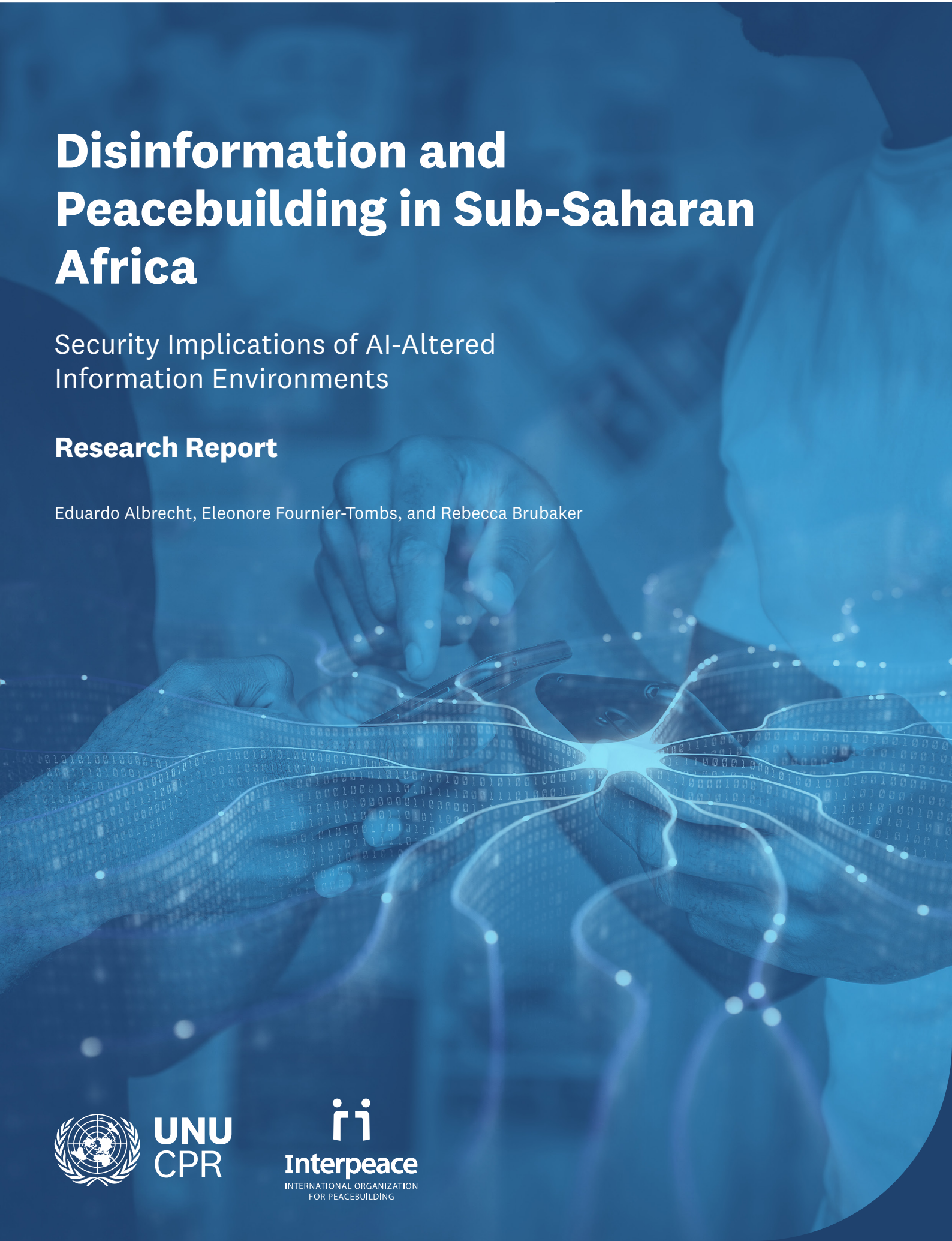
United Nations University
Centre for Policy Research

Disinformation and Peacebuilding in Sub-Saharan Africa

Security Implications of AI-Altered Information Environments

Research Report

Eduardo Albrecht, Eleonore Fournier-Tombs, and Rebecca Brubaker



UNU CPR

Interpeace
INTERNATIONAL ORGANIZATION FOR PEACEBUILDING

About the authors

Eduardo Albrecht is a Senior Fellow (Non-Resident) at United Nations University Centre for Policy Research (UNU-CPR) and Associate Professor in the Department of Social Sciences at Mercy College in New York; Eleonore Fournier-Tombs is head of UNU-CPR's Anticipatory Action and Innovation research programme; and Rebecca Brubaker is Director of Policy, Learning, and Advisory Services at Interpeace.

Acknowledgements

This report is the result of a joint research project between Interpeace and UNU-CPR on the effects of Artificial Intelligence on peacebuilding and conflict in sub-Saharan Africa conducted between April 2023 and November 2023. The authors benefited from the invaluable support of Apna Balgobin, Kendal Gee, Charlie Plumb, and Maria Sabater Nuñez. The authors would like to thank all of the experts that agreed to be interviewed and contributed their insights and expertise to this report.

Disclaimer

The views and opinions expressed in this paper do not necessarily reflect the official policy or position of the United Nations University or Interpeace.

ISBN: 978-92-808-6618-6 ©United Nations University, 2024.

Photo credits: ktsdesign - stock.adobe.com / I_am_zews - Stock Photo ID 1769153516

Citation: Eduardo Albrecht, Eleonore Fournier-Tombs, and Rebecca Brubaker, *Disinformation and Peacebuilding in Sub-Saharan Africa: Security Implications of AI-Altered Information Environments* (New York: United Nations University and Interpeace, 2024).

Disinformation and Peacebuilding in Sub-Saharan Africa

Security Implications of AI-Altered Information Environments

Research Report

Eduardo Albrecht, Eleonore Fournier-Tombs, and Rebecca Brubaker

Contents

Glossary

Artificial Intelligence (AI): A branch of computer science that aims to create systems capable of performing tasks that would typically require human intelligence, including the ability to perceive, reason, problem solve, learn, and react.

Censorship: The suppression, prohibition, or alteration of speech, information, or content deemed unacceptable, harmful, or sensitive by a governing body, institution, or other authoritative entity.

Deliberative AI: A category of AI tools, enhanced by generative AI, which allows for virtual discussions and the exchange of ideas.

Digital Literacy: The proficiency in using and understanding digital technologies, encompassing the ability to find, evaluate, create, and communicate information in an increasingly digital society. It includes navigating digital platforms, in particular social media, with an informed awareness of digital safety, ethics, and the critical assessment of information accuracy and reliability.

Disinformation: False information spread with the intention to mislead or deceive.

Fact-checking: The process by which to investigate the veracity of a statement, traditionally conducted in the field of journalism.

Fake News: Often used as a 'catch-all' term that includes misinformation, disinformation, and other forms of propaganda or falsehood, regardless of deliberate intent to mislead.

Generative AI: A form of AI that learns the patterns and structure of its training data such that it can generate new content, including text, images, and other media, using a generative model.

Guardrails: Measures or safeguards implemented in AI systems to ensure that they operate within desired parameters and do not produce unintended or harmful outcomes.

Hate Speech: Any kind of communication in speech, writing, or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, including their religion, ethnicity, nationality, race, colour, descent, gender, or other identity factor.

Influencers: An individual or group that has gained a significant audience on social media platform(s) through content creation and/or dissemination, often in a specific theme. They leverage their online presence and credibility to impact their audience's opinions and behaviours.

Internet Troll: An individual who deliberately offends, attacks, or starts arguments with people online, particularly on social media platforms and forums, provoking other users through inflammatory or even derogatory comments or posts.

Media Literacy: A framework to access, analyse, evaluate, create, and participate with messages in a variety of forms, including print, video, and digital media. It requires an understanding of the role of media in society and the essential skills of inquiry and self-expression.

Misinformation: The spread of false information or falsehoods without the deliberate intent to mislead.

Peacebuilding: The field and practice encompassing the approaches, interventions, and strategies that seek to address underlying and/or proximate dynamics and causes of conflict and work toward the attainment of positive, sustainable peace.

Peacekeeping: A tool available to the United Nations to assist host countries navigating the path from conflict to peace by providing security and political and peacebuilding support.

Shadow-banning: A moderation tactic where a user's content is secretly made invisible or less prominent to the broader community without the user being aware.

Troll Farm: An organized network of trolls or automated bots that collaboratively engage in coordinated online campaigns, frequently linked to disinformation or propaganda. Their objective is to manipulate public opinion, often by amplifying divisive narratives across social media platforms and online forums.

Watermarking: A technique used to embed hidden information or patterns into AI models or their outputs, primarily to establish ownership, verify the authenticity of the model, or trace unauthorized use and distribution.

Executive Summary

Over the last five years, artificial intelligence (AI) capacity and use has developed very rapidly, followed by a growing focus by the international community on global governance of AI, aimed at harnessing its opportunities and mitigating its risks to global objectives and norms, such as human rights, peace and security, and sustainable development.

Concerns about the risks of AI to peace and security, in particular, have been increasing. In July 2023, the United Nations Secretary-General addressed the Security Council on the theme of AI, and expressed that: "Both military and non-military applications of AI could have very serious consequences for global peace and security."

This concern has been linked especially to recent developments in a subfield of AI - generative AI, which allows text, image, video, audio, and other types of content to be created by new AI tools. These tools, now widely available globally, have begun to be linked to increases in disinformation campaigns, cybersecurity threats, hate speech towards women and minorities, and other conflict drivers.

In this context, this report aims to further explore the way in which AI technologies as they currently stand impact peace and conflict, and what methods might be used to mitigate their adverse effects - through the development of better tools and the inclusion of peace and conflict considerations in AI governance frameworks. The report presents the following findings:

1. AI is a driver of disinformation which adversely impacts peacebuilding efforts.
2. There are several ways to mitigate AI-powered disinformation.
3. AI can also be used to promote peacebuilding, although these mechanisms are still in their early stages.

In relation to these findings, the report proposes the following recommendations:

1. More funding and support should be provided to civil society organizations' efforts to expand media literacy and fact-checking initiatives using AI tools to enhance capabilities.
2. Governments need to work with civil society to develop and implement comprehensive, transparent legal frameworks combating disinformation. These legislative measures need to support digital and media literacy campaigns and fact-checking organizations.
3. Social media companies need to expand investment and research into understanding local information environments, so they can better identify and respond to instances of disinformation in all contexts in which they operate and enhance transparency.
4. Peacebuilding organizations need to carefully consider local media ecosystems and information environments when conducting conflict analyses, and factor these dynamics into their projects' frameworks.

Currently, disinformation exacerbates existing grievances and polarization, can act as a trigger for violence, leads individuals to make decisions based on false information, and creates or intensifies mistrust of institutions and international peacebuilding efforts. The efforts of journalists and civil society organizations to combat these effects, primarily through media literacy initiatives and fact-checking, cannot keep up with the scope and scale of 'fake news.' Generative AI will increase the rate at which false content is produced and disseminated, and the quality or 'believability' of such content. Governments, civil society, and social media companies must work in tandem to combat the spread of disinformation and the impact it can have on peacebuilding efforts and security in sub-Saharan Africa.

1. Introduction

Over the last year, significant changes in the capability and reach of artificial intelligence (AI) have disrupted many industries, including healthcare, government services, and education. Generative AI tools, in particular, that are able to create diverse and convincing text, audio, video, images, and code on demand, have been made widely available globally. Tools such as ChatGPT, for example, which allow users to generate text and images in response to prompts, gained 180 million users worldwide within eight months of its launch.

These recent developments have raised significant concerns about the impact of AI on disinformation, especially as it relates to conflict globally. Although the last decade has been characterized by studies on the impact of AI on disinformation and conflict, the mechanisms by which this took place have changed significantly. Until very recently, it was mostly ranking algorithms and recommendation systems used by social media platforms that concerned policymakers. These subcategories of AI have been known to promote polarization and hate speech, and have been manipulated in order to increase the reach of disinformation campaigns. Today, however, there is an additional subcategory of AI – generative AI – which allows vast disinformation campaigns to be deployed with much less effort than previously required.

These changes in AI capacity also come at a time of global polycrisis, where the effects of climate change also coincide with a post-pandemic economic downturn and new conflicts emerging in Ukraine and the Middle East. Political actors at various levels, from civil society organizations to private sector companies, national governments, and multilateral agencies, are all involved in addressing these threats on various fronts.

Recent reports highlight increasing concerns regarding peace and stability in sub-Saharan Africa. According to the *Global Peace Index 2022* by the Institute for Economics and Peace, sub-Saharan Africa is "less peaceful than the global average on the Safety and Security and Ongoing Conflict domains."[1] The region has been impacted by political violence and conflict, with the report noting "a rise in civil unrest and *political instability* across the region, resulting in an average deterioration across the region in the *political terror* indicator of 6.9 per cent."[2] According to the Armed Conflict Location & Event Date Project's *Global Disorder in 2022* report, the region had a notable spike in political violence, with an increase of 1000 more events from 2021 to 2022.[3] Although the Global Peace Index, which measures the state of peace according to various indicators such as economic impact of conflict and incidence of violence, finds a wide variation in the state of conflict in different countries of the region, there have been growing reports of disinformation campaigns, not only from local actors but also from foreign governments who use countries in the region as geopolitical proxies.

Peacebuilding activities in the region, which have worked as long-term processes to support dialogues, reform political institutions, and respond to conflict have already been impacted by new AI developments. There is a significant risk of a continuing increase in polarization, eroding attempts to reconcile groups and work towards peace. Conversely, there may also be opportunities to support dialogues and improve narratives using AI.

While there have been increasing efforts to regulate social media companies globally, these tend to focus more on the question of privacy protection than one addressing disinformation. However, social media has been linked to political violence for some time, such as in Myanmar, where the Facebook platform was condemned by the United Nations for its role in fostering hate speech against the Rohingya. In parallel, the World Health Organization declared in 2020 a second pandemic – the "infodemic" – and has focused its efforts on tackling misinformation online. This is in a context of increasing efforts to legislate AI at the national, regional, and international levels. In 2019, the EU published guidelines on ethics in AI, taking what is referred to as a 'human-centred' approach,[4] and has since adopted the EU AI Act, which proposes pre-deployment certification of high-risk uses of AI and should come into effect in 2026. In the same year, UNESCO launched a two-year consultation on the issue, which culminated in the adoption of the first ever global set of recommendations on

1 Institute for Economics & Peace, *Global Peace Index 2022: Measuring peace in a complex world* (Sydney: Intsitute for Economics & Peace, 2022), p. 21. Accessible at: https://www.visionofhumanity.org/wp-content/uploads/2022/06/GPI-2022-web.pdf.

2 Ibid.

3 Timothy Lay, *ACLED Year in Review: Global Disorder in 2022* (Grafton, WI: The Armed Conflict Location & Event Data Project, 2023). Accessible at: https://acleddata.com/2023/01/31/global-disorder-2022-the-year-in-review/.

4 Tambiama Madiega, *EU guidelines on ethics in artificial intelligence: Context and implementation* (European Parliamentary Research Service, 2019). Accessible at: https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2019)640163.

ethics in the use of AI.[5] These standards touch on important issues such as the impact of AI growth on inequality, the gender gap, and social cohesion, as well as AI's potential to help tackle global environmental and health crises. What many of these efforts lack, however, is an understanding of the impact of AI on peacebuilding, and what these regulatory efforts can do to address it.

