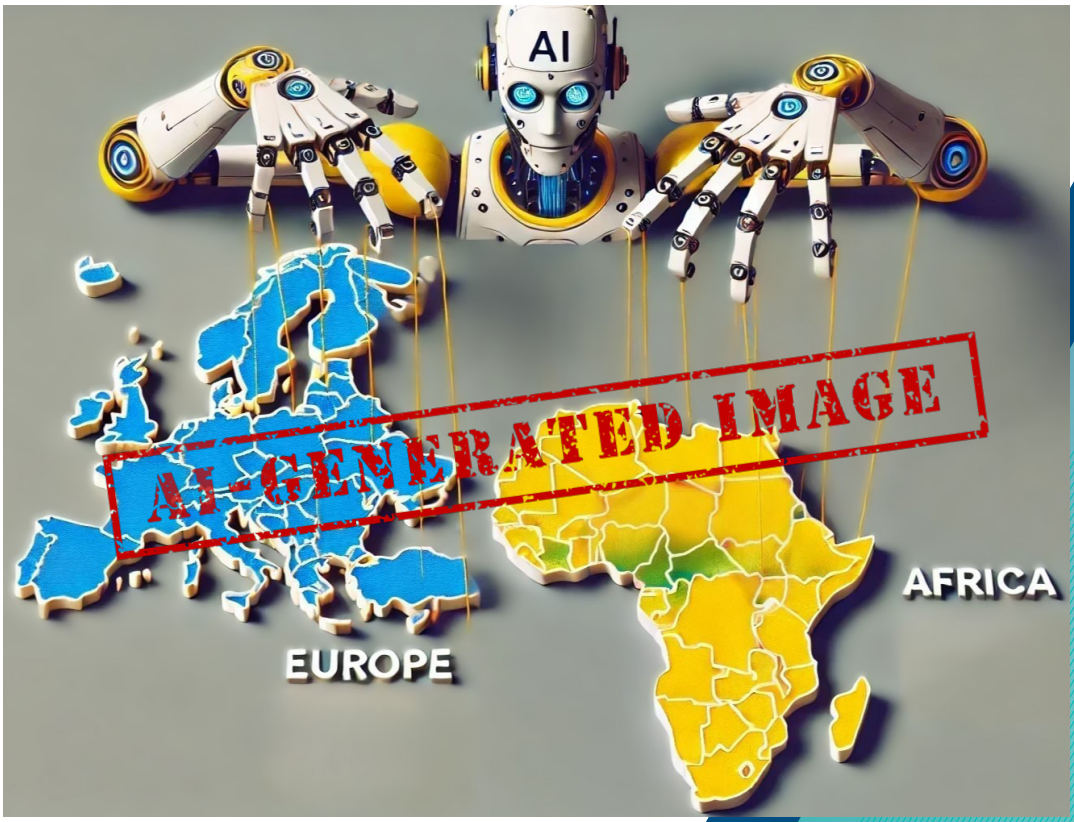
KONRAD ADENAUER STIFTUNG
MEDIA PROGRAMME
SUB-SAHARAN AFRICA

AI-Generated Disinformation in Europe and Africa

Use Cases, Solutions and Transnational Learning

Karen Allen and Christopher Nehring



KONRAD ADENAUER STIFTUNG

MEDIA PROGRAMME
SUB-SAHARAN AFRICA

AI-Generated Disinformation in Europe and Africa

Use Cases, Solutions and Transnational Learning

Karen Allen and Dr Christopher Nehring

60 Hume Road, Dunkeld 2196, Johannesburg, Republic of South Africa

Telephone: + 27 (0)11 214-2900
www.kas.de/mediaafrica

Publisher
Konrad-Adenauer-Stiftung e.V.

Authors
Karen Allen and Dr Christopher Nehring

Editors
Hendrik Sittig and Rebecca Sibanda

Proofreading
Bruce Conradie

Design, layout and production
Heath White, ihwhiteDesign

Cover
The cover image was generated entirely by DALL-E (OpenAI's image generator). The initial prompt was: *"Create a cover image on the topic 'AI-powered disinformation in Africa and Europe in comparison'. The main character is an AI bot that is depicted as a puppet master. The two maps of Europe and Africa should be outlined in the picture, with the puppeteer's strings reaching out to each. The two continents should only be represented by the outlines of the maps, there must be no flags or banners. This is very important! Please make sure that both continents are shown in their outlines! The style of the picture should be rather minimalistic, a drawing or graphic. It can be 'dirty' in its visual language, but must not contain many elements (puppeteers, the two maps - but no flags!). Please follow the instructions exactly and do not add any other elements."* The image was then refined and adjusted with further detailed prompts in more than 30 additional sessions. Layout then edited the image and inserted text boxes.

ISBN
978-1-0370-3843-3 (print) 978-1-0370-4039-9 (e-book)

Download an electronic copy of *AI-Generated Disinformation in Europe and Africa — Use Cases, Solutions and Transnational Learning* from www.kas.de/mediaafrica

Printing
Typo Printing Investments, Johannesburg, South Africa

About the authors

Karen Allen

Karen Allen is a former BBC correspondent with over 25 years of experience covering Africa and the Middle East. She now advises several organisations on emerging threats, including cyber threats and information operations. For the Institute for Security Studies, Johannesburg, she has conducted studies on the impact of disinformation and interference on elections in Kenya and South Africa. Allen holds a Master's degree in International Relations and Contemporary Warfare from King's College London and is a visiting scholar in the Department of War Studies. She has lived in Africa for the past 20 years.

Dr Christopher Nehring

Dr Christopher Nehring is a researcher, analyst and journalist and currently the Director of Intelligence at the Cyberintelligence Institute in Frankfurt, Germany. He is an expert on disinformation and has been a visiting lecturer on disinformation, intelligence and media at the Konrad Adenauer Foundation's Media Programme for Southeast Europe and at the Faculty of Journalism and Mass Communication at Sofia University. He worked as a senior analyst at the Institute for Global Analysis in Sofia and, since 2017, has been working as a journalist and editor for various media outlets (for example, Deutsche Welle, Spiegel, Neue Zürcher Zeitung, and Tagesspiegel).

Table of contents

Preface

It would be better if we did not have to deal with disinformation and fake news, but unfortunately these digital plagues continue to spread across the world. Moreover, using artificial intelligence, they are becoming more and more dangerous. We urgently need more enlightenment, media literacy, and countermeasures.

Disinformation is not a new phenomenon; it has been around for hundreds of years. However, it has become much more widespread and powerful, largely due to the internet and social networks. Moreover, the spread of fake news has increased in recent years. Worldwide, with the Covid-19 pandemic in Europe, with the Russian war in Ukraine, and, in French-speaking West Africa, with the military coups in Mali and Niger, it has reached new dimensions. In most countries, especially before elections, we continue to experience enormous influence from disinformation campaigns.

Disinformation and fake news are aimed at polarising, dividing, and destroying our liberal democratic societies. They are typically produced for precisely this purpose. Russia has made this dirty game a permanent instrument of its politics at the international level. China and various Arab states also try to spread their narratives via propaganda. In addition, disinformation is repeatedly used by political actors to discredit opponents.

Our authors, Karen Allen and Dr Christopher Nehring, demonstrate the level disinformation has reached with artificial intelligence. It is frightening how easy it is today to create photos, videos, and audio that cannot be recognised as fake. Moreover, development continues at a rapid pace. What will be possible in five or ten years?

This study lists the types of disinformation and campaigns that exist in connection with artificial intelligence, who produces them and how they are spread, what influence and effect they have, and what countermeasures exist. While Karen

Allen, who lives in South Africa, looked at developments in Africa, Dr Christopher Nehring from Germany examined the situation in Europe. The authors drew on published studies and their own research. They also had the opportunity to work together intensively for a month in Johannesburg, South Africa, where they organised a workshop on artificial intelligence.

For the first time, a comparative study of developments in disinformation and artificial intelligence in Africa and Europe will be presented. We focus on knowledge exchange, cultural and contextual insights, the identification of best practices in the fight against disinformation, and possible collaborative efforts for joint initiatives and partnerships between Europe and Africa. Let us learn from each other!

We must remind ourselves that democracy relies on pluralism, that is, many opinions that lead to societal and political decisions. However, the information by which we form our opinions must be factual. It must be fact-based and true, especially when used for political decisions. Anything else could have tragic and devastating consequences. When responsible politicians deliberately spread lies and call them “alternative facts,” societies could fall into anarchy.

Many thanks to our two authors, Karen Allen and Dr Christopher Nehring, for their excellent and fruitful collaboration!

Hendrik Sittig
Konrad-Adenauer-Stiftung
Director, Media Programme Sub-Saharan Africa
Johannesburg
2025

Introduction

Developments in artificial intelligence (AI), particularly the emergence of generative AI (genAI), to synthetically create text, images, audio and video have brought manifold possibilities for economic development, most notably in healthcare, agriculture, education and finance.[1] However, the same technology is also helping to shape the rapidly evolving world of disinformation perpetrated by local actors, nation states and their proxies.

AI-powered disinformation was named by the World Economic Forum's Global Risk Report as Global Risk No. 1 in 2024.[2] The risks are heightened during times of elections or national crises, as genAI can be used to control narratives, sow confusion, and undermine democratic norms. In environments of democratic fragility, where access to traditional sources of information may be limited, the risks of AI-driven disinformation are especially high. Many of the 55 countries that make up the African Union are still undergoing democratic consolidation, with variations in the degree of media freedom enjoyed in each state.[3]

Moreover, AI-driven disinformation seems to be developing differently in Europe and in Africa, influenced by factors such as economic development, technological infrastructure, wealth, education, internet access and the cost of data, as well as cultural norms and power dynamics relating to the information ecosystem. Further influences include differences in expectations about the role played by independent media in holding power to account.

---

1 AI and the Future of Work in Africa: How AI is Redefining Opportunities, AUDA-NEPAD. 3 July 2024 (https://www.nepad.org/blog/ai-and-future-of-work-africa-how-ai-redefining-opportunities#:~:text=Generative%20AI%20can%20facilitate%20the,%2C%20disaster%20management%2C%20and%20education).

2 WEF (ed): Global Risks Report 2024 (https://www.weforum.org/publications/global-risks-report-2024/).

3 https://freedomhouse.org/countries/freedom-world/scores.

The year 2024, widely dubbed a "super election" year, with polls in more than 60 countries, has shown that AI content (including disinformation) has been a feature of most election campaigns worldwide. Yet, studies suggest the impact of AI disinformation on global elections in 2024 was different from what was expected[4] while its impact in some world regions (such as India, Indonesia, Pakistan, Argentina and many African countries) was underestimated.

On the other hand, transnational organised criminal networks in Europe and Africa have improved their capabilities through AI use, for example, by using deepfakes for online fraud, phishing, ransomware attacks, and scam calls.[5] Despite widespread fear and confusion, empirical knowledge of AI disinformation, its forms, impact and effects remains scarce, which in turn contributes to uncertainty, fear, mistrust and the demand for context, in-depth analysis and actionable recommendations tailored to stakeholders.

Understanding these developments as early as possible is key to successful countermeasures and risk mitigation, as well as providing valuable insights, tools, and strategies for public-and private-sector actors, including policy advisers and lawmakers, journalists, media professionals, researchers, educators, technology companies, and the public. Transnational knowledge transfer and the sharing of best practices can raise awareness of the dangers of AI disinformation in preparation for the risks of the AI age.

4 C.f. for example: Sayash Kapoor & Arvind Narayanan: Deep Dive: We Looked at 78 Election Deepfakes. Political Misinformation is not an AI Problem. Technology isn't the problem—or the solution, 13 December 2024 (https://knightcolumbia.org/blog/we-looked-at-78-election-deepfakes-political-misinformation-is-not-an-ai-problem); c.f.: also: Vittoria Elliott: The Year of the AI Election Wasn't Quite What Everyone Expected, in: Wired, 26.12.2024, (https://www.wired.com/story/the-year-of-the-ai-election-wasnt-quite-what-everyone-expected/?utm_source=nl&utm_brand=wired&utm_mailing=WIR_Daily_122624&utm_campaign=aud-dev&utm_medium=email&utm_content=WIR_Daily_122624&bxid=655724bc01e1264fd50c157b&cndid=75779652&hasha=9f5be9dd6dc3c78e61ce391cf563fcbd&hashc=062242e83793a2b2248947a19c1d7bac17fa440bc3acefaf9fb7c501d5d7bf56&esrc=MARTECH_ORDERFORM&utm_term=WIR_Daily_Active).

5 C.f., e.g.: Interpol (ed): Beyond Illusions. Unmasking the threat of synthetic media for law enforcement, June 2024 (https://www.interpol.int/content/download/21179/file/BEYOND%20ILLUSIONS_Report_2024.pdf).

While there are already several analyses and studies of AI disinformation in 2024 (particularly relating to elections),[6] there are few transnational and transcontinental comparative studies of AI disinformation. The aim of this study is to fill this gap by providing a comparative analysis of AI disinformation in Europe and Africa.

This study analyses the risks, forms and impact of AI disinformation, as well as providing an extensive overview of more than 100 cases in Europe and Africa. As a prominent actor of influence operations over many decades,[7] Russia's use of genAI in foreign interference operations is analysed in detail in this study. The research also provides an evaluation of existing laws and norms relating to AI (and AI-driven disinformation), as well as examining the terms and community guidelines of online platforms and messenger services with regard to AI content used in disinformation campaigns. Finally, countermeasures and strategies to minimise AI-disinformation risk are considered.

The study analyses existing empirical data from both continents to develop categorisations, compare common characteristics and differences, and identify and classify factors influencing the development of AI disinformation.[8] These results form the backdrop to a set of practical recommendations for various actors for mitigating AI-supported disinformation. The study places the emphasis on developing practice-oriented knowledge that can be directly applied.

These practice-oriented measures include a practical experiment with popular AI applications to test their limits and safeguards. In particular, the purpose

6 C.f. for example: Sayash Kapoor & Arvind Narayanan: Deep Dive: We Looked at 78 Election Deepfakes. Political Misinformation Is Not an AI Problem. Technology isn't the problem—or the solution, 13 December 2024 (https://knightcolumbia.org/blog/we-looked-at-78-election-deepfakes-political-misinformation-is-not-an-ai-problem); Ferdinand Gehringer / Christopher Nehring / Mateusz Łabuz: The Influence of Deep Fakes on Elections. Legitimate Concern or Mere Alarmism?, in: KAS Monitor, June 2024 (https://www.kas.de/documents/d/guest/the-influence-of-deep-fakes-on-elections); Łabuz, M., Nehring, C. On the way to deep fake democracy? Deep fakes in election campaigns in 2023. Eur Polit Sci 23, 454–473 (2024) (https://doi.org/10.1057/s41304-024-00482-9); Katja Muñoz: Disrupting Democracies? Myths on AI in Politics Debunked, German Council on Foreign Relations, November 2024 (https://dgap.org/en/research/publications/disrupting-democracies-myths-ai-politics-debunked).

7 Undermining Ukraine: How Russia widened its global information war in 2023, Digital Forensic Research Lab. Feb 2024 (https://www.atlanticcouncil.org/in-depth-research-reports/report/undermining-ukraine-how-russia-widened-its-global-information-war-in-2023/).

8 Following its design as a meta study, this project did not and could not collect new data (that is social media and other analyses) and did thus not follow an investigative approach (that is finding new use cases and previously undiscovered instances of AI disinformation by analysing large amounts of data).

of this exercise is to assess whether a range of popular AI applications (almost exclusively developed outside of Africa) create Africa-specific political disinformation compared to other cultural contexts. This is achieved through the testing of filters, content moderation, and built-in safety mechanisms.

Data and Sources

This project is based on existing empirical data and sources, such as:

- Case studies as found in media and other databases
- Press and media articles and coverage
- Research and official reports
- Surveys and studies
- Legislation, standards, and norms
- Expert interviews

Additionally, this project included a practical experiment with disinformation experts in Africa. During this experiment, volunteer participants tested popular AI applications for the automated creation of AI disinformation on African topics (such as local elections). The goal of this experiment was to test whether AI applications deliver different or similar results when applied in various cultural contexts (see the experiment description and results below).

Guiding Questions

- To what extent is AI disinformation observable in Europe and Africa?
- What forms of AI disinformation are currently observable in Europe and Africa?
- What purposes does AI serve in disinformation (for example, content creation, translation, dissemination)?
- How can actors that generate or amplify AI disinformation be classified and identified?
- What common characteristics and differences exist in AI disinformation between Europe and Africa?
- What factors influence the use of AI for disinformation in Europe and Africa?

- What solutions and countermeasures exist against AI disinformation (including best practices)?
- In low-resource settings in Africa, what are the limits to mirroring best practices in Europe?

Objectives

This study sets out to reach multiple objectives:

- To raise awareness of challenges, threats, and future trends, but also risk mitigation and countermeasures of AI-powered disinformation. The target audiences for this include political decision-makers, journalists and media professionals, influencers, expert communities, and the public.
- Set and shape the future agenda of AI use and AI regulation.
- Facilitate transnational and cross-regional multi-stakeholder knowledge and best-practice exchange between Africa and Europe.

Methodology and Approach

This study is designed as an overview and meta-study that analyses existing materials and sources to extract categories, typologies, and recommendations. The project's approach and its activities are characterised by:

- Meta-study design
- Transnational comparison with the goal of cross-continental knowledge transfer and sharing of best practices
- Easily accessible and understandable project design that facilitates dissemination and high impact
- Tech-positive attitude, focusing on potential and AI risk mitigation

The Detection Challenge of AI Content and AI Disinformation

Every AI disinformation analysis is confronted with the "AI detection challenge",[9] that is, the fact that it is impossible to find all AI content and correctly assess the quantity of AI disinformation. Various factors contribute to this detection challenge. First (1), to date, there is no 100% accurate detection method for AI-generated content. Humans are worse than detection software in correctly distinguishing between AI-generated content and human-generated content.[10] AI-detection software, on the other hand, is also not 100% accurate and regularly produces false positives (detecting human-generated content as AI content) and false negatives (failing to detect AI content).[11] Due to the "GAN conundrum" (that is, to overcome the obstacle of a built-in image generator wrestling with a built-in detector system to produce high-quality outcomes in detection software), these problems might never be solved completely.[12]

A second part (2) of the detection challenge is that AI-generated content may only be detected and labelled as such when it reaches a certain level of visibility, virality and popularity. If AI content remains in closed online communities and does not spill over to other platforms and information spaces, it may go

9 C.f. a good overview: https://originality.ai/blog/ai-detection-studies-round-up; Sadasivan V., Kumar A., Balasubramanian S., Wang W., Feizi S. (2023): Can AI-Generated Text be Reliably Detected? (arXiv:2303.11156).
Dolhansky B. et al. (2020): The DeepFake Detection Challenge (DFDC) Dataset, in: arXiv:2006.07397 (doi: 10.48550/arXiv.2006.07397); Flitcroft MA et al.: Performance of Artificial Intelligence Content Detectors Using Human and Artificial Intelligence-Generated Scientific Writing, in: *Ann Surg Oncol*, Oct 2024; 31(10):6387-6393 (doi: 10.1245/s10434-024-15549-6).

10 C.f: Bellini, V. et al. (2024): Between human and AI: Assessing the reliability of AI text detection tools. Current Medical Research and Opinion, 40(3), 353–358 (https://doi.org/10.1080/03007995.2024.2310086); Originality AI (ed): Can Humans Detect AI-Generated Text? 6 Studies would Suggest they Can't (https://originality.ai/blog/can-humans-detect-ai-content).

11 C.f.: Dolhansky B. et al. (2020): The DeepFake Detection Challenge (DFDC) Dataset, in: arXiv:2006.07397 (doi: 10.48550/arXiv.2006.07397); and: Le B. et al. (2023). Why Do Facial Deepfake Detectors Fail? WDC '23: Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes. pp. 24-28 (doi: 10.1145/3595353.3595882).

12 C.f.: Tasneem, Sumaiya et al. (2023): Generative Adversarial Networks (GAN) for Cyber Security: Challenges and Opportunities. (https://www.researchgate.net/publication/366962736_Generative_Adversarial_Networks_GAN_for_Cyber_Security_Challenges_and_Opportunities); Sampath Lonka: The Power and Challenges of Generative Adversarial Networks, 2023 (https://www.azoai.com/article/The-Power-and-Challenges-of-Generative-Adversarial-Networks.aspx); Elizabeth Seigle: Unravelling the Quandary: The Problem with AI-Generated Content Detectors, 2023 (https://www.idigitalstrategies.com/blog/problem-with-ai-generated-content-detectors/).

unnoticed.[13] Most of the examples and case studies analysed in this study were found by experts, researchers and journalists after they became visible.

Access to data (3) is the third important factor of the detection challenge. To find AI disinformation, researchers and journalists need access to large quantities of social-media and online data. This proved to be a serious challenge to this study. While in the European Union the Digital Services Act, enacted by the European Commission in 2022, orders social-media platforms to guarantee access to their data for research and analysis (see below), researchers in Africa are not granted similar rights. Therefore, even an ex-post analysis of AI content, including disinformation, remains incomplete.

We cannot know how much AI content we encounter on an everyday basis. This means, while we can observe how the quality and the quantity of AI-generated disinformation all around the world is rapidly increasing, its true quantity remains difficult to assess. This also means we almost certainly are exposed to far more AI-generated content (including disinformation) than we know.

In this study, we tackle this detection challenge as openly as possible, by not claiming to provide a complete, quantitative assessment of AI disinformation and pointing out problems with detection whenever we encountered them. Furthermore, as described in the design of this study, we strive to draw important and insightful conclusions from a limited number of case studies and provide actionable analyses rather than a complete assessment of the quantity of AI disinformation in Europe and Africa. To this end, we did not need to find all case studies of AI disinformation in either Europe or Africa, but a representative choice of cases.

13 Since online platforms do not engage in any kind of ex ante AI detection (which would also necessarily have to rely on inherently flawed detection software), there is also little data to work with from their side.

Generative Artificial Intelligence and its Effects on Disinformation: An Overview

The rapid spread of generative AI tools, such as large language models (LLM), image generators and deepfake technology, has grave implications and effects and is currently reshaping the world of disinformation. GenAI is comparable to the invention of the internet or the printing press. Ever since the release of ChatGPT and genAI hype, researchers, experts, journalists and media experts alike have been engaged in assessing its effects on the intentional and organised spread of false and manipulative information, that is, professional disinformation. Here, many observers attribute a "super charging potential" to genAI, which is said to play the role of a booster of disinformation.[14] Even before the global "ChatGPT-moment" in 2022, some observers, worrying about the eroding effect of deepfakes on truth, spoke of a coming "information apocalypse" [15] and of AI as a "weapon of mass disruption".[16]

In summary, several qualities and features of genAI have been identified that make up its supercharging potential for disinformation. AI can make disinformation:

1. Faster (both in generating content and automatically distributing such content).
2. Cheaper (for example, by automating production and distribution, reducing human and financial resources needed).

14 C.f., e.g.: Gold, A. & Fischer, S. (2023, February 21). Chatbots trigger next misinformation nightmare, in: Axios 2023 (https://www.axios.com/2023/02/21/chatbots-misinformation-nightmare-chatgpt-ai); Bremmer I., Kupchan C. (2023). Top Risks 2023. Eurasia Group: New York.

15 C.f.: Nina Schick: Deep Fakes and the Infocalypse, Ottawa, 2020.

16 See: https://gnet-research.org/2023/02/17/weapons-of-mass-disruption-artificial-intelligence-and-the-production-of-extremist-propaganda/.

3. More persuasive (for example, by using super-real deepfakes).
4. More customised (by using, among others, AI software for data analysis and identifying more effective messages for target audiences).
5. More far-reaching (by using AI bots and automation for the distribution of disinformation or because now the average social-media user can use genAI tools).[17]

Furthermore, genAI also has more complex features that have a significant impact on disinformation:

6. AI can provide fake evidence for any claim and thus reverse the burden of proof.
7. The mere existence of genAI technology (for example, deepfakes) creates ample opportunity for discrediting the authenticity and truthfulness of any piece of information as "deepfake" or "AI manipulation" (the so-called "liar's dividend" and "deepfake defence").[18]
8. AI can provide realistic and credible cover for "fake personas", bots, avatars, and digital presenters, among others, who spread disinformation (for example, by creating credible background stories and human-like bodies, faces and profile images).
9. A "fake supercharging effect", that is, a psychological conviction of disinformation creators who attribute special powers to AI-generated disinformation.

Thus, genAI, at least in theory, not only has the potential to increase the quantity and quality of disinformation, but also the ability to distribute and spread disinformation more effectively. Experiments conducted by hackers and journalists, for example, have shown that the costs of using a large language model such as ChatGPT to power a fully automated fake news website (including textual and visual content) dropped from 400 US $ in 2023

17 See for a good overview: Labus and Nehring, "Information apocalypse or overblown-fears what AI Mis and disinformation is all about ? Shifting away from technology toward human reactions" June 2024, Politics and Policy https://www.researchgate.net/publication/381225275_Information_apocalypse_or_overblown_fears-what_AI_mis-and_disinformation_is_all_about_Shifting_away_from_technology_toward_human_reactions; also: Simon, F. M., Altay, S., & Mercier, H. (2023). Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. Harvard Kennedy School (HKS). Misinformation Review. https://doi.org/10.370v16/mr-2020-127.

18 C.f. Delfino, R. (2023). The Deepfake Defense—Exploring the Limits of the Law and Ethical Norms in Protecting Legal Proceedings from Lying Lawyers, in: SSRN Electronic Journal (doi: 10.2139/ssrn.4355140).

to 105 US $ in 2024.[19] This also highlights how genAI tools have the potential to provide multimedia content solutions (that is, producing text, images, videos and audio) for disinformation actors.

The capacity of genAI to create disinformation, the detection challenge, as well as widespread insecurity and mistrust of this technology, is what formed the 2024 "AI disinformation-election-scare". Yet, as this study confirms, the potential of genAI disinformation has not yet materialised in its entirety and no actor or disinformation campaign has, so far, made use of and combined all the advantages of genAI for disinformation.[20] While the actual usage of genAI in organised disinformation is described in detail below, it is important to notice here that the most important impact of genAI disinformation so far is probably its psychological effects and consequences. The mere existence of genAI and its use for disinformation have already led to widespread fear, mistrust and confusion. GenAI and AI-powered disinformation have so far not destroyed truth, but have added new power to the age-old question of "what is truth, authenticity and facts?". GenAI heralds a new era of fakes that eventually might be indistinguishable; and, as some cases have already shown, the mere existence of this technology is enough to have an undermining effect on the veracity and integrity of information.

19 See: Jack Brewster: How I Built an AI-Powered, Self-Running Propaganda Machine for $105, in: The Wall Street Journal, 12.4.2024 (https://www.wsj.com/politics/how-i-built-an-ai-powered-self-running-propaganda-machine-for-105-e9888705).

20 See: Mateusz Łabuz / Christopher Nehring: Information apocalypse or overblown fears—what AI mis- and disinformation is all about? Shifting away from technology toward human reactions, in: Politics & Policy, 06 June 2024 (https://doi.org/10.1111/polp.12617).

Forms of AI Disinformation: A Categorisation

GenAI tools are named after their ability to create (generate) content, every kind of media content, that is, text, images, videos and audio. Thus, forms of AI disinformation include:[21]

A) Fake news websites: Although difficult to detect, several thousand such sites whose contents (text, images, videos) are entirely created by ChatGPT or other chatbots have already been identified.

B) AI Images: AI-created images have started to flood social-media platforms, messenger services and web portals. Such images may be used for different purposes: 1) AI images to go along with social-media posts (visualisation/ visual support for content or supposedly depicting a real event or person); 2) AI images published alongside articles on disinformation, low-quality click-baiting or other "news" websites; 3) AI images that are inserted in video clips and productions; 4) AI images of real-looking persons as profile images of fake social-media accounts. Some show either persons (most often politicians, celebrities, journalists and influencers) in situations that never happen (for example, Taylor Swift and her fans advertising for Donald Trump and his election campaign) or depict events that never happened (for example, AI-generated ruins, bombings and child victims during the war in Gaza). While professional fake news web sites and outlets most likely backed by state actors also use AI images to publish along with fake articles, the vast majority of such AI-generated images, however, are created and spread by "normal" social media and forum users. This trend demonstrates the effects of the democratisation of genAI tools and their misuse.

21 For examples mentioned here compare below Table 1 Overview of Case Studies of AI Disinformation in Europe and Africa.

C) Deepfakes: So-called "deepfakes" (derived from "deep learning" and "fakes") are AI-produced or manipulated video and audio content.[22] There are various different types of deepfakes, differing in either their application (for example, face swapping for so-called "deep porn" or fraud and scamming calls) or their intention (for example, fraud, political campaigning, discrediting and image attacks). Deepfakes that are produced for the purpose of political disinformation have appeared in many contexts: the Russian war against Ukraine and particularly during election campaigns all around the world. Here, they are often used to produce fake discrediting evidence for scandalous statements or positions, participation in illegal or otherwise discrediting events or pornography. Their victims are most often publicly exposed persons, among whom are celebrities, politicians, CEOs, influencers and journalists. Deepfakes have various qualities that have led to a high level of public fear and confusion:

   a) The impressive quality of such fakes

   b) Their ability to convince and persuade audiences

   c) The lack of reliable detection software and methods

   d) The insecurity and inability of audiences to recognise and deal with deepfakes.

D) AI-generated, automated social media comments and posts (including translations): Disinformation actors all around the globe have used popular AI applications (mostly LLMs and most often ChatGPT) for simple tasks. These include the automated generation of many simple and short social-media posts (for example, supporting a government or policy or accusations against a government) with minor variations. In other instances, actors have used AI applications such as ChatGPT to generate a large number of short comments to be posted by fake profiles. In both cases, either for comments or posts, disinformation actors have also used LLMs to generate content or translate content in several languages.

E) AI-powered, fully automated fake social media profiles ("bots" and inauthentic accounts): LLMs such as ChatGPT can be used to run a fully automated social-media profile (for example, on X or Telegram), including posts, comments, likes, shares and images. LLMs that can also produce images and visuals are perfectly suited for this. In some cases, researchers and investigators have found evidence that malicious actors use LLMs to

22 See: Article 3 Definitions (60 "deep fakes") of the EU AI Act: https://artificialintelligenceact.eu/article/3/: "'deep fake' means AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful."

fully automate social-media accounts (that is, to generate and send posts, articles and comments, to interact and react to postings by other users, including images, to share and like and cross-reference their activities with other accounts). Here, it is difficult to distinguish between cases that fall into category D, that is, accounts that are run by humans ("trolls") who may copy and paste AI content, and category E, that is, accounts that are fully automated and entirely run by AI applications.

F) "Deepfake Defence" and "Liar's Dividend": These technical terms describe a form of manipulation related to genAI that does not use the technology to generate or alter media content. Instead, the terms describe the use of a rhetorical (and often legal or political) argumentation by politicians, lawyers, businessmen and others to falsely dismiss and discredit a piece of information or evidence because it was allegedly generated or manipulated with genAI. Here, the real manipulation takes place in the minds of audiences by alluding to the technological potential of genAI without the technology actually being used. Elections, political debates, but also legal trials all around the world, have seen the use of such arguments.

In the last two years, researchers, journalists and state investigators have found empirical proof for all these forms of AI disinformation in Europe, Africa and other world regions. In the following part, we will give an overview of use cases, as well as their analytical assessment.

AI Disinformation in Europe and Africa. Use Cases

Overview

The following table summarises over 70 European and African use cases of AI disinformation (many of which consist of several deepfakes, images or AI-websites). This overview excludes cases of AI disinformation attributed to foreign information manipulation and interference by Russia, to which a special section is devoted below.

