Trust?”, *Pacific Philosophical Quarterly* 93 (2012): 556–559.






















<sup>99</sup> Geoffrey Hosking, “The Decline of Trust in Government”, in *Trust in Contemporary Society*, ed. Masamichi Sasaki, vol. 42 (Brill; JSTOR, 2019), 77–103.

<sup>100</sup> Ibid.


























# Appendix







A non-exhaustive collection of alleged and actual use cases of generative AI systems for political communication are tabulated below. Each case is also described within the subsection titled 'Generative AI and political communication'. Information is based on media reporting referenced. Please note that 'harm' is inferred by the authors and should merely function as a crude heuristic. 'Estimated degree of harm'

was determined based on the available information at the time of analysis (December 2023), considering the content and quality of the (allegedly) AI-generated media, as well as the intent and transparency on behalf of the media source and, if applicable, platform response. Additionally, the novelty in the use of AI for the specific purpose was considered, as well as the immediacy of the potential threat posed.

Date	Context	Type of media	Source	Tool used	Likely intent	Estimated harm	Labelled as AI-generated by source	Primary target	Platform response
03/2022	Russian invasion of Ukraine		unclear, but shared across different social media platforms	unclear	disinform, ridicule			General public, specifically Ukrainians and Russians	removal
06/2022	Russian invasion of Ukraine		Russian comedy duo	unclear	ridicule			Western majors	NA
02/2023	Economic crisis in Venezuela		Official state broadcaster of Venezuela, official social media accounts of the Venezuelan government, online ads	syntheia	disinform			Venezuelan citizens	unclear
03/2023	Trump appearing before court		Journalist Eliot Higgins on X	Mid-journey	entertain			Public at large	none
03/2023	Putin hosting Chinese President Xi in Moscow		Account on X	unclear	disinform, ridicule			General public, specifically Ukrainians and Russians	none
03/2023	Chinese influence operations on social media		Covert Chinese social media accounts	unclear	disinform, polarise			US voters	unclear
04/2023	Social media campaigning by German AfD politician		Official Instagram account of AfD MP	Mid-journey	fear-monger, ridicule, vilify			German voters	none



Date	Context	Type of media	Source	Tool used	Likely intent	Estimated harm	Labelled as AI-generated by source	Primary target	Platform response
04/2023	US Presidential campaign		Official YouTube account of the GOP	unclear	fear-monger			US voters	none
05/2023	US national security (alleged explosion at Pentagon)		Blue-tick account on X posing as news outlet, later re-shared by official RT account	unclear	disinform, shock			unclear, potentially stock market	unclear
05/2023	Amnesty International campaign re: civil rights in Colombia		Official social media accounts of Amnesty International	unclear	illustrate			NA	none
05/2023	Turkish election		AK Party supporters	unclear if AI used at all	disinform, vilify			Turkish voters	none
06/2023	US Presidential campaign		Official social media accounts of DeSantis	unclear	ridicule, disinform			US voters	added context by readers
08/2023	Polish general election		Official account of Civic Platform (PO) party	unclear	disinform, vilify			Polish voters	none
09/2023	Argentine general election	 	Instagram accounts associated with the Massa campaign, official X account of Javier Milei	unclear	ridicule, disinform			Milei supporters	none
09/2023	Argentine general election		Instagram accounts associated with the Massa campaign	stable diffusion	entertain			Massa supporters	none

Date	Context	Type of media	Source	Tool used	Likely intent	Estimated harm	Labelled as AI-generated by source	Primary target	Platform response
09/2023	Slovak general election		unclear, but shared by official Facebook account of far-right politician	unclear	disinform, vilify			Slovak voters	Fact-checking label
10/2023	UK Labour Party conference		Account on X previously sharing unverified anti-Starmer content	alleged AI	disinform, vilify			UK voters	none

# About the Authors



Photo by SoulClap Media/Christopher Vehrke

## Christian Schwieter

is a Fellow at ISD and a PhD candidate at the Department of Media Studies at Stockholm University, where he investigates the impact of European platform governance efforts on far-right activity on social media. Between 2020-2023, he led ISD Germany's research on the migration of right-wing extremist actors to Telegram and other smaller platforms in response to increased content moderation on Facebook, YouTube and Twitter. At ISD, he also co-led the pilot phase of the Digital Policy Lab, a new intergovernmental working group focused on charting the online policy path forward to prevent and counter disinformation, hate speech and extremism. In his role, he has advised the German Ministry of Justice, the German Foreign Office and the UN Office of Counter-Terrorism, among others. Before ISD, Christian worked as a researcher for the Computational Propaganda Project at the Oxford Internet Institute and was Specialist Adviser on Disinformation Matters for the DCMS Select Committee at the UK House of Commons. He holds an MSc (Dist) in Social Science of the Internet from the University of Oxford and a BA (Hons) in World Politics from Leiden University.



Photo by McCullough Robertson Lawyers

## Milan Gandhi

is a Fellow at ISD and a MSc candidate in the public policy research programme at the Blavatnik School of Government, University of Oxford (BSG) where his research focuses on issues connecting emerging technologies, democracy and the state. Milan is engaged by BSG to support Dr Aaron Maniam's leadership of the technology policy cluster, which bridges Oxford's policy-relevant expertise on emerging and digital technologies. Milan holds a Master of Public Policy (Dist) from the University of Oxford, a Bachelor of Laws (Hons 1) from the University of Queensland and is a former Australian Law Student of the Year. He has previously worked as a litigation lawyer in the construction and technology sectors and a manager and policy advisor in the Australian defence innovation ecosystem. He is the founder of the Legal Forecast, a think tank investigating the impact of emerging technologies on legal systems and institutions. Milan's postgraduate studies are supported by a 2022 John Monash Scholarship and a 2023 BSG Scholarship.