Given the speed with which international regulation is moving forward, voices from the AI community are calling for partnerships to ensure that these regulatory and ethical frameworks include a stream on AI in relation to peacebuilding, particularly when it comes to the production and dissemination of disinformation and misinformation. As a result, those who act in this space are taking action largely outside of any overarching regulatory framework. The longer social media platforms operate without cohesive guidance, the harder it will be to overcome regulatory fragmentation. In short, as the global community moves forward on AI regulation, it is important to ensure that these initiatives are sensitive to the unique effects of AI in conflict and post-conflict contexts, and the potential of AI to be a force for good in peacebuilding.

This research was therefore conducted as a response to a complex and evolving landscape in sub-Saharan Africa, to better understand the intersection of new AI technologies, disinformation, and conflict in the region. The objective of the research is to inform national and global policy efforts in AI governance and in peacebuilding.

This report primarily used a qualitative, interview-based approach. Fifteen respondents working in this region were interviewed in order to collect the data informing the report. The interviewees were selected intentionally by the researchers through their past work, as well as through recommendations from other research participants.

In addition, the researchers conducted a thorough literature review of works on disinformation, peacebuilding, and AI, and conducted manual experiments to understand the current capacities of new generative AI tools. This report, written by United Nations University Centre for Policy Research and Interpeace, is the product of this thorough analysis and the insights provided by interviewees. It focuses on three countries: Democratic Republic of Congo (DRC), Côte d'Ivoire, and Kenya.

5 "Recommendations on the Ethics of Artificial Intelligence," UNESCO, last accessed on 16 January 2023, https://en.unesco.org/artificial-intelligence/ethics.

2. Context

Imagine it is election day. You are planning to go vote and suddenly you see a post on social media declaring that all the polling stations in your precinct are closed. You are doubtful, but then you see the same post pop up somewhere else, and then another, and yet another. You are now convinced and give up – perhaps you were not too excited about either candidate anyway.

You know the system will go on without you, and you change your plans for the day. What you don't know, and likely will never find out, is that none of those social media posts were real. They were generated and promoted by ill-intentioned actors with the express purpose to mislead people like you. They did not want you to vote because they knew that people in your precinct tended to vote in a way they did not prefer. You have been hacked.

Different geographies present different challenges. According to a recent article by Idayat Hassan, director of the Abuja-based Centre for Democracy and Development, countries in West Africa (where Internet penetration is high) are facing an increase in coordinated disinformation campaigns. Disinformation in Africa is "undermining democracy" as governments are unable to adequately track its influence and impact.[6]

An important point she makes is that in West Africa, "social media content is not confined online," as people typically share the things they read online via word-of-mouth networks offline.[7] By this mechanism, disinformation spreads exponentially. Hassan highlights Côte d'Ivoire's 2020 presidential elections, where many were tricked into not voting by "fake information."[8] She also emphasizes the power of outside influences, noting that "pro-Russia operatives" were responsible for a campaign to "spread disinformation on social media with the intention of defaming political opponents of the ruling party."[9]

In East Africa, research by Mozilla fellows Odanga Madung and Brian Obilo explores the "shadowy world" of "for hire" disinformation campaigns in Kenya.[10] They interviewed a number of "disinformation influencers" to gain insight into the inner workings of this in-demand industry and found that the campaigns are well-coordinated, utilizing WhatsApp groups for content distribution and cash transactions (influencers are paid around $10 to $15 per campaign).[11] The practice, they say, is "beginning to border on incitement and advocacy of hatred, which is against Kenyan Law."[12]

They also found that disinformation operations in Kenya are increasingly targeting specific individuals, particularly journalists, members of the judiciary, and civil society activists. A pernicious effect of these targeted campaigns is that they can lead to self-censorship, as individuals find it "pointless" to post when they are constantly being attacked.[13] Furthermore, it can be off-putting to participate in online environments where trying to decide what is true and false is very difficult.

The scenario briefly painted above is about to undergo a radical transformation. To date, disinformation peddlers had to write their own text. Today, with generative AI tools like Chat GPT and others, it is possible, via a few prompts, to generate thousands of 'fake news' texts that can then be automatically posted by bots to thousands of fake social media accounts or as replies to real ones. AI-based technologies can also be employed to manipulate recommendation systems and bypass content moderation and censorship measures, enabling the spread of disinformation with reduced scrutiny.

Hassan calls this "computational propaganda" and finds that "the automation of content is a growing feature of Africa's online disinformation industry."[14] Indeed, the

6 Idayat Hassan, "Disinformation is Undermining Democracy in West Africa," Centre for International Governance Innovation, July 4 2022, https://www.cigionline.org/articles/disinformation-is-undermining-democracy-in-west-africa/.
7 Ibid.
8 Ibid.
9 Ibid.
10 Odanga Madung and Brian Obilo, *Inside the Shadowy World of Disinformation-for-hire in Kenya*, Mozilla, 2 September 2021, https://foundation.mozilla.org/en/blog/fellow-research-inside-the-shadowy-world-of-disinformation-for-hire-in-kenya/.
11 Ibid.
12 Ibid.
13 Ibid.
14 Idayat Hassan, "Disinformation is Undermining Democracy in West Africa."

information environment is entering a new phase in which "botnets, groups of bots, and coordinated groups of trolls — called troll farms — promote specific narratives and are deployed to generate online conversations and get stories trending."[15]

Researchers at Georgetown University's Center for Security and Emerging Technology collaborated with OpenAI and the Stanford Internet Observatory on a workshop that resulted in a report entitled *Forecasting Potential Misuses of Language Models for Disinformation Campaigns*.[16] The workshop sought to unpack the potential impact of generative AI on influence operations and disinformation campaigns. The authors identify critical unknowns regarding the use of AI, such as the emergence of new capabilities and new actors. The availability of easy-to-use text generation tools, and the difficulty of developing norms that discourage AI-enabled influence operations, pose a significant risk. To address this risk, the authors propose a framework for mitigations that address three different dimensions of impact: actors, behaviour, and content, or ABCs of influence operations.[17]

The United Nations recognizes the grip that disinformation has on the 'digital ecosystems' of many nations. Disinformation has "enabled the rapid spread of lies and hate, causing real harm on a global scale."[18] In a policy brief accompanying the UN Secretary-General's *Our Common Agenda*, the impacts of disinformation are found to be particularly dangerous for youths, as well as people in "low-income tiers."[19]

John Villasenor, a senior fellow at Brookings, emphasized in his paper, *How to Deal with AI-Enabled Disinformation*, that "public policy will play a central role in both the human and technological aspects of the response to rapid disinformation attacks."[20] He proposes specific policy considerations, such as developing common intervention guidelines and building systems that can effectively detect disinformation.[21]

Villasenor explains that due to the rapidity with which information spreads online, detecting and dealing with disinformation campaigns presents a huge difficulty for social media companies.[22] Acting in a matter of minutes demands a high degree of confidence and knowledge about the attacking accounts, which is challenging to acquire in such a short amount of time. Delaying action for several hours could do severe damage. Choosing which accounts to ban becomes more difficult when trustworthy accounts unintentionally aid in the propagation of disinformation. Furthermore, it would be prohibitively costly to hire a sufficient number of people to monitor each of the practically infinite number of circumstances in which disinformation might develop. Organizations like Facebook and Twitter operate internationally; there are billions of accounts in 200 nations that may be used to spread misinformation.[23]

Given this, experts agree that the only effective way to combat automatically generated disinformation is with similarly automated tools.[24] In other words, it takes bots to fight bots. The creation and dissemination of disinformation are both facilitated by AI-based technology and countered by it. AI-powered technologies that can create fake narratives, deceitful texts, and changed pictures – like natural language generation and image generation algorithms – produce such voluminous quantities of disinformation and at such high speeds that combating it requires automated tools to detect and curb it just as quickly.[25]

---

15 Ibid.

16 Josh Goldstein et al., "Forecasting Potential Misuses of Language Models for Disinformation Campaigns and How to Reduce Risk," Stanford University, Internet Observatory: Cyber Policy Center, 11 January 2023, https://cyber.fsi.stanford.edu/io/news/forecasting-potential-misuses-language-models-disinformation-campaigns-and-how-reduce-risk.

17 Ibid.

18 United Nations, *Our Common Agenda: Policy Brief 8: Information Integrity on Digital Platforms* (United Nations, 2023), p. 3. Accessible at: https://indonesia.un.org/en/236014-our-common-agenda-policy-brief-8-information-integrity-digital-platforms.

19 Ibid., p. 11

20 John Villasenor, "How to deal with AI-enabled disinformation," The Brookings Institution, 23 November 2020, https://www.brookings.edu/articles/how-to-deal-with-ai-enabled-disinformation/.

21 Ibid.

22 Ibid.

23 Ibid.

24 Fátima Carrilho Santos, "Artificial Intelligence in Automated Detection of Disinformation: A Thematic Analysis," *Journalism and Media* Vol 4 Issue 2 (2023): 679–687. Accessible at: https://doi.org/10.3390/journalmedia4020043.

25 Linda Slapakova, "Towards an AI-Based Counter-Disinformation Framework," The Rand Blog, 29 March 2021, https://www.rand.org/pubs/commentary/2021/03/towards-an-ai-based-counter-disinformation-framework.html.

From a technical point of view, however, there are still many critical unknowns regarding the calibration and mode of deployment of these anti-disinformation tools. There are also issues of accountability, governance, and transparency. Who has authority over the process, and in whose name do they operate? Is the private sector, government, or civil society best placed to address this issue? How can they best be adapted to context, language, and culture? What role, and indeed what risks, do organizations operating in different regions face? The following sections explore these questions and others that arise around the impact of AI on peacebuilding, primarily through a manipulation of narratives or disinformation.

Box: A Primer on AI and Disinformation

AI primarily interacts with disinformation in three ways: (i) content creation; (ii) content dissemination; and (iii) content moderation. Each of these uses very different AI techniques. While content creation and dissemination are used to support disinformation campaigns, content moderation can be a way of fighting against disinformation.

AI for creating disinformation: This area primarily covers the field of generative AI. Generative AI is a subfield of AI which creates content based on prompts, principally text, images, videos, code, and other related artefacts such as logo designs, architectural drawings, and so on. In recent months, these tools have grown to be impressively powerful and accessible. Generative AI tools work by predicting the desired content based on massive amounts of data that they have been trained on: texts and books available on the Internet, photobanks, movies and online videos, and open-source code bases. The key concern of generative AI is the way in which it can be used to create disinformation content at scale. With a one-sentence prompt, it is now possible to create thousands of tweets, for example, with interconnected hashtags. Potentially of more concern, is the ability to create convincing artificial photos or videos, which make the doctoring of image evidence used in many disinformation campaigns that much easier and difficult to monitor.

AI for disseminating disinformation: This area considers the subfield of AI which deals with recommendation systems and ranking algorithms. These have been discussed at length in research and commentary about social media and disinformation. Recommendation systems work by predicting which types of content an Internet user would be most likely to interact with. Typically, research has shown that polarizing content, such as hate speech or disinformation, tends to have far higher rates of interaction than the average post. When content receives more interaction, whether likes, retweets, or comments, it is then placed higher in the recommendation list and is then presented in the recommended content feeds for more people. This means that generally, because disinformation has higher rates of interaction, it has more reach than factual information.

AI for fighting disinformation: There are various types of AI systems that can be used to moderate online disinformation. These can include text analysis algorithms and image analysis or recognition, which aim to flag or monitor inappropriate content, and also behavioural analysis and anomaly detection that might flag trolling, or even bots. Such tools are not perfect; models may be able to determine if content is false, but have greater difficulty making more nuanced assessments, such as potential to cause harm, malicious intent, or satiric nature.[26] AI tools are used by social media companies to moderate inappropriate content, such as harassment, hate speech, and disinformation, often as tools used by human moderators. Digital content provenance is another technique being developed that can present information on the origin of content, particularly image and video, that enables platforms and fact checkers to flag manipulated media.[27] However, as shown below, fact-checking organizations in sub-Saharan Africa more rarely use AI for disinformation monitoring, preferring to rely on moderators using manual checking methods.

26 Alla Katsnelson, "Identifying Misinformation's Intent," Columbia Engineering, 2023, https://topics.engineering.columbia.edu/identifying-misinformation-intent/intro/.

27 The Royal Society and BBC, "Generative AI, Content Provenance and a Public Service Internet," Royal Society, last accessed 16 January 2024, https://royalsociety.org/-/media/policy/projects/digital-content-provenance/Digital-content-provenance_workshop-note.pdf.

Case Study 1: Civil Society Responses to Disinformation in Kenya

Kenya faces significant and varied challenges when it comes to disinformation. Though the country has the Computer Misuse and Cyber Crimes Act – which explicitly outlaws the deliberate spread of false information online – the law has not stopped the rapid spread of disinformation through social media platforms and other digital channels, becoming a major concern for the country.[28]

Disinformation-linked political tensions in Kenya tend to spike during election cycles. Even before the widespread use of social media, during and after the 2007 election, radio stations in Kenya were found to be used as a vehicle to stoke existing political and ethnic tensions. According to Lilian Olivia Orero, Founder of Safe Online Women Kenya: "1,500 people were killed, at least 900 men, women, and children were treated for sexual violence, and hundreds of thousands more were displaced until a power-sharing agreement under a new constitution brought the violence to an end."[29] In that crisis, "violence was mostly incited using inflammatory language broadcast through vernacular radio and other local media."[30] A month-long ban on reporting ensued.[31]

The 2017 election was plagued with similar problems but with new communication technologies.[32] Today, election times continue to have a marked rise in disinformation campaigns, but they tend to occur on platforms such as Twitter. Going forward, disinformation will continue to pose a challenge to peaceful elections, particularly as the creation of large amounts of disinformation content will be made exponentially easier with the advent of generative AI tools like ChatGPT.

Some, therefore, propose that social media companies should do more to halt the spread of disinformation in "highly volatile political landscapes" like Kenya. TikTok, Facebook, and Twitter have all begun taking measures to attempt to reduce the spread of disinformation in general, but activists and rights groups believe they are not doing enough in developing countries.[33]

Maintaining Peace through Early Warning, Monitoring, and Analysis (MAPEMA), a consortium of Nairobi-based non-profits and social ventures, was recently established in order to proactively address this failure. Backed by the United Nations Development Programme (UNDP) and the Office of the United Nations High Commissioner for Human Rights (OHCHR), the consortium experimented with using AI-enabled systems to "combat toxic content and manipulation in online political spaces" during Kenya's 2022 general election.

This election saw a significant amount of disinformation circulating online, fueling division between supporters of the main candidates. The MAPEMA team used an array of tools to monitor social and digital media, developing and utilizing a machine-readable database called 'hatelex' and employing open-source tools to monitor conversations and identify networks involved in electoral disinformation campaigns and information operations on Facebook, Instagram, and Twitter. Furthermore, the team used its inbuilt media monitoring solution to track digital media, analyse trends, and generate real-time reports for key stakeholders and decision makers. Of the more than 550,000 toxic Facebook posts identified as a result, over 800 cases of hate speech content were flagged and shared with social media platforms for action.[34]

---

28 Olivia Lillian, *Disinformation was Rife in Kenya's 2022 Election* (London: The London School of Economics and Political Science, 2023).

29 Gillian McKay, *Disinformation and Democratic Transition: A Kenyan Case Study* (Washington, DC: Stimson Center, 2022). Accessible at: https://www.stimson.org/2022/disinformation-and-democratic-transition-a-kenyan-case-study/.

30 Ibid.

31 Ibid.

32 Jacinta Mwende Maweu, "Fake Elections? Cyber Propaganda, Disinformation, and the 2017 General Elections in Kenya," *African Journalism Studies* Vol. 40 Issue 4 (2019): pp 62-76.