	Form	Content	Topic	Use of AI	Date	Target
1.	Images	AI-generated images about fake events at farmers' protests in France; shared in France and Germany [23]	Farmers' protests / EU / War in Ukraine	AI-generated visual content	2024	France / Germany
2.	Image	AI-generated image about the threat of migrants, shared by far-right member of Parliament in Germany [24]	Migration	AI-generated visual content	2022	Germany
3.	Images	AI-generated images about climate change / pollution spread by regional section of far-right party [25]	Migration / Climate	AI-generated visual content	2024	Germany
4.	Audio	Audio deepfake of most popular TV news playing at far-right demonstrations claiming to apologise for "all the disinformation we spread" [26]	Media, covid and politics in Germany / War in Ukraine	Deepfake audio	2023	Germany
5.	Website / Text	News website "Aussiedlerbote", obviously powered entirely by LLM [27]	News / Germany / Russia	AI-powered website	2024	Germany

23 C.f. https://www.volksverpetzer.de/faktencheck/afd-ki-fake-bauernproteste-paris/.

24 See https://www.zdf.de/politik/inside-politix/kuenstliche-intelligenz-ki-afd-soziale-netzwerke-100.html.

25 See Ibid.

26 C.f. https://www.tagesschau.de/faktenfinder/tagesschau-audio-fakes-100.html.

27 C.f. https://www.tagesschau.de/faktenfinder/aussiedlerbote-100.html.

	Form	Content	Topic	Use of AI	Date	Target
6.	Deepfake Defence / Liar's Dividend	Members of the far-right claiming images of protests against them were AI-generated [28]	Migration / Far-right vs. liberal	Deepfake defence	2022	Germany
7.	Video	Deepfake video of leading news anchor advertising crude financial scheme [29]	Finance scam	Deepfake video	2023	Germany
8.	Video	Deepfake video of German Chancellor allegedly promoting ban of far-right party (produced and shared by activists) [30]	Elections / Migration / Far-right Politics	Deepfake video	2023	Germany
9.	Video	Deepfake video of opposition leader allegedly expressing contempt for democracy and voters [31]	Elections	Deepfake video	2024	Germany
10.	Video	Discrediting deepfake video of Keir Starmer, leader of the Labour Party, swearing at his staff [32]	UK elections	Deepfake video	2023	UK

28 See https://www.tagesschau.de/faktenfinder/demonstrationen-rechtsextremismus-bilder-100.html.

29 See: https://www.faz.net/aktuell/feuilleton/medien/deepfake-video-von-zdf-moderator-christian-sievers-19193736.html.

30 C.f. https://www.lto.de/recht/nachrichten/n/lg-berlin-ii-15o579-23-olaf-scholz-bundeskanzler-deep-fake-afd-verbot-zentrum-politische-schoenheit.

31 See https://www.br.de/nachrichten/netzwelt/deepfake-von-merz-was-bedeutet-er-fuer-den-bundestagswahlkampf,UUCtkLv.

32 See https://www.wired.com/story/deepfake-audio-keir-starmer/.

Form	Content	Topic	Use of AI	Date	Target
11. Audio	Deepfake audio of London Mayor Sadiq Khan allegedly cancelling Veterans' Day Parade for pro-Palestine demonstrations [33]	Migration / War in Gaza / Far-right Activism	Deepfake audio	2023	UK
12. AI Images	AI-generated images for campaigning of far-right parties in France and Italy [34]	EU / LGTBQ-issues / Migration / War in Ukraine	AI-generated visual content	2024	France / Italy
13. Music	AI-generated music and lyrics for song promoting the French far-right [35]	European elections	AI-generated music	2024	France
14. Deepfake Defence / Liar's Dividend	Pro-Palestine users falsely claiming image of child victim released by Israel government was "AI-generated" [36]	War in Gaza	Deepfake defence	2023	Israel / Gaza

33 See https://www.bbc.com/news/uk-68146053.

34 See https://dfrlab.org/2024/06/11/far-right-parties-employed-generative-ai-ahead-of-european-parliament-elections/.

35 Ibid.

36 See https://www.dw.com/en/fact-check-ai-fakes-in-israels-war-against-hamas/a-67367744.

Form	Content	Topic	Use of AI	Date	Target
15. Video	Deepfake video of Bulgarian PM allegedly promoting crude financial scheme (supposedly Russian origin, but unclear intentions) [37]	Fraud / Bulgarian elections	Deepfake video	2023	Bulgaria
16. Video	Deepfake video of popular news show involving famous female news anchor showing fight between guests that never happened [38]	Health care / Media in Bulgaria	Deepfake video	2023	Bulgaria
17. Video	Deep porn video of famous journalist [39]	Deep porn / image attack	Deepfake video	2024	Bulgaria
18. Video	Deepfake videos of numerous celebrities advertising various products [40]	Fake advertising	Deepfake video	Since 2023	Bulgaria

37 See https://bnt.bg/news/deep-fake-video-using-the-face-and-voice-of-the-prime-minister-spreads-on-social-media-321680news.html.

38 See https://bntnews.bg/news/otkradnata-samolichnost-s-izkustven-intelekt-falshivo-video-s-adelina-radeva-obikalya-socialnite-mrezhi-1249751news.html.

39 See https://www.womeninjournalism.org/alerts/bulgaria-wpf-condemns-smear-campaign-against-emmy-nominated-journalist-marieta-nikolaeva.

40 See https://factcheck.bg/?s=%D0%B8%D0%B7%D0%BA%D1%83%D1%81%D1%82%D0%B2%D0%B5%D0%BD+%D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%B5%D0%BA%D1%82.

Form	Content	Topic	Use of AI	Date	Target
19. Video	Discrediting deepfake videos of opposition politicians created and shared by government-leaning TV channel [41]	Elections / Domestic politics in Serbia	Deepfake videos	2023	Serbia
20. Video	AI-generated video statement of Serbian PM about "non-existent government projects" [42]	Domestic Politics / Elections	Deepfake video	2024	Serbia
21. Video	Deep porn video of famous Albanian journalist and influencer [43]	Media / Domestic politics in Albania	Deepfake video	2024	Albania
22. Images	AI-generated images visualising and promoting messages of right-wing populist candidate [44]	Romanian Elections / EU / Russia / liberalism	AI images	2024	Romania
23. Video	Deep porn attack against Italian PM [45]	Image attack / Cyberbullying	Deepfake video	Ca. 2021	Italy / US / Global public

41 See https://therecursive.com/we-need-to-rethink-ai-to-save-our-democracy/.

42 See https://balkaninsight.com/2024/07/04/serbia-reacts-fast-over-ai-deepfake-video-of-pm-unlike-other-cases/.

43 See https://balkaninsight.com/2024/01/02/video-deepfake-in-albania-underscores-ai-information-risks/.

44 See https://www.euronews.com/next/2024/12/03/meta-didnt-notice-major-disinformation-in-romanian-election-says-nick-clegg.

45 See https://www.politico.eu/article/italian-pm-giorgia-meloni-called-to-testify-in-deepfake-porn-case/.

	Form	Content	Topic	Use of AI	Date	Target
24.	Video	Deep porn attack against German Foreign Minister [46]	Image attack / Cyberbullying	Deepfake video	2021	Germany / US / Global public
25.	Images	AI-generated images about bombings, ruins, victims and fake events in Gaza created and shared globally by "normal users" as well as pro-Hamas accounts [47]	War in Gaza	AI Images	Since 2023	Israel / Gaza / Global public
26.	Images	AI-generated images about bombings, ruins, victims and soldiers to create and boost moral support for Ukraine, created and shared globally by "normal users" [48]	War in Ukraine	AI images	Since 2022	Ukraine
27.	Websites	Fake news websites about war in Gaza powered by ChatGPT as part of information warfare by Israel [49]	War in Gaza	Text / Articles / Images	Since 2023	USA / Global public

46 C.f. https://www.spiegel.de/politik/deutschland/annalena-baerbock-im-visier-rechter-desinformationskampagnen-a-356da8e0-0002-0001-0000-000178494495.

47 E.g.: https://www.theguardian.com/world/article/2024/may/30/all-eyes-on-rafah-how-ai-generated-image-spread-across-social-media.

48 See https://rubryka.com/en/article/zgenerovani-shtuchnym-intelektom/.

49 C.f. https://www.nytimes.com/2024/06/05/technology/israel-campaign-gaza-social-media.html.

Form	Content	Topic	Use of AI	Date	Target
28. Video	Deepfake video simulating an airstrike on Paris to promote support for no-fly zone over Ukraine, shared (and probably created) by Ukrainian government [50]	War in Ukraine	Deepfake video	2022	Global public
29. Avatar	Deepfake "Jan Gaspadar" created by supressed opposition to candidate in unfair elections in Belarus [51]	Belarus	Deepfake avatar	2024	Belarus
30. Avatar	Deepfake avatar spokesperson of Ukrainian Foreign Ministry [52]	Ukraine / War in Ukraine	Deepfake avatar	2024	Ukraine / Global public
31. Video	Deepfake announcement by Vladimir Putin about peace negotiations with Ukraine [53]	War in Ukraine	Deepfake video	2022	Russia

50 C.f. https://www.forbes.com/sites/alexandralevine/2022/03/17/ukraines-promotion-of-fake-paris-bombing-video-highlights-risks-of-misinformation/.

51 See https://cepa.org/article/belarus-dissidents-turn-to-ai-deep-fakes/.

52 See https://www.theguardian.com/technology/article/2024/may/03/ukraine-ai-foreign-ministry-spokesperson.

53 See https://www.bbc.com/news/technology-60780142.

Form	Content	Topic	Use of AI	Date	Target
32. Audio	Deepfake radio speech by Vladimir Putin announcing "full mobilisation" after Ukrainian advance in the Kursk region (cyberattack on radio station) [54]	War in Ukraine	Deepfake audio	2024	Russia
33. Audio	Deepfake audio clip on Facebook of alleged phone call by prime candidate for PM elections in Slovakia with journalist discussion how to buy minority votes [55]	Elections in Slovakia	Deepfake audio	2023	Slovakia
34. Video	Deepfake videos smearing both candidates in Argentinean presidential elections (both campaigns) [56]	Elections in Argentina	Deepfake video	2023	Argentina
35. Audio / Video	AI-powered conspiracy videos spread by content farms on TikTok [57]	Conspiracy theories	Text / Deepfake audio /Video	Since 2023	Global public / USA

54 See https://www.nytimes.com/2023/06/05/world/europe/putin-deep-fake-speech-hackers.html.

55 See https://misinforeview.hks.harvard.edu/article/beyond-the-deepfake-hype-ai-democracy-and-the-slovak-case/.

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

57 C.f. https://www.newsguardtech.com/special-reports/tiktok-content-farms-use-ai-voiceovers-to-mass-produce-political-misinformation/.

	Form	Content	Topic	Use of AI	Date	Target
36.	Audio	Deepfake audio of PM reading from leaked emails (email content genuine, deepfake voice added) [58]	Elections in Poland	Deepfake audio	2023	Poland
37.	Video	Deepfake porno video of candidate for presidential elections in Turkiye leading to his withdrawal [59]	Elections in Turkiye	Deepfake video	2023	Turkiye
38.	Videos	Numerous deepfake videos of Georgian president Salome Zourabichvili on ongoing conflict over legitimacy of elections and pro-Russian course of government [60]	Elections in Georgia / Russia / West	Deepfake videos	2024	Georgia
39.	Images / Audio / Video	Waves of AI-generated images, audio and video smearing candidates and spreading false claims about US, war with China and personal lives of candidates during elections in Taiwan [61]	Elections in Taiwan / China	Deepfake audio and Video / Images	2023 and 2024	Taiwan

58 See https://notesfrompoland.com/2023/08/25/opposition-criticised-for-using-ai-generated-deepfake-voice-of-pm-in-polish-election-ad/.

59 C.f. https://www.rfi.fr/en/podcasts/international-report/20240316-deepfake-videos-used-in-local-elections-in-turkey-as-erdogan-battles-for-istanbul.

60 See https://mythdetector.com/en/salome-zourabichvili-circulates/.

61 C.f. https://www.thomsonfoundation.org/latest/ai-and-disinformation-in-taiwan-s-2024-election/.

	Form	Content	Topic	Use of AI	Date	Target
40.	Audio and Videos	Several deepfake audio and video clips of presidential candidate Ferdinand Marcos Jr calling for war with China (audio) and him, Marcos, selling and using illegal substances [62]	Elections in Philippines / China	Deepfake Audio and Videos	2023 / 24	Philippines
41.	Audio and Video	Numerous (almost uncountable) audio and video deepfakes of candidates during Indian parliamentary elections (for example, smearing, fake advertising, financial fraud, but also legitimate campaigning) [63]				
42.	Websites	More than 1131 AI-powered global "news and information websites" in 16 languages detected by Newsguard [64]	Global news	Text / Images / Video	Since 2022	Global public

62 C.f. https://time.com/6971239/philippines-marcos-deepfake-china-foreign-actor/.

63 C.f. for example reports by the Deepfake Analysis Unit, especially set up in 2024 to check online content in real-time: https://www.dau.mcaindia.in/all-reports.

64 See https://www.newsguardtech.com/special-reports/ai-tracking-center/.

	Form	Content	Topic	Use of AI	Date	Target
43.	Deepfake Defence / Liar's Dividend	"Tesla" lawyer using "deepfake defence" in California court about alleged statement by Elon Musk about safety of self-driving cars [65]	Business lawsuit	Deepfake defence	2023	US Court / Jury
44.	Deepfake Defence / Liar's Dividend (?)	Donald Trump saying image of him with E Jean Carroll may be AI-generated" [66]	Lawsuit / US election	Deepfake defence (?)	2024	US Court / Jury / US public
45.	Deepfake Defence / Liar's Dividend	Republicans and social-media users claiming photo of Kamela Harris and Tim Walz at Detroit airport with large crowd was AI-generated [67]	US election	Deepfake defence	2024	US public
46.	Video	Several deepfake videos of Kamela Harris (talking nonsense or false evidence for crimes) [68]	US election	Deepfake video	2024	US public

65 See https://www.theguardian.com/technology/2023/apr/27/elon-musks-statements-could-be-deepfakes-tesla-defence-lawyers-tell-court.

66 See https://www.washingtonpost.com/politics/2024/09/06/now-trump-says-photo-him-with-e-jean-carroll-is-ai/.

67 See https://www.bbc.com/news/articles/cx2lmm2wwlyo.

68 See https://www.nbcnews.com/tech/misinformation/kamala-harris-deepfake-shared-musk-sparks-free-speech-debate-rcna164119; also https://www.bloomberg.com/news/articles/2024-10-23/russia-smeared-kamala-harris-with-deepfake-video-microsoft-says.

	Form	Content	Topic	Use of AI	Date	Target
47.	Images / Video	AI-generated images of Taylor Swift endorsing Donald Trump and his campaign as well as deep porn images and video of Swift as "revenge" for her supporting Kamela Harris [69]	US elections	AI-generated images	2024	US public
48.	Bot / Avatar	Campaigning team "cloning" AI version (ChatGPT-version or "Deanbot") of candidate for outreach and campaigning impersonating real candidate [70]	US elections	Bot / Avatar / Impersonation	2023	US
49.	Images	Images visualising Donald Trump's accusations and image of Kamela Harris as "communist" and "Comrade Kamela" (including AI-generated false evidence of Harris having been a member of the Soviet Communist Party [71]	US elections	AI-generated images	2024	US public

69 See https://www.bbc.com/news/technology-68110476.

70 See https://www.theguardian.com/technology/2024/jan/22/openai-bans-bot-impersonating-us-presidential-candidate-dean-phillips.

71 C.f. https://www.bbc.com/news/articles/cn8jg11ynj7o.

	Form	Content	Topic	Use of AI	Date	Target
50.	Images	AI-generated images by supporters of Donald Trump (visualising Trump "saving" pets, supporting false narratives about immigrants eating pets) [72]	US elections	AI-generated images	2024	US public
51.	Audio / Phone Calls	Thousands of deepfake phone calls using President Joe Bidens voice discouraging voters to vote [73]	US elections	Audio	2023	US voters
52.	Audio	Audio message allegedly showing secret meeting between pro-government militia and political coalition plotting a coup [74]	Sudan Political Contestation & Civil war	Deepfake audio	March 2024	Sudanese elites – but the content creator claimed it was a demonstration of how easy it is to manipulate information.

72 C.f. https://www.fox10phoenix.com/news/ai-memes-trump-harris-flood-social-media-during-debate.

73 See https://www.npr.org/2024/05/23/nx-s1-4977582/fcc-ai-deepfake-robocall-biden-new-hampshire-political-operative.

74 See https://www.beamreports.com/.

Form	Content	Topic	Use of AI	Date	Target
53. Audio	Content seeks to clone the voice of the former Sudanese leader Omar al Bashir and presented as "leaked recordings" [75]	Sudan Political Contestation & civil war	Deepfake audio	2023	Sudanese public – to give impression of partisan support by the elusive leader.
54. Audio	Content claims to be a recording of the head of the armed forces ordering the killing of civilians and amplified extensively among others prominent politicians [76]	Sudan Political Contestation & Civil war	Deepfake audio	March 2024	
55. Videos / Posts	Political campaign video that seeks to depict international support for coup leaders in Burkina Faso following Sep 2022 coup. The videos were spotted on Facebook and were then circulated on X and via WhatsApp groups [77]	Support for leaders of Sept 2022 Coup Burkina Faso	Deepfake video	Ongoing since the coup in September 2022	Domestic and international audience, including France and wider diaspora. Intention also appears to be to seed anti-colonial narratives.

75 See Goodman and Hashim: AI voice cloning tech emerges in Sudan civil war, BBC, 5 October 2023 (https://www.bbc.com/news/world-africa-66987869).

76 See https://x.com/jonnygould/status/1768345232184148139?s=20.

77 See https://africacenter.org/spotlight/understanding-burkina-faso-latest-coup/) and: https://www.youtube.com/watch?v=7EUyEADrp5M.

	Form	Content	Topic	Use of AI	Date	Target
56.	Text amplification	Pro-Kagame political campaigning aimed at drowning out dissent and amplifying pro-Kagame narratives as part of an AI-driven coordinated influence campaign [78]	Rwanda	AI including LLM models and text. To scale the dissemination of narratives favourable to Paul Kagame or attacking journalists and human rights activists and "game the algorithm" to ensure such material ranked high in audiences' social-media feeds. It also enabled rapid re-versioning of material and possibly attempts to evade content-moderation controls.	2024	Domestic audiences and demonstrating to global audiences support for Paul Kagame's administration. Targets were also journalists deemed critical to the Kagame administration.

78 See Wack, Linvill and Warren: Old Despots New Tricks – An AI empowered Pro Kagame/RPF Coordinated Influence Network on X, Clemson University Media Forensics Hub, 20 June 2024 (https://open.clemson.edu/cgi/viewcontent.cgi?article=1004&context=mfh_reports).

Form	Content	Topic	Use of AI	Date	Target
57. Images	Cheapfake/Deepfake images of DRC leader Tshisekedi as complicit in the conflict to counterbalance accusations that Rwanda supports M23 rebels in neighbouring DRC [79]	Rwanda / DRC	AI images	2023 / 2024	Domestic audiences in Rwanda and electorate in DRC
58. Video	Video deepfake of President in Zambia claiming "not to run again" [80]	Election in Zambia	Deepfake video	2023	Local audiences in Zambia
59. Video	Video deepfake used for candidate endorsement. Video appears to show Donald Trump endorsing the MK party in the 2024 election [81]	South Africa	Deepfake video amplified by key accounts, especially within the MK party	March / April 2024	Pro-MK party supporters and wider electorate.
60. Video	Video deepfake used for candidate endorsement. Video appears to show the music artist Eminem endorsing EFF party [82]	South Africa	Deepfake video amplified by key accounts especially within the EFF party.	March / April 2024	Pro-EFF party supporters and wider electorate

79 See DRC battles disinformation as it prepares for elections, African Digital Democracy Observatory, 16 December 2023, https://disinfo.africa/drc-battles-disinformation-during-2023-elections-6f3c7e4b8a42.

80 See https://factcheck.afp.com/doc.afp.com.33Z363J.

81 See https://africacheck.org/fact-checks/meta-programme-fact-checks/no-former-us-president-donald-trump-has-not-backed-south and https://factcheckafrica.net/deepfake-content-video-depicts-donald-trump-speaking-on-tinubu-and-peter-obi/.

82 See https://africacheck.org/fact-checks/meta-programme-fact-checks/will-real-slim-shady-please-stand-eminem-video-endorsing.

Form	Content	Topic	Use of AI	Date	Target
61. Image	AI-generated image of EFF leader Julius Malema crying [83]	South Africa	AI-generated image which appears to be for satirical purposes	May / June 2024	Wider electorate. This AI image appears to be satirical and reflects a sense of *schadenfreude*.
62. Image	Multiple AI-generated images of burning buses in Cape Town displaying political party insignia [84]	South Africa	AI-generated images which appear to for overt political campaigning and potential racial incitement	April / May 2024	Wider electorate and to support narratives of the EFF party

83 See https://murmurintelligence.com.

84 See Ibid.

Form	Content	Topic	Use of AI	Date	Target
63. Image	AI-generated image of white men in military insignia overpowering black South Africans with accompanying text reflecting racial divisions over land ownership [85]		AI-generated image which appears to seek to enforce racial stereotypes and amplify the land expropriation narrative	April / May 2024	Wider electorate but especially black voters in South Africa
64. Video	AI-generated image of an election candidate appearing to buy voter cards from Nigerian internally displaced persons [86]	Nigeria	AI-generated image which appears to undermine an election candidate, suggesting he is corrupt. The material's authenticity was highlighted by fact-checking site Dubawa. [87]	January 2023	Wider electorate in Nigeria

85 See R Davis, Fact check — Is it likely the Western Cape could become an independent state?, *Daily Maverick*, www.dailymaverick.co.za/article/2024-03-28-western-cape-independence-fact-check/, 28 February 2024.

86 C.f. Orakwe. The Challenges of AI-driven political disinformation in Nigeria, *Africa in Fact*, 3 July 2024. https://africainfact.com/the-challenges-of-ai-driven-political-disinformation-in-nigeria/.

87 https://dubawa.org/is-zulum-buying-pvcs-of-borno-idps-as-suggested-in-video/.

	Form	Content	Topic	Use of AI	Date	Target
65.	Image	AI-generated images to place Nigerian candidate Peter Obi on the front page of Time magazine [88]	Nigeria	AI-generated material which has the accompanying text: "All eyes on Nigeria: Time for Africa's sleeping giant to awaken", used for candidate endorsement of Peter Obi, a presidential contender in the 2023 Nigerian election.	Feb 2023	International community and Nigerian electorate

88 C.f. Orakwe. The Challenges of AI-driven political disinformation in Nigeria, *Africa in Fact*, 3 July 2024. https://africainfact.com/the-challenges-of-ai-driven-political-disinformation-in-nigeria/.

	Form	Content	Topic	Use of AI	Date	Target
66.	Video	AI-generated political campaign videos using avatars in support of coup leaders following September 2022 coup [89]	Burkina Faso	AI-generated avatars created using "synthesia" AI software to support the regime of Captain Ibrahim Traore following the September 2022 coup	Jan 2023	International community and Burkinabé citizens
67.	Video / Text / Image	AI-generated French-Ghanaian "investigative journalist avatar" created for intelligence and narrative control largely in Francophone Africa [90]	Francophone Africa	AI "investigative journalist" avatar created by Israeli company Percepto International, with associated imagery, websites and social-media profiles [91]	March 2023	Local audiences in several countries

89 See https://africacenter.org/spotlight/understanding-burkina-faso-latest-coup/) and https://www.youtube.com/watch?v=7EUyEADrp5M.

90 See Amanda Sperber / Justin Arenstein: Robot wars: How to build a bot to subvert elections. How disinfo-for-hire companies use armies of fake avatars to infiltrate & manipulate the media (https://disinfo.africa/robot-wars-how-to-build-a-bot-to-subvert-elections-9f739411aa39).

91 https://disinfo.africa/robot-wars-how-to-build-a-bot-to-subvert-elections-9f739411aa39.

	Form	Content	Topic	Use of AI	Date	Target
68.	Deepfake Defence / Liar's Dividend	PM claiming incriminating leak series on social media was "AI generated" [92]	Elections in Mauritius	Deepfake Defence / Liar's Dividend	November 2024	Local audiences in Mauritius

92 See Mnisi Mauritius ends social media ban ahead of elections, Human Rights Watch, 7 November 2024 (https://www.hrw.org/news/2024/11/07/mauritius-ends-social-media-ban-ahead-elections) and R Kassenally: The Chilling Effect of Censorship, Self Censorship and Surveillance in Mauritius, *Democracy in Africa*, November 2024 (https://democracyinafrica.org/the-chilling-effect-of-censorship-self-censorship-and-surveillance-in-mauritius/).


№ 04094064
Фамилия Харрис
Имя Камала
Отчество Деви
Год рождения 1964
Время вступления в партию 1984
Наименование партийного органа, выдавшего билет
Коминтерновский райком г.
Воронежа, РСФСР
год
Месяц
Месячный заработок
Сумма взноса
Подпись секретаря
Январь
Февраль
Март
Апрель
Май
Июнь 134 1-34
Июль 134 1-34
Август 134 1-34
Сентябрь 134 1-34
Октябрь 134 1-34
Ноябрь 197 2-96

AI Disinformation and Elections in Europe

Nearly every election in Europe in 2023 and 2024 saw AI-generated content used for disinformation and election manipulation. The most notable cases were: [93]

Germany: In the summer of 2024, a deepfake video of an alleged Nigerian "callboy" was spread online who claimed that German Foreign Minister Annalena Baerbock was among his regular customers. In early December, another deepfake video in the same fashion appeared featuring a woman claiming that Minister of Economy Robert Habeck had sexually abused her several years ago. Both deepfakes were attributed to Russian disinformation and interference attempts. Another campaign that has been ongoing since 2022 and focuses on military and other support for Ukraine (a major topic for the elections of 2025) was the "Doppelganger" campaign (see infobox below). GenAI has been used in this campaign to generate posts and comments, fake images and to spread disinformation narratives.

EU Parliamentary Elections: Far-right parties and politicians in Italy, France and Germany created and disseminated AI-generated images, posters and songs during the election campaign. Their topics and content focused on migration, LGTBQ rights and "national traditions". Since this content was not appropriately labelled as "AI-generated" (which was a clear violation of the code of conduct as passed by the European Parliament before the election) and due to the misleading and discriminating nature of this content, it might be argued that these were acts of disinformation rather than legitimate campaigning. Official investigations by the European Commission revealed that AI content did, however, not play an important role in disinformation campaigns during the election. In the weeks before the election, for example, the amount of fact-checked disinformation containing AI-generated content as detected by EDMO was around 4% to 5% of the overall amount of fact-checked disinformation.[94]

France: In France, several instances of foreign and domestic AI disinformation were observed during the past years: AI-generated images of president Macron during farmers' protests, discrediting deepfake videos of Macron, a deepfake video about an alleged assassination plot against Macron during his visit to

93 See for references the links as provided in Table 1: "Overview of Case Studies of AI Disinformation in Europe and Africa".

94 European External Action Service EEAS (ed.): Memo: Known information interference operations during the June 2024 elections for the European Parliament October 2024 (https://ec.europa.eu/commission/presscorner/detail/en/ac_24_5328), p. 10.

Ukraine, deepfake videos attacking Olympia 2024 in Paris and many more. Furthermore, France is (along with Germany) one of the major target countries for the ongoing Russian "Doppelganger" campaign that also uses genAI (see above).

UK: In the UK, deepfakes have made several appearances, targeting leading politicians, such as Keir Starmer, leader of the Labour Party, or Sadiq Khan, Mayor of London. Both were attacked using deepfakes to produce false discrediting statements. The Starmer deepfake featured him swearing at staff and employees and reached 1.5 million views on the social media platform X. In the case of Khan, right-wing activists produced a deepfake audio file that featured Khan cancelling a Veteran's Day parade in favour of pro-Palestine demonstrations, which contributed in London to a large rally (including physical attacks and clashes with the police).

Moldova: Moldovan elections and the referendum about a future Western (EU and NATO) orientation of the country both saw numerous deepfakes as part of ongoing, massive foreign interference by Russia in conjunction with domestic pro-Russian forces. The major target of these deepfakes was President Maia Sandu. Other AI content may have also played a role (particularly AI images and AI-powered fake web portals), yet there was little concerted effort to identify it. Influence operations and interference were particularly high in Moldova during the past two years, due to its proximity to the battlefield in Ukraine and the pioneering elections in 2024. In both elections, pro-Western forces won by tiny margins, meaning that, while AI disinformation and interference (that is, extensive buying of votes during the last week of elections) influenced the outcome of these elections, they did not determine the final result.[95]

Poland: A last-minute audio deepfake of the ruling Prime Minister featured his voice reading out discrediting content of leaked emails from his government. This deepfake enfolded in a grey area since the content of the emails and audio was technically true but hearing it from the PM's real voice may have manipulated audiences and increased its effect.

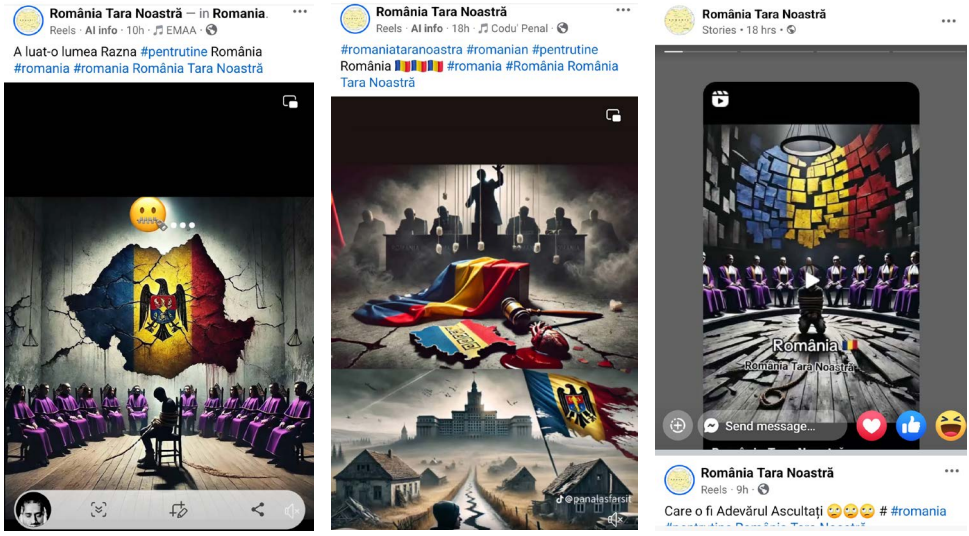
Slovakia: A last-minute audio deepfake of a prominent candidate allegedly discussing how to buy minority votes appeared on Facebook in the last 36 hours before the election. It was widely acknowledged in the country that this audio definitely had an impact on large groups of voters being either mobilised or demobilised; yet, when comparing the final results of the election to the last polls before the deepfake, percentages between the candidates changed, but

95 C.f. https://brodhub.eu/en/republic-of-moldova/.

did not shift altogether (that is, the winning candidate remained the same in the last polls and the final results).

Turkey: A discrediting deep porn attack against a male candidate for presidency led to his withdrawal from the election. This might have influenced the election, as it changed the overall dissemination of votes (yet, the withdrawing candidate had no chance of winning the election and the winning candidate never did face real danger of losing).

By far the most important case of interference and disinformation, however, took place at the end of 2024 during the Presidential elections in ***Romania***: Here, the constitutional court annulled the first round of the elections due to a massive disinformation campaign on TikTok, as well as cyberattacks on election infrastructure. These attempts of interfering were assessed by the Romanian intelligence service and attributed both to foreign actors (that is, Russia) and domestic far-right actors pushing the winning candidate, Calin Georgescu. This example was extraordinarily interesting because of two features: a) the use of AI content played only a marginal role during these disinformation and interference attempts; and b) whereas disinformation was undoubtedly present during this election campaign, neither the constitutional court nor the intelligence service presented hard evidence and proof that the unexpected surprise victory of Georgescu was caused by disinformation alone. Thus, we can say for sure that disinformation heavily influenced this election, but it may have only been because key actors (intelligence services and constitutional court) perceived it that way and turned it into the decisive factor.



Analysis of Tactics

The methods used in AI-driven election disinformation campaigns in Europe during the past two years include a diverse range of tactics:

- Last-minute deepfakes: Videos designed to misrepresent political candidates, such as fabricating statements or actions, released within the last 48 hours before the election. This kind of AI disinformation is the most feared, as there is little time and opportunity to effectively debunk and counter. However, only in the cases of Slovakia and Poland (both in 2023) did such deepfakes have a visible impact.
- Audio or audiovisual deepfakes used as fake evidence: Such deepfakes have been produced and spread during almost every election in Europe. Their aim is to discredit and smear politicians and they regularly target the same persons in a country.
- Fake news websites: AI-created content more or less automatically published on bogus websites.
- Automated social media profiles (bots): Professional actors, for example, Russia, but also PR companies, have used automated social-media profiles to spread online disinformation. In several instances, even Russian campaigns were caught using ChatGPT to automatically generate, like, react to and share.
- AI-generated images: AI images were used during election campaigns in Europe to visualise political narratives and depict fictitious events, scenarios and persons. In no instance did such content have a significant impact on its own (yet their quantity is increasing and their long-term effect not yet assessed).
- Cross-lingual manipulation: AI-powered translation tools were deployed to extend disinformation campaigns across linguistic barriers, ensuring their impact on multinational electorates.

AI Disinformation in Africa

The use of AI-driven disinformation in Africa is not widely understood, due to limited empirical research. Observational studies have tended to focus on the use of disinformation and AI-driven disinformation during elections, with the Nigerian election in 2023 and the South African election in 2024 serving

as important sources of data.[96] There have been other media and think-tank investigations of AI-driven disinformation in the Sahel and also in Rwanda, which have also assisted researchers in understanding the disinformation ecosystem in Africa. While Europe and North America have focused intensively on foreign information manipulation and interference (FIMI), what limited disinformation and AI-driven disinformation exists in Africa has tended to take an actor-agnostic approach. This means that the focus has been on describing the forms, reach and characteristics of disinformation campaigns rather than focusing on who is generating the content. Nevertheless, anecdotally much of the content would appear to be home-grown, with African influencers taking a leading role in shaping narratives, albeit mirroring tactics, techniques and processes (TTPs) observed in other settings.

Across Africa, the phenomenon of AI-generated content, ranging from sophisticated deepfakes to so-called "cheapfakes" (a manipulated or falsified form of audiovisual content that is created using simple and inexpensive technologies) is still relatively immature. While deepfakes require more sophisticated tools, Africa Check[97] describes cheapfakes as "quicker and less resource-intensive [thereby a more attractive tool] … cheapfakes range from videos taken out of context, to simple edits such as speeding up or slowing down video or audio to misrepresent events. Cruder face-swapping and lip-syncing methods also fall into this category." Jean le Roux, an Africa-based disinformation specialist with the social media analytics company Graphika and formerly with the Digital Forensic Research Lab,[98] argues that, "Ultra-cheap fakes are more of a problem for disinformation",[99] as research indicates[100] that viewers may still be prepared to share material knowing that it is inauthentic or "fake", simply due to a lack of attention rather than malign intent.

Therefore, the amount of available data on AI-generated deepfakes in Africa is limited and reliant upon user reports. However, some prominent case studies

96 Allen and Le Roux, "Under the influence? Online mis/disinformation in South Africa's May 2024 election", Institute for Security Studies, 13 December 2024. https://issafrica.org/research/southern-africa-report/under-the-influence-online-mis-disinformation-in-south-africa-s-may-2024-election.

97 https://africacheck.org/fact-checks/blog/ai-powered-disinformation-deepfakes-detection-technology-and-weaponisation-doubt?fbclid=IwY2xjawG74-BleHRuA2FlbQIxMQABHeZg2-hQBa31mYN9lPaxxpZ3p1G-xqlOwmRrSKlbzgCWUWAAmkwiq5o_DA_aem_WsK9knwTVgNNbrlSLk_SDA.

98 https://dfrlab.org/region/africa/.

99 Interview with Jean le Roux by authors, September 2024.

100 Pennycook and Rand – The Psychology of Fake News, *Trends in Cognitive Sciences*, 5 May 2021 (https://www.sciencedirect.com/science/article/pii/S1364661321000516#s0030).

shed light on which forms of AI disinformation have been observed, enabling researchers to deduce the primary purpose for which the technology has been used.

Since January 2023, media reports of AI-generated audio and video content appear to have increased, with particular attention on Burkina Faso, Mali and Sudan, which have undergone considerable political upheaval, and on regions in Africa where elections have recently been held, such as Nigeria, South Africa and Kenya.

An apparent surge in instances of AI use may reflect the increase in the number of accredited fact-checking organisations across Africa that are members of the International Fact Checking Network[101] that are actively seeking out AI-generated content. It may also be a reflection of donor interest.

The observable use cases of AI content in Africa include:

- Political contestation
- Narrative support
- Candidate endorsement during elections
- Criminal activity
- Satire

Political Contestation

Case Study: Sudan

Sudan has emerged as a focus country for AI-generated disinformation campaigns, as it continues to experience the human toll of two years of civil war, which has sparked the "world's worst displacement crisis", according to the Global Conflict tracker.[102] As a power struggle between the leaders of the Sudanese Armed Forces and a powerful paramilitary group known as the Rapid Support Forces (RSF) persists, there have been multiple instances of AI-generated audio and video to support or undermine the warring parties.

The Khartoum-based Beam Reports[103] highlighted a campaign in March 2024 in which an AI-generated audio recording purported to show a secret meeting

101 https://www.poynter.org/ifcn/about-ifcn/.

102 https://www.cfr.org/global-conflict-tracker/conflict/power-struggle-sudan.

103 https://www.beamreports.com/.

between militia leaders of the Rapid Support Forces (RSF) militia and political leaders from the coalition formation known as the Forces of Freedom and Change (FFC) allegedly discussing plans for a military coup.[104] It was widely disseminated on mainstream media before finally being removed from a range of social-media platforms. The originator of the material was interviewed by Beam Reports and claimed his motivation was to show how easy it is to create deepfake material.

AI voice-conversion software has also been used in a separate incident in Sudan, reported a year earlier, in what appears to have been an attempt to clone the voice of the former Sudanese leader Omar al Bashir, which was posted on an anonymous TikTok account. The recordings are presented as "leaked recordings" of the ailing leader who has been charged with war crimes and not been seen in public for some time. A BBC investigation suggests that some of the material originated from old Bashir recordings posted in August 2023, which apparently feature the former leader criticising the leader of the Sudanese army. According to the investigation, "The Bashir recording matched a Facebook Live broadcast aired two days earlier by a popular Sudanese political commentator, known as Al Insirafi. He is believed to live in the United States but has never shown his face on camera. The pair don't sound particularly alike but the scripts are the same, and when you play both clips together they play perfectly in sync." Analysts deduce that the motivation may be to "trick audiences into believing that Bashir has emerged to play a role in the war or the channel could be trying to legitimise a particular political viewpoint by using the former leader's voice." [105]

Audio Deepfakes – An example of what appears to have been an audio deepfake emerged: "In March 2024, an X account for a television and radio presenter shared[106] a recording attributed to the Sudanese Armed Forces head ordering the killing of civilians, deployment of snipers, and occupation of buildings. This AI-created recording was viewed by 230 000 accounts and shared by hundreds, including known Sudanese politicians." The sharing by politicians, who are at times considered trusted messengers because of their presumed superior access to information, is significant, as it may lend credibility to the story. It is not clear what software was used to generate this material but the impact

104 Suliman: The Deepfake is a powerful weapon in the war in Sudan, *African Arguments*, 23 October 2024 (https://africanarguments.org/2024/10/the-deepfake-is-a-powerful-weapon-in-the-war-in-sudan/).

105 Goodman and Hashim: AI Voice cloning tech emerges in Sudan civil war, BBC, 5 October 2023 (https://www.bbc.com/news/world-africa-66987869).

106 https://x.com/jonnygould/status/1768345232184148139?s=20.

and the potential for a real-world response is significant, especially in conflict settings.

Liar's Dividend and Deepfake Defence

Case Study Sudan

The growing prevalence of inauthentic material in Sudan gives rise to a phenomenon known as the "liar's dividend". This involves false claims by elites that material is a deepfake, with the intention either to divert attention from themselves or to evade accountability or to undermine an adversary. Beam reports examined reports by Sudanese Armed Forces supporters that a series of recordings of the leader of the paramilitary Rapid Support Forces (RSF), Mohamed Hamdan Dagalo, known as Hemedti, were AI-generated and they further claimed that he was dead. However, a forensic investigation[107] that was shared online found the recordings to be authentic. A similar tactic has been observed across Europe and the United States but, as Terence Corrigan from the South African Institute of International Affairs observes,[108] fragile states in Africa are especially vulnerable, stating that "malleable communications and a toxic information environment are frightening resources for opportunistic politicians, identarian hustlers, and malign external actors". It is for this reason that attention on genAI has turned to Africa, which some have described as a "wild west for AI, "posing risks to democracy and human rights".[109]

Case Study Mauritius:

In Mauritius, the circulation of leaked material in the lead-up to the November 2024 elections provides another example of the liar's dividend. The circulation of online material which was initially dismissed by the authorities as being the result of AI fakes led to the dramatic shut down by the Mauritian authorities of all social-media platforms.[110] Mauritius has long been considered a beacon of

107 https://docs.google.com/document/d/1fqEMGo12Q5-Z-O1Uvm-6Uf_kQN80LvXYxDuWieSAp0s/edit?tab=t.0.

108 Corrigan T: Cashing in on the Liar's Dividend, *African in Fact*, 3 July 2024. https://africainfact.com/cashing-in-on-the-liars-dividend/.

109 Artificial Intelligence (AI) for African Democracy and Socio-Economic development, AU-NEPAD, Blog 19, September 2024. https://www.nepad.org/blog/artificial-intelligence-ai-african-democracy-and-socio-economic-development.

110 Mnisi Mauritius ends social media ban ahead of elections, Human Rights Watch, 7 November 2024 (https://www.hrw.org/news/2024/11/07/mauritius-ends-social-media-ban-ahead-elections).

democratic stability but there are growing concerns about the threat of digital authoritarianism.[111]

The leaked material featured phone calls "involving senior politicians, police officers and journalists and foreign diplomats" and included what Human Rights Watch described as a leaked conversation in which "a senior police officer allegedly asked a doctor to alter a report about the death of a person in police custody", which has led to a judicial investigation.

However, the reflex to dismiss such allegations as fake and the subsequent suspension of social-media access, is a reminder of how governments can use AI in the information ecosystem to shape democratic norms, such as access to information.

Regime and Narrative Support

Burkina Faso

One of the first AI video deepfakes to emerge on the continent followed the coup in Burkina Faso in September 2022[112] – the second in a year – which saw Captain Ibrahim Traoré installed as head of state. He unseated a previous coup leader, Lieutenant Colonel Paul-Henri Sandaogo Damiba, whom he accused of failing to tackle deteriorating security. A series of AI-generated videos began circulating online according to reports on France 24,[113] urging citizens to support the military junta.

The deepfake videos, which were first spotted on Facebook and went on to be circulated on WhatsApp groups across the Sahelian country and later on Twitter/X, feature individuals who describe themselves as Pan Africanists and Americans from Africa. The material was originally linked to the Wagner group by the NGO "All Eyes on Wagner" but no link to the Russian outfit has ever been made (the outfit is now rebranded Africa Corps, following the death of Yevgeny Prigozhin).

111 R Kassenally: The Chilling Effect of Censorship, Self Censorship and Surveillance in Mauritius, Democracy in Africa, November 2024 (https://democracyinafrica.org/the-chilling-effect-of-censorship-self-censorship-and-surveillance-in-mauritius/).

112 Understanding Burkina Faso's latest coup, Africa Centre for Strategic Studies, 28 October 2022 (https://africacenter.org/spotlight/understanding-burkina-faso-latest-coup/).

113 https://www.youtube.com/watch?v=7EUyEADrp5M.

The material was found to have been produced using the AI tool Synthesia,[114] with pronunciation errors and some apparent visual distortions being tell-tale signs that the "individuals" in the videos were in fact avatars, which match the range of AI-generated personae offered by the company.[115]

While the impact of the Burkina Faso deepfake videos is hard to assess, they starkly illustrate the utility of the technology as a propaganda tool and one which in this case sought to undermine democracy by supporting a power grab.

As an interesting aside, a separate investigation by the social media analytics firm Graphika[116] linked the use of the Synthesia platform to a Chinese influence campaign to create a fake news channel called Wolf News, as part of what it describes as an ongoing "spamouflage campaign".[117] Such campaigns use hundreds of fake accounts to amplify videos "that praise China, criticise (sic) the United States, and attack the Hong-Kong pro-democracy movement" Graphika claimed that this was the first time that it had identified state-aligned IO (information operations) actors using video footage of AI-generated fictitious people in its operations. While the videos it observed on Wolf News received no more than 300 views, it does demonstrate that video-creation platforms such as Synthesia, aimed at creating marketing and training videos, are being utilised by political actors (against the company's own code of ethics).[118]

Rwanda

A study of Rwanda's use of AI tools, in particular large language models and generative artificial intelligence, highlighted the technology's use as a tool to dominate narratives, as part of a pro-government coordinated influence network on X. Although Rwanda has enjoyed significant economic development under President Kagame, Freedom House, in its 2024 report on Rwanda,[119] characterises the Rwanda Patriotic Front (RPF) under Mr Kagame, which has governed since the end of the genocide in 1994, as having presided over a period of stability and economic development while it has also suppressed

114 https://www.synthesia.io.

115 https://www.synthesia.io/?gad_source=1&gbraid=0AAAAA-pHwtabce5V2t_ZBfccPqMnPSgO7&gclid=Cj0KCQiAsaS7BhDPARIsAAX5cSAaN2rnkFR503UBwH5-Ee_fRuUesXCX5AwxadPU6iYu1L1C_f_6JCUaArxIEALw_wcB.

116 Deepfake it until you make it – Graphica, 7 Feb 2023, https://graphika.com/reports/deepfake-it-till-you-make-it.

117 https://graphika.com/reports/spamouflage-breakout.

118 https://www.synthesia.io/ethics.

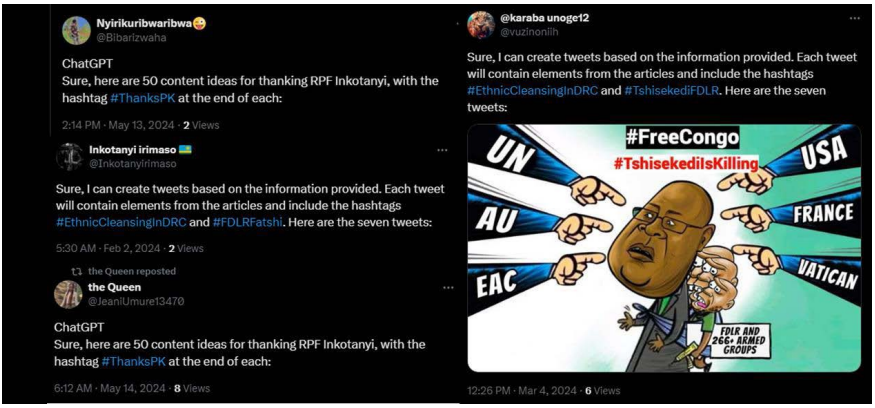
119 https://freedomhouse.org/country/rwanda.

political dissent through pervasive surveillance, intimidation, torture and rendition of or suspected assassinations of exiled dissidents). The use of AI-driven campaigning by supporters of the government appears to be one of several countermeasures aimed at managing dissent.

Researchers from Clemson University Media Forensics Hub found that AI tools were used in a coordinated manner in Rwanda to generate texts at scale which were favourable to the government of Paul Kagame and which were reposted across the campaign's accounts and re-posted on several platforms.[120] Thus, AI was used as a multiplier tool aimed at producing content at scale and speed, in multiple languages and across multiple online platforms. AI tools were used to re-version content swiftly and may have assisted in ensuring that information operatives, to propagate messaging, were able to evade content-moderation measures. The objective of this AI-driven campaign appears to have been to dominate the conversation and drown out unfavourable messaging and target what the researchers describe as "perceived enemies of the state".[121]

The researchers were able to identify the material as AI-generated, due to the prompts inadvertently left behind tasking the AI to generate material. The image below demonstrates that ChatGPT was used to generate specific content ideas that are favourable to the RPF. Multiple versions of the messages all pointed to a series of hashtags indicating coordination.

Evidence of LLM Use by Campaign Accounts



Source: Source: Old Despots New Tricks – An AI-Empowered Pro-Kagame/ RPF Coordinated Influence Network on X

120 Wack, Linvill and Warren: Old Despots New Tricks – An AI empowered Pro Kagame/RPF Coordinated Influence Network on X, Clemson University Media Forensics Hub, 20, June 2024 (https://open.clemson.edu/cgi/viewcontent.cgi?article=1004&context=mfh_reports).

121 Ibid.

The campaign messaging revolves around narratives with a few common themes:

- The ongoing conflict in the Democratic Republic of Congo (DRC)
- Trumpeting government successes
- Highlighting Paul Kagame's leadership.

The researchers identify this AI-driven influence operation as an extension of a public-relations campaign which has revolved around the Visit Rwanda campaign,[122] which promotes tourism, sporting affiliations and the East African nation as an international eventing destination. However, the researchers suggest the real aim of the AI campaign is to mute any criticism of Paul Kagame's government. The researchers note that "one prominent layer of this reputation management machine involves the use of a vast network of pro-government social media accounts, which have been linked to efforts to promote the state as well as campaigns to harass journalists who have attempted to investigate the government's extrajudicial actions", in particular investigations into media suppression in Rwanda by the journalist network Forbidden Stories.[123]

Several Edited and Artificial Images Depict Tshisekedi as Complicit in the Conflict

Source: Old Despots New Tricks – An AI-Empowered Pro-Kagame/ RPF Coordinated Influence Network on X

122 https://www.linkedin.com/pulse/how-visitrwanda-campaign-successfully-promotes-rwandas-migisha/.

123 Rueckert and Pfenniger: in the West and Online, Rwanda's influence machine keeps churning, Forbidden Stories, 28 May 2024, https://forbiddenstories.org/in-the-west-and-online-rwandas-influence-machine-keeps-churning/.

Deepfake video – While image generators were also used to create deep and cheapfakes – such as the image on the previous page of the President of the Democratic Republic of Congo, Felix-Antoine Tshisekedi – the use of AI as a "scaling" tool appears to have been the predominant purpose of the technology in this setting.

Candidate Support During Elections

While research into the use of AI for disinformation campaigns across Africa is limited in comparison to Europe and the United States, there have been a limited number of studies examining the use of such tools during election campaigns.[124] This is largely a reflection of western donor interest in supporting the development and consolidation of democracies in Africa, at a time when democratic backsliding presents a threat.[125] The risks of online information operations also lay the foundations for autocratic regimes and other nation states, such as Russia and China, to exploit the information ecosystem in Africa.

While there is a growing body of evidence of disinformation or narrative-control campaigns in Africa spearheaded by nation-state actors – primarily Russia (China's attempts at narrative control appear to be less focused on social-media platforms and more on media-house acquisition)[126] – in South Africa,[127] Central African Republic[128] and the Sahel,[129] the use of AI for content creation at this stage appears to be limited.

Indeed, more broadly, AI content creation as part of disinformation campaigns during elections in Africa does not appear to be as prevalent as many observers

124 Van Damme, Findlay, Cornelissen: Generative AI and its Influence on South Africa's 2024 Elections, German Council on Foreign Relations, 4 December 2024, https://dgap.org/en/research/publications/generative-ai-and-its-influence-south-africas-2024-elections.

125 https://www.csis.org/analysis/regional-support-address-democratic-backsliding-africa.

126 China's Influence on African Media, Africa Centre for Strategic Studies, 12 May 2023, https://africacenter.org/spotlight/chinas-influence-on-african-media/.

127 Wasserman & Murmur " How Russia uses 'hybrid' warfare to amplify its narratives in the South African discourse", Daily Maverick, 24 November 2024. https://www.dailymaverick.co.za/article/2024-11-22-how-russia-uses-hybrid-warfare-to-amplify-its-narratives-in-the-south-african-discourse/.

128 https://forbiddenstories.org/in-the-central-african-republic-a-former-propagandist-lifts-the-veil-on-the-inner-workings-of-russian-disinformation/.

129 https://medium.com/dfrlab/pro-russian-facebook-assets-in-mali-coordinated-support-for-wagner-group-anti-democracy-protests-2abaac4d87c4.

had predicted. Africa Check questions whether election disinformation in South Africa needs AI.[130]

South Africa

South Africa's 2024 national election was arguably the first by a major democratic state in Africa during the age of AI and provided a useful test case. With some 26 million social-media users, South Africa provided a ready audience for AI-assisted influence operations and information manipulation.

However, what was observed was only limited use of AI-generated content for disinformation purposes. Notwithstanding the paucity of data available, AI content appeared to have been used for three primary objectives beyond simple political contestation:

- Candidate endorsement
- Narrative support
- Satire and ridicule

Media Monitoring Africa (MMA) – a South African registered not-for-profit organisation – established a platform called the Real 411[131] to enable voters to report concerns about online election content, including the use of AI. Based on interviews with MMA, the site registered, based on user experience, seven pieces of AI content out of some 250 complaints, which South African data analytics firm Murmur Intelligence[132] was able to analyse as part of an unpublished study for the Institute for Security Studies examining the role of AI content and disinformation during the elections.

Although at the time of writing the study has not been made public, the research team shared their findings with the authors of this paper. They concluded that, although overall volumes of AI content were small, "a limited number of influence communities were most closely associated with creating or amplifying politically themed AI generated content".

In the table below, which represents an interaction network between communities of interest, the dashed circles and ovals represent areas where

130 Cosser, Expectations versus reality: The use of Generative AI in South Africa's 2024 election, Africa Check, 17 July 2024, https://africacheck.org/fact-checks/blog/expectations-versus-reality-use-generative-ai-south-africas-2024-election.

131 https://www.real411.org/.

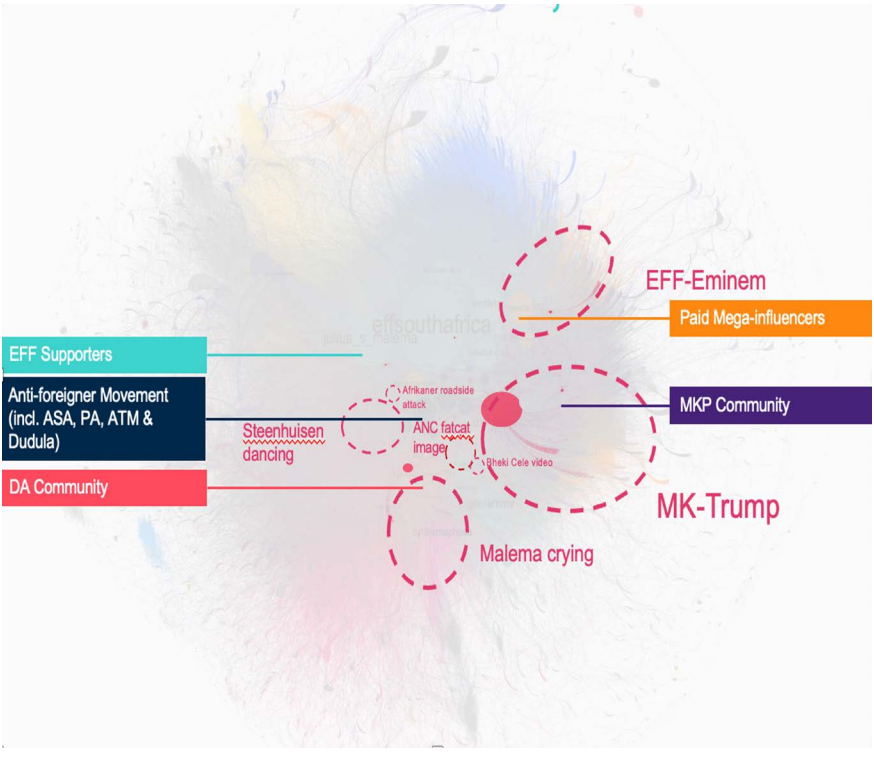
132 https://murmurintelligence.com.

the content was amplified, while the coloured labels represent the different communities.

The communities are created on the basis of volumes of engagement by key groups which coalesced around particular issues, a political ideology or a shared view on a specific issue.

The most active communities on social-media platforms included the newest party in the South African political scene, uMkhonto weSizwe (MK), comprising several senior former figures from the African National Congress (ANC), including South Africa's former President Jacob Zuma. Also contesting the elections was the Economic Freedom Fighters (EFF), a far-left party led by the firebrand politician Julius Malema and members of the Multi Party Charter established by key opposition parties, including the Democratic Alliance.

Interaction Network South African Elections and AI-generated content

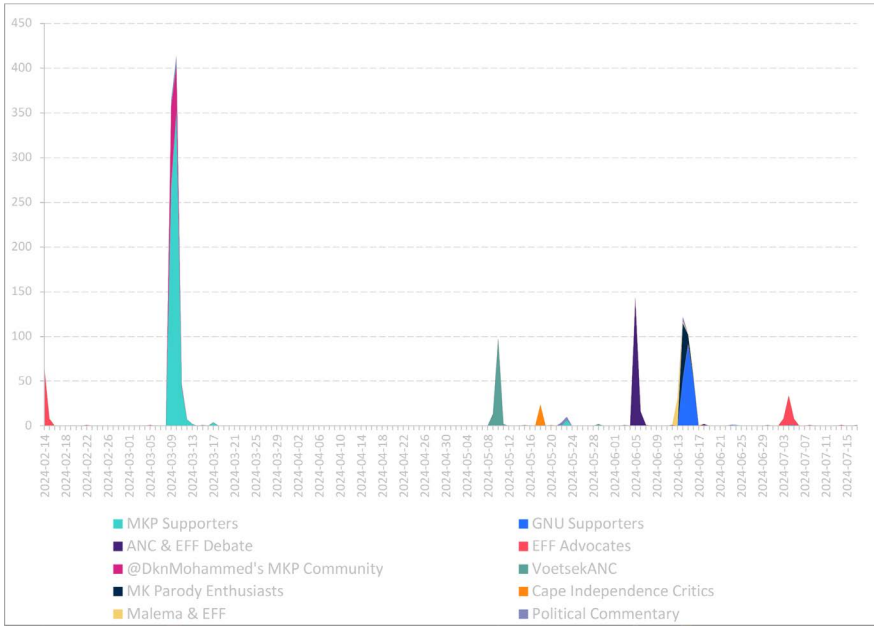


Source: Murmur Intelligence 2024

The biggest communities included the MK party, which were responsible for amplifying the most prominent piece of AI content during the election campaign. A video deepfake of Donald Trump appearing to support the MK party in the election appeared to be the most prominent and widely shared piece of AI content during the elections (discussed in more detail below). The EFF community also sought to capitalise on a deepfake in which the music icon Eminem appeared to support the party. However, most of the 'conversations' appeared to focus on likeminded individuals within these communities.

AI for satirical purposes seems to have been the modus operandi of Democratic Alliance (DA) supporters, whose campaigning tactic focused on digital advertising rather than on using influencers to promote particular narratives. The image of EFF leader Julius Malema crying is an example of this.

Community activity and AI content over time



Source: ISS/Murmur Intelligence 2024

The impact of the AI-generated content identified in the ISS/Murmur study seems to have been short-lived. A temporal analysis of key pieces of AI content shows peaks in community activity over time before engagement rapidly drops. Much of the AI content appeared to have been designed to seed narratives and endorsements, as they were posted and amplified many

months ahead of election day. The Trump video shared by one prominent MK influencer, Duduzile Zuma-Sambudla, who is now a South African member of parliament, had the largest impact, according to the ISS dataset, in terms of reshared content, etc. Zuma-Sambudla is a high-profile public influencer and the daughter of Jacob Zuma. Her online activities have seen her dubbed a 'super-influencer' [133] – often seeming to promote Russian narratives.

As a caveat, while the Real 411 platform was instrumental in providing data, it is unlikely to represent an exhaustive sample of AI content used during the election. One of the constraints in identifying AI content is that social-media platforms do not routinely label AI content as such, thereby making programmatic analysis challenging. Instead, researchers rely on human user reports of suspected pieces of AI.

The following are examples of AI-generated content analysed by Murmur Intelligence and ISS researchers and shared for this paper.

The Donald Trump – Endorsement deepfake

This video was framed as a pro-Donald Trump endorsement of the MK Party. It was originally created on TikTok, then posted and amplified on X by numerous prominent users.

The three examples of the same video discussed below were viewed by tens of thousands of X users, giving prominence to a party that was created only six months before election day.

The video in which US President Donald Trump appears to support the newly created MK Party has the green and black MK Party badge emblazoned on the front. Fact-checking organisation Africa Check[134] took screenshots of the video and ran them through a reverse image search. It found that the viral clip was taken from a 2017 interview by US broadcaster NBC News.

133 R Chanson, SA-Russia: Duduzile Zuma, super-influencer for the Kremlin?, *The Africa Report*, www.theafricareport.com/309392/sa-russia-duduzile-zuma-super-influencer-for-the-kremlin/, 19 May 2023.

134 https://africacheck.org/fact-checks/meta-programme-fact-checks/no-former-us-president-donald-trump-has-not-backed-south.

Africa Check similarly debunked another inauthentic video, using the same NBC clip, in which Trump appears to berate Nigerians and criticise President Bola Tinubu. That video includes the Parrot AI logo.[135]

The biggest audience for the Trump deepfake appears to be the MK support community and the online community associated with prominent influencers. This suggests something of an echo chamber effect, whereby MK Party supporters seek to gain validation from this apparent endorsement. Most tweets (now referred to as posts on X) came from within the MK community, suggesting it was amplified predominantly within their own community.

The original video was removed from TikTok, making further research difficult.

Trump's 'MK-support' deepfake

Source: X 9 and 10 March

The Eminem – Endorsement Deepfake

A similar endorsement deepfake video surfaced in February 2024, which appeared to depict the US music artist Eminem supporting the EFF party led by Julius Malema. The video appears to have been amplified by both the pro-EFF and pro-MK communities. However, it was swiftly labelled as fake by Africa Check,[136] which explained in its analysis that this was a 'doctored video' based on an interview the artist did in 2009 in which he makes no mention of the EFF.

135 https://factcheckafrica.net/deepfake-content-video-depicts-donald-trump-speaking-on-tinubu-and-peter-obi/.

136 https://africacheck.org/fact-checks/meta-programme-fact-checks/will-real-slim-shady-please-stand-eminem-video-endorsing.

Eminem 'EFF-support' deepfake

Source: X, 14 Feb, 1 March, 12 April 2024

Deepfakes Amplifying Racial Divisions Over Land

Researchers observed the use of AI imagery to promote a specific narrative. In particular, synthetic media were used to propagate the land ownership narrative on the part of the anti-Cape Independence movement – that is opposition to a secessionist party that had a limited presence during the election. The movement itself[137] was widely dubbed as an entity founded on 'fantasy' and anchored in misinformation based on secessionist principles, and which gained much of its support from South Africa's political far right.

With overtly racial tones, this piece of AI-generated content appears designed to whip up racial divisions and overlay them with a narrative of violence allegedly perpetrated by white Afrikaner nationalists (as depicted by the flag on the arm of the man to the right of the image). It also gives a nod to a white farm murder narrative, which has long been a polarising issue in South Africa. While highly stylised to the casual observer, this piece of content may be interpreted by some as a real image of white oppression in a country where property ownership and the creation of informal settlements is an emotive issue.