33 Nita Bhalla, "Online Disinformation Stokes Tensions as Kenya Elections Near," *Thomson Reuters*, 27 June 2022, https://www.reuters.com/article/idUSL4N2Y22HF/#:~:text=NAIROBI%2C%20June%2027%20(Thomson%20Reuters,stoking%20tensions%20around%20the%20vote.

34 It is interesting to note that as part of the effort, the team also employed "public perception surveys" that gathered data "from over 3,900 citizens in seven hotspot counties aimed at gaining a comprehensive understanding of the key challenges facing citizens during elections." This shows how qualitative information remains an important "knowledge base," enabling the production of effective media monitoring tools on "critical issues like hate speech, misinformation/disinformation, and electoral conversations." See: Code for Africa, "Unmasking Hate Speech in Kenyan Elections with AI and Collaboration," *Medium*, 12 June 2023, https://medium.com/code-for-africa/unmasking-hate-speech-in-kenyan-elections-with-ai-and-collaboration-576e37d4ccb5.

3. AI and Disinformation Countering Peacebuilding Efforts

The findings for this research report are divided into two parts. First, a discussion of the way in which AI-powered disinformation can currently exacerbate conflict drivers and counter peacebuilding efforts in the region, especially in the three case study focus countries. The proceeding section considers the solutions to AI-powered disinformation in the region, such as some digital tools, as well as analog or 'human-powered' methods. This includes examining opportunities of AI for peacebuilding and the role of civil society, States, and multilateral organizations.

AI-powered disinformation, that is, disinformation content that is generated by AI and/or disseminated by AI, can have powerful adverse effects on peacebuilding efforts by creating or contributing to narratives that drive conflict. To understand this exacerbating effect on conflict dynamics, it is necessary first to examine contextually-relevant conflict drivers.

Conflict and its Drivers

The UN@75 Report, *A New Era of Conflict and Violence*, provides a present-day snapshot on the nature of conflict globally, in comparison to when the United Nations was founded 75 years ago. The report highlights that:

1. The regionalization of conflict has resulted in conflicts becoming longer, more protracted, and less responsive to transitional forms of resolution because of the interlinked political, socioeconomic, and military issues across borders.
2. The increase in organized crime and gang violence across regions points to a breakdown in the rule of law, which contributes to climbing homicide rates and growing global political instability.
3. Conflict and violent extremism are drivers of one another "with more than 99 per cent of all terrorism-related deaths occurring in countries involved in a violent conflict or with high levels of political terrorism."
4. Technological advances are contributing to the changing nature of conflict with concerns about the potential of AI and machine learning to enhance cyber, physical, and biological attacks.[35]

Interpeace defines a driving factor as "a dynamic or element, without which the conflict would not exist, or would be completely different."[36] Factors are variables, and can increase or decrease in scope, scale, salience, or quality, and through perceptions of the same. The nature of such factors can drive conflict by contributing to people's grievances, thereby creating or contributing to the conditions for violent conflict.[37]

The same UN report names five "dominant drivers of conflict" in the modern age, each explored in the box below. This does not constitute an exhaustive list, but highlights key conditions that can lead to conflict, and that can be exacerbated by disinformation. Furthermore, these drivers cannot be understood in isolation; factors and causes of conflict are deeply interconnected. Local stakeholders may also disagree on what role, if any, a driver plays in the given conflict dynamic.[38]

35 "A New Era of Conflict and Violence," United Nations, last accessed on 16 January 2024, https://www.un.org/en/un75/new-era-conflict-and-violence.

36 Interpeace, *Peacebuilding How? Systems Analysis of Conflict Dynamics* (2010), p. 4. Accessible at: https://www.interpeace.org/wp-content/uploads/2010/08/2010_IP_Peacebuilding_How_Systems_Analysis_Of_Conflict_Dynamics.pdf.

37 "Conflict-sensitive Approaches to Development, Humanitarian Assistance and Peace Building Tools. A Resource Pack," Saferworld International Alert, last accessed on 16 January 2024, https://www.saferworld.org.uk/resources/publications/148-conflict-sensitive-approaches-to-development-humanitarian-assistance-and-peacebuilding.

38 Ibid.

Box: Dominant Drivers of Conflict

Unresolved inter-communal tensions: This involves deep-seated animosities and mistrust between different community groups, which often stem from historical grievances, ethnic or religious differences, or socioeconomic disparities. These tensions can lead to cyclical violence, hinder reconciliation efforts, and impede peacebuilding processes.[39] Disinformation can intensify these tensions by, for example, disseminating false reports (including doctored images or videos) of inter-communal violence, leading to heightened mistrust, fear, or hatred, and sometimes cycles of reprisal attacks before the original incident can be fact-checked.

Breakdown in the rule of law: This refers to situations where legal systems are either weak, corrupt, or non-functional. It results in a lack of justice and accountability, fostering an environment where human rights abuses can occur unchecked. This breakdown can lead to a loss of public trust in institutions, escalating conflicts as communities or groups take matters into their own hands.[40] Disinformation campaigns can create or intensify public perceptions of such breakdowns by, for example, spreading false stories of instances of corruption or injustice by State actors.

Absent or co-opted State institutions: In this scenario, State institutions either fail to effectively serve the population or are manipulated for the benefit of a few, often leading to widespread corruption, nepotism, and patronage. Such situations can exacerbate inequalities, fuel grievances, and create power vacuums that can be exploited by violent actors.[41] Disinformation campaigns by political actors to maintain or gain power can delegitimize State institutions and erode trust in them, making peacebuilding and the prevention of conflict more difficult.

Illicit economic gain: This driver relates to the exploitation of conflict for personal or group enrichment through illegal means, such as arms trafficking, the drug trade, or illegal resource extraction. These activities can provide financial incentives for continuing conflicts and can empower non-State armed groups or corrupt officials.[42] A recent study found that Southeast Europe, since the beginning of the war in Ukraine, has seen an increase in illicit economic activity in parallel to disinformation and 'fake news' campaigns.[43] Although it did not pinpoint a mechanism by which disinformation contributes to illicit economic gain, it does establish a correlation between disinformation and situations of conflict.

Resource scarcity, exacerbated by climate change: Climate change can intensify resource scarcity, particularly in regions dependent on natural resources for survival. Droughts, floods, and other climate-related events can lead to competition over water, land, and other scarce resources. This competition can turn violent, especially in areas where governance is weak, and adaptive capacities are low. The impact of climate change on resource availability can exacerbate existing tensions and contribute to new conflicts.[44] Disinformation campaigns about the reality of climate change make it more difficult to tackle its increasing urgency because groups continue to question its legitimacy and debate its existence, which fragments opinions and responses to it.

---

39 Stefan Döring and Katariina Mustasilta, "Spatial Patterns of Communal Violence in Sub-Saharan Africa," *Journal of Peace Research* (2023): 1–16. Accessible at: https://doi.org/10.1177/00223433231168187.

40 "Blueprint for Transformative Change through the Rule of Law and Human Rights," UNDP, last accessed on 16 January 2024, https://www.undp.org/sites/g/files/zskgke326/files/2022-08/Blueprint%20for%20Transformative%20Change%20through%20the%20Rule%20of%20Law%20and%20Human%20Rights%202022-2025%20lv.pdf.

41 Ibid.

42 Summer Walker and Mariana Restrepo, *Illicit Economies and Armed Conflict* (Geneva: Global Initiative Against Transnational Organized Crime, 2022). Accessible at: https://globalinitiative.net/wp-content/uploads/2022/01/GMFA-Illicit-economies-28Jan-web.pdf.

43 Center for the Study of Democracy (CSD), "Illicit Financial Flows and Disinformation in Southeast Europe," *Policy Brief No. 126* (Sofia: CSD, 2023 https://csd.bg/publications/publication/illicit-financial-flows-and-disinformation-in-southeast-europe/.

44 UN Interagency Framework Team for Preventive Action, *Renewable Resources and Conflict* (2012). Accessible at: https://www.un.org/en/land-natural-resources-conflict/pdfs/GN_Renew.pdf.

These conditions have emerged as primary drivers of conflict globally that disinformation, as explored below, exacerbates in the sub-Saharan context. Disinformation also impedes peacebuilding efforts at a fundamental level. It is well demonstrated that stakeholders in this context must mutually acknowledge key conflict factors, often as a necessary condition for progress toward peace.[45]

While individuals do not need to agree on the causes of conflict, there must be a level of shared understanding of the major issues in the conflict context. When these factors are not mutually recognized or acknowledged, and particularly when there is denial or 'undiscussability' of key drivers, it can inhibit the possibility of establishing a shared commitment to address them, significantly undermining peace processes, and/or freezing progress beyond a mere absence of violence.[46]

As explained by multiple interview respondents, disinformation more often targets and exploits pre-existing sentiments and beliefs rather than creating completely novel narratives. As articulated by Jamie Hitchen, an independent research analyst and Honorary Research Fellow at the University of Birmingham, UK:

> The mis- and disinformation is designed to … exacerbate that pre-existing belief and feed into it and … then reiterate and reaffirm it, and then it becomes more and more difficult or becomes more and more easy for those people that have a pre-existing belief to continue to hold those beliefs and even becomes more difficult for people [who] don't hold them to push back against them … You'll see mis- and disinformation narratives, piggybacking onto existing beliefs, and really often accentuating and driving those further.[47]

Foreign Affairs Analyst Chris O. Ogunmodede similarly notes that "there's the occasional outright fabrication. But very often, it's more of the mixture of truth, conjecture, guesswork mixed with some outright fabrication." He goes on to explain that "there's a lot of disinformation that tends to fly around already contentious, real issues, such as partisan politics, regional conflicts, and inter-communal conflicts."[48]

In addition to exacerbating structural and proximate causes of conflict, disinformation can also serve as a trigger for violence to break out. Jamie Hitchen describes examples:

> [V]ideos depicting attacks between individuals or ethnic groups can rapidly incite reprisal violence before any fact-checking response is possible. Often, divisive rhetoric is interpreted as a call to action, fueling violence based on the belief that one's group is being marginalized by another. In contexts where violent clashes are frequent, such misinformation only exacerbates the risk of these confrontations escalating in intensity and frequency.[49]

Disinformation exploits and reinforces existing issues and grievances, with the effect of confusing objective conditions, deepening beliefs in conflicting or polarizing narratives, spreading fear, mistrust, and/or hatred of different sociocultural groups and/or institutions, and serving as a call to action to commit violence in high-pressure contexts. As Beatrice Bianchi, Political Analyst and Sahel Expert at Med-Or Leonardo Foundation explains:

> Disinformation on social media is mainly on political and conflict-related issues … One of the biggest [examples of disinformation] was this year, February 2023, when there was disinformation saying that the army of Niger found the French military together with terrorists … The situation became so serious that the presidency of Niger had to make a public statement.[50]

Multiple interviewees identified elections as a common high-pressure situation in which disinformation is leveraged for various ends. Alpha Daffae Senkpeni, Executive Director at Local Voices Liberia Media Network, describes how domestic politicians "use disinformation as a means of persuasion" in the run-up to elections:

---

45 Diana Chigas and Peter Woodrow, *Adding up to Peace: The Cumulative Impacts of Peace Initiatives* (Cambridge, MA: CDA Collaborative Learning Projects Inc., 2018). Accessible at: https://www.cdacollaborative.org/wp-content/uploads/2018/04/ADDING-UP-TO-PEACE-The-Cumulative-Impacts-of-Peace-Initiatives-Web-Version.pdf; "Strategy 2021–2025: A Resilient Peace," Interpeace, last accessed on 16 January 2024, https://www.interpeace.org/strategy-2021-2025-resilient-peace/.

46 Chigas and Woodrow (2018) *Adding up to Peace*.

47 Jamie Hitchen, Independent Research Analyst and Honorary Research Fellow at the University of Birmingham, UK. Interview conducted via videoconferencing technology, September 2023.

48 Chris O. Ogunmodede, Foreign Affairs Analyst. Interview conducted via videoconferencing technology, September 2023.

49 Jamie Hitchen, Interview conducted via videoconferencing technology, September 2023.

50 Beatrice Bianchi, Political Analyst Sahel Expert, Med-Or Leonardo Foundation. Interview conducted via videoconferencing technology, September 2023.

> They craft all kinds of distorted information aimed primarily at gaining political support, utilizing social media and occasionally mainstream media. In promoting their agenda, they often mix disinformation with factual content, and sometimes include hate speech, all as strategies to persuade the public and control the narrative.[51]

"Disinformation is often at the root of acts of violence" because, in addition to fuelling uncertainties, tensions, divisions, and grievances, it can intensify the extent to which various stakeholders have radically different perceptions and understandings of the context and factors of the conflict, and can lead people "to use false or incomplete information as a basis for their decisions."[52]

In addition to identifying this new era of conflict, the UN@75 report provides examples of the ways in which AI-powered disinformation drives conflict. Violent extremist groups use disinformation to disseminate xenophobic speech and incite violence, giving social media a "crucial role" in driving conflict. Moreover, AI helps facilitate the "more efficient and effective" spread of disinformation, speeding up the rate at which extremist groups can recruit, incite violence, and share propaganda.[53] This increased efficiency is made even more dangerous by both State and non-State actors using "AI-enabled deep learning to create deep fakes."[54] These new technologies, without proactive management, jeopardize the digital information landscape and, as the report explains, drive and contribute to conflict.[55]

AI-Powered Disinformation as Proof of Harmful Narratives

One of the important mechanisms by which AI can counteract peacebuilding efforts is through the creation of false proof for false narratives, particularly those that involve violence. Alphonce Shiundu, Kenya Editor at Africa Check, substantiates this, stating: "The prevalent forms of misinformation are still rooted in rumours, hate speech, and conspiracy-laden false narratives."[56] He identifies the real danger of AI-generated content as the potential to create convincing proof for these narratives, which then become difficult to disentangle from reality.[57] Videos or images containing graphic violence are already regularly disseminated on social media. A simple mechanism for image doctoring involves taking a real photo, for example one portraying persons killed in a violent incident, and relabeling it with a claim that this incident took place in a different country, or with different perpetrators and victims than those really involved.

One of the prominent features of disinformation campaigns in the three focus countries is the way in which they target humanitarian interventions and peacekeeping operations. The UN has a number of peacekeeping missions in sub-Saharan Africa – MINURSO in Western Sahara, UNMISS in South Sudan, UNISFA in Abyei, MINUSCA in the Central African Republic, and MONUSCO in the Democratic Republic of the Congo (DRC). MINUSMA, in Mali, ceased operations on January 1, 2024.

Populations in these countries have had mixed feelings in relation to the presence of UN forces, especially in DRC and Mali, where this has led to attacks on peacekeeping forces.[58] Some have argued that disinformation campaigns are to blame for these aggressions, especially foreign actor-led campaigns which aim to undermine the activities of the UN more broadly.

On the other hand, certain State and non-State actors have continued to claim that dissent towards UN peacekeeping is legitimate, as it is linked to a desire for emancipation from former colonial powers in addition to the perceived failures of peacekeeping operations to provide protection and increased security.[59] It is often tied to a desire for self-determination and autonomy in the management of natural resources and local economies, rather than being tributary to various foreign actors.

---

51 Alpha Daffae Senkpeni, Executive Director at Local Voices Liberia Media Network. Interview conducted via videoconferencing technology, September 2023.
52 Fondation Hirondelle, *Annual Report 2022* (2023). Accessible at: https://rapportannuel.hirondelle.org/en.
53 United Nations, *A New Era of Conflict and Violence*.
54 Ibid.
55 Ibid.
56 Alphonce Shiundu, Kenya Editor at Africa Check. Interview conducted via videoconferencing technology, September 2023.
57 Ibid.
58 Albert Trithart, *Local Perception of UN Peacekeeping: A Look at the Data* (New York: International Peace Institute, 2023). Accessible at: https://www.ipinst.org/wp-content/uploads/2023/09/2309_Local-Perceptions.pdf.
59 Ibid.