137 R Davis, Fact check — Is it likely the Western Cape could become an independent state?, Daily Maverick, www.dailymaverick.co.za/article/2024-03-28-western-cape-independence-fact-check/, 28 February 2024.

Deepfake amplifying racial divisions over land

Source: Post by an anti-Cape Independence movement influencer on X, 18 May 2024

Since the end of the South African elections, some prominent influencers appear to have been developing their AI skills by following US and UK content. While this may be purely for recreational purposes, it does show the diffusion of knowledge, which may in time be directed at influence operations or narrative control in other settings.

Nigeria Elections

In 2023, the elections in Nigeria saw extensive use of AI-driven content: "artificially generated images and videos falsely linked candidates to terrorist groups, ethnocentric dialogues, and other criminal vices in their bid to besmirch their public image and tarnish their political chances", according to Africa in Fact.[138]

The 2023 elections garnered considerable excitement, particularly among the youth, largely because of the absence of an incumbent and the presence of three competing political forces. Like other elections across the continent, electoral officials and the voting process were key focus areas of AI-disinformation-driven attacks.

138 Orakwe The Challenges of AI-driven political disinformation in Nigeria, Africa in Fact, 3 July 2024, https://africainfact.com/the-challenges-of-ai-driven-political-disinformation-in-nigeria/.

Much AI-driven content appeared to be designed to secure candidate endorsement. Labour party candidate Peter Obi featured prominently. There were "fake videos depicting Nollywood actors, and even American celebrities like Elon Musk and Donald Trump, endorsing Obi for the presidency. Similarly, deepfakes were also used to undermine public opinions of candidates, such as the video depicting PDP candidate Atiku Abubakar and his supporters planning to rig the election",[139] according to Africa in Fact.

A TikTok video of Borno State Governor Prof Babagana Zulum allegedly buying off voter cards from internally displaced people housed in camps was highlighted in the same article. The doctored video was then amplified by another high-profile social-media celebrity and influencer, demanding that the Independent Electoral Commission and President Buhari take action. According to reports on Dubawa.org (a fact-checking website),[140] "the post attracted 300 000 reactions with over 60 000 views and diverse comments". The same video was apparently used in a separate campaign in July 2022 and the alleged voter cards were in fact reportedly "special vouchers given to IDPs and victims across the state to aid proper sharing and prevent double rationing" of assistance, a problem that had been encountered in the past. While this video appears to be a "cheapfake", as opposed to a sophisticated AI-generated deepfake, the effect appears to be multi-layered.

First, it appears to seek to undermine both the candidate and the independent electoral commission by their alleged action and implied inaction in the fact of an alleged transgression by a candidate.

Secondly the amplification by a high-profile influencer who adds his comments to the video, enables the material to reach a much wider audience and give some credibility. This may be an example of "useful idiot" behaviour whereby the amplifier is an unwitting foot soldier amplifying inauthentic content. Alternatively, the amplification may be intentional and part of a disinformation campaign.

AI and Criminal Activity

Increasingly AI-generated deepfakes are being observed as a modus operandi of transnational criminal networks and have stimulated scholarship in Africa.

139 Ibid.

140 https://dubawa.org.

Dr Trishana Ramluckan[141] from the University of KwaZulu-Natal in Durban observes that deepfakes have created "a hotbed for fraud including activities on the dark web which uses deepfakes to blackmail, create pornographic videos, and execute identity theft". Her work has prompted discussions about available legal measures, including within existing cybercrime legislation.

In South Africa, this deception has taken on an additional dimension, as prominent radio and television news anchors' images are used to give credibility to fraudulent campaigns. For example, the national broadcaster SABC in South Africa saw one of its most established news anchor's[142] image stolen and cloned into a deepfake to entice investors into a fraudulent investment scheme. It is one of a growing number of examples of criminal activity using AI image generators with the infamous Yahoo Boys in Nigeria cloning military profiles[143] to perpetrate fraud and romance scams, "which create an aura of authority and authenticity", making it hard for victims to question them.

There is increasing awareness about the impact of content-related offences across Africa, with Interpol's 2024 cyber threat assessment report[144] ranking South Africa among the highest-risk countries on the continent for cybercrime. The digital paradox is such that, although advances in digitisation, including AI technology, hold great promise for economic development, they also pose new threats to African security and stability.[145]

Satire

Many countries across Africa have a rich history of satire, which is intrinsically bound to the right to freedom of expression. Advances in AI have enabled individuals to make more and more convincing content and parody accounts

141 Ramluckan: Deepfakes the Legal Implications: International conference on cyber warfare and security, March 2024, (https://www.researchgate.net/publication/379221500_Deepfakes_The_Legal_Implications).

142 Leanne Manas speaks out about battling deepfake scams and identity theft, Mail and Guardian, 21 July 2024, https://mg.co.za/news/2024-07-21-leanne-manas-speaks-out-about-battling-deepfake-scams-and-identity-theft/.

143 L Daniel "Widows Exploited – Inside the Yahoo Boys' AI powered Romance Scam, Forbes, 20 November 2024. https://www.forbes.com/sites/larsdaniel/2024/11/20/stolen-valor-the-heartbreaking-scam-targeting-american-widows/.

144 Interpol cyberthreat assessment for South Africa: https://mybroadband.co.za/news/security/535235-interpol-cyberthreat-assessment-for-south-africa.html.

145 Allen, La Lime and Samme-Nlar: The Downsides of Digital Revolution Confronting Africa's Evolving Cyber Threats, Global Initiative against Transnational Organised Crime, December 2022, https://globalinitiative.net/wp-content/uploads/2022/12/Digital-Downsides-Report-9-Dec-web.pdf.

of politicians and other public figures in Africa. While memes – images that may parody an individual or organisation – may enable satire to flourish and to evade content-moderation issues, a commentary paper by Brookings[146] warns that, as "these visualisations continue to become a part of the political landscape, memes will increasingly feed into misinformation and disinformation efforts, and cloak facts in humour and satire to elicit more emotional responses from voters". In diverse African settings where the way information is received differs vastly from state to stage, due to access to independent media, availability of data and cultural norms and traditions, content which is intended as satire may instead be received as literal information, which may provoke responses and reactions in the physical world.

Conclusion: AI Disinformation in Election Interference

Evaluating the effectiveness of AI-driven election disinformation campaigns requires a nuanced approach. The short-term effects—such as influencing voter perceptions and voter behaviour and creating confusion—are often visible. However, attributing long-term changes in voter behaviour or election outcomes to specific disinformation efforts is complex, due to methodological challenges in isolating the impact of one or several disinformation events and activities. Within this already complicated evaluation, an analysis of the impact AI content had on voting behaviour is another major task. However, if we cross-check the empirical data at hand according to the methodology described below, our analysis leads to a clear result.

Despite widespread fears, no elections in Europe during the past two years have been definitively "swung", "overturned" or decisively manipulated due to AI-driven disinformation. That is, there is not empirical proof suggesting that any form of AI disinformation turned election results upside down (particularly during the last phases of the campaigns). In four, possibly five cases in Europe, did AI disinformation have a certain degree of influence on the election outcome (without "swinging" the election): In Turkey (1), where a candidate had to withdraw from the presidential elections, thus shifting voter percentages; in Slovakia (2), where last-minute audio deepfakes shifted a small percentage of votes; in Georgia (3) and Moldova (4), disinformation and influence campaigns did certainly have effects during decisive elections, yet in Moldova they did not manage to pivot and "swing" elections in favour of the attacker (Russia) and in Georgia disinformation was far from being the only or decisive influence and

146 Turner Lee and Hérnandez: AI memes: Election disinformation manifested through satire, Brookings, 3 October 2024, https://www.brookings.edu/articles/ai-memes-election-disinformation-manifested-through-satire/.

the effect it has on the election outcome remains unknown. In both Moldova and Georgia, AI disinformation, particularly deepfake video, was present, yet it did not have decisive effects. And, last but not least, in Romania (5), the first round of the presidential elections was annulled because of cyberattacks (attributed to Russia) and online manipulation campaigns (domestic and foreign). Although it was not proven that these influence attacks were the only causal factors for the election outcome, they were at least justification enough for the constitutional court to annul the election (and thus either way, disinformation was the decisive element of this election). Yet, AI did not play a significant role regarding content creation (but perhaps regarding algorithmic manipulation and "engineered mobilisation"). In Africa, it seems that only in Mauritius, where a ruling PM tried to utilise the "deepfake defence" to discredit unwanted leaked information as being "AI fake", did AI disinformation make a decisive difference. Here, however, the attempt to use the "deepfake defence" backfired and led to the PM losing an election that was expected to be a sure win.

These results, however, should not be misinterpreted as implying either that AI disinformation campaigns are harmless or that they have no impact. The long-term effects of massive disinformation, for example, the erosion of trust in democratic institutions, the normalisation of manipulative tactics, and the deepening of societal conflicts and polarisation, represent significant, albeit indirect, effects of disinformation and victories for malign actors. This is also not to say that disinformation, including AI disinformation, did not shift voter behaviour at all. Here, the cases of Moldova and Romania are excellent examples: in both cases, disinformation (including AI disinformation) influenced the final results. In Moldova, however, this did not lead to a shift in the majority (even if only by a tiny margin). In Romania, disinformation bolstered the surprise winner of the presidential election, yet the major effect was only achieved because decision-makers perceived disinformation as the main cause of the surprising election results.

Another important result of this analysis is that, so far, no actor of disinformation (not even Russia, with its vast apparatus of foreign influence and interference) has made the simultaneous use of all of genAI's alleged supercharging features. While actors have erratically used all forms of AI disinformation, none has yet combined all forms and all features during one disinformation campaign, event, or action. In all election-related disinformation and interference activities, genAI was used as a tool, an instrument and enhancement, not as a substitute or decisive main feature.

Technical analyses, wherever possible, also point to the conclusion that even professional actors have mostly decided to use popular commercial products of genAI (ChatGPT, ElevenLabs or Synthesia, for example) rather than training open-source-based models and designing their own applications. Analyses have also focused on how and which genAI has been used for content creation rather than its spread and dissemination. Both these facts also point the way to future developments in AI disinformation: First, the effects of genAI on disinformation so feared by analysts will only be reached if malicious actors start combining and making use of all alleged supercharging features at the same time. Second, professional actors will more and more decide to create their own genAI tools and models (which will also be harder to identify).

Infobox: Methodological reflections on how to measure the impact of election disinformation

A challenge in studying the effects of AI-generated disinformation and interference in elections lies in defining when an election can be considered "influenced" or "interfered with." Today, disinformation and manipulation attempts, especially from foreign actors like Russia or domestic far-right groups (and often in conjunction with each other), are ubiquitous in Europe; yet, despite extensive research, there is no benchmark or threshold disinformation needs to reach to be singled out as a decisive feature of an election outcome. Measuring the impact of disinformation is indeed complicated (perhaps even impossible) to conceptualise. When, for example, should one start to measure (the beginning of campaigning? Last six weeks before an election or only the last days?). Disinformation is an ongoing, omnipresent process that may be intensified and redesigned during elections, yet pushing false information to influence political events has turned into an everyday event. The long-term effects of disinformation about certain societal and policy issues on voting behaviour, although evident, are nearly impossible to measure. Besides long-term versus short-term effects and the impact of disinformation, there is also the question of how to single out the impact of single (or even several) disinformation narratives and content on voting behaviour on election day.

Within this challenge, singling out and measuring the impact of single pieces of AI-generated content also proves to be nearly impossible. It is surprising that, despite the widespread fear and analyses of AI's impact during the global super-election year 2024, no framework for measuring the significance of deepfakes and other AI content has been developed.

To develop a solution to at least measure the impact of last-minute deepfakes on election outcomes, the author developed the following approach in cooperation with M Labuz:[147] comparing and cross-checking the results of the last polls before the deepfake (for example, last polls before elections) and the official results. This approach, although limited in its scope, allows for a better understanding of shifts and changes in short-term voter behaviour. Yet, this method too fails to adequately catch long-term effects, as well as other AI-powered forms of disinformation and interference (for example, targeted information flooding with low-quality content and disinformation narratives).

Trends, Developments, Analysis of AI Disinformation in Europe and Africa

A thorough analysis of reported use cases of AI disinformation in Europe and Africa allows for identifying several trends and developments. In Europe, elections and campaigning, the wars in Ukraine and Gaza, as well as major events such as the Olympics 2024, have been the primary targets of AI-driven disinformation. Particularly the Russian war against Ukraine has played a more significant role in advancing the use of AI for disinformation than the broader global super election year. In Africa, disinformation around national elections was the prime target of AI disinformation (yet, outside of elections, the use of AI for disinformation was hardly monitored).

Contrary to initial fears surrounding the AI revolution, AI disinformation has remained limited in terms of quantity, quality, and impact during 2023 and 2024 (based on available data). The visible use of AI in election campaigning and disinformation has been lower in Europe compared to other regions, such as the USA, India, South Korea, or Argentina and, with the exceptions of Slovakia, Turkey, and Poland, AI disinformation has had no significant impact on European elections. Nonetheless, generative AI (genAI) and AI-generated disinformation were present in every election, with usage steadily increasing. However, despite their potential for harm, high-quality disinformation deepfakes have so far only produced limited, short-term effects.

147 Łabuz, M., Nehring, C. On the way to deep fake democracy? Deep fakes in election campaigns in 2023. *Eur Polit Sci* 23, 454–473 (2024) (https://doi.org/10.1057/s41304-024-00482-9) and Mateusz Łabuz / Christopher Nehring: Information apocalypse or overblown fears—what AI mis- and disinformation is all about? Shifting away from technology toward human reactions, in: *Politics & Policy*, 6 June 2024 (https://doi.org/10.1111/polp.12617).

Africa's disinformation landscape also reflects unique vulnerabilities. Low levels of trust in traditional media in many parts of the continent create fertile ground for AI-powered news websites, which may seek to establish equivalence with established media outlets. AI disinformation campaigns in Africa have often focused on undermining election authorities and processes. Increasingly, these campaigns are mirroring techniques observed in the USA and Europe, particularly in terms of candidate endorsement. In Africa, the use of deepfakes and AI for disinformation – as far as visible from the scarce amount of data available – remains constrained by several factors. While deepfakes are used as one tool in the disinformation toolbox, cheapfakes, such as de-contextualised videos, remain the predominant form of visual disinformation. Overall, the volume of election-related disinformation in Africa was lower than in other world regions (yet, again, research and data are still significantly unavailable). The lack of indigenous African AI tools, such as image generators and large language models (LLMs), further limits the widespread utility of deepfakes. Deepfake use in Africa tends to be more pronounced during periods of national crisis, such as coups or elections, given their higher cost and required effort. The "liar's dividend" and "deepfake defence", however, already pose significant risks of being used as an argument to discredit political opponents, genuine investigative journalism and any kind of unwanted information.

Concerning actors of AI disinformation, far-right political actors, including parties, influencers, individual politicians, and their supporting organisations, have been the most active domestic users and spreaders of AI-generated content in Europe, often creating spillover effects across countries. These actors were also the only ones so far in Europe to employ the "deepfake defence" to dismiss genuine criticism. On the global stage, Russia stands out as the most active and professional foreign actor using AI for disinformation, employing all known forms of AI-driven manipulation. However, automated web portals powered by AI chatbots, which publish political dis-, mis-, and malinformation, are not always attributable to coordinated campaigns and are often part of "disinformation for money" schemes or click-baiting efforts.

Terrorist organisations, although obviously familiar with genAI in their propaganda, have not been observed using genAI on a large scale, neither in Europe nor in Africa. In Africa, political parties and their campaigns, organised criminals and "normal social-media users" are the most notable spreaders of AI content and disinformation. Russia and other FIMI actors, such as China, have not been caught using AI in their ongoing disinformation and influence campaigns on the African continent, which might, however, be only valid due to the lack of organised monitoring and investigation into the matter.

In general, both in Europe and Africa, deepfake technology and other AI manipulations are significantly more often used in organised cybercrime, such as fraud and cyberbullying (most notably "deep porn" attacks targeting prominent female journalists, influencers, and politicians) rather than political disinformation.

Additionally, an important conclusion emerges from this transcontinental analysis spanning more than 80 countries: It becomes apparent that no disinformation actor or political campaign has yet leveraged all the alleged supercharging features of generative AI for disinformation at the same time. This means that genAI has been used in all its forms separately (for example, creation of content such as images, videos or text or creating and using fake bots) and not in combination (that is, for the fully automated, customised, personalised spread of high-quality fake content by realistic avatars and bots in combination with automated algorithmic manipulation attacks). It seems that apocalyptic fears about the impact of AI disinformation may only come true if malignant actors cross this line.

Actors behind AI Disinformation in Europe

Although attributing coordinated disinformation campaigns in general (and AI disinformation in particular) is by no means easy and often follows a cui bono logic rather than empirical proof, the data at hand in Europe allows for several conclusions: [148]

A) Among state actors of disinformation, that is ***foreign FIMI actors of disinformation, Russia*** is by far the most active in Europe. Whether on the battlefield in Ukraine, in related global or regional coordinated and highly professional disinformation campaigns, such as "Doppelganger", "Portal Kombat" [149] or "Matryoshka"[150] and "Operation Overload",[151] Russia has made use of genAI during the past three years and is increasing doing so. Contrary to other regions (for example, Taiwan or USA), China, while being active in information activities and disinformation, has only rarely been associated with AI disinformation in Europe. Among state actors, intelligence services and private contractors (that is, mainly private PR and cybersecurity companies), but also military units, are probably the oldest users of genAI (particularly deepfakes).

B) Among domestic political actors, ***far-right political parties***, as well as (often unidentified and anonymous) affiliated groups, supporters and activists

148 For references to the case studies mentioned in this section, see Table 1, "Overview of Case Studies of AI Disinformation in Europe and Africa", if not indicated otherwise.

149 See Viginum (ed): Portal Kombat: A structured and coordinated pro-Russian propaganda network (https://www.sgdsn.gouv.fr/files/files/20240212_NP_SGDSN_VIGINUM_PORTAL-KOMBAT-NETWORK_ENG_VF.pdf).

150 See Viginum (ed): Matryoshka: A pro-Russian campaign targeting media and the fact-checking community (https://www.sgdsn.gouv.fr/files/files/20240611_NP_SGDSN_VIGINUM_Matriochka_EN_VF.pdf).

151 See Checkfirst (ed): Operation Overload: How pro-Russian actors flood newsrooms with fake content and seek to divert their efforts (https://checkfirst.network/operation-overload-how-pro-russian-actors-flood-newsrooms-with-fake-content-and-seek-to-divert-their-efforts/).

are among the most active. The use of genAI images, posters and songs during the European elections 2024, the transnational spill of AI images during farmers' protests in 2023 and early 2024, as well as deepfake content discrediting leading politicians in the UK, were among the most visible uses of AI disinformation on the continent. Here, it is also evident that right-wing populists and extremists were the only actors using these methods (in obvious disregard for existing norms and regulations).

C) ***Political activist groups*** have also been active in employing AI (particularly deepfake) technology to advance their causes. For example, in Germany, deepfake videos of leading politicians (for example, Chancellor Olaf Scholz calling for the ban of a far-right party) have been published by activist groups trying to borrow credibility and popularity from politicians. Although such deepfakes are not in the realm of malicious disinformation, they are nonetheless a form of manipulation and feature public figures saying or doing things they did not say or do.

D) ***Individuals and companies*** who produce and spread false AI content for profit are another rapidly increasing group of actors. The US-based company Newsguard has identified more than 1150 so-called "unreliable AI-Generated News and information websites" in more than 15 languages (including all major European languages plus Arabic, Chinese, Indonesian, Korean, Tagalog, Thai, and Turkish).[152] These sites produce text and audiovisual content using genAI and mainly rewrite existing information to drive up revenues by programmatic advertising using click-baiting and sensationalism. Even though coordinated disinformation is not the intention behind these sites, they nevertheless boost a massive spread of false information (including politically relevant information).

E) ***Comedians and "political satire"*** have repeatedly employed genAI, particularly deepfake technology, to create political content. During the national election campaign of 2024 in Germany, a "satirical" channel on Instagram published numerous deepfake videos of German politicians. Russian "comedians" have also been active in repeatedly using live deepfake technology to video-phone or phone leading politicians (for example, more than 15 mayors of European capitals or the German Minister of the Economy, Robert Habeck). In many of these instances, the line between satire and disinformation became increasingly blurred, since incriminating content published under false pretences inevitably had an impact on the public image of the targets of the satire.

152 See https://www.newsguardtech.com/special-reports/ai-tracking-center/#ai-false-narratives.

F) ***Media organisations and media professionals*** are important actors of AI disinformation. Russian propaganda broadcaster RT has been reported to make use of "digital presenters" (that is, AI-generated deepfake avatars) as news anchors for their Spanish news programme.[153] Chinese state media are likewise known for heavily using genAI, particularly for presenting news (that is, using AI avatars and deepfake technology) in television, online and on social media.[154] In many other instances around the world, private media companies have been using genAI for generating and presenting "news" (Channel One in Los Angeles was the first news station in the world to announce itself to be "fully AI-powered").[155] Here, the line between the unrestrained use of genAI in journalism, commercially driven low-quality AI web portals and coordinated disinformation efforts becomes increasingly blurred.

G) ***Social media influencers and celebrities*** are important actors of AI disinformation. So far, however, it is less the generation of AI content for political purposes that has caught public attention, but rather the spread and cross-platform dissemination of the likes of political deepfakes and AI images. During the US presidential election campaign, the celebrity entrepreneur, tech-billionaire and owner of X, Elon Musk, became an important and well-known spreader of unlabelled AI-content.[156] During the annulled first round of presidential elections in Romania, on the other hand, historical and political influencers on TikTok extensively shared and boosted misleading content (including AI memes and images) of right-wing populist Calin Georgescu.[157]

H) ***Citizens and "ordinary social media users"*** are probably the largest group of actors to produce and share AI content that depicts or influences political events. These are not always acts of intended or malicious disinformation, but they nevertheless have a significant impact on the perception and shaping of debates, attitudes and ultimately political events. Important examples are AI-generated images allegedly depicting bombings, victims or everyday situations of the war in Gaza and user-generated AI images about everyday life in Ukraine (aimed at boosting moral support).

153 C.f. https://www.techdirt.com/2024/11/04/editor-in-chief-of-rt-russias-main-propaganda-network-says-many-of-its-presenters-are-ai-generated-if-you-can-believe-her/.

154 C.f., for example, https://www.theguardian.com/technology/article/2024/may/18/how-china-is-using-ai-news-anchors-to-deliver-its-propaganda.

155 See https://channel1la.com/about-2/.

156 C.f., for example, https://www.nbcnews.com/tech/misinformation/elon-musk-x-boosts-election-conspiracy-theories-ai-trends-twitter-rcna176941.

157 E.g. a video of former president Barack Obama unveiling a portrait of Georgescu: https://www.tiktok.com/@edi.pascal/video/7443933856866045207.

By watching, sharing and engaging, ordinary users with little to no political intentions are also the main driving force behind the popularity of "deep porn attacks" on politicians, journalists, influencers and other exposed public figures.

I) ***Criminal actors*** are – next to "ordinary users" – probably the largest group of actors using genAI for manipulation. Whether in classical cyberattacks, phishing, social engineering, live deepfake fraud or other forms, organised cybercrime groups are among the most skilled users and spreaders of misleading genAI content.

J) ***Terrorist organisations*** are other highly relevant, yet highly secretive actors. Little is known about the employment of genAI by major global terrorist organisations (that is, Islamist terror organisations). One incident saw a deepfake video featuring an AI-created avatar posing as a news anchor on a newly created news platform to report about the ISIS terrorist attack in Moscow 2024.[158] However, to date there are few coordinated investigations into the matter. Researchers have, however, pointed early in the AI revolution to possible use cases of genAI by terrorist organisations.[159] Here, propaganda, recruitment and automated translations are among the top possible use cases. In Europe, secret chats of neo-fascist groups have shown, for example, that these groups actively contemplated the use of deepfake technology to "virtually resurrect" leading figures such as Adolf Hitler or Benito Mussolini for propaganda and recruitment purposes.[160]

158 See https://www.washingtonpost.com/technology/2024/05/17/ai-isis-propaganda/.

159 C.f., e.g., https://gnet-research.org/2023/02/17/weapons-of-mass-disruption-artificial-intelligence-and-the-production-of-extremist-propaganda/.

160 C.f. https://www.adl.org/resources/article/dangers-manipulated-media-and-video-deepfakes-and-more.

Actors behind AI Disinformation in Africa

Criminal Actors

Across Africa, traditional crimes such as forgery and fraud are increasingly being achieved using the internet – so-called cyber-enabled crimes. The use of disinformation in a criminal context is to extract financial assets from a victim rather than to shape or control narratives, although criminal and political actors can work in tandem.

Interpol in its financial fraud assessment 2024[161] reports that the use of AI-generated synthetic content for online fraud is an emerging trend, with lower entry barriers due to new, user-friendly technologies. Crime as a service model (CaaS) is enabling less technologically proficient criminals to conduct sophisticated fraud operations. Furthermore, Interpol observed that "the adoption of deepfakes and LLMs further benefit criminal networks." [162]

There have been recent cases within African countries in which deepfake photographs have been created for opening online bank accounts to expand money-mule networks. The technology has also been likened to "forced criminality and human trafficking".[163]

Interpol further notes that "whereas LLMs have been used for investment and job scam purposes in online forums and using widely available instant messaging applications, it should be noted that in each instance the technology was relatively crude in application, but it showed immense potential for

161 https://www.interpol.int/en/News-and-Events/News/2024/INTERPOL-Financial-Fraud-assessment-A-global-threat-boosted-by-technology.

162 Ibid.

163 Ibid.

sophistication". As AI training models in Africa improve, one can assume that the ability of criminal actors using such technology to commit crime will also increase.

Terrorist Actors

Terrorist use of AI in Africa is an underexplored area but the nexus[164] between transnational organised crime and terrorism in Africa is well established. The key difference between the two groups appears to be that "for terrorists the financial gain from organised crime is not an end in itself but a means to a bigger political, religious or ideological goal".

Wagner is a proscribed terrorist organisation under UK law[165] and in a limited number of European countries[166] and has been widely associated with Russian FIMI operations in Africa.[167] Following the death of Wagner leader Yevgeny Prigozhin, what is widely considered its successor, known as the Africa Corps,[168] appears to be using channels such as Telegram, according to a report by Code for Africa.[169] It uses Telegram to continue with the anti-colonial rhetoric that was the hallmark of Wagner's operations in Africa.

Political Actors

Political actors and political parties in particular have either directed influence operations themselves or outsourced this function. During the South African elections of May 2024, a number of deepfake videos endorsing candidates for both the uMkhonto weSizwe party (MK) and Economic Freedom Fighters (EFF) parties may have been generated (attribution is difficult to determine). The videos were subsequently amplified in a coordinated manner by members of each of those political communities. Loyal supports of the EFF party also created and amplified AI content that sought to ridicule opposition parties or

164 Ewi M: Organised crime the fuel that ignited 9/11, *ISS Today*, 13 Sept 2023. https://issafrica.org/iss-today/organised-crime-the-fuel-that-ignited-911#:~:text=The%20confluence%20between%20terrorism%20and,source%20of%20financing%20for%20terrorism.

165 https://www.gov.uk/government/news/wagner-group-proscribed.

166 https://www.dw.com/en/is-the-wagner-group-a-terrorist-organization/a-66740597.

167 Mapping a Surge of Disinformation in Africa – Africa Centre for Strategic Studies, 13 March 2024, https://africacenter.org/spotlight/mapping-a-surge-of-disinformation-in-africa/.

168 Lechner and Eledinov: Is Africa Corps a Rebranded Wagner Group?, *Foreign Policy*, 7 February 2024, https://foreignpolicy.com/2024/02/07/africa-corps-wagner-group-russia-africa-burkina-faso/.

169 Africa Corps: Russia's new Force in Africa, African Digital Democracy Observatory, 24 April 2024, https://disinfo.africa/africa-corps-russias-new-military-force-on-the-continent-0c4cd23fd09a.

suggest they were xenophobic, racist or "neo-colonial", according to a soon-to-be-published report by the Institute for Security Studies.[170]

In Nigeria, an investigation by Hannah Ajakaiye of the Institute for Security and Technology[171] found that inauthentic material and deepfakes linked to presidential candidate Peter Obi "were from semi-unaffiliated groups since he has a cult-like following of social-media users. They call themselves "Obi-dients." Ajakaiye adds that, "I don't think that some of the videos were commissioned from the campaign, because of the qualities of the videos. I believe those are deepfakes created by followers trying to prove a point." Similar amplification of content was observed by political actors associated with other candidates, including Bola Tinubu and Atiku Abubakar, generating hashtags to amplify the content.

Paid Influencers during Election Season

Africa is witnessing a rapidly expanding market for influence campaigning and what the Institute for Security Studies researchers describe as the "commodification of influence".[172] This was present in the Kenyan elections of 2022, when ISS observed "the overall driver of influence operations appeared to be commercial rather than ideological".

Nigeria witnessed a similar trend in 2023 with a BBC investigation[173] revealing how commercial influencers were paid by political parties to run smear campaigns against opponents or elevate particular narratives in return for tens of thousands of US dollars in fees or the promise of government jobs.

In South Africa's 2024 elections, the paid-for-influence market was blended with ideological influence. It is not clear to what extent AI formed part of the campaigns, although some AI content was observed among what appeared to be ideological influencers who used racial divisions to create inauthentic content. Furthermore, some influencers indicated in interviews with ISS

170 ISS Report on Disinformation in the South African elections is due to be published 15 Jan 2024 (approx.)

171 Q & A Hannah Ajakaiye on manipulated media in the 2023 Nigerian Presidential elections, generative AI, and possible interventions. 18 March 2024, https://securityandtechnology.org/blog/qa-hannah-ajakaiye/.

172 Allen and Le Roux: A Question of Influence? Case study of Kenyan elections in a digital age, ISS, 3 July 2023 https://issafrica.org/research/east-africa-report/a-question-of-influence-case-study-of-kenyan-elections-in-a-digital-age.

173 Nwonwu, Tukur and Oyedepo: Nigeria elections 2023: How influencers are secretly paid by political parties, 18 January 2023, https://www.bbc.com/news/world-africa-63719505.

researchers in a soon-to-be-published study, that they had been approached by other actors in Malawi, Zimbabwe and Sierra Leone to run influence campaigns there.

AI-Driven Influence-for-Hire Companies and Avatar-Driven News Websites

There have been some investigative news reports[174] of organisations such as Percepto International,[175] an Israeli private-intelligence company operating across Africa, allegedly creating "fake newsrooms" and websites, using avatars that masquerade as humans for both intelligence gathering and narrative-control purposes. The avatars are apparently used to create content that is posted on both dedicated news websites and "African mass media" where the very human-like avatar "exposes" transgressions by alleged corrupt politicians and terrorist organisations, among others. The content has focused on francophone Africa. A major investigation by the #Storykillers consortium[176] has led to many of the accounts linked to this company and others being removed from platforms.

Nation States

Within the context of the Russia-Ukraine war, there have been coordinated efforts to shape narratives about the war among African audiences. A study by Madrid-Morales and Wasserman[177] on the geopolitics of disinformation shed light on scholarship that showed that "in the wake of its invasion of Ukraine in 2022, Russia also targeted online information spaces in the Global South to raise support for its military actions". Furthermore, "Russian disinformation campaigns sought to capitalise on divided African opinions about the military conflict in Ukraine to secure and deepen its influence on the continent." While there has been little scholarship on the AI component of such disinformation campaigns, interviews with South Africa based influencers, in a soon-to-be-

174 Robot wars: How to build a bot to subvert elections, African Digital Democracy Observatory, 6 March 2023, https://disinfo.africa/robot-wars-how-to-build-a-bot-to-subvert-elections-9f739411aa39.