To understand and address the dissent against peacekeepers, one must consider the propagation of false narratives against them. It is a source of the dissent, though not the only one; therefore, to address the dissatisfaction, it is necessary to identify these false narratives. For example, false claims in DRC accuse UN peacekeepers of selling weapons to armed groups, supporting foreign troops, and participating in illegal natural resource extraction.

The digital sphere has been the primary platform for the dissemination of "blended disinformation," where claims that are negative but potentially valid are mixed with false claims, adding to an atmosphere of tension and distrust. Often, disinformation content can be presented as a type of proof to general population discomfort. For example, the mislabeling of images when shared on social media, has been quite common.[60]

Mechanisms for Diffusion of Disinformation Campaigns

It is important to note that disinformation campaigns might operate across multiple platforms, including traditional media, and in some cases, word of mouth. Each of these mediums reach different audiences, which can accelerate the impact of a disinformation campaign as it leaps across the digital divide.[61] In Côte d'Ivoire, for example, and many other countries in the region, the majority of the rural population does not have an Internet connection or a device, and therefore many people do not obtain information online. These differences in access impact how people receive disinformation. Certain people with social media access might therefore receive disinformation and then share it with their community verbally, or by more traditional mediums such as radio.[62] The table below outlines the main mechanisms for the diffusion of disinformation observed in this research.

Mechanism	Examples
Mainstream Media	
Radio	During and after the 2007 elections in Kenya, local radio stations, such as KASS FM, Eldoret's popular Kalenjin radio station, stoked existing ethnic tensions and used inflammatory language.[63] These radio stations shared what has since been characterised as hate speech, sharing disinformation in the process.[64] In 2019, influencers began to share pro-Russia messaging on a radio station in the Central African Republic (CAR),[65] following the military cooperation deal between the two States.[66]
Television	Afrique Média, a popular Cameroon-based television channel, aligns itself with a pro-Russia narrative, disseminating pro-Russian propaganda on its network and social media channels. The network's media executive was connected with the Wagner Group and reported extensively on its operations. Code for Africa, a Kenyan non-profit, identified the television network as being partially responsible for the amplified reach of Russia and Wagner through its propaganda campaigns.[67]

60 Emile Beraud and Erin Flanagan, "Posts use old video in misleading claim about UN drone crashing in DRC with weapons and gold," *AFP Fact Check*, 28 November 2023, https://factcheck.afp.com/doc.afp.com.34689T4.

61 Jeffrey Conroy-Krutz and Joseph Koné, "AD410: Promise and peril: In changing media landscape, Africans are concerned about social media but opposed to restricting access," *Afrobarometer Dispatch No. 410* (Ghana: Afrobarometer, 2020). Accessible at: https://www.afrobarometer.org/publication/ad410-promise-and-peril-changing-media-landscape-africans-are-concerned-about-social/.

62 Elena Gadjanova et al., *Misinformation Across Digital Divides: Theory and Evidence from Northern Ghana* (New York: Columbia University, 2022). Accessible at: https://doi-org.ezproxy.cul.columbia.edu/10.1093/afraf/adac009.

63 "Ballots to Bullets: Organized Political Violence and Kenya's Crisis of Governance," Human Rights Watch, 16 March 2008, https://www.hrw.org/report/2008/03/17/ballots-bullets/organized-political-violence-and-kenyas-crisis-governance#_ftn113.

64 Gillian McKay, *Disinformation and Democratic Transition*.

65 "Mapping Disinformation in Africa," Africa Center for Strategic Studies, 26 April 2022, https://africacenter.org/spotlight/mapping-disinformation-in-africa/.

66 "Russia Signs Military Deal with the Central African Republic – Agencies," *Reuters*, 21 August 2018, https://www.reuters.com/article/us-russia-centralafrica-accord-idUSKCN1L60R2/.

67 Gretel Kahn, *A Kremlin Mouthpiece at the Heart of Africa: How Afrique Média Helps Putin Court Audiences in their Own Language* (Reuters Institute and University of Oxford, 2023). Accessible at: https://reutersinstitute.politics.ox.ac.uk/news/kremlin-mouthpiece-heart-africa-how-afrique-media-helps-putin-court-audiences-their-own.

Mechanism	Examples
Traditional news outlets	In 2019, Russian State media retweeted expansion plans in Africa, which was picked up as normal sources and shared by over 600 African news websites.[68]
Digital Platforms	
Fake web pages	Fake accounts, supporting the Government of the United Republic of Tanzania, targeted the Twitter accounts of leaders of Tanzanian civil society. A report by the Internet Observatory at Stanford University found that in at least three instances, these accounts created websites with falsified evidence to legitimize their targeted reporting for the removal of the leaders from Twitter.[69]
Social media platforms	Actors spreading disinformation on these platforms have many techniques, including shadow-boxing (using influencers for specific campaigns), sponsored content (such as pro-Russia messaging targeting audiences in Côte d'Ivoire), inauthentic social media networks (such as the fake Kampala Times Facebook page), bots and fake accounts (such as impersonated accounts in Ethiopia in 2020), inciting strong emotions by appealing to intercommunal divisions (such as demonizing different cultural groups in the DRC), and transnationally coordinated click-to-tweet campaigns (such as coordinated networks targeting countries in the Sahel region, promoting anti-Western narratives). Established political and community networks on social media platforms such as WhatsApp have far reaching effects, as noted by Jamie Hitchen: > In 2019, groups in Nigeria, linked to different political campaigns, established a WhatsApp group for each of the 774 local government areas. Each group was managed by two or three individuals with administrative authority. These groups, open for people to join, could reach the maximum capacity allowed by WhatsApp. Additionally, there were 36 State-level groups, which supplemented the local government ones. This structure formed a pyramid-like system where information was centrally distributed within the 36 State groups and then disseminated to the 774 local government groups. From there, it became a free-for-all, with the information being shared broadly to reach as wide an audience as possible.[70]
Offline Networks	
Word of mouth	Disinformation, once initiated on the previously stated mechanisms, can permeate into broader, offline communities, by trusted local voices with influence, such as religious or community leaders. As discussed by Jamie Hitchen: > The critical aspect of mis- and disinformation I always stress is understanding the overlap between online and offline. For instance, a piece of false information originating on Twitter, where only 10 per cent of Mali's population might be active, can escalate into a wider public discourse. It may be picked up by a local radio station, reaching a larger audience through broadcasts in the local language. This discussion about the fake news then permeates into public spaces like markets and local communities, further spreading through offline networks. This transition from online to offline significantly amplifies the reach and impact of the misinformation.[71]

68 "Mapping Disinformation in Africa," Africa Center for Strategic Studies.

69 Shelby Grossman et al. "The New Copyright Trolls: How a Twitter Network Used Copyright Complaints to Harass Tanzanian Activists," Stanford University, Internet Observatory: Cyber Policy Center, 2 December 2021, https://cyber.fsi.stanford.edu/io/publication/new-copyright-trolls.

70 Jamie Hitchen. Interview conducted via videoconferencing technology, September 2023.

71 Ibid.

Actors in AI-Powered Disinformation Campaigns

Although many respondents discussed the role of foreign actors in spreading disinformation, some minimized this issue, claiming that blaming foreign actors distracted from the responsibility of local governments and politicians. In fact, there are several categories of actors involved in creating and spreading disinformation in the region. These actors were involved in disinformation even before new developments in AI technologies. However, their AI adoption rates differ, with foreign actors most likely to display mastery of more recent AI tools.

Local influencers, intellectuals, and media figures: In 2021, a scholar with government connections in Eritrea published a pseudo fact-checking report, validating its position.[72] Also in 2021, in Kenya, verified social media users were paid to rent their accounts in order to conduct disinformation campaigns.[73] In 2022, the social media accounts of authentic Nigerian journalists were hacked, and used to post pro-Russian propaganda.[74] Jamie Hitchen explains the significance of targeting journalists this way:

> [Their voice] is going to resonate and be more relevant to its audience than a Russian TV channel trying to push its narrative. It's better if they're able to work through local people who have influence, be that religious leaders, community leaders, people online who are listened to, that are able to then push their narrative for them. There is an industry around this.[75]

Private firms: There are also many private companies participating in disinformation in the region. For example, MintReach, which was active in Nigeria and Egypt in 2019, promoted positive content about the United Arab Emirates while criticising Qatar, Türkiye, and Iran.[76] Its page was removed by Facebook due to false news claims.[77] In 2016, Bell Pottinger, a British PR firm, was retained by the Gupta brothers in South Africa to amplify disinformation websites.[78] The Israeli company Archimedes Group was also hired to promote certain politicians and denigrate others, influencing various elections in sub-Saharan Africa.[79]

Foreign actors: Many foreign countries have been accused of using disinformation campaigns to meddle with politics in the region, including Russia, China, the United Arab Emirates, Türkiye, Saudi Arabia,[80] and France.[81] Lilian Olivia, Advocate for the High Court of Kenya and Founder of Safe Online Women Kenya, provided more details on AI-specific methods used by these foreign actors:

> A few months ago, in Zimbabwe's recent election, China's introduction of mass surveillance and facial recognition technology has had a significant impact. This technology enables access to individuals' images through facial recognition and, through generative AI, gathers extensive data on political candidates and citizens. Before you know it, this will spread a lot of disinformation.[82]

72 Digital Forensic Media Lab, "Eritrean Report Uses Fact-checking Tropes to Dismiss Evidence as 'Disinformation,'" *Medium*, 23 June 2021, https://medium.com/dfrlab/eritrean-report-uses-fact-checking-tropes-to-dismiss-evidence-as-disinformation-385718327481.

73 Odanga Madung and Brian Obilo, *Inside the Shadowy World of Disinformation-for-hire in Kenya*.

74 "Disinformation and Russia's War of Aggression Against Ukraine," OECD Ukraine Hub, 3 November 2022, https://www.oecd.org/ukraine-hub/policy-responses/disinformation-and-russia-s-war-of-aggression-against-ukraine-37186bde/.

75 Jamie Hitchen. Interview conducted via videoconferencing technology, September 2023.

76 Amos Abba, "Facebook confirms removal of accounts peddling fake news in Nigeria, UAE and Egypt," International Centre for Investigative Reporting, 4 October 2019, https://www.icirnigeria.org/facebook-confirms-removal-of-accounts-peddling-fake-news-in-nigeria-uae-and-egypt/.

77 Nathaniel Gleicher, "Removing Coordinated Inauthentic Behavior in UAE, Nigeria, Indonesia and Egypt," *Meta*, 3 October 2019, https://about.fb.com/news/2019/10/removing-coordinated-inauthentic-behavior-in-uae-nigeria-indonesia-and-egypt/.

78 Adriaan Basson, "The End of White Monopoly Capital," *News24*, 2 November 2018, https://www.news24.com/news24/opinions/columnists/adriaanbasson/the-end-of-wmc-20181101.

79 Digital Forensic Media Lab, "Inauthentic Israeli Facebook Assets Target the World," *Medium*, 17 May 2019, https://medium.com/dfrlab/inauthentic-israeli-facebook-assets-target-the-world-281ad7254264.

80 Idayat Hassan, "Disinformation is Undermining Democracy in West Africa."

81 Graphika and the Stanford Internet Observatory, *More-Troll Kombat: French and Russian Influence Operations Go Head to Head Targeting Audiences in Africa* (2020). Accessible at: https://graphika.com/reports/more-troll-kombat.

82 Lilian Olivia, Advocate of the High Court of Kenya and Founder of Safe Online Women Kenya. Interview conducted via videoconferencing technology, September 2023.

Non-State actors: There have also been incidences of non-State actors conducting disinformation campaigns, including diaspora groups, as Foreign Affairs Analyst Chris O. Ogunomodede revealed:

> Significant support for these troll farms comes from countries with large diaspora communities. In the case of Cameroon, there is notable involvement from the United States and several European countries, including France, Belgium, the Netherlands, Germany, and the UK, where there are substantial Cameroonian diaspora populations. This pattern of diaspora involvement in troll farms is a recurring theme in regions experiencing conflict.[83]

In addition, Alphonce Shiundu shared that "M23 [a Tutsi-led rebel group] has the capability" to launch disinformation campaigns.[84] Although it is not possible to confirm if they have, it is notable that they can do so.

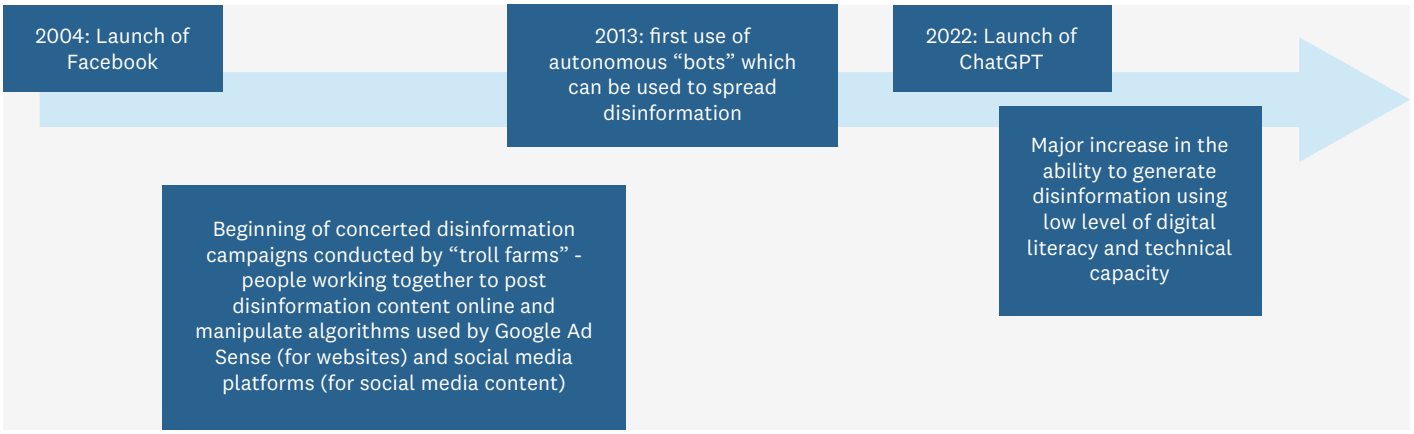
Audiences: Broadly, a general public is targeted, especially one that would either vote or take action, including violent action, against disinformation targets. Youth in the region are especially active on social media and likely to be influenced by disinformation, and often in turn spread misinformation in wider communities. Young people are also often specifically targeted by manipulative disinformation campaigns that aim to exploit their vigour and naivety of recent history.[85]

Evolving Methods in AI-Powered Disinformation

In the region, disinformation has been spreading regardless of its source, or even the sophistication of the false content. Doctored images, even if they easily appear to be fake, are often shared, as are deep fakes using older technologies, which more clearly show gaps and glitches. In this sense, the true impact of AI-generated disinformation will not be observed before some time, as the new tools become more broadly adopted in the region.

However, respondents cited technical skills most often as the main barrier to AI-generated disinformation. This may no longer be the case, as AI-generated content, whether text, image, or even audio and video, are quite easily accessible. In mid-2023, a few weeks after demonstrations against MONUSCO broke out in Goma, DRC, deep-fake videos went viral of President Emmanuel Macron appearing to call the population to mobilize against the regime of President Félix Antoine.[86]

In fact, disinformation techniques have evolved very rapidly over the last decade. The diagram below displays an approximate timeline for this evolution. The box describes some characteristics of troll farms as they have developed in recent years.



83 Chris O. Ogunmodede. Interview conducted via videoconferencing technology, September 2023.
84 Alphonce Shiundu. Interview conducted via videoconferencing technology, September 2023.
85 Ethical Productions ltd, "Misinformation Great Lakes V2" [Interview with Christophe Hamisi], video, 28 October 2023, https://www.youtube.com/watch?v=91W-IFePa_A.
86 Ibid.

Box: What is a Troll Farm?

Troll farms have their roots in the early 2010s and are predominantly associated with disinformation campaigns. These organized entities employ individuals, known as 'trolls,' to manipulate public opinion online. Using fake profiles and automated bots, they disseminate disinformation in a coordinated way and amplify divisive narratives across social media platforms and online forums. Kwami Ahiabenu II, Director at Penplusbytes, highlighted an important feature in the scalability of automated bots in that their "deployment in disseminating disinformation is not only efficient but also cost-effective … [Therefore,] these bots can manipulate narratives and public opinion on a large scale with minimal resource expenditure."[87]

Their tactics often involve exploiting existing societal fissures, political polarizations, or contentious events to create confusion or deepen divisions. The most notable example is the alleged interference of troll farms in the 2016 US presidential election, which brought significant international attention to the phenomena. Since then, the awareness of troll farms has grown, and their tactics have been adopted by various State and non-State actors worldwide to influence public sentiment and political landscapes.

It should be noted that troll farms are often distributed networks of individuals collaborating across international borders, rather than physically co-located groups. As Chris O. Ogunomodede notes: "A lot of the troll farms are based in these countries that, in many instances, are based in EU Member States among diaspora groups of those countries in Europe and the United States or Canada."[88] This highlights the international reach of troll farm activities and underscores the importance of multinational efforts, like those of the UN and EU, in addressing their impact. These networks are also varied in nature. Foreign Affairs Analyst Chris O. Ogunomodede adds further:

> Troll farms are basically networks of people connected to a cause. They are often very tech-savvy people, recruited by ideologically-driven individuals or groups for their technical skills … You will have groups linked to political campaigns and funding troll farms to send out information and amplify [their cause] on platforms like Twitter, Facebook, Instagram, Telegram, [and] WhatsApp.[89]

Other respondents suggest that foreign actors are often responsible for using troll farms to conduct disinformation in Africa. These entities are highlighted here because they have been critical players in exploiting AI for disinformation practices. Originally manipulating recommendation systems, developing social media bots, and hijacking accounts, they are now very likely to begin using more sophisticated methods of generating disinformation.

87 Kwami Ahiabenu II, Director at Penplusbytes. Interview conducted via videoconferencing technology, September 2023.
88 Chris O. Ogunmodede. Interview conducted via videoconferencing technology, September 2023.
89 Ibid.

Case Study 2: Healthy Information Ecosystems for Peacekeeping in DRC

The DRC faces numerous challenges to political stability, including corruption, poor governance, leadership disputes, and armed conflicts over control of mineral resources. This has resulted in endemic instability and has left a significant portion of the population at risk. At least 24.6 million people in the DRC are at risk of food insecurity, and an additional 6 million citizens have been internally displaced – the highest number on the African continent. Violence has recently flared in Ituri and Kivu provinces, leading to the death of at least 1,300 people.

Observers have found that online disinformation plays a key role in fueling this conflict. As Sammy Mupfuni, Managing Director at CongoCheck, shared: "In the DRC, fake news and misinformation can be lethal. They contribute to a climate where misinformation can directly lead to death."[90] Social media platforms spread false stories as various groups are taking to using bots to manipulate public opinion and raise tensions. Legacy media outfits struggle to keep up with the speed at which disinformation spreads on social media and are, therefore, unable to counter the false claims. Low media literacy among the population further amplifies the impact of disinformation.[91]

In July 2022, there was a resurgence of violence in eastern DRC due to clashes between the Tutsi-led rebel group M23 and the Congolese army. This conflict led to a surge of online activity where disinformation played a large role in creating even more division and tension, which in turn led to more conflict. The DRC has accused Rwanda of supporting the rebellion and fostering instability. By late July, the hashtag #RwandaIsKilling started trending on social networks, with posts accusing Rwanda of engaging in warfare in eastern DRC. However, many of these posts contained false information that only exacerbated divisions.

CongoCheck, a non-profit organization made up of independent journalists that fact-check articles and other media sources, emerged as one of the dedicated organizations working to verify facts and contribute to a de-escalation of tensions. CongoCheck works with Facebook through its 'Third Party Program' that provides fact-checking resources to independent partner organizations that "review and rate the accuracy of stories."[92] One limitation of CongoCheck is that it does not seem to employ any automated AI-based method for filtering large volumes of content, as may soon come to be created via generative AI.

A report developed by Insecurity Insight highlights a particular case of disinformation in DRC. Following the assassination of the Italian ambassador to the DRC, which occurred while the ambassador was travelling with a World Food Programme (WFP) and UN Organization Stabilization Mission in the DRC (MONUSCO) convoy, disinformation started to spread online accusing the UN and international agencies of stealing resources from DRC.

90 Sammy Mupfuni, Managing Director CongoCheck. Interview conducted via videoconferencing technology, September 2023.

91 Nadine Temba, *Disinformation and Hate Speech Continue to Fuel the Conflict in Eastern DR Congo* (Collaboration on International ICT Policy for East and Southern Africa - CIPESA, 2023).

92 France 24 Observers, "#RwandaIsKilling: Tensions Between Rwanda and DRC Fuel Misinformation," video, 29 August 2022, https://observers.france24.com/en/tv-shows/truth-or-fake/20220829-rwandaiskilling-tensions-between-rwanda-and-drc-fuel-misinformation; See: CongoCheck, https://congocheck.net/; and Facebook Third Party Program: https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking/partner-map.

A post entitled "When MONUSCO Plunder Us" was shared widely on Facebook groups in North Kivu, claiming that "locals found sacks of cobalt minerals being transported by a MONUSCO vehicle," and accusing the "UN mafia of stealing natural resources and turning their base into a warehouse." The post makes a direct link with the incident involving the WFP convoy and the Italian ambassador, "posing questions on the involvement of UN agencies in the continuation of the war in Congo."

Increasingly, multilateral agencies and foreign actors are becoming aware of the importance of this kind of disinformation. Disinformation may decrease trust in (and access for) aid and security agencies operating in an area. It also highlights the importance of maintaining healthy information ecosystems for multilateral entities to maintain trust and ensure the safety of aid workers.[93] The head of MONUSCO's office, Leila Bourhil, has stated in this regard: "There was a campaign of disinformation, misinformation, and false rumours which created a breach of trust that MONUSCO and the population shared. It is important that we rebuild that trust, and in this regard, we need to engage and communicate much better. Initiatives [must] seek to establish more open lines of communication between MONUSCO and the Congolese people."[94]

93 Insecurity Insight, "Disinformation Targeting the United Nations Presence in the DRC," (2021). Accessible at: https://reliefweb.int/report/democratic-republic-congo/social-media-monitoring-democratic-republic-congo-disinformation.

94 "North Kivu: MONUSCO Encourages Communities to Foster Social Dialogue with a view to Fighting Misinformation," UN Peacekeeping, 25 May 2023, https://peacekeeping.un.org/en/north-kivu-monusco-encourages-communities-to-foster-social-dialogue-with-view-to-fighting; See also Albert Trithart, "Disinformation against UN Peacekeeping Operations," International Peace Institute, November 2022, https://www.ipinst.org/wp-content/uploads/2022/11/2212_Disinformation-against-UN-Peacekeeping-Ops.pdf; and "DR Congo Mission Chief Leads Proactive Fight Against Deadly Misinformation," *UN News*, 31 March 2022, https://news.un.org/en/audio/2023/03/1135227.

4. Mitigation Methods to Address AI-Powered Conflict Drivers

Grassroots and Civil Society Initiatives

Overwhelmingly, the fight against disinformation in the region is conducted by grassroots initiatives such as Local Voices Liberia Media Network (LVL), Africa Check, and CongoCheck. Fact checkers in these organizations all tend to value manual fact-checking over AI-powered fact-checking, with Alpha Daffae Senkpeni, Executive Director at LVL, saying: "[W]e use our human ability to make a decision as to whether or not this is worth spending time to fact-check or not."[95] Ernest Dukuzumuremyi, Programme Manager of Interpeace's Rwanda Programme also notes a tendency not to rely on AI-powered fact-checking: "[O]ur approach is mainly participatory research, involving dialogues with diverse groups, including young men and women and ordinary citizens."[96]

These organizations are largely staffed with journalists, who fight disinformation using fact-checking methods that would be used pre-publication in professional journalism more broadly. An illustration of this fact-checking process is included in the box below.

In fact, independent journalists have a very specific skill set in fighting disinformation that make them uniquely suited for the job. In addition, as various actors wield disinformation as a force for persuasion, actors that are vested in sharing accurate information have the upper hand when it comes to credibility.

In addition, although it might seem efficient to use AI-powered tools to fight disinformation, this should be done very carefully. A model's ability to accurately identify salient disinformation requires layers of analysis, not all of which are clear-cut. "Just checking facts is hard enough," said one Computer Science Professor at Columbia University, "but gauging intent is significantly harder because determining whether something is propaganda, or clickbait – and whether it's intended to cause harm – can be very subjective."[97]

However, as Naomi Miyashita, Project Manager, Addressing Mis/Disinformation, at the United Nations, notes:

> AI-generated images, commonly utilized by disinformation agents, aren't yet so advanced that they can't be detected with the right tools. Photos fabricated using AI technology can still be identified computationally. Therefore, it's important and useful for platforms, fact-checkers, and organizations, including the UN, to employ these detection tools.[98]

For the successful implementation of these tools, it is critical for organizations to have financial and material support, and, as Albert Trithart, Editor and Research Fellow at the International Peace Institute, highlights, the growing number of African fact-checking organizations "don't necessarily have the level of resources they need to really make a meaningful difference." He stresses the importance of not just addressing individual falsehoods but also tackling the broader false narratives they stem from.[99] In this sense, a more critical path forward is to provide financial and material support to journalistic fact-checking initiatives, so that they can more effectively conduct their various disinformation dismantling and digital literacy activities.

As explained by Ernest Dukuzumuremyi: "[W]hile our competence in AI is limited, it is a potential area for future work."[100] Another approach is to focus resources on increasing populations' media literacy and building capabilities in communications. By strengthening individuals' knowledge of disinformation tactics and indicators, and enhancing fact-checking and critical thinking skills, some organizations aim to limit the degree to which 'fake news' is believed and shared.[101]

95 Alpha Daffae Senkpeni. Interview conducted via videoconferencing technology, September 2023.
96 Ernest Dukuzumuremyi, Programme Manager of Interpeace's Rwanda Programme. Interview conducted via videoconferencing technology, November 2023.
97 Alla Katsnelson, "Identifying Misinformation's Intent."
98 Naomi Miyashita, Project Manager, Addressing Mis/Disinformation, United Nations. Interview conducted via videoconferencing technology, September 2023.
99 Albert Trithart, Editor and Research Fellow at International Peace Institute. Interview conducted via videoconferencing technology, September 2023.
100 Ernest Dukuzumuremyi. Interview conducted via videoconferencing technology, November 2023.
101 Ethical Productions ltd., "Misinformation Great Lakes V2" [Interviews with Fred Mfuranzima and John Paul Habimana], video, 28 October 2023, https://www.youtube.com/watch?v=91W-IFePa_A.

Social Media Uses of Content Moderation and Guardrails

There are many social media companies operating in sub-Saharan Africa including older players such as Facebook, Twitter, WhatsApp, LinkedIn, and Instagram; relatively new players such as TikTok and Telegram; and regional players such as 2go and Nairaland. These companies have been increasingly made aware of the importance of addressing disinformation and have developed a variety of techniques for doing so.

Content moderation procedures: A set of guidelines, processes, and practices employed by online platforms to monitor, review, and manage user-generated content. This ensures that the content aligns with the platform's policies and community standards. Common procedures include the removal of harmful content, flagging or labelling misinformation, blocking users who violate terms of service, and implementing automated systems to detect and handle inappropriate material. Three approaches to content moderation are described below.

Censorship: The suppression, prohibition, or alteration of speech, information, or content deemed unacceptable, harmful, or sensitive by a governing body, institution, or other authoritative entity. Censorship can occur in various forms, such as media, literature, and online platforms, and can be driven by political, moral, religious, or commercial reasons. Social media companies can moderate content on their platforms by overtly banning users that do not comply with their community guidelines.

Shadow-banning: This is another moderation tactic, used on online platforms, where a user's content is secretly made invisible or less prominent to the broader community without the user being aware. The user can still post and interact normally, but their content does not appear or has reduced visibility in search results, feeds, or other discovery mechanisms on the platform.

Partnerships with civil society organizations: These are collaborative relationships between private companies, online platforms, or governmental bodies and non-governmental organizations (NGOs), community groups, or other non-profit entities that represent the interests of the public. These partnerships aim to address societal challenges, promote transparency, and ensure that policies and practices consider a wider range of perspectives, especially in areas like content moderation and digital rights.

Unique Challenges to Mitigation in the Region[102]

The sub-Saharan Africa region has many characteristics which make fighting disinformation uniquely challenging. For example, countries in sub-Saharan Africa suffer from a pervasive lack of public trust in institutions which impairs the quality and effectiveness of institutions and government programmes, as well as social trust in government policies.[103] Highlighting this challenge, Jamie Hitchen remarked: "Often the narratives are out there. And people don't trust these institutions."[104] This lack of trust complicates efforts to establish credible and effective communication channels necessary for fighting disinformation. However, as shown below, many of these can be addressed with concerted efforts from social media platforms, AI companies, national governments, and international organizations. The unique challenges in the region can be summarized as below (see also the 'Unique Challenges' visualization.)

Under-resourced languages: Many African languages are under-resourced in content moderation tools, leading to increased susceptibility to harmful content online. Additionally, language diversity makes it difficult to monitor, moderate, or verify information. As Jamie Hitchen explains: "The capacity of AI in non-English languages raises questions, particularly for fact-checking and creating information in local languages beyond English and French. What is the capacity of AI in this context?"[105]

Poorly understood information ecosystem: There is a high demand for more evidence, data, and deeper research to understand the scale and nuances of disinformation. It is particularly difficult to track covert operations, such as hidden financing or digital operations that leave minimal traces.

---

102 These challenges exist in other areas around the world as well. We identify them as unique to the sub-Saharan African region because of their severity and pervasiveness, in addition to a lack of resources dedicated towards addressing and supporting local and regional organizations working to address these challenges. Additionally, it is important to note that countries, such as the United States, have similar levels of distrust in their government as many countries in the region. See: "2023 Edelman Trust Barometer: Global Report," Edelman, last accessed on 16 January 2024, https://www.edelman.com/sites/g/files/aatuss191/files/2023-03/2023%20Edelman%20Trust%20Barometer%20Global%20Report%20FINAL.pdf.

103 "CPI 2022 for Sub-Saharan Africa: Corruption Compounding Multiple Crises," Transparency International, 31 January 2023, https://www.transparency.org/en/news/cpi-2022-sub-saharan-africa-corruption-compounding-multiple-crises.

104 Jamie Hitchen. Interview conducted via videoconferencing technology, September 2023.

105 Ibid.

Technological and platform limitations: Different social media platforms operating in the region have inconsistent content moderation policies and may rely on automated systems that might not be attuned to local nuances or cultural context.