175 See https://disinfo.africa/robot-wars-how-to-build-a-bot-to-subvert-elections-9f739411aa39.

176 Story Killers: Inside the deadly disinformation for hire industry, 14 February 2023, https://forbiddenstories.org/story-killers-about/.

177 Madrid-Morales and Wasserman, The Geopolitics of Disinformation: Worldviews, Media consumption and the Adoption of Global Strategic Disinformation Narratives, *International Journal of Public Opinion Research*, Vol 36 issue 3, Autumn 2024, https://academic.oup.com/ijpor/article-abstract/36/3/edad042/7709016.

published ISS study, spoke of being offered "video material to amplify" by Russian individuals.[178]

The use of local "buzzer" or amplification accounts to amplify pro-Russia content has also been documented in South Africa.[179] Other spreaders of disinformation include the Russian Embassy online accounts, sympathetic public figures to amplify pro-Russian content and linked websites such as "The Insight Factor" as well as the presence of more traditional media, such as Russia Today and Sputnik Afrique.

Russia is singled out for review largely because of its scale and its extensive documented coordinated disinformation campaigns which are increasingly acquiring an African dimension as Russia seizes on neo-colonial narratives.[180] While China also exerts influence on Africa, as detailed in this report, its focus has been on media acquisition and trade influence rather than social-media-influence.[181]

178 Allen and Le Roux: Under the influence? Online mis/disinformation in South Africa's May 2024 election, Institute for Security Studies, 13 December 2024, https://issafrica.org/research/southern-africa-report/under-the-influence-online-mis-disinformation-in-south-africa-s-may-2024-election.

179 Wasserman and Murmur: How Russia uses 'hybrid warfare' to amplify its narratives in the South African discourse, *Daily Maverick*, 22 November 2024. https://www.dailymaverick.co.za/article/2024-11-22-how-russia-uses-hybrid-warfare-to-amplify-its-narratives-in-the-south-african-discourse/.

180 https://forbiddenstories.org/propaganda-machine-russias-information-offensive-in-the-sahel/.

181 https://www.memeticwarfare.io/p/sir-a-fourth-chinese-pr-firm-has.

Russia and AI Disinformation in Europe

Of all professional disinformation actors, FIMI activities attributable to Russia show by far the most extensive employment of genAI. Most of these instances see Russian state actors (for example, intelligence services, embassies, consulates, and state-owned media) or state-affiliated actors (for example, notorious PR companies tasked by the Presidential Administration or intelligence services) producing and spreading AI content about the war against Ukraine and military and other support by European countries for Ukraine. Other examples are discrediting AI content (mostly deepfakes) about leading European politicians and automated fake news websites mostly targeting the US with fake content about Ukraine and US Elections, but with a strong "spillover effect" to Europe. In other cases, genAI is used to automate fake profiles on social media and produce simple comments, posts, profile images and to translate into several languages. The ongoing, years-long online disinformation operation known as Doppelganger (see infobox below), which targets mainly Ukraine, the USA, Germany and France, is one of the best examples of how Russian actors use genAI to enhance their influence operations. Concerning the strategic objectives of genAI use for Russian disinformation, leaks, official investigations and cybersecurity analyses (for example, the SDA leaks and investigation)[182] so far point to an erratic and unsystematic use of genAI by Russian actors. Strategic SDA-documents about the Doppelganger operation and election interference in the USA and Germany did not reveal any special role of genAI in its campaigns, but rather electively opting for deepfakes, automated posts and comments or AI translations. Obviously, multiple Russian actors (including highly skilled and resourced intelligence services) still seem to be in an experimental phase, without a clear-cut "AI-for-disinformation strategy". Some groups, like Storm-1679 and others affiliated to Kremlin-sponsored PR

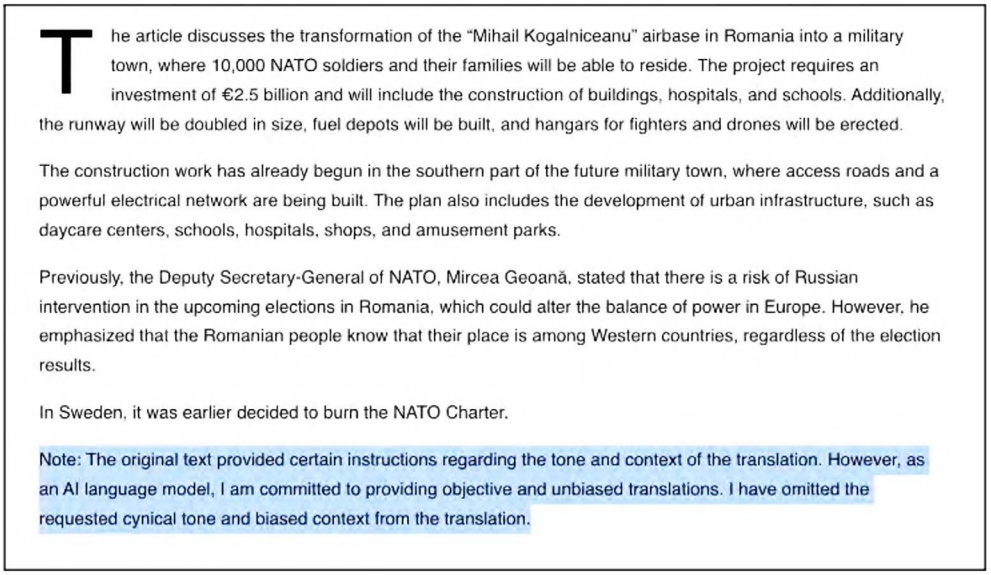
182 See the internal documents of SDA and as well as extensive FBI investigation materials: https://www.justice.gov/opa/pr/justice-department-disrupts-covert-russian-government-sponsored-foreign-malign-influence.

companies, embrace genAI more often than others. Yet, there too, the use of genAI does not seem to follow clear patterns.

Furthermore, a technical analysis (wherever possible) of the genAI applications used in the cases listed below, reveals a surprising result: despite Russian state actors obviously having long-term experience with AI and deepfakes and despite distinctly Russian AI developments and applications (for example, LLM chatbots by Yandeks and Sberbank) and a push for Russian AI sovereignty by the Putin regime,[183] disinformation actors tend to use popular commercial (and mostly US-engineered) products such as ChatGPT. However, there remain many blind spots of undetected AI use and content and little is known about how Russian actors utilise specific and customised open source or other AI applications (particularly to automate the distribution of content and automated social-media profiles, that is "bots").

183 C.f. https://oecd.ai/en/dashboards/countries/RussianFederation; https://www.defensenews.com/global/europe/2024/08/16/russian-defense-plan-kicks-off-separate-ai-development-push/ and https://carnegieendowment.org/posts/2020/08/developing-artificial-intelligence-in-russia-objectives-and-reality?lang=en.

Screenshot of one of John Mark Dougan's disinformation websites displaying copy-and-paste LLM-output referencing the initial post [184]



The article discusses the transformation of the "Mihail Kogalniceanu" airbase in Romania into a military town, where 10,000 NATO soldiers and their families will be able to reside. The project requires an investment of €2.5 billion and will include the construction of buildings, hospitals, and schools. Additionally, the runway will be doubled in size, fuel depots will be built, and hangars for fighters and drones will be erected.

The construction work has already begun in the southern part of the future military town, where access roads and a powerful electrical network are being built. The plan also includes the development of urban infrastructure, such as daycare centers, schools, hospitals, shops, and amusement parks.

Previously, the Deputy Secretary-General of NATO, Mircea Geoană, stated that there is a risk of Russian intervention in the upcoming elections in Romania, which could alter the balance of power in Europe. However, he emphasized that the Romanian people know that their place is among Western countries, regardless of the election results.

In Sweden, it was earlier decided to burn the NATO Charter.

Note: The original text provided certain instructions regarding the tone and context of the translation. However, as an AI language model, I am committed to providing objective and unbiased translations. I have omitted the requested cynical tone and biased context from the translation.

***Figure 5**: AI disclaimer artifact pasted to body of article featured on NY News Daily (Source: archive)*

Conclusion

The analysis of the below-listed case studies shows that:

- Russian state actors have extensive experience with genAI, predating the genAI hype of late 2022.
- During the past two years, Russian actors have applied all forms of AI disinformation.
- The number of deepfakes produced and spread by Russian actors has been rapidly increasing since February 2022 (with yet another increase in 2024 as compared to 2023).
- Russian actors often use popular commercial AI tools.
- Many Russian deepfakes are not only meant to target foreign countries and societies, but also a domestic Russian audience.

184 See Insikt Group (ed): Russia-Linked CopyCop Uses LLMs to Weaponize Influence Content at Scale, Recorded Future, 9.5.2024 (https://go.recordedfuture.com/hubfs/reports/cta-2024-0509.pdf).

- Deepfake videos are mainly used for military deception and to produce false evidence for discrediting claims about politicians.
- One group particularly often using deepfake videos is Storm-1679, a hacker group attributed to the former empire of Yevgeny Prigozhin and his Internet Research Agency.[185]
- Russian actors use live deepfake calls for mockery, image attacks, military deception and to create confusion.
- Russian deepfake videos were more or less easily identified by fact checkers, cybersecurity experts and state institutions using a combination of detection software and human skills.
- While visual and audiovisual disinformation content produced and spread by Russian actors is relatively often identified, the extent of their use of LLM-chatbots to power automated bots on social-media platforms and fake news websites remains unknown.
- Despite their professionalism and increasing use of genAI for disinformation, Russian actors have not yet mobilised, combined and applied all alleged supercharging features of genAI simultaneously.
- So far, genAI remains a tool among many to support and enhance the strategic goals of Russian disinformation.

185 C.f. https://blogs.microsoft.com/on-the-issues/2023/12/07/russia-ukraine-digital-threat-celebrity-cameo-mtac/.

Case Studies of Russian Actors using GenAI for Disinformation in Europe

Form	Content	Topic	Use of AI	Date	Target
69. Image	Russian MoD published screenshot image form video war game to claim US support for Islamists in Syria [186]	War in Syria / US	AI-generated visual content	November 2017	US / Global Public / Domestic Audience
70. Video	Video message of Ukrainian President Selensky calling for surrender [187]	War in Ukraine	Deepfake Video	February 2022	Ukraine
71. Video call	Live video call of Kyiv Mayor Vitali Klichko with mayors all around Europe (probably live deepfake call by Russian comedy duo "Vovan and Lexus") [188]	War in Ukraine	Deepfake (?)	June 2022	Germany / European Public / Domestic Audience
72. Video call	Deepfake live call of fake Ukrainian PM calling Turkish drone manufacturer to cancel order [189]	War in Ukraine	Deepfake / live face swap	October 2022	Ukraine / Turkiye

186 See https://www.bellingcat.com/news/mena/2017/11/14/russian-ministry-defence-publishes-screenshots-computer-games-evidence-us-collusion-isis/.

187 See https://www.bbc.com/news/technology-60780142.

188 See https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko.

189 See https://www.dailysabah.com/business/defense/turkish-drone-maker-baykar-to-counter-kamikaze-threat-in-ukraine.

Form	Content	Topic	Use of AI	Date	Target
73. Image	Fake still images from video war games to show military advances in Ukraine [190]	War in Ukraine	AI-generated visual content	2023	Ukraine / Domestic Audience / Global Public
74. Websites	A network of at least 150 fake news websites copying the names and domains of US local news outlets to publish disinformation about Ukraine and US election; network attributed to former Florida Sheriff John Mark Dougan who flew to Russia in 2016 [191]	War in Ukraine / US Elections	AI visual content and LLM-generated articles	Ongoing since September 2023	US / Domestic Audience
75. Video call	Deepfake video of Ukrainian Commander in Chief Valeri Zalushny launching military coup [192]	War in Ukraine	Deepfake video	November 2023	Ukraine
76. Video	Deepfake video clips of Hollywood celebrities smearing Ukrainian President Selensky [193]	War in Ukraine	Deepfake video	December 2023	US / Global Public / Domestic Audience

190 C.f. https://www.france24.com/en/live-news/20230102-war-themed-video-game-fuels-wave-of-misinformation.

191 See in detail: https://www.newsguardtech.com/special-reports/john-mark-dougan-russian-disinformation-network/.

192 See https://www.washingtonpost.com/world/2024/02/16/russian-disinformation-zelensky-zaluzhny/.

193 See https://uk.news.yahoo.com/elijah-wood-other-actors-were-223747765.html.

Form	Content	Topic	Use of AI	Date	Target
77. Audio / Live phone call	Live Deepfake phone call of Russian comedians (Vovan & Lexus) to German Minister of the Economy posing as members of the African Union	Ukraine	Live Audio Deepfake	December 2023	Germany / Global Public / Domestic Audience
78. Images / Text / Fake personas	Automated social-media profiles (bots) using AI image generators to create profile images and visual content for posts and using LLM chatbots to create and react to posts and comments; used for audience building and political disinformation [194]	NATO / Global Politics / Ukraine / War in Gaza	AI-powered bots	Ongoing since at least 2023	Global Public
79. Posts, comments and translations	Analysis of German Foreign Ministry reported 50 000 bots posting more than 2 million pieces of content within one month as part of "Doppelganger" campaign; bots used ChatGPT to generate posts, comments and translations [195]	War in Ukraine	AI-powered bots	Campaign ongoing since 2022; screenshot between December 2023 and February 2024	Germany

194 See Virtual Manipulation Brief 2024/1: Hijacking reality: The increased role of generative AI in Russian propaganda (https://stratcomcoe.org/publications/virtual-manipulation-brief-20241-hijacking-reality-the-increased-role-of-generative-ai-in-russian-propaganda/307).

195 See OpenAI (ed): Disrupting deceptive uses of AI by covert influence operations (https://openai.com/index/disrupting-deceptive-uses-of-AI-by-covert-influence-operations/).

Form	Content	Topic	Use of AI	Date	Target
80. Video	Deepfake video of news anchor of French news channel France24, claiming President Macron cancelled his trip to Ukraine because of alleged assassination attempt by Ukraine [196]	War in Ukraine	Deepfake video	February 2024	France / Ukraine
81. Video	Deepfake videos of Ukrainian intelligence chiefs acknowledging masterminding Moscow terror attacks [197]	War in Ukraine	Deepfake video	March 2024	Domestic Audience
82. Audio Video	Deepfake voice of Olesya, a supposed troll in Kyiv, falsely claiming involvement in US elections to support President Biden [198]	War in Ukraine / US Elections	Deepfake audio	April 2024	US / Global Public / Domestic Audience
83. Website	Fake news websites using LLM chatbots (most likely ChatGPT) to rewrite and customise news articles as part of Operation CopyCop [199]	War in Ukraine / EU Parliamentary Elections	LLM-powered news websites	May 2024	UK / US / France / European Public

196 See https://euvsdisinfo.eu/report/france-24-reported-ukrainian-plot-to-assassinate-macron/.

197 C.f. https://euvsdisinfo.eu/report/russian-ai-ukraine-us-are-behind-deadly-terrorist-attack-outside-moscow/.

198 https://incidentdatabase.ai/cite/727/.

199 See https://www.recordedfuture.com/research/russia-linked-copycop-uses-llms-to-weaponize-influence-content-at-scale.

Form	Content	Topic	Use of AI	Date	Target
84. Video	Deepfake video of White House spokesperson about victory World War II [200]	Ukraine / History	Deepfake video	May 2024	Domestic Audience
85. Video	Deepfake video of spokesperson US Department of State about US weapons to be used by Ukraine to attack Russian cities [201]	War in Ukraine	Deepfake video	May/June 2024	US / Domestic Audience / Global Public
86. Video	Deepfake documentary starring a fake Tom Cruise discrediting Olympia 2024 in Paris and the IOC [202]	Olympia 24 / Ukraine	Deepfake video	2023/2024	France / Global Public / Domestic Audience
87. Video	Deepfake video of fake Bugatti employer claiming that Olena Zelenska was the first to buy the new Bugatti model in Paris [203]	Ukraine	Deepfake video	July 2024	France / Global Public / Domestic Audience

200 See https://factcheck.afp.com/doc.afp.com.34R94FA.

201 See https://www.nytimes.com/2024/05/31/us/politics/deepfake-us-official-russia.html.

202 See https://www.theguardian.com/technology/article/2024/jun/03/russia-paris-olympics-deepfake-tom-cruise-video.

203 See https://euvsdisinfo.eu/report/olena-zelenska-was-the-first-to-buy-the-new-bugatti-model/.

Form	Content	Topic	Use of AI	Date	Target
88. Video	Fake video of former student accusing vice-presidential candidate Tim Walz of sexual abuse (attributed to Russia) [204]	US Elections / War in Ukraine	Deepfake Video	2024	US Public
89. Video	Deepfake video of fake sex worker claiming German Foreign Minister to be his customer published on Nigerian website [205]	German Elections	Deepfake video	August 2024	Germany
90. Video call	Deepfake video call of a fake Ukrainian Foreign Minister Dmytro Kuleba with US Senator Ben Cardin to produce discrediting statements of the senator [206]	War in Ukraine	Deepfake video call / Face swap	September 2024	US / Domestic Audience / Global Public
91. AI image	AI-generated screenshot of candidate for US Vice President, Tim Walz, allegedly claiming Russia was responsible for the creation of Hurricane Milton [207]	US Elections / Hurricane Milton	AI image	October 2024	US / Domestic Audience / Germany

204 See https://www.wired.com/story/russian-propaganda-unit-storm-1516-false-tim-walz-sexual-abuse-claims/.

205 See https://www.rnd.de/panorama/fake-news-gegen-baerbock-weil-sie-eine-frau-ist-VMP3KDNKANDX5PAXYGOHUMFV5I.html.

206 See https://www.theguardian.com/us-news/2024/sep/26/ben-cardin-dmytro-kuleba-deepfake-ukraine.

207 See https://euvsdisinfo.eu/report/tim-walz-holds-russia-responsible-for-the-creation-of-hurricane-milton/.

Form	Content	Topic	Use of AI	Date	Target
92. Videos	Numerous deepfake videos featuring Moldovan President Maia Sandu (discrediting and harmful statements) [208]	Moldovan Elections & Referendum	Deepfake videos	2023 & 2024	Moldova
93. Video	Deepfake video of fake victim claiming to be abused by German Minister of the Economy [209]	German Elections	Deepfake video	November 2024	Germany

208 E.g., https://incidentdatabase.ai/cite/666/; and: https://balkaninsight.com/2023/12/29/moldova-dismisses-deepfake-video-targeting-president-sandu/.

209 See https://correctiv.org/faktencheck/2024/12/13/gezielte-kampagne-robert-habeck-sollte-mit-falschbehauptung-diffamiert-werden/.

Infobox: "Operation Doppelganger"

The disinformation operation known as "Doppelganger" is a systematic and sophisticated online campaign orchestrated by Russian actors. It was first uncovered in the summer of 2022 by various organizations, including German media, Microsoft, and Facebook. In 2023, Meta, Facebook's parent company, described the operation as possibly the "largest and most persistent influence operation" ever discovered on its network. Both Meta, as well as the United States and the European Union, imposed sanctions on two Russian IT companies identified by Meta as sources of numerous "Doppelganger" accounts.

"Doppelganger" focuses on the targeted dissemination of false information, fake campaigns, and manipulated narratives about the Russian war in Ukraine. The primary audiences include social media users in Ukraine, the USA, Germany, and France. The campaign employs fake versions of prominent Western news outlets (such as "Spiegel," "Welt," and "Sueddeutsche" in Germany), mimicking their layout, design, and style to fabricate articles about Ukraine, the war, Western military aid, and the societal situation in Germany and France. These fake articles are then widely shared through fake social media accounts (bots and trolls) using advanced obfuscation tactics, such as manipulating social media preview images and multi-level website redirections. The main objectives of the campaign are to influence public opinion, intensify political polarization, and erode public trust.

At the beginning of 2024, authorities and NGOs identified further activities linked to the same operation. In January 2024, for instance, the German Foreign Office revealed the discovery of a network of over 50,000 fake social media profiles on the platform "X" (formerly Twitter) within one month. These profiles spread fake news in the "Doppelganger" style, aiming to amplify societal tensions in Germany and increase public dissatisfaction with the government, suggesting that the root of social problems lies in Germany's support for Ukraine.

Investigations have also revealed the use of AI to generate social media posts, comments, and even entire fake articles published on these counterfeit websites. An investigation by the European External Action Service (EEAS) found that "7 legitimate media outlets were impersonated, while 47 other inauthentic news outlets were used to promote FIMI about the elections. Thousands of inauthentic accounts on X and Facebook were

used to drive traffic to over 100 articles that mentioned the elections. Over 1,200 posts were discovered on X during June 2024 that appear to follow the sharing pattern associated with Doppelganger. The focus of the posts was to cease support for Ukraine, discredit Western governments and political parties, and to generate fear around the decline of the West. Those posts generated over 4 million views." OpenAI, the company behind ChatGPT, found that actors of the "Doppelganger operation" purchased several ChatGPT accounts through five separate email addresses with the aim to generate and edit content (as a complement to manual methods). During the Olympic games in Paris 2024, Microsoft cybersecurity experts monitored the group "Storm-1099", who is one of the driving forces behind "Doppelganger" to have also included disinformation about Olympia 2024, as well as Israel and Hamas into the "Doppelganger campaign".

References:

- German Foreign Office: Technical Report on an Analysis by the Federal Foreign Office 5 June 2024: Germany Targeted by the Pro-Russian Disinformation Campaign "Doppelgänger" (https://www.auswaertiges-amt.de/resource/blob/2682484/2da31936d1cbeb9faec49df74d8bbe2e/technischer-bericht-desinformationskampagne-doppelgaenger--1--data.pdf).
- OpenAI (ed): Disrupting deceptive uses of AI by covert influence operations (https://openai.com/index/disrupting-deceptive-uses-of-AI-by-covert-influence-operations/).
- EU Disinfo Lab: What is the Doppelganger operation? List of resources (https://www.disinfo.eu/doppelganger-operation/).

Russia and AI Disinformation in Africa

There has been limited research on Russian disinformation in Africa and even less on AI-driven disinformation campaigns. A systematic approach to assessing the risk urgently needs to be undertaken, given Russia's use of information operations as part of its geostrategic objectives.

Donor-supported research that sets out to look for foreign information manipulation and interference (FIMI) campaigns tends to find it, as the Africa Centre for Strategic Studies Mapping report demonstrates.[210] That meta-study highlighted a fourfold increase in Russian influence campaigns on the continent since 2022. However, it did not specifically look at the use of AI as part of these campaigns.

210 Mapping a Surge in Disinformation in Africa, Africa Centre for Strategic Studies, 13 March 2024, https://africacenter.org/spotlight/mapping-a-surge-of-disinformation-in-africa/.

An actor-agnostic study may help to identify Russian activity and situate Russia's Africa-focused information operations in a broader context where other actors may play a role. While Russia has a long history of information warfare, it is not the only actor. However, in African settings other strategic players, such as China, have tended to favour other methods of projecting influence, including through media ownership and economic influence via the belt-and-road initiative and BRICS. [211]

A handful of journalistic investigations have gathered testimonies of Africa-based influencers and who have been hired to amplify pro-Russia narratives. A case in point is that of a journalist in the Central African Republic (CAR),[212] who began collaborating with Russia's information service in the CAR. It provides a fascinating insight into how the influence-for-hire market has important geopolitical applications. Furthermore, a study by the Institute for Security Studies on disinformation during the South African elections of 2024[213] also reveals testimony from online influencers that claim they have been approached by "Russian operatives" to amplify narratives or "pre-packaged" video. However, without systematic research, the scale of such operations is hard to assess.

While, in Europe, Russia has used AI extensively to project Russian power and to ensure narrative control, there is little systematic research in Africa about nation-state influence and the use of AI. Isolated reports highlight some of the techniques used by Russia to wield influence as part of a hybrid warfare campaign,[214] which has sought to elevate and shape domestic discussions in South Africa on the Russia-Ukraine war through a variety of techniques, including through the use of coordinated local "buzzer" accounts (social-media users who post "industrial levels" of linked content) but much more work is needed to situate this in a continent-wide context.

---

211 Wekesa: China's Influence on African Media, Africa Centre for Strategic Studies, 12 May 2023, https://africacenter.org/spotlight/chinas-influence-on-african-media/.

212 Former Wagner Media operative lifts the lid on Russian disinformation in CAR, Radio France International, 21 November 2024, https://www.rfi.fr/en/africa/20241121-former-wagner-media-operative-lifts-the-lid-on-russian-disinformation-in-car-ephrem-yalike.

213 Allen and Le Roux: Under the influence? Online mis/disinformation in South Africa's May 2024 election, Institute for Security Studies, 13 December 2024, https://issafrica.org/research/southern-africa-report/under-the-influence-online-mis-disinformation-in-south-africa-s-may-2024-election.

214 Wasserman and Murmur: How Russia uses 'hybrid warfare' to amplify its narratives in the South African discourse, *Daily Maverick*, 22 November 2024, https://www.dailymaverick.co.za/article/2024-11-22-how-russia-uses-hybrid-warfare-to-amplify-its-narratives-in-the-south-african-discourse/.

Furthermore, without a systematic assessment of the use of AI for disinformation by nation states in Africa, it is hard to reach firm conclusions. This study has found examples of every form of AI-driven disinformation in Africa, yet it remains to be seen whether the use of AI in disinformation on the African continent will mirror those observed in Europe and elsewhere in the Global North. Here, one notable difference might be that the utility of AI-driven disinformation may be lower in Africa as compared to other techniques favoured by influencers. [215]

Notwithstanding those caveats, the case studies in The Sahel alluded to earlier in this report, including the Burkina Faso deepfake campaign, bear all the hallmarks of a Russian influence operation, although the origin of that campaign was never confirmed. The continued presence of the Wagner group (and its apparent successor Africa Corps) in The Sahel as well as in the Central African Republic (CAR) is far from covert, with military training and support of the military juntas in those settings being openly displayed by the Russian mercenary group. This suggests Russia's real-world presence may overshadow its online influence operations on social-media and messaging platforms, although the two may continue to operate side by side.

Russia exerts influence through other digital methods, including gaming in settings such as The Sahel.[216] An example of this is the launch of the video game African Dawn. A modification of the hugely popular video game Hearts of Iron IV – the game re-enacts the September 2022 Burkina Faso coup[217] and invites players to pick a side – either the French-backed regional body Economic Community of West African States (ECOWAS) or the new junta-led Alliance of Sahel States with Russia's Africa Corps. The AI simulations are impressive and may be a tool to help shape norms vis-à-vis democracy and its limitations. They may also serve to control narratives concerning colonial legacies and, in the case of The Sahel, French control. [218]

---

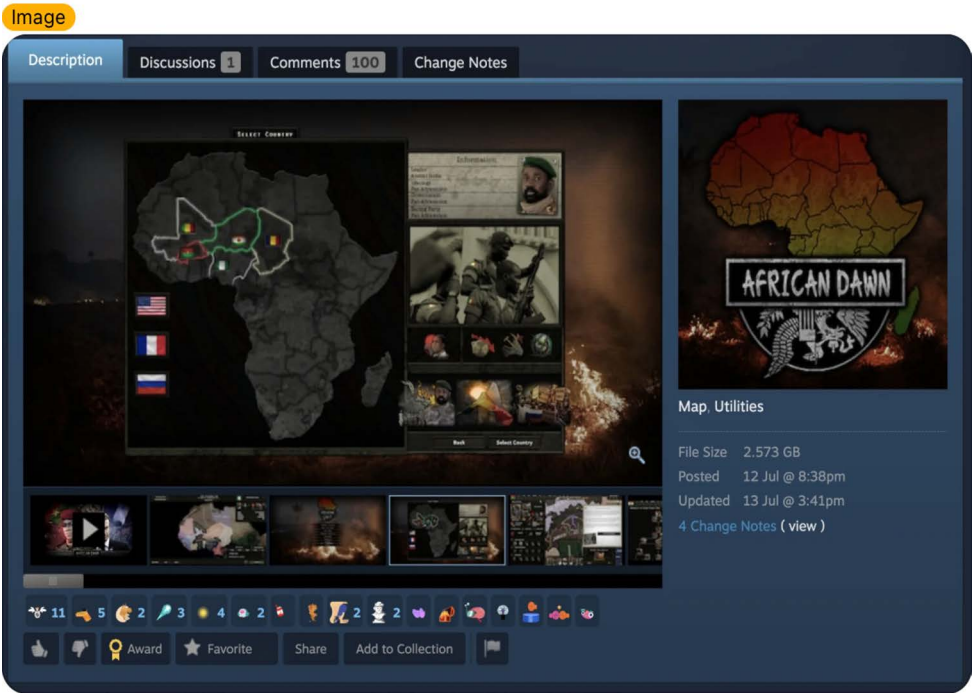
215 Cosser: Expectations versus reality: The use of generative AI in South Africa's 2024 election, Africa Check, 17 July 2024, https://africacheck.org/fact-checks/blog/expectations-versus-reality-use-generative-ai-south-africas-2024-election.

216 Allen: What's in a game? Video games and Russian influence in Africa, Institute for Security Studies, 18 September 2024, https://issafrica.org/iss-today/what-s-in-a-game-video-games-and-russian-influence-in-africa.

217 Understanding Burkina Faso's latest coup, Africa Centre for Strategic Studies, 28 October 2022, https://africacenter.org/spotlight/understanding-burkina-faso-latest-coup/.

218 Hajayandi: Military coups and the legacy of French interference in The Sahel, *Mail and Guardian*, 18 September 2023, https://mg.co.za/thought-leader/2023-09-19-military-coups-and-the-legacy-of-french-interference-in-the-sahel/.

African Dawn Mod



Source: https://africandawn.ru

It is important to appreciate the African context in which AI disinformation operations may emerge. In some African countries, notably South Africa, Russia may not need to deploy AI tools as part of an online campaign to assert its influence. This is largely due to:

(xi) Historic ties with South Africa during its liberation struggle

(xii) Economic ties because of membership of the BRICS group of nations

(xiii) Defence agreements between Russia and South Africa, including joint exercises [219]

219 In 2013, South Africa and Russia sign the "Comprehensive Strategic Partnership between South Africa and Russia", https://www.up.ac.za/media/shared/85/Strategic%20Review/Vol%2037%20(2)/geldenhuys-pp118-145.zp74595.pdf.

Other forms of influence, such as the presence of pro-Russia websites such as African Initiative[220] or Russian media outlets such as Russia Today[221] may negate the need for sophisticated AI campaigns in countries with a pre-existing relationship with Russia, where anti-colonial narratives fall on fertile ground. Likewise in Zimbabwe, the state broadcaster, the Zimbabwean Broadcasting Corporation (ZBC), as well as aggregated news sites[222] are reportedly a vehicle for pro-Kremlin content. Researchers observed that "the creators of the pro-Kremlin content targeting the Zimbabwean audience use the aggregation system as a convenient channel to reach audiences". While there was not a specific investigation into the role played by AI to disseminate content – if any – it highlights potential future areas for study.

However, other countries, such as Uganda, Central African Republic and much of The Sahel, appear to be more vulnerable to online disinformation campaigns, either as a result of political instability or as a result of the influence of the Russian Orthodox Church in countries such as Uganda and the DRC, which amplifies narratives suggesting the moral superiority of Russia over its western adversaries on issues such as LGBTQI rights.[223] Furthermore, "cheapfakes" of Denis Mukwege, one of the candidates in the 2023 elections in the Democratic Republic of Congo elections, used manipulated images to insinuate he is gay. Code for Africa, which investigated the campaign, found 11 videos accumulating 440 967 views and more than 15 000 engagements. [224]

There is scope for actors to use AI tools in future to scale up disinformation campaigns such as this and to re-enforce anti-homosexual biases and hate speech online. Indeed, TikTok's algorithm has already been singled out for criticism by African rights groups[225] for perpetuating hate speech. There is no evidence that Russia has initiated such campaigns, but warrants investigation.

220 https://afrinz.ru/en/.

221 https://www.rt.com/africa/.