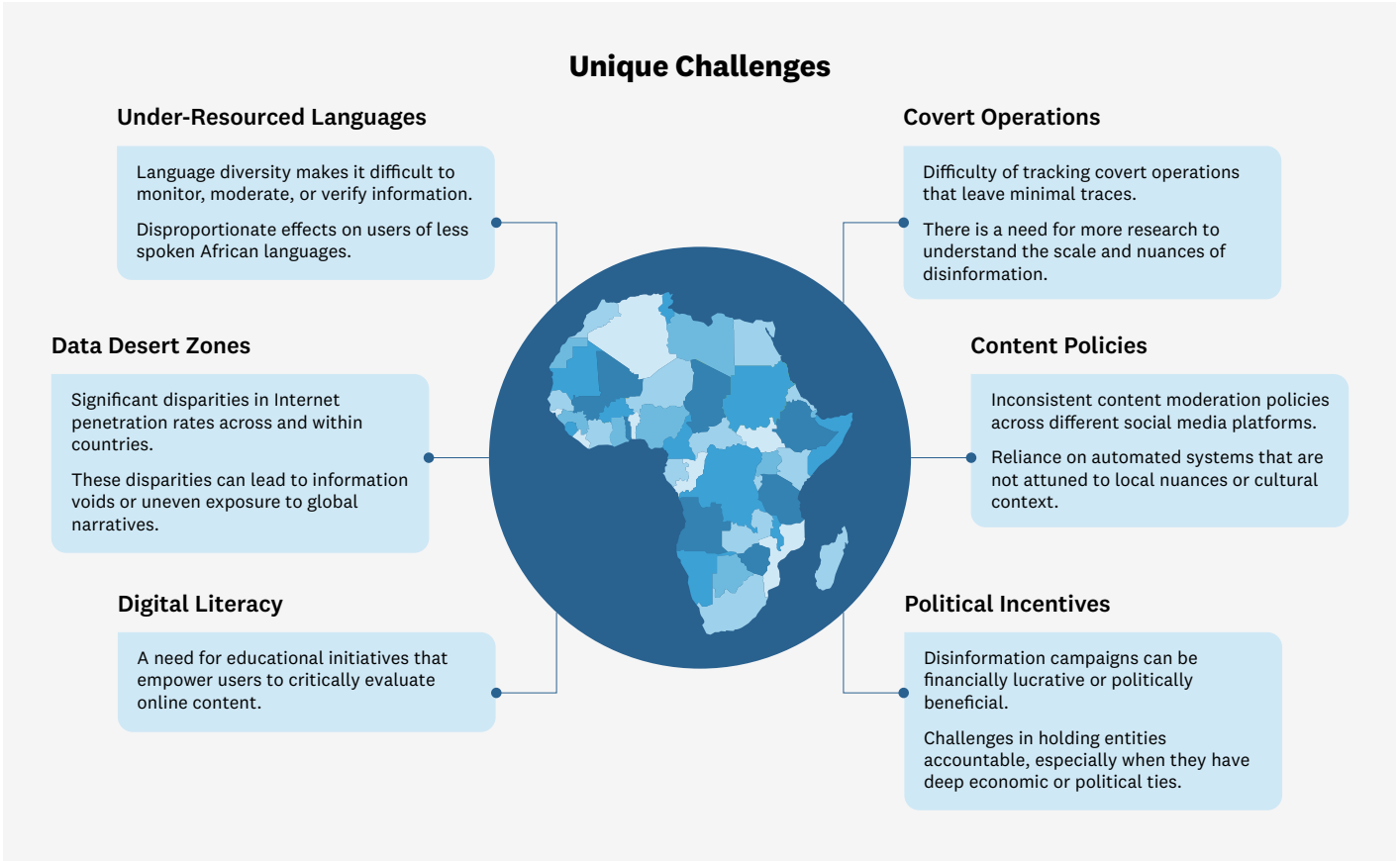
Lack of digital literacy: A significant portion of the population might not be equipped with the skills to discern credible sources from false ones, making them more susceptible to disinformation. Sammy Mupfuni shared an example from DRC:

> In our society, information from both social media and traditional sources, for example radio and television, is often accepted as true without critical scrutiny. People generally do not verify whether the content is fact-based. So, the primary issue is the lack of widespread media and information literacy education.[106]

A need has therefore been expressed for educational initiatives that empower users to critically evaluate online content.[107]

Political and economic incentives: Disinformation campaigns can be financially lucrative or politically beneficial, driving actors to continue these efforts despite the societal harm. It is challenging to hold powerful entities accountable, especially when they have deep economic or political ties.

Data desert zones and Internet penetration inequalities: There are significant disparities in Internet penetration rates across countries, and they are on the basis of several factors, including gender, socioeconomic status, and age. For instance, as of January 2022, Morocco, the Seychelles, and Egypt had rates of over 70 per cent for Internet penetration, in contrast to the CAR which had a rate of only 7 per cent. These disparities can lead to information voids or uneven exposure to global narratives, amplifying the effects of misinformation in areas with limited connectivity.



106 Sammy Mupfuni. Interview conducted via videoconferencing technology, September 2023.

107 A caveat, however, is that many countries with higher levels of digital access and familiarity are also vulnerable to disinformation campaigns. While digital literacy can certainly help individuals increase their discernment, it shouldn't replace tackling disinformation from a systemic perspective.

Opportunities for AI to Promote Peacebuilding

Using AI to promote peacebuilding is still relatively underexplored in the region, or even globally, especially if compared to the amount of resources dedicated towards exploring AI-powered disinformation. Nevertheless, there are still several promising avenues for AI in peacebuilding which may be harnessed, especially as the capability of AI technologies further evolve.

As explained by several respondents, much of the efforts to counter AI-powered disinformation have relied on traditional methods, such as journalistic fact-checking and enhancing media literacy capabilities. While valuable, these methods seem outmatched by the speed, virulence, and volume of AI-powered disinformation. However these methods can be enhanced through using these same tools. Below is an overview of how AI-technologies can enhance these processes.

Automated fact-checking: The implementation of AI in synthetic media detection harnesses the power of machine learning and classification algorithms, such as convolutional neural networks, recurrent neural networks, and natural language processing.[108] These sophisticated tools categorize data and identify various types of disinformation, streamlining the early detection process and managing the vast amounts of content online. This emerging technology could bolster the work of fact checkers by enhancing their efficiency in addressing the proliferation of fake news.

Fact-checking is a strategy already leveraged by many regional civic organizations. Alpha Daffae Senkpeni, asserts: "Fact-checking remains the most potent antidote to disinformation. Coupled with media literacy, it empowers the public to discern truth from falsehood."[109]

Lillian Olivia highlights the potential of AI in this process, noting: "In Kenya, technological aids like the Uchaguzi app were crucial during elections, enabling citizens to report inconsistencies." Automated synthetic media detection could significantly bolster such technologies, providing a more robust and scalable solution for real-time reporting and verification during critical events like elections. However, Olivia points out that integrating automated detection faces hurdles, particularly due to the digital divide: "Addressing disinformation, especially in electoral contexts, necessitates a focus on deep-fake detection and attribution. However, these technologies are not widely recognized in sub-Saharan Africa, demanding significant investments and educational efforts for effective deployment," while also competing with urgent issues like "food security and poverty."[110]

The linguistic diversity in the region poses an additional challenge for AI detection systems for disinformation, which struggle with local dialects and nuances. The understanding of these nuances is essential for accurate detection. Finally, these tools are subject to a number of technical challenges such as false positives, and a recent study found that "as language models become more sophisticated and better at emulating human text, the performance of even the best-possible detector decreases."[111]

Strengthening digital provenance: Provenance-based authentication utilizes cryptographic methods to trace digital content back to its source, employing metadata to inform users about the origins and changes to the content. This method offers a definitive evaluation of authenticity, fostering transparency and trust among media consumers.[112]

These efforts, which are enabled through the development of technical standards for certifying the source and history (or provenance) of media content are paramount, with major tech companies backing the initiative to encode the origin of content transparently. This approach could greatly assist in verifying the origin of content and combating the 'liar's dividend,' where genuine information is mistakenly dismissed as fake. However, for this approach to be effective, it must be implemented at scale and coupled with comprehensive media literacy initiatives. Users must understand that while watermarking can indicate content authenticity, it does not necessarily confirm factual inaccuracy.

108 Esma Aïmeur, Sabrine Amri, and Gilles Brassard, "Fake News, Disinformation and Misinformation in Social Media: A Review," *Social Network Analysis and Mining* Vol 13 Issue 1 (2023), p. 30.
109 Alpha Daffae Senkpeni. Interview conducted via videoconferencing technology, September 2023.
110 Lillian Olivia. Interview conducted via videoconferencing technology, September 2023.
111 Aounon Kumar et al. "Can AI-Generated Text be Reliably Detected?" *ArXiv*, 28 June 2023, https://arxiv.org/abs/2303.11156.
112 Imani Sherman, Elissa Redmiles, and Jack Stokes, "Designing Indicators to Combat Fake Media," ArXiv, 1 October 2020, https://arxiv.org/abs/2010.00544.

For sub-Saharan Africa, the public's ability to critically engage with content provenance could be challenged by a lack of media literacy programmes. Countries, such as Côte d'Ivoire, Kenya, and Nigeria, have integrated media literacy into State-run school curriculums but not substantively. They are the exception, with most countries including little to no media literacy in their curriculum.[113] Furthermore, Chris O. Ogunmodede warns of the technical challenges to this strategy, noting: "I have seen a lot of video clips where people put it to prevent its use for purposes of disinformation … I've already seen people removing such watermarks."[114]

Finally, if the institutions and news organizations that use provenance labels for content are not trusted by general audiences, watermarking and other forms of content provenance may have limited impact as labels could be perceived as an authoritarian or 'establishment' tool.[115] As previously noted, government regimes and political actors are active participants in the spread of disinformation for some countries in the region. "[W]e see politicians exploit stereotypes and prejudices for their gain," explained Ernest Dukuzumuremyi.[116]

Similar to this risk is that some information sources might be better equipped with the resources to implement a system of digital provenance than others. This risk could be exacerbated if provenance capability is used as a mechanism for filtering out content, or to influence ranking results in search engines.[117] As explained by Matthew Gentzel, Program Officer at Longview Philanthropy: "If you intentionally promote mainstream media sources, demote smaller sources, you might also take out independent investigators [and/or] local news."[118]

Cross-platform monitoring: Cross-platform monitoring by civic organizations provides a comprehensive outlook on social media to detect and counteract disinformation. This approach is crucial for spotting early signs of malevolent campaigns and keeping pace with the dynamic nature of AI technologies.

AI algorithms could enhance the efficiency and scope of cross-platform monitoring by quickly scanning and analysing vast amounts of data across different platforms, also known as automated content analysis. AI systems can also operate in real-time, providing instant alerts and updates on emerging trends, viral content, or sudden changes in discourse.[119]

These algorithms can identify patterns, keywords, trends, and sentiments in the content, providing a comprehensive overview that would be impractical for humans to achieve due to the sheer volume of data – enabling organizations to detect and analyse disinformation campaigns across different social media platforms and digital forums more effectively. Yet, Albert Trithart, expresses concerns regarding the suitability of the tools available: "Currently, a lot of the tools for monitoring mis- and disinformation come from the private sector, which can be problematic. They often require adaptation to fit very different mandates."[120] These private sector organizations are often Western-based digital marketing firms with their own interests. Although not inherently bad, they often do not develop AI tools specifically for peacebuilding or even with it in mind. This impacts the suitability of these tools for this field and problematizes their usage.

It is important to note that strategies using AI for peacebuilding and countering disinformation all focus on the digital realm and do not take into account online/offline overlaps. Muthoki Mumo, sub-Saharan Representative at the Committee to Protect Journalists observes: "There's a crossover between the online to offline media environments."[121] In places like Mali, with its rich oral culture, online disinformation frequently transitions to community discussions and local radio broadcasts. The majority of information environments require specialized AI tools developed for both peacebuilding and deep understanding of the local information environments.

113 Peter Cunliffe-Jones *et al.* (2021) *Misinformation Policy in Sub-Saharan Africa: From Laws and Regulations to Media Literacy* (London: University of Westminster, 2021). Accessible at: https://www.uwestminsterpress.co.uk/site/chapters/m/10.16997/book53.a/.
114 Chris O. Ogunmodede. Interview conducted via videoconferencing technology, September 2023.
115 The Royal Society and BBC, "Generative AI, Content Provenance and a Public Service Internet."
116 Ernest Dukuzumuremyi. Interview conducted via videoconferencing technology, November 2023.
117 The Royal Society and BBC, "Generative AI, Content Provenance and a Public Service Internet."
118 Matthew Gentzel, Program Officer at Longview Philanthropy. Interview conducted via videoconferencing technology, September 2023.
119 Eleonore Pauwels, "Artificial Intelligence and Data Capture Technologies in Violence and Conflict Prevention," Policy Brief (The Global Center on Cooperative Security, 2020). Accessible at: https://www.globalcenter.org/wp-content/uploads/GCCS_AIData_PB_H.pdf.
120 Albert Trithart. Interview conducted via videoconferencing technology, September 2023.
121 Muthoki Mumo, Sub-Saharan Africa Representative at the Committee to Protect Journalists. Interview conducted via videoconferencing technology, September 2023.

Large-Scale Digital Dialogues

In addition to changes in the way in which peacebuilders fight disinformation, a new area of opportunity has been deliberative AI. Deliberative AI is a category of AI tools, enhanced by generative AI, which allow for virtual discussions and the exchange of ideas. Recent advances in AI have allowed developers to include unique features to deliberative tools, such as determining areas of convergence and divergence, grouping contributions by theme, and allowing for a scaling up of one-to-one dialogues.

The UN Department of Political and Peacebuilding Affairs (DPPA) has been at the forefront of implementing large-scale digital dialogues in support of peace efforts. Reflecting on the history of peace polls, such as those conducted in Northern Ireland, the DPPA has identified operational challenges and opportunities in digital dialogues. The experience in Northern Ireland, where peace polls played a pivotal role in the peace process by including public opinion in negotiations, serves as a foundational model for these initiatives.

The role of AI in these dialogues is to facilitate large-scale, real-time interactions among diverse groups. For instance, the Remesh AI platform, used in collaboration with UN DPPA, allows up to 1000 participants to engage anonymously in digital dialogues. This platform not only accommodates multiple-choice polling but also invites open-ended responses, offering a comprehensive view of public opinion on various peace-related issues. The AI algorithms process these responses to identify themes and patterns, helping to shape a nuanced understanding of the conflict dynamics.

In Libya, the UN Support Mission leveraged these digital dialogues to gather insights on key issues like the civil war, foreign intervention, economic challenges, and human rights concerns. These dialogues provided a platform for Libyans to voice their opinions, contributing to a more inclusive peace process.

The success of these digital dialogues in Libya, marked by broad participation and real-time engagement, demonstrates the potential of AI in peacebuilding. These dialogues offer a template for future peace efforts, where technology can facilitate more inclusive and comprehensive discussions on critical issues. However, the DPPA also acknowledges the limitations of digital dialogues, such as cybersecurity risks, the need for internet access, and cultural barriers to technology adoption.[122]

122 Daanish Masood Alavi, Martin Wählisch, Colin Irwin, and Andrew Konya, "Using Artificial Intelligence for Peacebuilding," *Journal of Peacebuilding & Development* Vol 17 Issue 2 (2022): 239–243.

Box: How to Fact-Check Against Disinformation Using Journalistic Methods

The following process illustrates how grassroots and civil society organizations of journalists approach AI-powered disinformation.

Many of these steps draw from an interview with Joy Muthanje Mwaniki, Partnerships Account Manager with Shujaaz Inc, a network of social ventures based in Nairobi. Joy oversees programmes for Shujaaz, including a media for social change venture that uses the superpower of story to inspire, motivate, and mobilize millions of young people in Kenya and Tanzania to navigate their "digital and real-world community."[123]

Verifying the source:
- Checking where the information originally came from and whether it is credible.
- For first hand sources, ensuring their credibility and reliability. Are they a recognized expert in the field? Do they have firsthand knowledge of the event or data?
- For secondary sources, checking the reliability of the publication or outlet.
- Mwaniki offers a tip for source verification: "If you can clearly see an account has just been created, with very inflammatory messaging, that's what we would label as disinformation."

Cross-referencing with multiple sources:
- Always seeking multiple sources to confirm a piece of information.
- Being wary of echo chambers where one incorrect piece of information is repeated by multiple outlets.

Checking primary sources:
- Whenever possible, go to the primary source of information. This could be a research paper, an official document, a direct interview, or raw data.
- Ensuring that secondary reporting hasn't taken the primary source out of context or misrepresented it.
- Mwaniki advises caution in disseminating unverified information: "If you are unsure about the source of a post or information, then don't forward it to people within your group."

Evaluating the data:
- If the claim involves statistics or data, ensure the data is taken from a credible source, such as an established research institution or government database.
- Understand how the data was collected and what it represents.

Checking dates and timelines:
- Ensure that the information is current and relevant to the context in which it's being presented.
- Old data or quotes might not be applicable to current events.

Consulting experts:
- When dealing with specialized or technical topics, consult with experts in the field to verify claims and ensure accuracy.
- Mwaniki explains: "We worked with PesaCheck … If you see any mis- and disinformation or like a post that looks sort of odd, you can always send it to PesaCheck via WhatsApp, to determine the authenticity of the actual post."

Considering the motivation:
- Understand the potential biases or motivations behind a claim. Is there a political, financial, or personal gain involved?
- Be objective and avoid letting biases influence the fact-checking process.
- According to Mwaniki: "If a post seems designed to create a negative reaction from you, if it seems to be overly emotive, sparks fear, makes you question your security, [and/or] makes you angry. Then that is likely to be mis- and disinformation."

123 Joy Muthanje Mwaniki, Partnerships Account Manager at Shujaaz Inc (https://www.shujaazinc.com/). Interview conducted via videoconferencing technology, September 2023.

Case Study 3: The Growing Challenge of Fake News in Côte d'Ivoire

Côte d'Ivoire, following a nine-year conflict which ended in 2011, continues to face challenges in its peacebuilding process. One such challenge is the circulation of 'fake news,' a catch-all phrase that encapsulates both disinformation and misinformation, via word-of-mouth and social media platforms, such as Facebook and WhatsApp. This fake news contributes and catalyses political violence, which progressively deteriorates the trust between Ivorians and the Government of Côte d'Ivoire.[124]

Political actors have historically used political propaganda to bolster their power and to escalate ethnic tensions. Propaganda related to rhetoric about what is true Ivorian heritage was a contributing factor to the civil war.[125] Despite the role of disinformation in political violence, these actors continue to use it.

Disinformation campaigns increase during election cycles. In the 2020 presidential and 2021 legislative elections, political actors ran campaigns to sow seeds of fear and doubt amongst Ivorians, with rumours circulating about the Ivorian Government using youth gangs to target opposition supporters. The rumours resulted in violence between supporters of different political parties.[126]

A poignant example of the influence and danger of disinformation is the situation in M'battao. Following the 2020 presidential election, protests erupted in a predominantly Malinke area, the ethnicity of President Ouattara. Clashes resulted from these protests, and there were six deaths and approximately 40 people were injured. On Twitter and other social media platforms, rumours and other forms of fake news circulated. These included fake claims of ethnic killings, exacerbating existing tensions further.[127]

Although government officials directly spread disinformation, other actors do so as well. For example, anonymous social media accounts, known as 'avatars,' are trusted sources of information in the Côte d'Ivoire political sphere. In a report by the Centre for Democracy and Development, entitled *Côte d'Ivoire's Fake News Ecosystem: An Overview*, an interviewee shared the degree of influence of the well-known and very influential Chris Yapi, explaining that "people watch this guy on Twitter more than they watch the news or the government – even in villages people are always looking to see what Chris Yapi said." This avatar is associated with an opposition leader, Guillaume Soro, and functions as a political actor in the information sphere.[128]

A concrete example of this political actor's impact is when the former Prime Minister Hamo Bakayoko died. Chris Yapi spread a rumour that Téné Birahima, the President's brother, poisoned him in a plot to position Birahima as his brother's successor as President of Côte d'Ivoire. Therefore, when Birahima became defence minister soon thereafter, his reputation made it difficult for him to do his job because people believed the rumours spread by Yapi, even military personnel. Yapi contributed to the Government's instability and delegitimized it through their spread of disinformation, challenging the peacebuilding process and strengthening of institutional infrastructures.[129]

124 Jessica Moody, *Côte D'Ivoire's Fake News Ecosystem: Overview* (Centre for Democracy and Development, 2021).
125 Reuters, "Propaganda War Rages as Violence Escalates in Abidjan," *France 24*, 13 March 2011, https://www.france24.com/en/20110313-propaganda-war-rages-violence-escalates-abidjan-civil-war-press-newspapers.
126 Jessica Moody, *Côte D'Ivoire's Fake News Ecosystem: Overview*.
127 Jessica Moody, "The Genocide that Never Was and the Rise of Fake News in Côte d'Ivoire," *African Arguments*, 21 January 2022, https://africanarguments.org/2022/01/the-genocide-that-never-was-and-the-rise-of-fake-news-in-cote-divoire/.
128 Jessica Moody, *Côte D'Ivoire's Fake News Ecosystem: Overview*, p. 12.
129 Ibid., p. 16.

The Government's poor communication strategy creates a vacuum that avatars in Cote d'Ivoire can fill. At the start of the Covid-19 pandemic, the Government's communication strategy was so inadequate that journalist Israël Guébo created a radio show, WA FM, to combat the growing disinformation about the crisis and amplify the Government's message regarding health policies.[130] Despite civil society interventions, fake news continues to flourish.

To tackle fake news, the Government of Côte d'Ivoire has taken a legal approach through Article 173 of the penal code which introduces severe penalties associated with spreading misinformation. However, the Act is used to target opposition members and journalists, further politicizing the information sphere and exacerbating existing tensions.[131] Moreover, this intervention does not address the intervention of foreign actors, specifically the rise of pro-Russian propaganda, including a shadow-boxing campaign in April 2022, which suggested that religious extremist groups in the region had the support of the United States and France.[132]

Côte d'Ivoire has a layered, complicated information sphere that flows in and out of digital spaces and has increasing amounts of fake news. All of these factors negatively impact peacebuilding in the region. To address fake news, the Government and civil society organizations must consider its methods of dissemination, which include via social media platforms and 'word-of-mouth.' The Government must also go through a process of transparency and confidence-building, penalizing individuals from all political parties that spread disinformation. Ultimately, the political nature of fake news in Côte d'Ivoire requires the Government and other political parties to take responsibility for solving it.

130 Traoré, Kpénahi, "Une Webradio Contre les Fake News sur Whatsapp," International Center for Journalists, 14 May 2021, https://ijnet.org/fr/story/une-webradio-contre-les-fake-news-sur-whatsapp.

131 "Increased Harassment of Journalists in Côte d'Ivoire," Reporters Without Borders, 2016, https://rsf.org/en/increased-harassment-journalists-c%C3%B4te-d-ivoire.

132 Tessa Knight and Jean le Roux, *The Disinformation Landscape in West Africa and Beyond* (Washington, DC: The Atlantic Council, 2023), p. 8. Accessible at: https://www.atlanticcouncil.org/wp-content/uploads/2023/06/Report_Disinformation-in-West-Africa.pdf.

5. Conclusion

Throughout this report, the contributions of the interviewees and the literature review of existing work analysing the AI and peacebuilding space contextualized how AI and disinformation counter peacebuilding efforts and what methods exist to mitigate AI-powered conflict drivers. This conclusion focuses on the key stakeholders in sub-Saharan Africa's digital information landscape, offering guidance and recommendations for future initiatives.

Civil society organizations, as shown across the literature reviewed and interviews conducted, are taking the initiative and leading in the fight against disinformation. Their approach is more often journalistic and requires the traditional methods of manual fact-checking. Although AI can be useful in supporting this work, their discernment and understanding of facts within both the digital and broader context, makes their contribution invaluable in the fight against disinformation. It is important that civil society organizations continue these projects and receive financial and political support.

The understanding that civil society organizations have of local information spheres make them invaluable to the creation of digital literacy and media literacy educational programming. The development of an individual's critical thinking and fact-checking skills, in addition to content moderation and guardrails, provides an important support to AI tools and can help combat disinformation. The information sphere in Côte d'Ivoire exemplifies why approaches to disinformation require the support of both AI and human-led approaches. Word-of-mouth takes a leading role in the spread of disinformation, and interpersonal relationships are central to current fact-checking norms. In understanding these characteristics of the local information sphere, media and digital literacy programmes support the fact-checking done by each individual and in turn by the broader communities. It is this local approach to understanding information spheres that will make tools and policies most effective.

Governments must play a role in addressing disinformation. While some political actors, both internal and external, can drive the dissemination and creation of disinformation, a coordinated response benefits from public sector involvement. But if governments take the initiative to address disinformation, they must adopt a holistic and multifaceted strategy that is informed by the complex dynamics of their respective information spaces. This approach necessitates a deep understanding of various communication channels, including social media and traditional word-of-mouth networks, to tailor effective counter-disinformation strategies. Emphasizing transparency and accountability in governmental communications is critical, setting a standard of trust and reliability. A thoughtful communications strategy will decrease the chance of creating vacuums that disinformation could fill. Lastly, approaches to addressing disinformation should not further politicize the disinformation sphere. Governments, when establishing laws and regulations, should apply them consistently across all actors and not target opposition leaders, activists, journalists, and/or civil society leaders, while ensuring that they adhere to their human rights obligations, including freedom of speech.[133] As noted by Jonathan Rozen, Senior Researcher at the Committee to Protect Journalists: "Unintended consequences of regulation can be seen around the world, enabling authorities to control new information landscapes in ways that don't support freedom of the press."[134] These consequences can be seen in countries across sub-Saharan Africa, where regulations are often used against the media.[135] In sum, governments should commit to an active, inclusive, and transparent role in their information ecosystems.

Social media companies should develop and refine their AI tools for addressing disinformation. There exist multiple approaches towards addressing this, including common and automated content moderation, guardrails, shadow-banning, and censorship. Each of these requires a thoughtful and locally-specific approach, which necessitates companies' investment into expanding the resources behind languages in the sub-Saharan Africa region and deepening their understanding of the local information ecosystem.

Collaborative efforts between civil society organizations, governments, multilateral organizations, and private actors are crucial to amplify accurate information and mitigate the

133 United Nations General Assembly, *The Universal Declaration of Human Rights (UDHR)* (1948). Accessible at: https://www.un.org/en/about-us/universal-declaration-of-human-rights.

134 Jonathan Rozen, Senior Researcher at the Committee to Protect Journalists. Interview conducted via videoconferencing technology, September 2023.

135 Peter Cunliffe-Jones et al., *Misinformation Policy in Sub-Saharan Africa: From Laws and Regulations to Media Literacy*.

influence of unauthorized sources. Furthermore, an awareness of and resilience against foreign propaganda and external influences that can exacerbate misinformation are essential components of this strategy. These actors, especially civil society organizations and governments, must work together to create comprehensive, transparent legal frameworks. Moreover, civil society organizations should be integrated from the research, development, and implementation stages of any policy approach or AI tool for peacebuilding; their contextual knowledge is critical to success and harm mitigation. All actors must work together to create a transparent and ethical environment that prioritizes each individual's digital rights.

All actors have the potential to support peacebuilding efforts, whether it be through moderating content, strengthening digital and media literacy, or promoting innovative new approaches that integrate AI and peacebuilding, such as large-scale digital dialogues. It has become increasingly clear that AI plays a role in exacerbating or driving conflict, as shown through the UN@75 report and the research by Interpeace. However, peacebuilding organizations more broadly have to integrate AI into their conflict analyses and factor the role of disinformation into their frameworks. They should also be informed actors, understanding local information and media ecosystems.

The relationships between the different actors are sensitive and increasingly complex. There needs to be an international governing body that oversees and provides guidance on combating disinformation. Moreover, the body should work towards establishing global standards for combating disinformation while respecting cultural and regional differences. As a regulatory, international framework comes to fruition, it is essential that all actors work towards creating a coordinated, effective response towards disinformation and misinformation, recognizing that the consequences of maintaining the status quo will be increasingly detrimental to the stability of the region and beyond.

Annex A: Bios and Interview Summaries

1) Albert Trithart - Editor and Research Fellow at International Peace Institute

Albert Trithart is an Editor and Research Fellow at the International Peace Institute, with ten years of experience working on governance and peacebuilding. He is an expert in comparative politics, elections, and conflict resolution, with extensive experience in sub-Saharan Africa.

Albert discussed disinformation against the UN and its peacekeepers in Mali, CAR, and the DRC, including allegations that they are complicit with armed groups or exploiting natural resources. He considered the challenges of defining disinformation and separating it from other forms of harmful information, such as hate speech, and the ways in which disinformation spreads from online platforms (on Facebook and WhatsApp for example) to offline. He explained that although the actors involved in disinformation are often difficult to identify, in some cases they can be traced back to Russia or the diaspora. He explained a number of solutions for tackling disinformation at the grassroots/civil society, national, and multilateral/international levels, which includes providing training and funding to local media outlets and fact-checking organizations.

2) Alpha Daffae Senkpeni - Executive Director/Editor at Local Voices Liberia Media Network

Alpha Daffae Senkpeni is the Executive Director/Editor at Local Voices Liberia Media Network (LVL). He provides guidance to the network of journalists, based in the country's 15 counties, and helps them gain a wider audience by liaising with other news outlets. LVL reports on issues that are underreported in the mainstream media, and seeks to give a voice to local/rural communities.

Alpha explained that disinformation is a major problem in Liberia, both in the mainstream media and on social media, and often used by politicians to gain political support. He discussed the role of disinformation in contributing to conflict by creating tension and mistrust, and the role of foreign actors in propagating it. He believes, however, that most of the disinformation in Liberia is created by Liberians themselves. He explained how LVL fact-checks information on social media using online tools and human judgement. He believes that the best way to address disinformation is to support fact-checking organizations and to educate the public about how to identify and avoid disinformation.

3) Alphonce Shiundu - Kenya Editor at Africa Check

Alphonce Shiundu is the Kenya Editor of Africa Check, Africa's leading independent fact-checking organization. He oversees the Kenya Office's day-to-day operations and is its public representative. He introduced fact-checking research to Kenya's mainstream print and electronic media, including the BBC. Since 2012, Africa Check has fact-checked thousands of claims on topics ranging from crime and race in South Africa to population numbers in Nigeria and fake health cures in other African countries.

Alphonce explained that disinformation around elections is a key issue, but the actors involved can be difficult to identify, and are likely to be those who benefit from the disinformation, including Russia, militia groups, and political opponents. He discussed the challenge of identifying disinformation, arguing that deliberate disinformation is often well-crafted and believable, making it difficult to distinguish from regular information or bad quality information. He believes that currently AI-generated disinformation is often shallow and easy to identify, but has the potential to be used to generate and disseminate more credible and sophisticated disinformation in the future, which could have a devastating impact on peace and conflict. He argued that media literacy is essential to combat disinformation, but that even the most sophisticated tools may not be able to keep up with the pace of technological innovation.