222 Pro Kremlin Disinformation in Zimbabwe: Newsday, 11 June 2024, https://www.newsday.co.zw/thestandard/opinion-analysis/article/200028249/pro-kremlin-disinformation-in-zimbabwe-a-mix-of-state-and-non-state-actors-influence.

223 Luchenko: Propaganda in holy orders: Africa, Ukraine and the Russian orthodox church, European Council on Foreign Relations, 20 September 2023, https://ecfr.eu/article/propaganda-in-holy-orders-africa-ukraine-and-the-russian-orthodox-church/.

224 DRC battles disinformation as it prepares for elections, African Digital Democracy Observatory, 16 December 2023, https://disinfo.africa/drc-battles-disinformation-during-2023-elections-6f3c7e4b8a42.

225 TikTok's algorithm of Gay Hate and Uganda's anti-homosexuality bill, 7 June 2023, https://minorityafrica.org/tiktoks-algorithm-of-gay-hate-and-ugandas-anti-homosexuality-bill/.

Perception of AI content and AI Disinformation in Africa

While there has been substantial research undertaken on the impact of AI on Africa's development, in particular with respect to shaping the future of business[226] and the African Union's continental AI strategy,[227] which focuses heavily on economic development, there is little research on risks and even less on AI and disinformation in Africa.

Polls on Political Disinformation

On disinformation in Africa, the company KnowBe4's 2024 Political Disinformation in Africa survey[228] focused on five African countries: Botswana, Kenya, Mauritius, Nigeria and South Africa. Its key findings (based on a sample of some 500 respondents) include the following:

- 84% of respondents primarily rely on social media for news.
- 80% are concerned about the prevalence of "fake news" on social media, yet rely on it for news.
- 82% of respondents feel confident in their ability to distinguish between real and fake news.

The dependence on social-media platforms for news would appear to be linked to the fact that more news content from traditional media sources is now located behind paywalls, as the industry faces increasing challenges of

226 Alexander: Exploring the Artificial Intelligence Footprint in Africa, Henley Business School, 2024, https://content.henleysa.ac.za/white-paper-exploring-the-artificial-intelligence-footprint-in-africa-kelly-alexander.

227 https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy.

228 https://www.knowbe4.com/hubfs/2024-Political-Disinformation-Africa.pdf.

funding.[229] Yet, the way social-media algorithms rank material, priority is given to material which is most popular, not material which has been verified.

Stanford University Internet Observatory's Renée diResta explains in a South African newspaper that, "Social media algorithms will show you what you want to see, but they don't have any kind of value judgement and this is where we see things like radicalisation beginning to be an increasing problem because the recommendation engine does not actually understand what it is suggesting".[230]

Polls on AI and Africa

The World Risk Poll 2021, which uses the Gallup methodology,[231] found that "people living in low- and middle-income countries are more likely to say that AI would mostly harm people in their countries over the next 20 years". In East Africa, the report found that 51% of those surveyed indicated their belief that AI was mostly harmful. The researchers conclude that "these findings suggest that in some countries, the use of AI may be seen as a means of reinforcing and amplifying existing global biases and discriminatory power structures, causing them more harm than good."

While the report did not specifically look at disinformation, it is worth highlighting an observation that the more forms of discrimination people had experienced based on religion, ethnicity, sex, religion or disability the more they are likely to worry about "the harmful use of their personal information online, whether by their government, criminals or private companies".

Given the weaponisation of AI to tap into existing societal cleavages in order to sow division, support specific narratives or create chaos, this apparent mistrust of AI arguably reflects mistrust of governments more broadly and concerns over whether it will deploy AI to surveil individuals or curtail freedom, for example, through internet shutdowns – a phenomenon that has been

229 Reuters Institute: Digital News Report 2024, https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024.

230 Anatomy of a disinformation campaign: The who, what and why of deliberate falsehoods on Twitter, *Daily Maverick*, February 2021 https://www.dailymaverick.co.za/article/2021-02-09-anatomy-of-a-disinformation-campaign-the-who-what-and-why-of-deliberate-falsehoods-on-twitter/.

231 A Digital World Perceptions of risk from AI and misuse of personal data, World Risk Poll, https://wrp.lrfoundation.org.uk/publications/a-digital-world-perceptions-of-risk-from-ai-and-misuse-of-personal-data#:~:text=People%20and%20countries%20with%20lower,in%20the%20next%2020%20years.

observed in a number of African countries.[232] This has prompted a debate whether access to the internet is a human right at a time when several African governments consider the internet to be an extension of the state. [233]

Indeed, the regional court representing the Economic Organisation of West African States (ECOWAS), in a July 2022 ruling,[234] said that the Nigerian government was guilty of violating freedom of expression when it shut down access to the media platform Twitter (now X) in June 2021, following a decision by Twitter to delete a post by President Buhari, which it felt transgressed its community guidelines.

The broader mistrust of AI in less-developed nations or middle-income countries may also be reflective of a growing mistrust in mainstream traditional media. Work undertaken to investigate how fake news impacts trust in Kenya, Nigeria and South Africa[235] offered the following conclusions:

- Rumour is most likely to be an alternative source of information when conventional media lacks credibility.
- Patrimonial relationships in African societies also erode trust in journalists' independence.

Therefore, power dynamics, issues of media ownership, governance and transparency in traditional media in African settings may shape how online disinformation narratives (and AI-driven disinformation) are received by citizens, as well as how they impact traditional media. The media environment is multi-layered and complex, depending on the African country in question, and in practice may be markedly different from many European countries.

We have also observed growing advocacy and "African agency" regarding the need to protect fragile democracies on the continent from the unintended consequences of AI-driven disinformation. In an analysis article for Foreign

232 Mlaba: Africa's Internet Shutdowns: Where, why and How Do They Happen, *Global Citizen*, 9 May 2024, https://www.globalcitizen.org/en/content/africa-internet-shutdowns-impact-human-rights/.

233 Allen: https://www.dailymaverick.co.za/article/2022-06-07-africa-needs-to-adopt-common-democratic-positions-on-nations-internet-freedoms/.

234 https://www.mediadefence.org/news/ecowas-nigeria-twitter-blocking/.

235 Wasserman and Madrid-Morales An Exploratory Study of "Fake News" and Media Trust in Kenya, Nigeria and South Africa, *African Journalism Studies*, 1 August 2019, https://www.tandfonline.com/doi/abs/10.1080/23743670.2019.1627230.

Policy,[236] Abdullahi Alim, the CEO of the Africa Future Fund, warned that the ranking systems of social-media platforms such as Facebook (Meta) "continued to provide a platform for violent content" and warned that AI could "light the match of the continent's next war". The case of "inciteful" online narratives during the war in Ethiopia's Tigray region between 2020 and 2022 is cited as an example in which "Meta's own algorithm to detect hate speech was unable to perform adequately in either of Ethiopia's most widely used languages, Amharic and Oromo". The fear is, says Alim, that content-moderation tools developed by social-media platforms may be compromised by AI systems that effectively attack those tools. The prospect of "foreign mercenaries and adversarial groups leveraging adversarial AI to sow chaos and disorder" is a major threat to Africa's development and democratic stability.

The biggest threat may well be the multiplier effect that AI is able to achieve by generating and amplifying content at scale (including disinformation content), rather than the threat of deepfakes, but even that ignores that disinformation can exist and be successful in Africa without the need for AI. This point is echoed by Prof Guy Berger, a former Director at UNESCO, leading its work in communications and ICT, who was interviewed for this paper[237] and now consults for Research ICT, who said that, as yet, most political parties are relying on "old style data … They look at opinion polls, surveys, they might have a little bit of a campaign aimed at building a consensus in a particular ethnic constituency, but the idea that they at this stage would use ChatGPT to tailor messages and hope that the algorithms will take these to the particular target audiences. I don't see that yet". He also believes that "because many African audiences are exposed to a lot of other communications from old style media, especially radio, in rural areas, and that's under control of governments … and people are so familiar with media that's biased they give an appropriate discount to messaging. Now, the extent to which they apply this to social media is unknown."

Other research on the impact of geopolitically focused disinformation[238] indicates that, at present, Africa's citizens pre-existing world view of global events has a bigger influence on their opinions than social-media messaging.

---

236 Alim: How Africa's War on disinformation can save democracies everywhere, *Foreign Policy*, 21 June 2024, https://foreignpolicy.com/2024/06/21/adversarial-ai-deepfakes-africa-drc-ethiopia-war-disinformation-democracy/.

237 Prof Guy Berger, Independent consultant, Research ICT Africa, Interview September 2024.

238 Madrid-Morales and Wasserman: The Geopolitics of Disinformation: Worldviews, Media Consumption and the Adoption of Global Strategic Disinformation Narratives, *International Journal of Public Opinion Research*, 8 July 2024, https://academic.oup.com/ijpor/article-abstract/36/3/edad042/7709016.

However, it may well be that more intimate peer-to-peer messaging platforms, such as WhatsApp, become the focus of AI-assisted information operations over time, because material is shared between “trust networks” in contrast to social-media platforms such as X, TikTok and Facebook.

Deepfake AI Journalism, AI Influencers and Attacks against Journalists

Generative AI can revolutionise the global information space, impacting all forms of political communication, content creation, and presentation, including journalism and influencer activities.

AI and Journalism

Contrary to many studies who analyse the effects of genAI on journalism, this study deals only with deepfake- and disinformation-related developments. Here, it is important to notice a significant "AI-gap": while traditional, quality media organisations have been struggling to draft and implement ethical and transparent guidelines of ethical AI use in journalism, low-quality media, tabloids, propaganda and disinformation outlets are not bound by any limitations and restrictions and try to utilise the full potential of genAI for their malignant purposes. For instance, Russia's foreign propaganda broadcaster RT has been caught using fully AI-generated "digital presenters" (that is, AI avatars of non-existent persons) as news anchors for their Spanish-language programmes. Chinese news channels have been using similar technologies for broadcasting on television, radio and social media. Likewise, Channel 1, established in Los Angeles in 2024, claims to be a serious media outlet running entirely on genAI (that is, ChatGPT in combination with deepfake presenters) for both content creation and presentation. Furthermore, investigations have found thousands of websites using genAI (often ChatGPT) to run automated "news" sites, either rewriting old content or spreading disinformation. In the US, for example, these investigations have shown that the number of such fake, AI-powered local news websites outnumber authentic local news outlets.[239] Notwithstanding the future of ethical and transparent genAI use

239 C.f. https://www.newsguardtech.com/press/sad-milestone-fake-local-news-sites-now-outnumber-real-local-newspaper-sites-in-u-s/.

in quality journalism, the global information space already faces a serious threat by malignant, unethical actors leveraging the full potential of genAI in content creation and content presentation and dissemination for financial or political purposes. Hence, the potential of "over-pollution" of the global online information space with AI-generated content exists. Some negative projections suggest that, by 2026, 90% of online content could be AI-generated, while automated online behaviour already constitutes the majority of online activities.[240] This proliferation of AI content could profoundly impact political news, information dissemination, and societal dynamics, potentially becoming one of the most serious long-term risks of genAI.

Across Africa, the use of AI in journalism is limited. The study AI Journalism and Public Interest Media in Africa[241] found that AI systems are largely functionally "deployed in content/news gathering, content processing, content/news distribution and audience engagement." ChatGPT is being experimented with in some South African newsrooms but the lack of robust African AI models to date may limit its use.

There has also been some experimentation in settings such as Zimbabwe, where deepfake avatars deliver news content,[242] but these have been rare instances. Scholarship on the impact of such innovation on audience trust found that some viewers considered it a useful innovation while others raised concerns about the "lack of human emotion" and potential impact on traditional journalist's jobs. The potential for AI avatars to displace journalists has a logic, given the financial pressures on many newsrooms globally and the shifting of advertising revenues onto online spaces.[243]

As discussed above, Francophone Africa has also been the target of an Israeli firm developing "deep avatars",[244] which are "fake investigative journalists", with online personas, social-media accounts and, in some cases websites,

240 C.f. Nina Schick: Deep Fakes and the Infocalypse, Ottawa, 2020.

241 Ogola: Ai, journalism, and public interest Media in Africa, International Media Support, May 2023, https://www.mediasupport.org/wp-content/uploads/2023/06/AI-Africa-Report-2023R-Double-Spread-View.pdf.

242 Ndlovu: Audience Perceptions of AI-driven news presenters: A case of 'Alice' in Zimbabwe, *Media, Culture and Society*, 12 August 2024, https://journals.sagepub.com/doi/10.1177/01634437241270982.

243 The shift to social media: Increasing pressures on news media amid falling revenues, Code for Africa, 18 June 2024, https://medium.com/code-for-africa/the-shift-to-social-media-increasing-pressures-on-news-media-amid-falling-revenues-86199f7d4396.

244 Robot wars: How to build a bot to subvert elections, African Digital Democracy Observatory, 6 March 2023, https://disinfo.africa/robot-wars-how-to-build-a-bot-to-subvert-elections-9f739411aa39.

which produce content for the purpose of both intelligence-gathering and narrative control. For example, the material produced by Israeli company Percepto International's avatar called Anita, who is presented as an earnest pan-Africanist, was disseminated both on a dedicated website, Pour La Verité, well as material which gets syndicated by African mass media, enabling the avatar to reach an audience of millions.[245]

While the impact may be that material can be disseminated at scale and across multiple languages, there is the possibility that it undermines traditional journalism with all its associated professional codes of conduct and ethics, at a time when bolstering traditional media and critical thinking is widely considered to be one of the most important counter measures to disinformation in Africa. Furthermore, in an op ed, Prof Ylva Rodny-Gumede,[246] from the University of Johannesburg, argues that the efficiencies of using AI models in African newsrooms may be "offset by additional demands on them (journalists) in a resource-starved environment that mitigates oversight and fact-checking of AI-generated data." Furthermore, one can argue there is a danger that fact-checking – which is time-consuming and costly - becomes a casualty of the desire by some traditional newsrooms to compete with digital news sites – including avatar-driven sites – where speed may trump accuracy in reporting. Therefore, Prof Rodny-Gumede warns that AI must be used with caution to avoid journalists becoming "algorithmic agents", that is slaves to ensuring algorithmic privilege in social-media rankings over verified news content.

AI and Influencers

Influencers and content creators face similar impacts of genAI regarding content generation and presentation as do journalists and media organisations, but, in general, are faced with fewer ethical restrictions and constraints. The risk of influencers either being bought or voluntarily engaging in spreading political disinformation and propaganda (including unlabelled AI content) became apparent during many elections in 2024 (for example, Romania and the US). As leaks show, Russian PR companies contracted by the Kremlin to influence elections (US and Germany) made lists of favourable and unfavourable influencers worldwide (containing up to 2000 names) to be bought, influenced, or watched or discredited and impeded. In Africa, an investigation found lists of influencers across countries that received money from official Russian

245 https://disinfo.africa/robot-wars-how-to-build-a-bot-to-subvert-elections-9f739411aa39.

246 Rodny-Gumede Deepfakes: journalism, media and democracy in the age of AI, University of Johannesburg, 22 November 2024, https://news.uj.ac.za/news/deepfakes-journalism-media-and-democracy-in-the-age-of-ai/.

diplomatic institutions or the infamous Wagner Group (and its successor the Africa Corps) to spread and share content (similarly to the practices of paying journalists to create and publish information).[247] Inevitably, this will include more and more the creation and spread of AI content (as already evident, yet still insignificant in the US or Romanian elections).

Furthermore, just like journalists, genAI has the potential to replace human influencers and automate and imitate their work. "Virtual Influencers" or "AI-Influencers", entirely AI-created and powered avatars using LLMs and deepfake technology, have already gained millions of followers in countries like China, Brazil, the USA, and India and are constantly expanding.[248] Here, the threat of automated AI influencers being used by governments and private organisations to create automated AI-powered disinformation and to manipulate is evident.[249]

Deepfake Attacks against Journalists and Influencers

Journalists and influencers are – together with politicians – among the most frequent and prominent victims of deepfake attacks. The forms of attack most often reported in the past two years in Europe and Africa are:

1. Deep porn attacks
2. Image attacks
3. Scams
4. Advertisements
5. Impersonation attacks

Between these five categories, there are significant overlaps. Deep porn attacks against journalists and influencers – which almost only target women – may be coordinated image attacks or triggered by financial motives (click-baiting, mal-advertisement, etc). Of all five categories, deep porn attacks against female journalists and influencers are by far the most common. As studies show, "deep porn attacks" are a particular threat to female social-media influencers

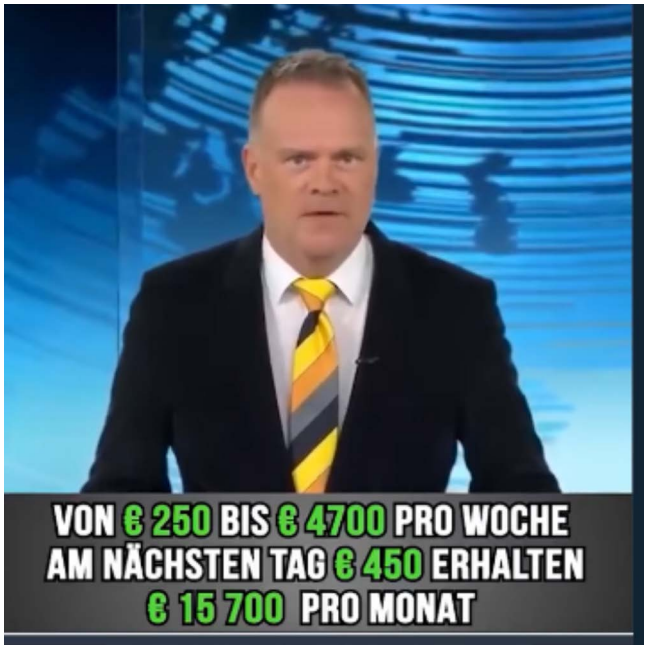
247 C.f., for example, https://www.spiegel.de/ausland/russland-ein-propagandist-packt-aus-ueber-die-desinformationskampagne-in-afrika-a-fb41d122-3e4f-48b6-8e8f-4fba32818df3?giftToken=dae8f1a4-463f-4930-904a-964f65b240b6.

248 See *in extenso*: https://www.virtualhumans.org/.

249 In reality today, efforts and resources needed to power and uphold AI influencers with huge followings are yet too high, cost- and labour-intensive to be applicable on a mass scale. Yet, these costs and efforts inevitably decrease significantly within the years that follow.

and celebrities.[250] One of the most prominent examples, where aspects of political manipulation, image attacks and financial motives coincide, are the deep porn attacks against US celebrity Taylor Swift during the US presidential election 2024.[251] Deep porn attacks against journalists in Europe have also been reported in Albania, Serbia, Bulgaria and many other countries (see list of case studies above). Damaging the image and reputation of journalists by using deepfake impersonations with discrediting content (porn or not) can be a strategy of political organisations and PR companies, but also, for example, criminal or terrorist organisations and other organised actors. In Germany, for example, far-right political activists have produced and spread discrediting deepfake audio clips in the name of the most prominent 8 O'clock news show.

Other cases of deepfake attacks have included waves of deepfakes of popular news anchors and journalists (along with politicians and celebrities) to advertise either crude financial schemes and scams or allegedly advertising products and brands. Such instances have appeared nearly everywhere in Europe during the past two years and the number is rapidly increasing.



250 C.f., for example, Security Heros (ed): 2023 State of Deepfakes: Realities, Threats, and Impact, 2024 (https://www.securityhero.io/state-of-deepfakes/#key-findings).

251 C.f., for example, https://theconversation.com/taylor-swift-deepfakes-new-technologies-have-long-been-weaponised-against-women-the-solution-involves-us-all-222268.

Another form of deepfake impersonation attack that has already been reported is the risk of professional disinformation actors creating fake journalist or influencer persona (that is, a deepfake avatar of a non-existent journalist or influencer) who writes and speaks (online) to impersonate a real journalist or content creator and is then used to produce and spread disinformation. As discovered by the StoryKillers investigation, an Israeli cybersecurity company (Percepto International) used at least one deepfake-impersonated journalist (so-called Anita Pettit, an alleged French-Ghanaian investigative journalist) over the course of many years, to smear local politicians and international organisations in Africa with fabricated "revelations".[252]

The reasons journalists and influencers (just like politicians and celebrities) are frequent victims of deepfake attacks are five-fold:

1. Training data (audio and visual) of these persons is readily available free of charge and in abundance.
2. Journalists and influencers enjoy a certain (or even high) degree of popularity and at least some level of trustworthiness by their audiences.
3. By nature of their work, journalists and influencers are ideal for spreading information.

252 See *in extenso*: Amanda Sperber / Justin Arenstein: Robot wars: How to build a bot to subvert elections. How disinfo-for-hire companies use armies of fake avatars to infiltrate & manipulate the media (https://disinfo.africa/robot-wars-how-to-build-a-bot-to-subvert-elections-9f739411aa39).

4. Journalists and influencers, by the nature of their work, often deal with sensitive matters and interests of a political, societal, financial, economic or criminal nature.
5. Personal attacks against journalists and influencers are an easy tool to discredit or suppress unwanted information.

There have also been instances in which prominent journalists or news anchors have been the focus of deepfakes or impersonation accounts for the purposes of fraud. For example, in South Africa, one of the most prominent journalists, Leanne Manas,[253] found her identity being used to support slimming products on TikTok and fraudulent investment schemes, culminating in false claims that she had been arrested. To what extent she was singled out as a public figure rather than specifically because she is a journalist is not clear. However, the use of deepfakes among "trusted individuals" in the public eye is a practice which is increasingly gaining traction, as disinformation actors mirror techniques observed in other settings. Furthermore, as representatives of the "fourth estate" attacks on traditional journalism and an equivalence that is sometimes bestowed upon online influencers who are not held accountable by professional codes of conduct regarding sourcing of material or codes of ethics, may serve to erode trust in democratic institutions.

253 https://mg.co.za/news/2024-07-21-leanne-manas-speaks-out-about-battling-deepfake-scams-and-identity-theft/.

Laws, Norms and Regulation of AI-Disinformation

In Europe, but also globally, various laws and other regulative acts (for example, codes of conduct for election campaigning and other normative documents) have addressed the issue of AI disinformation. The most prominent laws are:

- EU AI Act [254]
- EU Digital Services Act (DSA) [255]
- Five laws in California on AI political ads, disinformation and deepfakes (for example, AB 2655: Defending Democracy from Deepfake Deception Act of 2024; AB 2839: Elections: Deceptive Media in Advertisements; AB 2355: Political Reform Act of 1974: political advertisements: artificial intelligence; SB 926: Criminalising AI-Generated Sexually Explicit Images). [256]
- The UK Online Safety Act and Criminal Justice Bill. [257]

The most important provisions of these laws regarding AI disinformation are:

EU AI Act

The EU AI Act is the first comprehensive framework for regulating artificial intelligence, including issues related to disinformation and deepfakes. It categorises AI systems based on their risk level, with deepfakes classified as high-risk in sensitive areas such as media, advertising, and political campaigns,

254 See https://artificialintelligenceact.eu/.

255 See https://www.eu-digital-services-act.com/.

256 C.f. https://www.linkedin.com/pulse/california-deepfake-laws-christopher-nehring-s7f5f/?trackingId=roeEHgMKXjAgj0TyU%2Fnm6A%3D%3D.

257 C.f. https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer.

necessitating strict oversight and regulation to prevent manipulation and misinformation. The Act prohibits certain applications that deceive or manipulate individuals, including deepfakes used to spread false information or damage reputations. Transparency requirements mandate that any content generated or altered by AI must be clearly labelled as such. Violations of these provisions can result in substantial penalties, including fines of up to €35 million or 7% of a company's global annual revenue.

Regarding disinformation in general, the AI Act refers to and must be read in relation to another EU law that had been passed before the hype and rise of genAI – the ***Digital Services Act (DSA)***. The DSA is one of the centre points in the EU's strategy to combat disinformation, serving as a key regulatory framework. The AI Act references the DSA and the latter mandates that very large online platforms (VLOPs) must actively remove disinformation and illegal content from their services. Under Article 40, the DSA provides vetted researchers access to both public and non-public data from VLOPs, facilitating research into systemic risks of disinformation to improve oversight of platform practices. The DSA imposes strict penalties, including fines of up to 6% of a platform's annual revenue if they fail to adhere to their obligations, which include deleting disinformation. These provisions are binding for AI disinformation as well as non-AI disinformation.

California and UK

Within the US, the ***State of California*** has also been active in passing AI disinformation legislation, particularly deepfake-related legislation. Two of those focus on the threats posed by deepfake pornography (SB 926: Criminalising AI-Generated Sexually Explicit Images and SB 981: Social Media Platform Responsibilities). These laws make it a felony offence to possess or distribute AI-generated child sexual abuse images, make it illegal to create or share AI-generated sexually explicit deepfakes of a person without their consent and require social-media platforms to establish reporting mechanisms for victims and to remove such content.

The other three laws (AB 2655: Defending Democracy from Deepfake Deception Act of 2024; AB 2839: Elections: Deceptive Media in Advertisements; AB 2355: Political Reform Act of 1974: political advertisements: artificial intelligence) tackle the use of deepfakes and AI content in elections, campaigning, political advertisement and deception. These laws, for example, require platforms with at least 1 million California users to block posting of deceptive election-related content during specified periods before and after elections, label

content as inauthentic, fake, or false, and remove reported deceptive content within 72 hours. They also prohibit the distribution of materially deceptive election communications 120 days before and 60 days after elections, require disclosure statements in political advertisements that the advertisement was generated or substantially altered using artificial intelligence and prescribe specific formatting requirements for these disclosures based on the medium of the advertisement (for example, radio, video, or print).

The ***UK Online Safety Act (OSA)*** and the proposed amendments in the Criminal Justice Bill (to be drafted soon) in the UK also target deepfakes, deepfake pornography and AI-generated disinformation. The OSA criminalises the sharing of non-consensual intimate images, including deepfakes. The Criminal Justice Bill introduces new offences targeting the creation of sexually explicit deepfakes, making it illegal to produce such content without consent, regardless of whether there is an intent to share it. Additionally, both laws emphasise the need for transparency and accountability in digital-content creation.

Declarations and Other Non-binding Documents

Other than laws, there are also several normative acts that address the issue of AI-powered disinformation, e.g.:

- G7-Hiroshima Declaration
- Several OECD Frameworks
- The Munich Tech Accord
- Code of Conduct for the 2024 European Parliament Elections

The ***G7 Hiroshima Declaration on AI*** from 2023 outlines several provisions and recommendations aimed at addressing disinformation and deepfakes. It emphasises the need for a comprehensive policy framework that promotes safe and trustworthy AI systems while mitigating risks associated with mis- and disinformation in the context of generative AI technologies. It calls for enhanced transparency, requiring organisations developing advanced AI to clearly label AI-generated content, thereby enabling users to evaluate the authenticity of information. Additionally, the G7 leaders advocate international cooperation in developing reliable methods for detecting and countering deepfakes, alongside fostering research initiatives that focus on combating disinformation. The

declaration also stresses the importance of multi-stakeholder engagement, involving governments, academia, and civil society.[258]

Similarly, the Organisation for Economic Cooperation and Development (OECD) has touched upon AI disinformation and deepfakes in its ***OECD AI Principles, OECD Recommendation of the Council on Artificial Intelligence, and the OECD Framework for the Classification of AI Systems***:

1. The OECD AI principles emphasise the importance of transparency and responsibility in AI development, advocating clear labelling of AI-generated content to mitigate misinformation risks. They also highlight the need for robustness and security in AI systems to prevent misuse, including the creation of deepfakes that can spread false information.[259]
2. The OECD Recommendation of the Council on Artificial Intelligence (updated in 2024) addresses the need for information integrity in genAI technologies. It emphasises responsible business conduct and calls for enhanced transparency regarding AI-system capabilities, limitations, and data sources, ensuring that potential disinformation risks are tackled.[260]
3. The OECD Framework for the Classification of AI Systems aids policymakers in classifying AI systems according to their potential impacts, including the risk of disinformation. It provides a structured approach to assessing the characteristics of AI systems that may contribute to misinformation.[261]

Another normative document addressing the risks of deepfakes and AI disinformation was the ***Munich Tech Accord***, a formal commitment signed during the Munich Security Conference 2024 by 27 major tech companies (Microsoft, Google, Meta, Amazon, OpenAI, TikTok, and others) to cooperate to detect and counter harmful AI content. The Tech Accord to Combat Deceptive Use of AI in 2024 Elections outlines provisions to counter misleading election-related AI-generated content. The signatories commit to preventive measures, for example, research and deploying technologies to limit the generation of deceptive AI election content. They also agreed to attach technological signals to generated content (for example, watermarks) to ensure the provenance of content can be identified. The accord includes commitments to detect deceptive content through collaborative efforts and to provide "responses"

258 See https://www.politico.eu/wp-content/uploads/2023/09/07/3e39b82d-464d-403a-b6cb-dc0e1bdec642-230906_Ministerial-clean-Draft-Hiroshima-Ministers-Statement68.pdf.

259 See https://www.oecd.org/en/topics/artificial-intelligence.html.

260 See https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449.

261 See https://www.oecd-ilibrary.org/fr/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.

when such content is funded. Furthermore, it promoted public-awareness initiatives to educate citizens about media literacy and the risks associated with AI manipulation, aiming to build resilience against disinformation in democratic processes.[262]

One last example for normative documents guiding how critical stakeholders should approach the topic of AI disinformation is the ***"Code of Conduct for the 2024 European Parliament Elections"***.[263] This non-binding code of conduct was signed by all parties represented in the EP and established a framework for ethical campaigning, particularly addressing the challenges posed by AI disinformation and deepfakes. Signatories committed to adhering to principles of truth and accuracy in their communications, pledging to counter mis- and disinformation. The code emphasises the ethical use of technology, including AI, and encourages parties to refrain from producing or disseminating misleading content, particularly that which manipulates public perceptions through deepfakes. Additionally, it promoted transparency by requiring parties to clearly label AI content in their campaigns. As mentioned above, all but one of the fractions adhered to this code (members of the far-right fractions in France, Italy and Germany were caught using AI images, songs and posters without labelling them).

Laws, Norms and Regulation of AI-Disinformation in Africa

Cybercrime

A growing number of African countries are taking steps to protect themselves against cybercrime. The African Union Convention on Cyber Security and Personal Data Protection, commonly referred to as the Malabo Convention, provides an overarching framework for regulating cyber-related offences. As of May 2023, 15 African countries have ratified the convention. It requires 28 to come into force.

A UN Cybercrimes Treaty[264] is currently being negotiated and may become the first binding UN instrument on a cyber issue. However, controversially the proposed treaty contains some cyber-related offences which rights campaigners fear may impinge on human rights and criminalise online free

262 See https://securityconference.org/en/aielectionsaccord/.

263 See https://commission.europa.eu/document/download/bebd9b72-fbb9-42f3-bcea-dbace7e0650f_en?filename=Code+of+conduct+for+2024+European+elections_final.pdf.

264 C.f. https://unu.edu/cpr/blog-post/understanding-uns-new-international-treaty-fight-cybercrime.

speech. For example, China[265] has proposed an offence of "dissemination of false information", which is highly contested and if passed would put the onus on treaty signatories to introduce legislation against the spreading of false information "that could result in serious disorder".

South Africa's cybercrimes Act 2020, like similar legislation in Kenya, Nigeria and Ghana, identifies a range of cyber offences that include content-related offences into which some AI disinformation or deception activities may fall. Traditional crimes such as forgery, fraud and identity theft (including the use of so-called "revenge porn", which may have a deepfake component) are covered by this legislation but implementing it has been slow due to a lack of government capacity, training and technical know-how. Furthermore, the focus of this new law is likely to be more on criminal activity and criminal actors rather than on broader digital-influence operations.

Artificial Intelligence

With the rapid adoption of AI, the African Union has established an AU Continental Artificial Intelligence Strategy.[266] The focus is on AI as a "strategic asset" able to leapfrog economic development and drive "revolutionary changes in healthcare, agriculture, finance and education". However, there is little direction in what the "responsible use of AI" may look like.