4) Beatrice Bianchi - Political Analyst Sahel Expert, Med-Or Leonardo Foundation

Beatrice Bianchi is a visiting fellow on the Sahel and West Africa for the Med-Or Leonardo Foundation. She has over ten-years of research and professional experience working on political outreach, mediation, and assessment of fragility factors throughout Africa. She has extensive experience in the Sahel region and particularly in Niger, and from 2018 to 2022 was responsible for the Sahel regional program at the Tony Blair Institute for Global Change (TBI).

Beatrice explains that disinformation on social media in the Sahel and West Africa mainly focuses on political and conflict-related issues, which has become more serious in the last two years, coinciding with Russia's increased interest in the region. She provides examples of disinformation that promoted conflict and caused international concern. She discusses the use of software to

create disinformation, referring to three main categories: reusing old images and videos; manipulating recent images and videos; and creating new images with AI. She explains that the level of sophistication of disinformation in Africa is generally low, but that people are still susceptible to it because they often do not apply a critical lens and do not know what is possible with AI. She believes that national governments are not doing enough to stop the spread of disinformation. She argues that they need to start using the same methods of communication as the spreaders of fake news to effectively counter it, such as social media-ready short, easy-to-understand, messages. She also believes that international organizations should support governments in these efforts.

5) Chris O. Ogunmodede - Foreign Affairs Analyst

Chris Ogunmodede is a Foreign Affairs Analyst specializing in African governance, political economy, defence and security, trade, and regional integration. He has over ten years of professional experience working across three continents with governments, civil society organizations, multilateral organizations, think tanks, and the private sector.

Chris shared insights on how disinformation is often weaponized during high-stress moments such as elections, protests, and the passing of contentious legislation. He highlighted the role of platforms like WhatsApp in spreading disinformation. He discussed the role of disinformation in conflict situations in Côte d'Ivoire, Nigeria, and Ethiopia, and highlighted the use of AI technologies like deep fakes and the involvement of troll farms, which are often funded by diaspora groups or foreign governments. He also explained the challenges governments and international organizations face in regulating this issue, and potential solutions. He discussed the influence of political campaigns and their use of social media platforms to disseminate information, and how the rise of smartphone penetration, especially among young people, has made disinformation spread more effectively. He also touched on the issue of State-sponsored disinformation, particularly during election periods.

6) Jamie Hitchen - Independent Research Analyst and Honorary Research Fellow at the University of Birmingham, UK

Jamie Hitchen is an independent research analyst and Honorary Research Fellow at the University of Birmingham, UK. His recent work has focused extensively on the sociopolitical impact of social media in Africa and the risk posed by disinformation. He has conducted studies covering Nigeria, Sierra Leone, Uganda, and the Gambia for the University of Edinburgh, the Centre for Democracy and Development (Abuja), and the National Democratic Institute. Jamie has also co-authored articles for academic journals such as *Party Politics and the Journal of Democracy* on the political use of social media in Nigeria and contributed a chapter on Sierra Leone's 2018 election to an edited volume, *Social Media and Politics in Africa* (Zed Books, 2019). His co-edited volume, *WhatsApp and Everyday Life in West Africa: Beyond Fake News*, was published by Bloomsbury in 2022.

Jamie discussed the spread of disinformation in Africa and its impact on politics, elections, and public health. He explained the role of AI in spreading disinformation and the need for fact-checking and local community involvement in combating it. He explained the impact of disinformation and misinformation on elections and political violence, and emphasized the need for tech companies to take a more proactive role in content moderation and for States to demand these companies uphold their user agreements. He also highlighted the importance of civic education efforts and improving digital literacy to combat the spread of false information.

7) Jonathan Rozen - Senior Researcher at the Committee to Protect Journalists

Jonathan Rozen is an internationally experienced journalist and researcher. As Senior Researcher with the Committee to Protect Journalists (CPJ), he reports, conducts advocacy, and coordinates emergency responses for journalists across sub-Saharan Africa. Since joining CPJ in February 2017, Jonathan has led numerous reporting and advocacy trips, including to Ghana, Côte d'Ivoire, Senegal, Liberia, and Nigeria, where he covered national elections in 2019 and 2023. His investigations have tracked efforts to control information during conflicts, and the criminalization and censorship of expression online. He also managed a CPJ project that mapped the use of commercial spyware to target journalists and those close to them around the world.

Jonathan explained that there is a growing trend of using disinformation as a cover to target journalists and suppress critical reporting, including the use of cybercrime laws to prosecute journalists, and the surveillance of their communications. He explained that there are major impediments to journalists' ability to report freely and contribute positively to information landscapes, including the threat of violence and arrest. He calls on governments to decriminalize journalism and protect reporters from

surveillance and harassment. He also discussed the potential for new technologies, such as generative AI, to be used to create even more sophisticated and believable disinformation campaigns, and argued that these trends pose a serious threat to press freedom and democracy. However, he argued that regulation should be carefully considered and evaluated to ensure that it does not inadvertently harm freedom of the press, and urged the public to be critical of the information they consume and to support independent journalism.

8) Joy Muthanje Mwaniki - Partnerships Account Manager at Shujaaz Inc

Joy Muthanje Mwaniki is the Partnerships Account Manager at Shujaaz Inc where she manages grants, strategy, budgets, and relations with partners and donors, including UNDP, the German Embassy to Kenya, and the Hewlett Foundation to inform, inspire, and activate a generation to participate in and lead Kenya's democracy. She develops campaign strategies, manages project implementation, and ensures that campaigns are continuously measured for evidence impact. Shujaaz Inc is a social enterprise based in Kenya that works to break down barriers so young people can take control of their future.

Joy discussed the issue of disinformation and misinformation in Kenya, particularly in relation to elections and the Covid-19 pandemic. She identified the main platforms for spreading misinformation as Facebook, Twitter, and WhatsApp. She discussed solutions from a grassroots perspective, including training and fact-checking initiatives, and suggested that national governments need to understand AI tools and create a regulatory framework to address them. Additionally, multilateral and international organizations could help by providing expertise and support to fact-checking organizations in Africa. She explained the role that Shujaaz plays, focusing on building the agency, resilience, understanding, and knowledge of young people, and enhancing their skills to identify and respond to misinformation/disinformation.

9) Kwami Ahiabenu II - Director at Penplusbytes

Kwami Ahiabenu II is the Director of Penplusbytes, an organization that focuses on leveraging technology and knowledge to enhance good governance, and empower the media, civil society, and other stakeholders with cutting-edge digital tools and innovations. He has over 20 years of experience in business development, management, marketing, new media, and ICT.

Kwami discussed a number of key themes in disinformation on social media in West Africa: religious extremism, elections, the decline of democracy, the decline of French influence, and the rise of Chinese and Russian influence and their role in spreading disinformation, noting that China is more subtle in their approach than Russia. He identified local political actors, foreign actors, and social media influencers as involved in creating disinformation. He explained that the messages are spread through platforms like Facebook, WhatsApp, and Telegram, then picked up and disseminated further by traditional media outlets, and finally, amplified by word of mouth. He identified two main areas where AI is being used: in content creation for deep fakes, edited images, and text; and in content distribution using, for instance, social media bots and targeted advertising. He also discussed the use of AI in combating disinformation, and noted that AI-powered disinformation campaigns are becoming increasingly sophisticated and difficult to detect. He pointed out that social media companies are not doing enough to address the problem of disinformation on their platforms. He suggested that to reduce the effects of disinformation it was necessary to increase fact-checking, hold social media companies accountable, support research, educate the public, and develop better regulations.

10) Lilian Olivia - Advocate of the High Court of Kenya and Founder of Safe Online Women Kenya

Lilian Olivia is an Advocate of the High Court of Kenya and Founder of Safe Online Women Kenya (SOW-Kenya), providing digital literacy training programmes to young women and girls in Kenyan high schools and universities, empowering them with the skills needed to navigate online spaces confidently. SOW-Kenya raises awareness about online safety, cyberbullying, and gender-based violence, and conducts research and gathers data on digital threats and challenges faced by women and girls in Kenya.

Lilian discussed disinformation in sub-Saharan Africa related to: elections and the attempt to sway public opinion and votes; exploitation of religious, cultural, ethnic, and tribal tensions to incite violence; and Covid-19 – that led to vaccine hesitancy and other health problems. She identified the actors involved as politicians, citizens paid to spread disinformation, and foreign groups from Russia, the UK, and China. She explained that disinformation was spread through the following mediums: social media platforms such as Twitter, Facebook, and WhatsApp; websites and blogs that are designed to appear credible; and Tik Tok through the use of deep-fake and cheap-fake videos. She

discussed the use of AI in spreading disinformation in two ways. First, through Chat GPT and the like – spreading multiple versions of the same messages quickly, and second, through the limits of natural language processing techniques that cannot identify keywords in vernacular languages that are often used to incite violence and hatred. She also believes that AI could be used to combat disinformation in the future through deep-fake detection, and enhanced natural language processing that can detect hate speech in local and vernacular languages.

11) Matthew Gentzel - Program Officer at Longview Philanthropy

Matthew Gentzel is a Program Officer at Longview Philanthropy. He was previously a researcher at OpenAI, researching mitigations against AI-enabled influence operations and conducting case studies on government and civil society reactions to evolving military technologies.

Matthew discussed the potential threats and capabilities of AI-enabled influence operations. Using his technical background he provided insights into the various ways AI can be used for influence operations, including automated mapping, fake personas, reputation assassination, and more. He also touched on the potential for AI to be used in spearfishing and social engineering attacks. He explained the potential risks of deploying AI systems, particularly in the context of influence operations and misinformation, and noted the way AI systems can be used against each other, by, for example, polluting the open source data they rely on to make decisions. He highlighted the importance of understanding the potential for AI to be manipulated and the need for robust decision-making processes to mitigate these risks. He also noted the potential for AI-enabled censorship and the need for careful regulation around that risk.

12) Muthoki Mumo - Sub-Saharan Africa Representative at the Committee to Protect Journalists

Muthoki Mumo is Sub-Saharan Africa Representative at the Committee to Protect Journalists, based in Kenya. She has a master's in journalism and globalisation from the University of Hamburg. She previously worked as a journalist with the Nation Media Group, covering a variety of beats from East African Community integration and regional trade to technology and telecommunications.

Muthoki discussed disinformation on social media in the sphere of politics, elections, and protests, and also the sphere of conflict, where disinformation is used as an attempt to shape the narratives of war. She argued that disinformation is often coordinated by groups of people, sometimes with the support of politicians, or by politicians themselves, and notes that it can be used to push both false and true narratives, depending on the goals of the actors involved. She observed that social media platforms such as Facebook and Twitter, and messaging apps like WhatsApp, can be particularly effective at spreading disinformation because they allow people to share information quickly and easily with their trusted networks. She believes that AI could be used to create more sophisticated and believable disinformation campaigns in the future, and suggested that social media platforms should do more to remove harmful content, that governments should hold politicians accountable for spreading disinformation, and citizens should be more critical of the information they consume.

13) Naomi Miyashita - Project Manager, Addressing Mis/Disinformation, United Nations Department of Peace Operations

Naomi Miyashita is the Project Manager in charge of addressing misinformation/disinformation at the UN Department of Peace Operations. In 2022 she joined the University of Washington's Center for an Informed Public, as the multidisciplinary research center's first community fellow, to research social media-enabled influence operations in the CAR and Mali. She has served in a variety of roles at UN headquarters and in various conflict-affected countries, covering child protection in armed conflict, political affairs, demobilization and reintegration of former fighters, and peacekeeping policy development.

Naomi identified several key issues related to disinformation in sub-Saharan Africa: the role of political influencers, often in the diaspora; a shrinking civic space and limited access to information; the targeting of minority groups and ethnic groups; and the instrumentalization of hate and genocide narratives to attack opponents. She also noted that disinformation has contributed to increasing unrest, violence, and a loss of trust in peacekeeping missions, which in Mali had led to their withdrawal and further destabilization. She discussed the various mechanisms by which disinformation is created and spread, which include: social media platforms influencing offline channels, such as radio and word-of-mouth (for example local leaders may be getting talking points and certain narratives to promote); new dubious civil society organizations that are created to mobilize protest against the UN; AI-generated images and videos that can create the appearance of global support for certain governments or authorities; and fake accounts

created to automate the spread of certain messages. She explained some of the solutions being discussed to fight conflict-related disinformation at a grass-roots, national, and multilateral level. These include media literacy training and awareness-raising for journalists and bloggers, fact-checking associations and media literacy campaigns, and the protection of threatened journalists. She also emphasized the need to support civic spaces and responsible journalism, and the importance of raising awareness of the issue and its impact.

14) Sammy Mupfuni - Managing Director at CongoCheck

Sammy Mupfuni is the Managing Director at CongoCheck, a fact-checking organization that is part of the Africa Facts network. He has worked as a journalist in the DRC for over nine years, covering politics, security and armed conflict, humanitarian affairs, agriculture, and the environment. CongoCheck was launched in 2018 to fight against the disinformation and fake news that have become commonplace on social networks.

Sammy discussed the effects of disinformation on peace and conflict in the DRC. He argued that disinformation can exacerbate existing conflicts and lead to violence, even death. He provided examples of disinformation campaigns in the DRC, such as false news about the Ebola outbreak and the security situation in North Kivu. He concluded by identifying the key factors driving disinformation on social media in the DRC: a lack of media and information literacy among the population; the use of social media by influencers to spread propaganda; and the accidental sharing of misinformation by traditional media.

15) Ernest Dukuzumuremyi - Programme Manager of Interpeace's Rwanda Programme

Ernest Dukuzumuremyi is the Programme Manager of Interpeace's Rwanda Programme. Prior to joining Interpeace, he served as a Researcher and the Great Lakes Peacebuilding Programme Manager at Never Again Rwanda where he contributed to the Participatory Action Research processes and the establishment and facilitation of cross-border dialogue spaces for peace in Rwanda, Burundi, and the DRC.

Ernest explained that disinformation, exploiting stereotypes, and misinformation significantly impacts peace and conflict in Rwanda and its neighbouring countries, especially disinformation spread through social media platforms. Politicians misuse historical narratives during elections, damaging relationships and inciting violence. The spread involves various actors, including the media. Solutions, he argued, should focus on fact-based research, dialogue to combat stereotypes, and utilizing media for accurate information dissemination. Social media, the fastest propagator of disinformation, is countered through dialogue, publishing factual content, and facilitating authentic interactions. While AI-driven disinformation exists, strategies to counter it remain underexplored. Leveraging AI for accurate information dissemination is an important area to further explore for peacebuilding.

United Nations University
Centre for Policy Research

About UNU-CPR

United Nations University Centre for Policy Research (UNU-CPR) is a think tank within the United Nations that carries out policy-focused research on issues of strategic interest and importance to the UN and its Member States. The Centre prioritizes urgent policy needs requiring innovative, practical solutions oriented toward immediate implementation.

The Centre offers deep knowledge of the multilateral system and an extensive network of partners in and outside of the United Nations. The United Nations University Charter, formally adopted by the General Assembly in 1973, endows the Centre with academic independence, which ensures that its research is impartial and grounded in an objective assessment of policy and practice.

cpr.unu.edu

New York (Headquarters)
767 Third Avenue 35B
New York, NY 10017
United States
Tel: +1-646-905-5225
Email: comms-cpr@unu.edu

Geneva
Maison de la Paix
Chemin Eugène-Rigot 2E
Geneva, Switzerland
Tel: +1-917-225-0199
Email: comms-cpr@unu.edu