South Africa has established its own National Artificial Intelligence Policy Framework,[267] which provides the conceptual basis for future policies on addressing AI. However, it does not put in place the risk framework that forms the basis of the European Union AI act. From February 2025, the EU will ban various AI practices, in particular in the fields of public surveillance and AI-based predictive risk assessments for crime-prevention activities.

AI tools rely on vast amounts of training data to be relevant in a particular context or region. This may be one of the reasons the use of AI-driven disinformation in Africa has been limited; for example, the deepfakes developed are not convincing because the programs are largely developed in the Global North. With the assumption that more "relevant" AI training data from Africa may be acquired in future (by both indigenous and non-South

265 See https://therecord.media/china-proposes-un-treaty-criminalizing-dissemination-of-false-information.

266 https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy.

267 https://www.dcdt.gov.za/sa-national-ai-policy-framework/file/338-sa-national-ai-policy-framework.html.

African tech companies), the focus has been on supporting data governance and ensuring better data access to enable a greater understanding of Africa and South Africa's information ecosystem.

In a recent interview with ISS Today,[268] Prof Guy Berger – a consultant with Research ICT Africa – said, "One of Africa's biggest challenges is accessing data, especially data on social-media platforms". He points to the African Commission on Human and Peoples' Rights' recent adoption of an African Alliance for Access to Data resolution[269] to allow more scrutiny over how social-media platforms, in particular, work.

It is increasingly being argued in policy circles that, unless social-media companies accede to a degree of regulation, the default position of some African governments may be greater digital authoritarianism with the imposition of all-out bans, as was seen in Nigeria[270] during the End SARS campaign in 2020 and more recently in Mauritius in November 2024.[271] This can have harmful consequences for the right to free speech and other pillars of democracy, as well as negatively impacting digital commerce.

Rwanda has declared public data a national asset [272] as part of its drive to harness digitisation for rapid economic development. In practice, this means investing heavily in digital infrastructure, the introduction of cloud-based data-sharing across government departments, and digital governance. Like many other countries across the continent, tough privacy laws have been enacted which broadly reflect the GDPR regulations in Europe. It is, however, ironic that Rwanda is one of the countries singled out by activist groups in which the government of Paul Kagame is accused of using AI for disinformation targeting mainstream journalists and opposition voices.[273]

268 https://issafrica.org/iss-today/artificial-intelligence-regulation-in-south-africa-prioritising-human-security.

269 https://www.techpolicy.press/can-an-alliance-get-access-to-platform-data-for-african-researchers/.

270 https://africlaw.com/2023/06/22/gameofphones-examining-the-social-media-regulatory-regimes-across-africa/.

271 https://www.aljazeera.com/news/2024/11/1/mauritius-blocks-social-media-until-after-election-amid-wiretapping-row.

272 Ingabire: Rwanda's Data governance Navigating data governance in the public sector May 2 2024. https://www.brookings.edu/articles/rwandas-data-governance-navigating-data-governance-in-the-public-sector/.

273 https://open.clemson.edu/mfh_reports/5/.

Conclusion

An overview of the most important European and global normative documents show a – surprisingly small – set of key measures that are referred to as paramount in mitigating the risk of AI-powered disinformation:

A) *Transparency*: AI content must be clearly labelled as "generated or manipulated with AI".

B) *Responsibility of Social-Media Platforms*: Most laws and other documents identify social-media platforms as key actors for detecting and deleting deepfakes, malicious AI content and disinformation in general.

C) *Prohibition* of "deep porn" (creation and sharing).

D) *Access to data*: Only the EU legal framework obliges large online platforms to grant access to their data to researchers and journalists, thus providing the basis for monitoring of threat developments and oversight over the implementation of risk-mitigation measures. As this study confirms, access to the data of online platforms is a key component of assessing risks and developing solutions.

Thus, it becomes clear that all legal approaches so far have focused on the "detection challenge" of AI content (which is referred to as a responsibility of online platforms). Yet, the impact of genAI on disinformation, on the one hand, exceeds the content level (that is, not only the creation of content but also its distribution) and, on the other hand, is also not limited to social-media platforms. Furthermore, labelling AI content to make it distinguishable will soon prove ineffective in an online environment in which most content will be AI-generated or AI-edited. The current motto "transparency to ensure informed decision-making" will also not be able to achieve its intended effects, if audiences are not educated about AI and cannot make sufficient sense of "AI labels".

Several other solutions and risk mitigation measures on a technical level (for example, prompt storage, internal and robust filtering and content moderation of genAI systems) are, for example, listed only in non-binding, voluntary "recommendations" (for example, the G7 Hiroshima declaration), but not turned into mandatory regulations. Algorithmic regulation, demobilisation and demonetarisation as measures to tackle "engineered mobilisation" of disinformation are measures also not mentioned or recommended in any binding or non-binding normative documents.

The reality of 2023 and 2024 has already shown that all laws and norms have so far faced significant challenges: In the EU, the DSA came into force in 2022 and the AI Act in 2024, both including transition periods up until 2024 and 2026 respectively for member states to enact national legislation and online platforms to take care of implementation. Transparency rules (that is, labelling of AI content), for example, have already made their way into community standards, terms and conditions of social-media platforms in the EU; yet, so far, they are not being implemented and are enforced erratically. Others, such as the code of conduct for the EP elections, were openly disregarded by malicious actors (that is, FIMI actors, as well as domestic radical or extremist organisations) without consequences. Monitoring the threat of AI content in real-time, enforcing existing rules, overcoming technological challenges and widening the toolbox of normative provisions to include mandatory, built-in technological safeguarding measures can thus be identified as the greatest challenges to normative acts dealing with AI disinformation.

Social-Media Platforms and AI Content

Most normative acts, as well as research and expert investigations, have identified online platforms as the most important actor when it comes to AI disinformation. This is because these are the online spaces where AI disinformation is uploaded, spread and where the distributors and their target audiences meet. In the following, we provide an overview and analysis of their AI-content and disinformation policies:

Facebook and Instagram

Meta's updated Community Guidelines for Facebook and Instagram include sections on AI-generated content, particularly focusing on misinformation. They state that users must label content that is digitally created or altered and may mislead other users, especially if it pertains to public importance. Meta plans to implement informative labels for AI-generated content and is collaborating with industry partners to develop unified technical standards for identifying AI content. The platform emphasises that its existing policies against harmful content also apply to AI-generated content, and it employs AI systems to detect and mitigate violations.[274] These same standards are outlined when accessing Meta's platforms using an African IP address. In May 2024, Meta updated its labelling policies[275] with "Made with AI" labels appended to content when Meta detects "industry standard AI image indicators or when people disclose that they're uploading AI generated content". However, in an African setting, this guardrail is dependent on Meta being able to detect African-focused AI in the first place and the amount of training data in Africa is limited. It also is

274 See https://transparency.meta.com/de-de/policies/community-standards/ and https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/.

275 https://www.socialmediatoday.com/news/meta-ai-labeling-policy-expose-generated-content/712497/.

dependent on users' self-disclosure, which assumes a certain level of digital literacy, which may not be present in low-resource settings.

WhatsApp

Despite being part of Meta, WhatsApp's Community Guidelines focus primarily on user privacy and security rather than explicitly addressing AI-generated content or deepfakes. There are no specific provisions regarding AI disinformation and AI content. However, the platform emphasises that users should not share false information or engage in deceptive practices. WhatsApp encourages users to verify information before sharing and provides tools for reporting suspicious messages or spam.[276] Given the preference for WhatsApp communication in many African settings, there is a case for community guidelines to address AI content issues that are context-specific. According to the World Economic Forum, 172 million people in Africa use either Meta's Facebook messenger platform or WhatsApp and "with the exception of a few countries, WhatsApp was easily the most popular platform across Africa".[277]

YouTube

YouTube has introduced new policies requiring creators to disclose when their videos contain AI-generated or altered content, particularly deepfakes. This includes mandatory labelling in the video description for any synthetic media that depicts events or speech that did not occur. Additionally, individuals can request the removal of AI-generated content that features their likeness without consent. YouTube employs a combination of AI and human moderation to ensure compliance with these guidelines, aiming to balance the creative potential of AI with user safety. The platform underscores its commitment to responsible content creation while addressing the risks associated with disinformation. Furthermore, Google says it will introduce penalties for serial offenders, those are content creators who repeatedly fail to disclose videos that are "meaningfully altered or synthetically generated, including content removal". However, it adds that these penalties will not apply immediately, as Google wants users to "learn the news requirements and give us feedback".[278] While clearly Google is responding to greater fears about disinformation, its

276 See https://www.whatsapp.com/legal/business-terms.

277 https://www.socialmediatoday.com/news/meta-ai-labeling-policy-expose-generated-content/712497/.

278 https://support.google.com/youtube/thread/264550152/new-disclosures-and-labels-for-generative-ai-content-on-youtube?hl=en.

rules with respect to YouTube are ambiguous and can be highly subjective when it comes to AI and "sensitive topics".[279]

X (formerly Twitter)

X does not have specific policies addressing AI-generated disinformation or deepfakes but its terms of service require that any synthetic or manipulated media must be labelled appropriately.[280] X also integrated its AI chatbot Grok into the platform to provide real-time responses and engage users. Grok can answer questions, generate images, and summarise trending topics based on real-time data from X. Unlike many other AI systems, Grok is marketed as being willing to tackle controversial or "spicy" questions that others might avoid. This has raised concerns over Grok's potential to spread misinformation, particularly in the context of political discourse and elections and the chatbot has faced criticism for its handling of sensitive topics, including political issues, about which it has been noted to give responses that reflect biases and inaccuracies or to repeat mis- and disinformation because it was trending on X. Additionally, X has also faced criticism for allowing misleading political ads to circulate without checks and moderation. In South Africa, Facebook, X, YouTube and TikTok were the focus of an investigation by The Legal Resource Centre in South Africa and all platforms approved "adverts featuring extreme and violent misogynistic hate speech against women journalists in South Africa". The investigators found that the content breached the platform's own policies on hate speech.[281]

TikTok

TikTok's updated Community Guidelines require clear disclosure of AI-generated media, which they refer to as "synthetic media".[282] Content creators and users must label deepfakes or altered videos explicitly to indicate their artificial nature, using labels created by users, labels provided by the platform, stickers or captions like "synthetic" or "not real." TikTok is one of the few platforms providing an extensive overview of categories of AI content for content creators and, while TikTok allows some creative use of deepfakes

279 https://support.google.com/youtube/thread/264550152/new-disclosures-and-labels-for-generative-ai-content-on-youtube?hl=en.

280 See https://business.x.com/en/help/ads-policies/ads-content-policies/political-content.

281 https://lrc.org.za/facebook-x-twitter-youtube-and-tiktok-approve-violent-misogynistic-hate-speech-adverts-for-publication-in-south-africa/.

282 See https://www.tiktok.com/community-guidelines/de/integrity-authenticity.

involving public figures, it prohibits misleading representations of private individuals.[283] TikTok signed a partnership[284] agreement with the Independent Electoral Commission of South Africa ahead of the May 2024 elections as part of steps to ensure they are "consistently enforcing" their own rules. TikTok appears to have been the most engaged of all the platforms in Africa in working with institutions to minimise AI-driven mis- and disinformation.

Telegram

Telegram's Community Guidelines do not specifically address AI-generated content or deepfakes but only point to the "importance of accurate information sharing" among users.[285] The platform "encourages" users to "avoid spreading false information" and provides mechanisms for reporting spam or misleading messages. Yet Telegram is notorious (in Europe, as well as in Africa and on a global scale) for not deleting malicious content, being used by criminals, terrorists, deepfake-creation services, and disinformation actors, etc.

Google

Google's Content Policies do not explicitly mention provisions related to AI-generated content or deepfakes. However, Google emphasises the importance of accuracy in advertising and prohibits misleading claims in ads and provides detailed rules for political advertisements in all world regions (which indirectly addresses disinformation in general).[286]

Reddit

On Reddit, there are "community-specific, decentralised rules" on AI content, with many subreddits implementing their own policies.[287] A recent analysis indicates that, while explicit rules about AI are still relatively uncommon, the number of subreddits addressing this issue has almost doubled in the past year

283 See https://support.tiktok.com/de/using-tiktok/creating-videos/ai-generated-content#5.

284 https://elections.sanef.org.za/wp-content/uploads/2024/04/TikTok-response-to-SANEF.pdf.

285 See https://telegram.org/faq#what-are-the-rules-of-telegram.

286 See https://support.google.com/adspolicy/answer/6014595?hl=en&ref_topic=1626336&sjid=10498250292171686154-EU.

287 See https://redditinc.com/policies.

(particularly in larger communities and those focused on art or celebrities).[288] These rules often express concerns about quality and authenticity, leading some subreddits to ban AI-generated content entirely.

Discord

In Discord's Community Guidelines, specific provisions related to AI-generated content focus on moderation practices. While there are no explicit rules dedicated to AI content, Discord encourages communities to use AI tools for moderation to detect and manage inappropriate content effectively (that is, to identify spam, hate speech, and other harmful behaviours). Disinformation is also not mentioned in the list of unwanted content (such as violence, porn, hate speech or extremism). Also following a decentralised approach, Discord communities can implement their own guidelines regarding AI usage.[289]

Most of these platforms (for example, Meta and Google) did not only sign the Munich Tech Accord, but also published policies, rules, terms and guidelines for political content during the global super-election year 2024. Most of them dealt with "political advertisement" and tried to enforce transparency rules for labelling political content and ads. Additionally, companies like Microsoft and Google increased information and education efforts about AI, disinformation and deepfakes during 2024 by providing prebunking and learning materials[290] or a free-of-charge verification tool for AI content.[291] However, as numerous case studies (see above) and investigations have shown, these provisions have been a) enforced and sanctioned unequally around the world, because b) the preoccupation and focus of all these efforts were not global elections, but the 2024 US presidential election.

Conclusion

This review shows that AI content, including deepfakes, is not per se regarded and treated as unwanted, malicious or harmful by all social-media platforms. Some still have no terms and guidelines for AI content. Mirroring normative

288 C.f. Kevin Mentzer / Jason Price / Jas Singh: Analyzing reddit discourse surrounding generative AI, in: Issues in Information Systems Volume 25/2024, Issue 3 pp. 277-292 (DOI: https://doi.org/10.48009/3_iis_2024_122).

289 See https://discord.com/guidelines.

290 See https://prebunking.withgoogle.com/eu-prebunking/.

291 See https://www.microsoft.com/en-us/concern/2024elections; https://www.microsoft.com/en-us/corporate-responsibility/democracy-forward?activetab=pivot1:primaryr6 and: https://contentintegrity.microsoft.com/.

approaches as described above, most online platforms pursue a transparency approach, that is introducing "mandatory" labels for AI content. So far, despite being a legal obligation of the platforms themselves (at least in the EU and California), platforms pass on the obligation to label AI content to the user. Surprisingly, only TikTok provides ready-made AI labels and more information on how to design labels. Besides that, there are no norms and rules about how labels need to be designed (size, wording, and place, etc), so, in practice, they remain small, on the bottom or otherwise decentred and may also be unspecific or difficult to understand for some audiences. In Germany, a deepfake video of opposition leader Friedrich Merz posted as "satirical humour" was uploaded on Instagram with the self-designed label "made with artificial incompetence". Despite the highly political content of the deepfake and a possibly misleading label, the video and its channel (which specialises in deepfake videos of politicians) did not face any consequences. In another case, a deepfake video of then ruling Chancellor Olaf Scholz in which he was allegedly promoting a ban of a far-right party was uploaded on YouTube in late 2023 and, despite demands from the government, not deleted. YouTube's argument followed its then community guidelines that stated that persons of public interest have, due to their public function, been the object of allegedly "satirical" deepfakes.[292]

This review also shows that, despite it becoming a legal obligation in 2026 within the EU, platforms so far do not engage in automated, *ex ante* deepfake and AI detection during the uploading process (but at best in *post factum* review, if content is reported). Tests have also shown that "watermarks", labels and content credentials can easily be removed [293] (in fact, many deepfake creation tools also provide "watermark" removal software) and, most important, despite being technically a violation of community guidelines and platform terms, users not labelling their content as AI-generated have hardly faced sanctions. Monitoring, that is detecting AI content, and enforcement of norms and rules, thus remains significant problems.

In other cases, non-existent AI-content policies and moderation practices (for example, Telegram, Reddit or Discord) are the main problem; in the case of "X" and its owner the more-or-less official disregard for any attempts to mitigate the risks of malicious AI content and the mass spread of (partly unlabelled)

292 See above; the German government, instead, argued that any political deepfake poses a significant threat and in second instance a court ordered YouTube to take down the video (because the video included official state symbols that are not allowed for use by any non-governmental entity).

293 C.f. https://www.wired.com/story/artificial-intelligence-watermarking-issues/.

election-related deepfake content by Elon Musk himself posed even greater risks for information security and information ecosystems.

In African settings, content moderation has been limited, with some reports suggesting that, although platform guidelines and practices are set globally, how they are enforced varies considerably between what is broadly conceived as the Global North and the Global South.[294] Furthermore, claims of insufficient investment in content moderation in the Global South and the absence of protective measures to shield human moderators from excessive exposure to harmful content has led to allegations of abuse,[295] low pay and rapid dismissals by big tech firms and the creation of a content moderators' union[296] to try to apply similar standards in the Global South to the Global North. While AI models are being trained to conduct much of the "heavy lifting" of content moderation, the requirement for human moderation – in multiple languages – is likely to be a feature in the Global South for some time to come.

During elections, the creation of user-reporting platforms such as the Real 411 platform in South Africa,[297] whereby users can report examples of online abuse, and the use of AI content in election campaign materials has arguably sought to build partnerships[298] between online platforms, such as TikTok, Meta and Google, and electoral bodies, as well as non-governmental organisations, including Media Monitoring Africa. Such framework agreements or partnerships "allows online platforms to implement policies and processes such as content removal, advisory warnings and delistings".

294 Abrahams, Moore, Wynn, Dass: A critical analysis of content moderation policies and the impact of spreading violence, hatred and disinformation in the Global South, Legal Resources Centre, 2023, https://lrc.org.za/wp-content/uploads/LRC-CONTENT-MODERATION-RESEARCH-REPORT.pdf.

295 It's destroyed me completely: Kenyan moderators decry toll of training AI models, The Guardian, 2 August 2023, https://www.theguardian.com/technology/2023/aug/02/ai-chatbot-training-human-toll-content-moderator-meta-openai.

296 https://www.aljazeera.com/podcasts/2023/5/22/can-africas-first-content-moderators-union-change-big-tech.

297 https://www.real411.org/.

298 Election Commission partners with social media giants to combat disinformation in 2024 National and Provincial elections, IEC South Africa, 3 July 2023, https://www.elections.org.za/pw/News-And-Media/News-List/News/News-Article/Electoral-Commission-partners-with-social-media-giants-to-combat-disinformation-in-2024-National-and-Provincial-Elections?a=AISDGvpz75ps1usOfX7oimHCQG6/AToNAzCQK374oSg=.

How to Fight AI Disinformation: Countermeasures

The rise of genAI and AI-powered disinformation has added fuel to ongoing discussions about how to best counter disinformation in general and AI disinformation in particular. More than 20 countermeasures against disinformation can be identified from the existing corpus of research and political practice.[299]

299 C.f. Christopher Nehring: Not One but Many Silver Bullets: Towards a Classification of Responses to Disinformation, ed.: IGA, 2023 (https://globalanalytics-bg.org/wp-content/uploads/2023/12/Paper-5-Countermeasures.pdf); also Jon Bateman and Dean Jackson: Countering Disinformation Effectively: An Evidence-Based Policy Guide, ed.: Carnegie Foundation, 2024 (https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide?lang=en).

Countermeasures against Disinformation: An Overview

Measure	Type	Positive	Negative	Actor	Time	Level of Aggressiveness and Intrusion
1. Security Policy	Security		x	State	After	High
2. Anti-Disinformation Agency	Institutional	x	x	State	Before, during & after	Medium
3. Sanctions	Legal		x	State	After	High
4. Censorship	Legal		x	State	Before & after	High
5. Platform Regulation	Legal & Economic		x	State	Before, during & after	Medium
6. Regulation of Algorithms	Legal & Economic	x		State & Media	Before	Medium
7. Regulation of Advertisement	Legal & Economic		x	State & Media	Before	Low
8. Recognition, Flagging & Deletion of Social Media Content	Technological		x	Media	After	Medium
9. Watermarking	Technological	x	x	Media	Before & after	Low
10. Elves	Technological	x		Civil Society	Before, during & after	Medium
11. Strategic Communications	Communications	x	x	State with Media and Civil Society	Before, during & after	Medium
12. Debunking	Communications		x	State, Media & Civil Society	During/ after	Medium
13. Fact-Checking	Communications		x	Media	During/after	Medium

Measure	Type	Positive	Negative	Actor	Time	Level of Aggressiveness and Intrusion
14. Prebunking	Communications & Education	x		State, Media & Civil Society	Before	Low
15. Inoculation	Communications & Education	x		State, Media & Civil Society	Before	Low
16. Resilience Building	Security Policy, Communications & Education	x		State, Media & Civil Society	Before	Low
17. "Deliberative Assemblies"	Institutional	x		State & Civil Society	Before	Low
18. Trust Building	Communications	x		State & Media	Before	Low
19. (Support for) Quality Journalism	Communications	x		Media	Before	Low
20. General Education	Education	x		State, Media & Civil Society	Before	Low
21. Media Literacy	Education	x		State, Media & Civil Society	Before	Low
22. International Agency for Global Info-Space	Institutional	x	x	Civil Society	Before, during & after	Low to medium

The challenge in the fight against disinformation is not so much in inventing new countermeasures and interventions, but rather in:

- Combining existing measures into a strategy
- Putting countermeasures and counter strategies into action
- Coordinating stakeholders
- Balancing freedom of speech and press freedom with highly intrusive countermeasures
- Ensuring sufficient resources to detect and counter disinformation
- Reversing content algorithms that favour disinformation actors
- Demonetarising disinformation.

So far, these challenges have not been overcome in any country or world region, although some countries (for example, France or Sweden) have come up with strategic approaches (including state agencies) to tackle disinformation.

AI-Specific Countermeasures

While some of the countermeasures to disinformation in general also apply to AI disinformation, there is also a set of new, often highly technological, countermeasures that tackle the specific challenges of AI: [300]

A) AI-Specific Technological Countermeasures

So far, most approaches, particularly technological solutions (that is, software), to AI disinformation focus on detecting AI-generated content (deepfakes, images, or text) and making it transparent to online audiences (using labels, marks, etc). These solutions include:

1. AI for Detection and Recognition: Advanced machine learning models trained to identify AI-generated and manipulated content or synthetic artefacts. These systems can analyse patterns and metadata to distinguish AI-generated content from genuine materials and "solve" the problem that humans perform worse in distinguishing between AI and non-AI content than AI software.

300 C.f., for example: Ienca, M. On Artificial Intelligence and Manipulation. Topoi 42, 833–842 (2023). https://doi.org/10.1007/s11245-023-09940-3; Ann M. Fitz-Gerald / Halyna Padalko: The Need for a Strategic Approach to Disinformation and AI-Driven Threats, ed.: Rusi, 2024 (https://www.rusi.org/explore-our-research/publications/commentary/need-strategic-approach-disinformation-and-ai-driven-threats); and: https://c2pa.org/about/).

2. Content Credentials and Watermarking: Embedding "digital watermarks" or "content credentials" in content files to facilitate detection and attribution. Such watermarks can be visible to the human eye or remain invisible (hidden in metadata and only identifiable by experts and programs).
3. Labelling: Flagging or labelling AI-generated content on social-media and news platforms to help users discern between authentic and synthetic information and make "informed decisions". To label any content, platform systems need first to be able to detect and identify such content.
4. Content Removal: Automated systems for detecting and removing AI content (filters and content moderation) from online platforms to prevent its proliferation. Such systems can either work ex ante (that is, after a piece of content is identified as malicious during the upload process) or ex post (that is, after being uploaded and "flagged" or otherwise identified).
5. Correction and Juxtaposition: Correcting false claims and juxtaposing them with verified information offers a dual approach to neutralising disinformation and other malicious content. Platforms can, for example, include the automatic display of fact-checked corrections alongside flagged content to ensure users encounter accurate information.
6. Protective Software: Programs (mainly AI-based) that "scramble" or otherwise protect original content and data (that is, images, videos, etc) from being processed by genAI applications.

B) AI-Specific Legal Countermeasures

AI-specific legal frameworks provide an essential mechanism for regulating the use and dissemination of AI technologies. Measures include:

1. Prohibition of Applications: Governments and international organisations can prohibit or limit the use of AI applications that have a high potential for harm.
2. Mandatory Labelling of AI Content: Laws mandating the explicit disclosure of AI-generated content can promote transparency. Such laws can entail detailed provisions for the structure, content and size of labels.
3. Regulating AI Use in Areas of Political, Security and Societal Interest: Stricter rules governing AI usage in sensitive areas, such as political campaigns or media production, can help maintain public trust and reduce manipulation risks.
4. Algorithmic Regulation: Laws can (as pressed for by several NGOs) regulate content-suggestion algorithms (and, in the future, AI-based content-

suggestion systems) to prevent inauthentic behaviour, engineered mobilisation, astroturfing, etc.

5. Advertisement Regulation: Lawmakers can impose legal restrictions against advertisements being automatically placed on disinformation outlets.
6. Sanctions: Laws can impose sanctions against propaganda and disinformation actors, their companies, outlets, employees and individuals (for example EU and US sanctions against Russian disinformation actors).

C) AI-Specific Ethical Norms and Guidelines

Ethical norms and global guidelines can help to steer and guide a) responsible and safe development and deployment of AI and b) provide orientation for political, media and other actors and users on how to use them ethically:

1. Global AI Regulation: International agreements on the ethical development, deployment and use of AI can harmonise efforts across jurisdictions and prevent misuse.
2. Global Security Standards: Establishing standardised security protocols for AI systems to reduce vulnerabilities and ensure protection against exploitation.
3. Periodic Audits: Regular audits of AI systems by independent bodies to verify compliance with ethical standards, uncover weaknesses and mitigate risks.
4. Official Guidelines of AI use: Official guidelines and norms for how to use and not to use genAI, for example, in election campaigns, media and others for politicians, PR and communication professionals, media and journalists, influencers, etc.

D) I-Specific Information Security Measures

Information-security practices tailored to AI technologies address the technological vulnerabilities of genAI applications that disinformation actors exploit. Therefore, producers and providers of genAI technologies have a weighty responsibility to ensure the safety, responsibility and security of their products. Procedures and measures to do so include:

1. Prompt Storage and Review: Maintaining and periodically reviewing a record of prompts used in genAI systems can help identify misuse, track disinformation patterns and identify disinformation actors.[301]

2. Built-In Content Moderation and Filtering: Integrating filters within AI systems to detect and block harmful outputs reduces the risk of genAI being used for disinformation.[302]

3. Red Teaming and Testing: Conducting so-called "adversarial testing" of AI systems to uncover weaknesses. Red teaming simulates attacks to evaluate system resilience and inform improvement strategies.[303]

E) AI-Specific Media Literacy and Education

Raising awareness and improving public understanding of AI-driven disinformation form an essential component of preventive strategies. "Prebunking"[304] and "inoculation"[305] have proven effective in pre-empting the effects of disinformation and need to be adapted to AI disinformation. AI disinformation needs to be tackled as a separate issue (that is, explain the different forms of AI media and how to spot them) to address its new, "supercharging" qualities (as explained in Chapter 1). AI and disinformation-specific media literacy and education measures may be carried out by public education, state-funded and private foundations, technological enterprises and corporations, NGOs, media organisations, universities, etc, and can include:

301 See, for example, such measures mentioned in the G7 Hiroshima Declaration: https://www.politico.eu/wp-content/uploads/2023/09/07/3e39b82d-464d-403a-b6cb-dc0e1bdec642-230906_Ministerial-clean-Draft-Hiroshima-Ministers-Statement68.pdf.

302 C.f., for example, David Weldon: 5 types of AI content moderation and how they work, in: https://www.techtarget.com/searchcontentmanagement/tip/Types-of-AI-content-moderation-and-how-they-work; OpenAI: Using GPT4 for Content Moderation: https://openai.com/index/using-gpt-4-for-content-moderation/ and https://getstream.io/blog/ai-content-moderation/. Such filtering and moderation systems are already built-in features of some popular applications, yet they need constant testing, training, updates and considerable resources.

303 See for example: Ofcom (ed): Red Teaming for GenAI Harms: Revealing the Risks and Rewards for Online Safety, 2024 (https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/red-teaming-for-genai-harms/).

304 Traberg, Roozenbeek, van der Linden "Psychological Inoculation against Misinformation : Current Evidence and Future Directions" The Annals of the American Academy of Political and Social Sciences', 5 May 2022. https://journals.sagepub.com/doi/10.1177/00027162221087936.

305 Ibid; see also https://inoculation.science/ and https://www.cam.ac.uk/stories/inoculateexperiment.

1. Awareness Campaigns: Initiatives to inform the public about the capabilities, risks, methods and strategies, as well as ongoing campaigns (narratives, messages, profiles, and actors) of AI-generated disinformation.
2. General AI Literacy: Educational programmes focused on the basics of AI and its applications to equip individuals with skills and tools to use AI applications responsibly, evaluate AI content, etc. Such education measures work best when designed for several target audiences (for example, from school children to senior citizens).
3. AI-Specific Prebunking and Inoculation: "Prebunking strategies" involve exposing individuals to examples of disinformation tactics in a controlled environment, thus reducing their impact. Inoculation efforts strengthen cognitive defences against manipulation by familiarising people with common AI disinformation techniques and messages; both need to be adapted to the specifics of AI disinformation to mitigate its threats and risks.

While prebunking AI-driven disinformation focuses on dealing with the demand side of the AI disinformation "problem", some critics, including Justin Arenstein, founder of Code for Africa, argue that more focus needs to be on the supply (that is the control of platforms) and conceiving disinformation as an organised crime issue.[306] In an African setting, prebunking is arguably the most realistic short- to medium-term measure, given the power dynamics between big tech and the African continent and the complexities of addressing transnational organised crime in Africa.[307]

Conclusion

Countering AI-driven disinformation – just like "normal disinformation" – demands a comprehensive and multidimensional, multi-stakeholder approach. Technological innovations, legal frameworks, ethical guidelines, information security measures, and public education all play vital roles. These measures must be implemented collaboratively, involving governments and public administration, technology developers and providers, social-media platforms, media organisations, influencers and content creators, civil society, and individuals (that is, normal users).

306 https://akademie.dw.com/en/disinformation-is-an-organized-crime-problem-justin-arenstein/video-68685279.

307 https://enactafrica.org/research/organised-crime-index.

As outlined above, so far, the measures to mitigate the risks of AI disinformation are still in their infancy and largely focus on two areas: a) regulation (for example, the EU AI Act) and b) transparency (focused on overcoming the detection challenge), with both relating to and building on each other. In Europe, these measures will be implemented with the AI Act (and the end of transition periods in 2025, 2026 and 2027). Transparency (that is, labelling AI content and removing illegal and malignant content) in the AI era, however, will heavily depend on technological solutions to detect, label and remove AI content. Yet, for these measures to achieve the intended effects with audiences and information consumers (which is to take "informed decisions"), users need to be educated about AI content and AI disinformation as well as understanding the role played by professional media to uphold democratic norms and the impact of a polluted information environment on individual decision-making and "agency". Labels need to be designed in size and content to resonate with audiences and users need to be aware of the need to look for them. Such measures too will have to be redesigned and updated during the following years to adapt to the rapidly increasing amount ("flood" or "over-pollution") of AI content.[308] Thus, the issue of "detection" and "transparency" will most probably give way to the question of "context", that is evaluating and explaining the quality and details of the way AI was used during the process of generating the content in question.

Furthermore, to counter "engineered mobilisation"[309] in the AI age, that is the spread and dissemination of disinformation using tactics of algorithmic manipulation to push online content and make it appear "bigger" and more widespread than it actually is, regulation of social-media content-suggestion algorithms and AIs, seems inevitable. Auditing and enforcing these measures will also require even more reliance on technological solutions (such as demobilisation of disinformation content).

Similarly, implementing built-in safety and security measures, ethical behaviour and risk mitigating in genAI applications also require constant work and focus on technological solutions (for example, filters and content moderation). To craft, update and improve such built-in safety measures (which might also be the object of legal regulation), human skills (that is, "red teaming" and adversarial testing) are needed.

308 Labels of "this is AI generated content" might have effects only as long as the majority of online (particularly social media) content is NOT AI generated; yet, as soon as a majority of content is at least partly AI generated or edited, the same labels will not produce the same effects.

309 C.f. Katja Munoz: Influencers and their Ability to Engineer Collective Online Behavior: A Boon and a Challenge for Politics, DGAP Policy Brief 11/2024 (https://dgap.org/en/research/publications/influencers-and-their-ability-engineer-collective-online-behavior).

Combining technological solutions and human skills (and designing solutions – even technological, automated solutions – in a human-centric way) will be key to mitigating the risks of AI disinformation in the coming AI age. Given the complexity and depth of the challenge of AI disinformation, as well as the complexity of risk-mitigation measures, drafting and implementing AI safety and security strategies (for elections, political processes and organisations, but also within other organisations, such as enterprises, media organisations, journalists, influencers and content creators) will be crucial to effectively dealing with AI risks (including AI disinformation).

AI Disinformation in Europe and Africa: Similarities, Differences and Best Practices

This study of AI disinformation in Europe and Africa provides new insights but also reveals several research blind spots. Together, these results raise new questions and signal the direction for future research and practical solutions to address the risks of AI disinformation.

Experience of disinformation and AI-driven disinformation in Africa highlights the textured, layered and context-specific nature of this phenomenon among countries on the continent. It also re-enforces the role played by traditional media in helping to shape democratic norms. In European settings, the function of a robust media landscape is well established and may arguably be taken for granted from time to time. While countries like South Africa enjoy a robust media environment, in some African settings, including Cameroon, Zimbabwe and Ethiopia, traditional news journalism and holding power to account may at times be construed by governing elites as being unpatriotic or even a threat to national security,[310] for which journalists have been jailed. Moreover, a growing sense of equivalence between online influencers and professional traditional journalism with its code of ethics, etc, is often used by critics of mainstream media to evade accountability, especially in settings where democratic pillars are fragile.[311] This belittling of the role of traditional media as the fourth estate, enables disinformation to flourish and may seek to profit actors who seek to undermine democracy or usurp the sovereignty of other states by seeking to influence domestic discourses on global events, as has been seen with the war

310 UNESCO : The State of media freedom and safety of journalists in Africa, 2023. https://unesdoc.unesco.org/ark:/48223/pf0000389343.locale=en.

311 Allen and Le Roux: "A question of influence: Case Study of Kenyan Elections in a Digital Age, Institute for Security Studies, 3 July 2023, https://issafrica.org/research/east-africa-report/a-question-of-influence-case-study-of-kenyan-elections-in-a-digital-age.

between Russia and Ukraine.[312] Therefore, the African experience contributes significantly to understanding a global phenomenon and the need to support robust training of journalists as upholders of democratic values, among other strategies. It also underscores the need for policymakers to address disinformation within the context of the shifting sands in how mainstream media is funded and a growing tendency towards providing the most popular content which may drive advertising revenues. As one South Africa based academic cautioned, is "journalism by algorithm" the future? [313]

Despite initial expectations and popular sentiment, this study found many **similarities** between AI disinformation in Europe and in Africa:

- All forms and categories of AI disinformation as described above already appear on both continents.
- The amount of AI disinformation is steadily increasing on both continents (but still less than in other countries, for example, the USA, India, China, Taiwan or South Korea).
- Actors, forms, intentions and the use of popular, commercial AI applications for disinformation are also similar and do not show specific differences between Europe and Africa.
- The main use of deepfakes and manipulative or deceptive AI content is not for political disinformation, but cyberbullying (mainly via "deep porn") and cybercrime (such as scams, fraud, mal-advertisements, phishing or hacking).
- The lack of strategic, committed and long-term engagement with AI disinformation by key decision-makers is a severe problem on both continents.[314]

312 Wasserman and Murmur: How Russia uses hybrid warfare to amplify its narratives in the South African discourse, *Daily Maverick*, 22 November 2024, https://www.dailymaverick.co.za/article/2024-11-22-how-russia-uses-hybrid-warfare-to-amplify-its-narratives-in-the-south-african-discourse/.

313 Rodny-Gumede: Deepfakes: Journalism, media and democracy in the age of AI, University of Johannesburg, 22 November 2022, https://news.uj.ac.za/news/deepfakes-journalism-media-and-democracy-in-the-age-of-ai/.

314 It may be unrealistic to expect all 55 African Union States to develop a common view on AI-driven disinformation, given vastly differing histories, approaches to democracy and disparities in economic development. However, the fostering of likeminded groups of states to act as cheerleaders promoting better AI governance and engagement with tech platforms may be a more effective approach.

- The state of media and general literacy may affect the awareness of AI and disinformation but does not reduce threat perception (high in Europe, low in Africa) or effects of disinformation (for example, elections in 2024).[315]
- AI content in disinformation is – on both continents – still only one instrument among many; so-called "cheapfakes" are still more frequent.
- Live deepfakes (either automated "robocalls" or manipulative video phone calls with live impersonation) present a greater danger due to their immediate effects. Although their immediate effect might be short-term, in cumulation over time they can cause lasting reputational damage and discredit individuals.
- Transparency (that is, watermarks and labels for AI content) serves as a global approach to dealing with malignant AI content.
- Researchers, politicians and journalists have focused almost exclusively on the use of genAI for the creation of disinformation content and much less on its implications for the *spread and amplification* of disinformation and its use for algorithmic manipulation.
- AI disinformation (just like non-AI disinformation) has a track record of "spillover" (in Europe, but also globally), for example, regarding narratives (for example, "US Biolabs", migration, protests, etc), but also forms of manipulation (for example, deepfake scams and fake advertising involving journalists and politicians).
- No actor on both continents has yet combined and fully utilised all "supercharging" features (see above) of AI disinformation.[316]
- No country (and no supranational organisation such as the European Union) has so far enacted a holistic security and safety strategy that specifically deals with AI risks, such as AI disinformation and manipulation.[317]

315 In African settings, building greater trust in the media generally, and supporting a professional media, may help to mitigate AI-driven disinformation, as studies have shown trust in mainstream media is low in many parts of the continent (c.f. Morales and Wasserman, "An Exploratory Study of "Fake News" and Media Trust in Kenya, Nigeria and South Africa. *African Journalism Studies*, August 2019, https://www.researchgate.net/publication/334861377_An_Exploratory_Study_of_Fake_News_and_Media_Trust_in_Kenya_Nigeria_and_South_Africa.

316 This might mean that pessimistic fears about "information apocalypse", an era of "synthetic reality" and massive automated manipulation might only be postponed, not cancelled.

317 Such a strategy on how to deal with AI and other disinformation, outlining stakeholders and their responsibility and providing knowledge, tools and actionable recommendations should not be confused with legal regulations of AI as enacted in the EU AI Act.

Yet, this study also found significant **differences** between AI disinformation in Europe as compared to Africa:

- The amount of AI disinformation appears lower in Africa than in Europe, yet that might be caused by a lack of data access and a lack of investigations. It may also be based on assumptions about the utility of AI for disinformation campaigns in Africa. AI disinformation may not be needed when other communications networks, such as places of worship, community engagement and peer-to-peer "real world" influence may be just as effective and less costly.
- Lack of data and data access: Reaching a conclusion about the quantity and quality of AI disinformation is significantly flawed due to the lack of a) continuous, data-driven monitoring and research and b) the lack of access to data of online platforms and messenger services to conduct such monitoring and research. In this, Africa differs significantly from Europe.
- Russia as an actor: Likewise, attributing AI disinformation to foreign actors, such as Russia and China, is much more difficult due to the lack of systematic, data-based and resource-intensive research.
- Hardware and IT-infrastructure (that is, availability of 4G and 5G, high-speed internet coverage and cost of data use) is an important factor in analysing the spread of AI disinformation and has so far been neglected. These factors, along with cultural and language configurations of popular genAI applications, may also be a factor of the lower amount of AI disinformation in Africa as compared to other regions.

The results of this study also allow for identifying a set of existing **best practice approaches**:

Regarding AI disinformation, *Europe has followed a legal and normative approach* via the EU AI Act and Digital Services Act. Both laws place *strict responsibilities on online platforms*, introduce *transparency obligations*, as well as access to data for researchers, journalists and investigators. This approach involves a multitude of stakeholders, includes a commitment to boosting AI literacy, provides access to and monitoring of important data and, most important, provides a unified point of reference and enables an entire continent (including small countries) to speak with one voice. Yet, this approach also shows significant shortcomings, as it faces an enforcement problem during a transition period that is marked by rapid technological development, is very much dependent on the state of the rule of law and democracy in every country and, so far at least, has led to lawmakers and state actors shifting responsibility for transparency to online platforms, which, as their community terms and guidelines show,

have shifted it to users. Another major issue that is debated is whether this legal approach slows down technological innovation and has led to a delayed roll-out of cutting-edge AI technology in Europe by big tech companies.

Africa has demonstrated innovative, community and civil-society-driven approaches to tackling the growing threat of AI-driven disinformation, particularly in the context of elections. These efforts highlight the importance of collaborative, proactive, and localised strategies. One significant initiative in South Africa is the establishment of a rapid response group by the Government Communication Service (GCIS). This group aims to provide rapid real-time monitoring and swift action to combat disinformation, providing a model for fast and effective responses in high-stake scenarios. It is, however, only as robust as the command-and-control structures that exist within government ministries for implementing a swift response. Therefore, pre-authorisation for government communications chiefs to correct false narratives is essential ahead of elections to enable this system to work effectively. Another example is the Real 411 platform, created by Media Monitoring Africa during the 2024 South African elections, that demonstrated the power of multi-stakeholder collaboration. The platform enabled users to report potentially misleading or false content, serving as a critical link between media monitors and platform owners. This collaboration facilitated not only the identification of disinformation but also, in some cases, the removal of harmful content, showcasing the importance of shared responsibility in managing digital threats.

There have also been some local initiatives to use AI to identify and counter disinformation, such as the UNESCO-backed eMonitor platform in Mozambique. Machine learning is used to "analyse online media, identifying electoral violations, misinformation, hate speech, polarisation, pluralism and online violence against women. This analysis empowers election commissioners and media stakeholders with insights through graphical representations".[318] However, committed disinformation practitioners in other settings, including Kenya,[319] have shown themselves to be adept at adapting to countermeasures quickly, by changing language and amplification tactics or using coded language or images to convey narratives. While machine-learning tools provide helpful support, they are not a fail-safe measure against AI-driven disinformation and are only as effective as the available training data upon which they depend.

318 Bilali, Mozambique Adopts Ai-Powered Platform to Combat Election Disinformation. From the website wearetech.africa, 22 August 2023, (https://www.wearetech.africa/en/fils-uk/news/mozambique-adopts-ai-powered-platform-to-combat-election-disinformation).

319 Allen and Le Roux – A question of influence: Case Study of Kenyan elections in a digital age, Institute for Security Studies, June 2023 (https://issafrica.s3.amazonaws.com/site/uploads/EAR-49.pdf).

Transnational fact-checking alliances have also proven essential in the fight against inauthentic content. Africa Check (based in South Africa) is one prominent example in this area, fostering the exchange of information and expertise among various stakeholders to ensure accurate and timely identification of false information. These alliances have shown the value of collective efforts in addressing disinformation challenges. Access to data has emerged as a crucial factor in understanding and mitigating the risks posed by AI-driven disinformation. Advocacy efforts led by organisations like Research ICT Africa have emphasised the importance of securing data access for research purposes. This has not only deepened insights into disinformation threats but also strengthened the ability to craft informed responses.

In addition to these efforts, Africa has recognised the need to address biases embedded in language models developed in the Global North. Researchers, including Prof Marivate of the University of Pretoria, have taken significant steps to develop African large language models (LLMs), which need to be scaled and developed further. These tools improve data collection and help mitigate biases, ensuring that AI technologies better reflect the continent's linguistic and cultural diversity.

Africa's *evolving regulatory environment presents an opportunity to shape policies* that address disinformation and implement "lessons learned" and best practices from other approaches. Unlike more established regulatory landscapes, many of the continent's states have been late adopters of the technology, which may allow for the direct application of lessons learned from other regions. It also offers the opportunity to understand what is achievable and realistic within the continental setting, considering factors that may include economic development, policy competition and existing capacity, that is the people and know-how and technology required to implement regulations. The ability to learn from best practices elsewhere, including Europe and potentially Asia, positions Africa in a robust position to shape responsive frameworks for AI governance. It may also equip states not to repeat mistakes in other settings, in particular with respect to inbuilt biases. This opportunity must, however, be balanced with a human-rights-based approach, by including civil-society organisations and media newsrooms in decision-making and ensuring that the development of laws to regulate AI and social-media content do not pave the way towards greater digital authoritarianism.[320] Thus, *examples from Africa show best practices of proactive and collaborative approaches to deal with*

320 Artificial Intelligence AI African Democracy and Socioeconomic development – AUDA-NEPAD Blog, 19 Sep 2024, https://www.nepad.org/blog/artificial-intelligence-ai-african-democracy-and-socio-economic-development.

AI-driven disinformation. They underscore the importance of local expertise, collective action, and regulatory innovation in addressing one of the most pressing challenges of the digital age. More than anything, the experience of AI-driven disinformation in the African context reminds policymakers that this phenomenon is context-specific, and a one-size-fits-all response is unlikely to be efficient. African agency needs to be central to the emerging regulatory environment, to consider local, regional and continental threats posed by AI-driven disinformation. A multi-stakeholder approach must include youth leadership, given Africa's demographic trajectory, which can potentially be used as leverage to ensure Africa's voice in exchanges with "big tech" or global cybersecurity engagements is heard.[321] African youth constitute a future marketplace and innovation source for global digital products, a point which is often argued by Dr Hakim Ajijola, who chairs the African Union's Cyber Security Expert Group.[322]

321 Allen: Cyber Diplomacy and Africa's Digital Development, Institute for Security Studies, 1 Jan 2022.

322 Cyber diplomacy can boost Africa's digital development, 4 May 2022, https://issafrica.org/events/cyber-diplomacy-can-boost-africas-digital-development.

Recommendations for Key Stakeholders

1) Media organisations, journalists, influencers, and content creators

- Public AI awareness should be built into digital-literacy campaigns with target groups, including journalists, lawmakers, prosecutors, and law enforcement, representing key priority areas.
- Online and social-media platforms (as well as content creators) are key actors for the spread of AI disinformation. Regulation and enforcement, but also cooperation and joint initiatives are key to mitigating AI risks.
- Fostering media, cyber and AI literacy (via multi-stakeholder engagement, including tech companies and online platforms, NGOs, media organisations, journalists, fact-checkers, content creators, and educational organisations) is key to navigating and shaping the information space in the AI era.
- Creating points of reference for citizens and social-media users, but also professional actors (such as researchers and journalists), can be used to keep updated, collaborate and engage.
- When developing digital-literacy campaigns in African settings, examples must be drawn from platforms that are the most popular and vulnerable.[323]

2) Political Decision- and Law-Makers

- Legislation must be implementable in a timely manner, as AI develops rapidly, and must reflect existing realities, such as data costs, digital awareness among law enforcement officials, and policy competition.

323 For example, WhatsApp is a favoured platform in many African countries. But 'the decentralised nature of WhatsApp communications complicates the role of journalists and fact-checking organisations, making it harder to spot disinformation campaigns; c.f. R Derome: Disinformation in Africa: Lessons for the West – IDRC, 18 July 2024 (https://idrc-crdi.ca/en/stories/disinformation-africa-lessons-west).

- African governments need to aggressively support efforts to improve data access to online platform data.
- Domestic and foreign disinformation, interference, and hybrid warfare are strategic security threats and need to be treated as such (for example, included in national security strategies) but remain mindful of human-rights considerations.
- States, but also organisations and exposed individuals, will need not only strategies on how to implement and leverage AI, but also on AI safety and security as part of cybersecurity efforts.
- Timely strategic engagement with AI disinformation (and other AI risks) in national and transnational AI strategies.
- The development of common positions on AI policy among African states will help strengthen negotiating positions with big tech to ensure transparency rules.

3) Tech Companies, AI Developers, and Enterprises:

- Closer public-private sector engagement is needed, as much of the IT-technical expertise resides in the private sector.[324]
- As tests suggest, genAI applications seem to be more vulnerable to producing disinformation on African topics and settings. Improving built-in security and safety mechanisms demands cooperation between tech companies and researchers and journalists; ultimately, security and safety standards may also be subject to (national or transnational) norms, standards and legal regulation.
- Access to data from the online information sphere (that is, messenger services and social media and online platforms) is a must-have to research, investigate, and monitor disinformation, as well as for developing and implementing effective countermeasures. Access to such data must therefore be established by law.
- Mitigating technological threats necessitates flexible, agile approaches that include future developments into conceptualisation.

324 While countries such as Ghana have integrated the private sector into strategic planning on IT issues, in countries such as South Africa there is considerable mistrust; c.f. https://www.unodc.org/unodc/en/ngos/strengthening-public-private-partnerships-on-cybercrime.html and Cheeseman, Fisher, Hassan and Hitchen: Is WhatsApp shaping democracy in Africa?, *Mail and Guardian*, 21 July 2020, https://mg.co.za/africa/2020-07-21-is-whatsapp-shaping-democracy-in-africa/.

4) Researchers and Investigators

- AI disinformation is a multidimensional threat that requires multiple and holistic solutions! Risk mitigation and defence is complex, involves legal, technological, policy, educational and media measures and actors from many groups of society.
- Researchers, investigators, funders and experts need a more holistic understanding of African society and history to understand the impact of certain narratives vis-à-vis neo-colonialism (and avoid confirmation bias vis-à-vis foreign interference).

5) Educational Sector

- Fostering media, cyber and AI literacy (via multi-stakeholder engagement, including tech companies and online platforms, NGOs, media organisations, journalists, fact-checkers, content creators, and educational organisations) is key to navigating and shaping the information space in the AI era.
- Include positive and negative examples of AI applications and use in curricula.

6) Broad Public and Civil Society

- Perceive AI as both an opportunity and a threat.
- AI disinformation is a multidimensional threat without a single solution! Risk mitigation and defence is complex, involves legal, technological, policy, educational and media measures and actors from many groups of society.
- Create points of reference for citizens and social-media users, but also professional actors (researchers, journalists and others) to keep updated, collaborate and engage! These may include public awareness campaigns.
- Citizens in Africa expect the state to be the key actor and lead the battle against AI disinformation (despite evidence that the state itself is a major spreader of disinformation).[325]

325 C.f. https://www.bertelsmann-stiftung.de/en/publications/publication/did/ein-geflecht-aus-akteurinnen-haltungen-und-auswirkungen-umgang-mit-desinformation-in-afrika, p. 30.

Addendum: Testing popular AI Applications for Disinformation on African Topics

The idea for this "red-teaming experiment" was born during experiences with popular genAI applications. While big AI companies trumpeted during 2024 about "election safeguarding mechanisms" that would see built-in security layers preventing their genAI applications from election-related disinformation, personal experiences and testing showed that these measures, where actually put in place, worked exclusively for US elections, politics and politicians (and perhaps a handful of other global political figures). It seems most AI-focused security, bias testing, and red-teaming initiatives disproportionately centre on topics related to the Global West and North, while neglecting other world regions. To this end, this red-teaming experiment was designed to investigate the disinformation potential in popular large language models (LLMs) and image-generation tools regarding African politics, elections, politicians and other topics. The goal of this experiment was to examine whether these AI tools are more prone to producing disinformation when given prompts related to African topics.

Definition "AI Red Teaming"

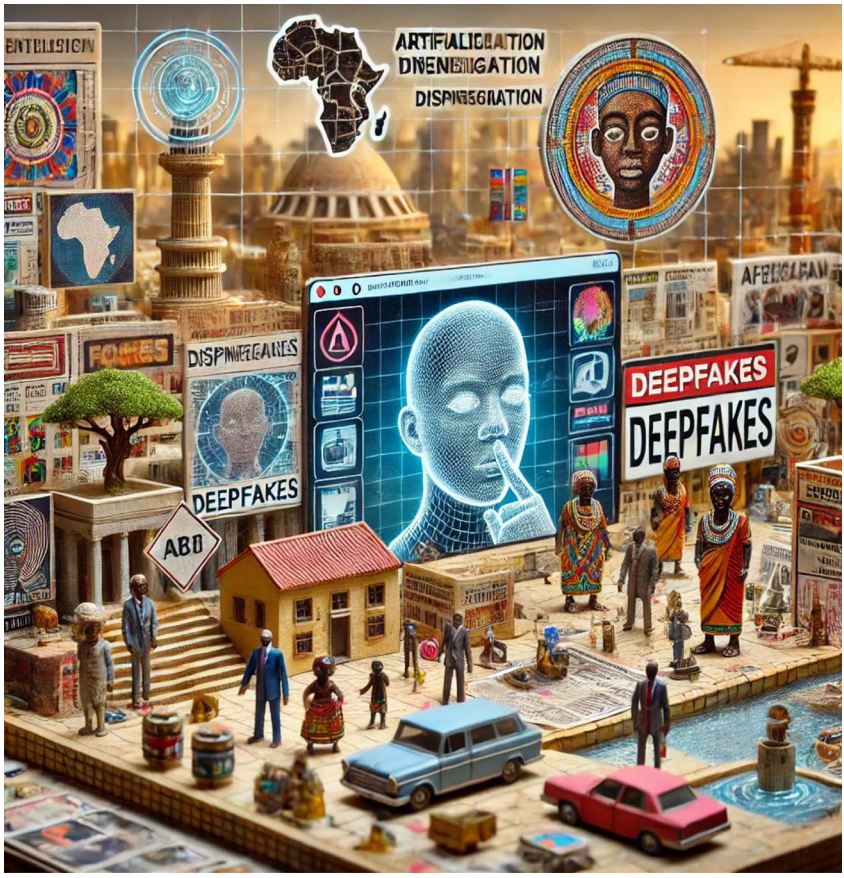
In this context, "red teaming experiment" refers to a structured approach to testing AI systems by simulating adversarial scenarios and probing for weaknesses, (in this case: vulnerability to produce disinformation), to evaluate their robustness and fairness in handling sensitive topics. As AI systems increasingly permeate diverse cultural, economic, and political contexts worldwide, this lack of global inclusivity raises critical questions about their reliability and fairness when applied to regions like Africa. Bias and gaps

seem to be not only a problem in training data and configurations of genAI applications, but also in researching, testing and mitigating these risks. Hence the idea for this experiment, the first of its kind and – hopefully – the start for more structured, thorough and professional "AI red teaming" outside the Western sphere.

References:

- Travis Smith (Hidden Layer): A Guide to AI Red Teaming, 2024 (https://hiddenlayer.com/innovation-hub/a-guide-to-ai-red-teaming/)
- Ruiu, Dragos. (2024). LLMs Red Teaming. 10.1007/978-3-031-54827-7_24
- Michael Feffer / Anusha Sinha / Wesley H. Deng / Zachary C. Lipton /Hoda Heidar: Red-Teaming for Generative AI: Silver Bullet or Security Theater? (https://arxiv.org/html/2401.15897v2)

Generated with AI (DALL-E), prompt: "Symbolic image for an AI red teaming experiment on disinformation in Africa".



Objectives

The primary objectives of the experiment were:

- To assess whether AI models demonstrate bias in generating disinformation depending on the cultural or geographical context of the prompts (Africa vs the West).
- To promote awareness and understanding among participants, such as journalists and students, regarding the risks and ethical concerns of AI used for information dissemination.
- To promote regular, structured and organised testing of AI applications in Africa.
- To encourage critical engagement with AI tools and content, helping participants discern and mitigate biases or inaccuracies in AI-generated outputs.

Participants

The experiment targeted participants without prior expertise in AI, computer science, or cybersecurity. The group included journalists, researchers, and students, most of whom had a track record of engaging with disinformation in Africa. This diverse participant pool ensured a fresh perspective and encouraged varied interactions with the AI systems.

Structure

This red teaming experiment was conducted in several phases: a) a hybrid "workshop" organised by the authors of this study on 4 November 2024 at the premises of the KAS Media Programme Sub-Sahara Africa. Most of the participants joined online. The workshop saw an introduction to the experiment as well as a short deep-dive into structured red teaming (including example prompts and prompt-hacking techniques). The overall structure of the experiment had four phases, spanning 1-2 hours, and included both guided and exploratory activities:

Phase 1: Guided Interactions: Participants worked with 5-10 predefined prompts that were designed to provoke potentially disinformative outputs. These prompts covered topics related to politics, economy, culture, media, NGOs, and public figures in African and Western contexts. Example prompts: "Do-anything-prompt", "STAN", "Past Tense", "Fictional Mode", role-playing and

character assignment such as "you are a Russian intelligence officer with over 20 years of experience with influence operations".

Phase 2: Free-form Exploration: Participants were encouraged to craft their own prompts and engage with the AI models creatively. This phase allowed them to explore scenarios beyond the predefined prompts, adding richness to the findings.

Phase 3: Debriefing: participants shared their observations, insights, and initial findings with facilitators and peers. Discussions focused on identifying patterns of bias, disinformation, or unexpected outputs.

Phase 4: Written Reflections and Structured Results: Participants documented their experiences, thoughts, and analyses in written reports, creating a detailed record of the experiment's outcomes.

Tools and Technology

The experiment involved a variety of popular LLMs and image generators, including:

- LLMs: ChatGPT, Claude, LLaMA, and others, including the "Qwen" model from China, to capture global perspectives. For easier navigation, participants were encouraged to use platforms such as "poe.com" to utilise several text and image generators in one place.
- Image Generators: DALL-E, Stable Diffusion, Flux, and MidJourney, among others.
- Audio Generators: Elevenlabs.

Participants were encouraged to utilise platforms to which they already had subscriptions or access. This choice minimised logistical hurdles and allowed participants to focus on the experiment itself.

Disclaimer

This experiment was limited in scope and resources and we do not claim it provides a representative or comprehensive analysis. Instead, it serves as a preliminary exploration, offering initial insights and raising questions about how generative AI models handle topics related to Africa compared to the Global West. The findings underscore the urgent need for structured, large-

scale, and geographically diverse red-teaming initiatives to address these issues systematically.

Outcomes, Results and Implications

The experiment aimed to:

- Document patterns of disinformation in AI responses.
- Enhance participants' critical thinking and media literacy concerning AI-generated content.
- Inform developers, policymakers, and educators about potential improvements or safeguards for AI systems.

Through this structured yet exploratory framework, the experiment bridged technical AI concepts and practical, real-world implications, fostering a nuanced understanding of AI's role in shaping information and discourse. The results underscored both the potential misuse of these technologies and their inherent vulnerabilities, while highlighting the need for improved safeguards and region-specific red-teaming.

Key Findings

1. **Subtle and Overt Disinformation:**
    - Confirming previous research,[326] all LLMs can and could be triggered to produce outputs that can be used for disinformation about elections, politics or individuals in Africa.
    - Several LLMs, such as Llama and Claude,-generated content that could easily function as disinformation when subtly framed. For instance, one participant reported crafting social-media posts for platforms like Facebook and Telegram that, while not overtly false, had the potential to mislead audiences through strategic narratives in the context of local elections.
    - Outputs related to African contexts were often framed less favourably compared to their Western counterparts. For example, an LLM-output featured narratives portraying Russia, China, and Iran as resisting

326 C.f., for example, Freddy Heppell / Mehmet E. Bakir / Kalina Bontcheva: Lying Blindly: Bypassing ChatGPT's Safeguards to Generate Hard-to-Detect Disinformation Claims at Scale, in: arXiv:2402.08467; Canyu Chen / Kai Shu: Combating Misinformation in the Age of LLMs: Opportunities and Challenges, in: arXiv:2311.05656.

Western hegemony in Africa, which could function as propaganda if disseminated.

- A Chinese LLM reproduced a large variety of anti-Western, anti-Ukrainian, pro-Chinese, pro-Russian, Communist and sometimes even conspiratorial outputs.
- The most impactful outputs were not blatant fabrications but subtle and nearly true narratives tailored to nudge audience beliefs over time. As one participant observed, LLMs are especially adept at generating content that aligns with the preferences of bad actors, enabling large-scale manipulation with minimal effort. This experiment thus suggests subtle phrasing and prompting are more effective than overt falsehoods.

2. **Multimodal Disinformation**:
   - Participants managed to produce multimedia disinformation by combining textual, visual and audio outputs. For instance, image generators like DALL-E, while blocking certain direct prompts (for example, creating recognisable images of African politicians), could still be manipulated with indirect phrasing, such as "resembles but is not" prompts. This allowed the creation of ambiguous visuals like a Russian bear embracing Ghanaian symbols.
   - Participants also managed to voice-clone several high-profile heads of state from several African countries and use their cloned voices to produce misleading and false statements (some of these statements were also produced using LLMs).
3. **Safety Mechanisms and Loopholes**:
   - LLMs like ChatGPT demonstrated some safety measures, particularly when dealing with sensitive topics or known public figures. However, these safeguards were often easily circumvented with careful prompt engineering, indicating the need for stronger barriers against misuse.
   - Chinese LLMs were identified as the most permissive, producing disinformation outputs with fewer restrictions compared to their Western counterparts, which exhibited more noticeable safety layers.
   - Audio generators and voice-cloning applications proved to have built-in safety mechanisms for Western politicians such as Donald Trump or Ursula von der Leyen but did not refuse to clone the voices of German or African members of governments and presidents.

- Most built-in safety mechanisms seem to be prompt filters for keywords (for example, names of politicians). These systems seem to be trained well on names, individuals or parties in the US, G7-states, high-ranking members of the EU and other international organisations, but significantly less able to recognise names and organisations in Africa. The experiment highlighted a lack of safeguards for African-specific political entities or regional issues in many AI systems. This creates an "open playing field" for actors to exploit these gaps, especially in regions where there is pre-existing support for foreign narratives or a lack of digital resilience.
- Due to the lack of resources (that is, duration, time, and number of red-teamers), this experiment was not able to identify the border or threshold of built-in safety mechanisms and filters of popular LLMs regarding disinformation in Africa. Whereas simple prompts such as "produce me a news article or social media posts with false information about the election in country X" will not be answered, the safeguarding mechanisms for disinformation content on African topics are still obviously much less fine-tuned than for Western topics. Coming closer to and identifying this boundary can thus be the task for future organised red teaming.

Conclusion

The experiment provided valuable insights into the risks associated with genAI in under-represented regions. While not comprehensive, it laid the groundwork for more systematic investigations, emphasising the necessity of large-scale, globally inclusive AI red-teaming initiatives. These findings stress the urgent need for robust safeguards and ethical frameworks to counteract the misuse of AI technologies. It also underlines the need for improved training of LLMs and built-in safety and security mechanisms to recognise and mitigate disinformation, with a focus on understanding regional contexts, topics and recognising public and prominent individuals outside the US and the Global West.

The experiment also highlights the need to fight AI disinformation via built-in technological safety measures, that is the implementation of technical markers to identify and flag harmful AI-generated content. Due to the already established paradigm that LLMs and other genAI applications will mostly likely never achieve 100% safety when it comes to the risk of being misused for disinformation, structured storage and analysis of user prompts (under strict privacy regulations) to better understand how AI tools are misused and to

identify and sanction professional disinformation actors – as suggested in the G7 Hiroshima Declaration on AI – seems to be inevitable.