**United Nations University**
Centre for Policy Research

POLICYBRIEF
September, 2023

# Artificial Intelligence-Powered Disinformation and Conflict

## Exploring Recent Developments in Artificial Intelligence and their Impact on Disinformation-Fuelled Conflict

Eleonore Fournier-Tombs, Rebecca Brubaker, Eduardo Albrecht

### Recommended policy actions:

- Disinformation-related efforts should be connected to global, regional, and national initiatives governing Artificial Intelligence and digital spaces, especially those working to sustain global peace;
- Social media platforms should prioritize efforts to address disinformation, particularly disinformation that can lead to conflict or prevent peace, as explained in a recent policy brief by the United Nations Secretary-General, 'Information Integrity on Digital Platforms';
- Governmental organizations, civil society, and private sector companies should commit further funding to fact-checking initiatives, run by both journalists and social media platforms;
- National governments and other bodies in Sub-Saharan Africa should develop digital literacy programmes to help people identify disinformation more easily.

### Introduction

Disinformation, a type of false or misleading information that is inaccurate, intended to deceive, and shared in order to do serious harm,[1] appears in a multitude of different forms, including rumours or counternarratives shared by influencers, doctored images or videos of politicians, and photos with false captions. As online actors become more adept at manipulating social media platforms, the spread of posts has increased dramatically, as has their effect on real-life political events.

In the last two decades, disinformation on social media has fuelled political conflict in Sub-Saharan Africa.[2] Some phenomena, such as interference in elections, have been observed globally; whereas others, such as disinformation related to humanitarian interventions, may have elements of regional specificity, while also being pushed by international actors.

Until recently, the mechanism by which disinformation was spread on social media was largely through the manipulation of recommendation systems powered by Artificial Intelligence (AI) and ranking algorithms, which allowed false and dangerous content not only to be spread but also to spread *more rapidly* than other content.[3] In recent years,[4] an impres-

1 United Nations General Assembly, "Resolution A/RES/76/227," adopted by the General Assembly at the 76th Session, United Nations, 10 January, 2022, S/RES/76/227. Accessible at: https://digitallibrary.un.org/record/3955093.
2 Maggie Dwyer and Thomas Molony (eds), *Social Media and Politics in Africa: Democracy, Censorship, and Security* (London: Zed Books, 2019).
3 Kris Shaffer, *Data Versus Democracy: How Big Data Algorithms Shape Opinions and Alter the Course of History* (New York: Apress, 2019), pp. xii, 17, 44, 88–89.
4 See, for example, an early example of the risks of generative AI for global discourse: Joseph Bullock and Miguel Luengo-Oroz, "Automated Speech Generation from UN General Assembly Statements: Mapping Risks in AI Generated Texts," *arXiv preprint arXiv:1906.01946*.

sive rise in the accuracy and accessibility of large language models and other types of generative AI has led to even more potentially dangerous mechanisms for the spreading of disinformation.

In this policy brief, we discuss the potential impact of generative AI on disinformation in Sub-Saharan Africa and propose recommendations designed to inform policies on AI and security globally.

## Key Issues and Priorities

### Fake Violence Leading to Real Violence: Media Polarization to the Extreme

One of the most common ways of fostering conflict in Sub-Saharan Africa and other conflict-affected regions is to invent false violence, falsely attribute actual violence, or accuse actors of violent intent, inflaming pre-existing tensions. With approximately one quarter of the region's population on social media, false claims can spread extremely quickly, in part due to the transfer of disinformation online to analogue mediums, such as the radio or even word-of-mouth.[5] This allows disinformation to reach those who do not have Internet access. Although false flags have always been a tactic in conflict, online disinformation in Sub-Saharan Africa may be exacerbated by existing tensions. A particular trend has been re-captioning images taken in different countries and different contexts, and misleadingly attributing them to a false conflict.[6] In the Democratic Republic of Congo (DRC), for instance, tensions with Rwanda have heightened due to social media users re-captioning violent images and videos from other countries, such as a church massacre in Nigeria, and using this as false proof that Rwandans were killing Congolese, and vice-versa.[7] The website CongoCheck has been painstakingly fact-checking this and other conflict-related disinformation, and calling for more digital literacy training for citizens in both countries.[8]

In a similar vein, there have been reports of videos and images taken of women and children coming into the north of Côte d'Ivoire from Burkina Faso, with captions accusing men of staying behind to join extremist groups.[9] This disinformation was fanned by fear of extremism in Côte d'Ivoire, which led to the content spreading through a variety of traditional and non-traditional media.

### From Foreign Intervention to the Discrediting of Traditional Media: How Governments Contribute to Disinformation

Globally, there have been many cases where governments appear to have spread false information for political purposes. The Africa Center for Strategic Studies documented 16 cases of Russian-sponsored disinformation in Africa alone, including in Kenya in 2021, where a network of 3,700 accounts spread 23,000 tweets on various issues, including the distortion of public opinion about the release of the Pandora Papers, and discrediting journalists and activists. There was also a campaign in both DRC and Côte d'Ivoire in 2018 to fan anti-French sentiment and promote Russian interests, all with the objective of political destabilization.[10]

In a sampling of countries from the region, there have also been reports of national governments discrediting traditional media, sometimes eroding trust in journalists in favour of social media influencers, which has pushed people to more readily accept news from less reputable web platforms or even AI-powered bots.

### False Claims About Peacekeeping and Humanitarian Interventions: Campaigns Against International Organizations

The UN Secretary-General's June 2023 policy brief, 'Information Integrity on Digital Platforms,' notes that 75 per cent of surveyed UN Peacekeepers reported that misinformation or disinformation had impacted their safety and security.[11] This finding echoes research in the region on humanitarian and peacekeeping operations that have been compromised by online disinformation. Bintou Keita, Head of the United Nations Organization Stabilization Mission in the Democratic Republic of the Congo (MONUSCO),[12] for instance, has argued that this is one of the biggest challenges facing the peacekeeping operation, and may have contributed to violence, including the killing of a military peacekeeper and two UN police personnel in July 2022.[13]

### Prompt Engineering for Disinformation: Easier Than Ever

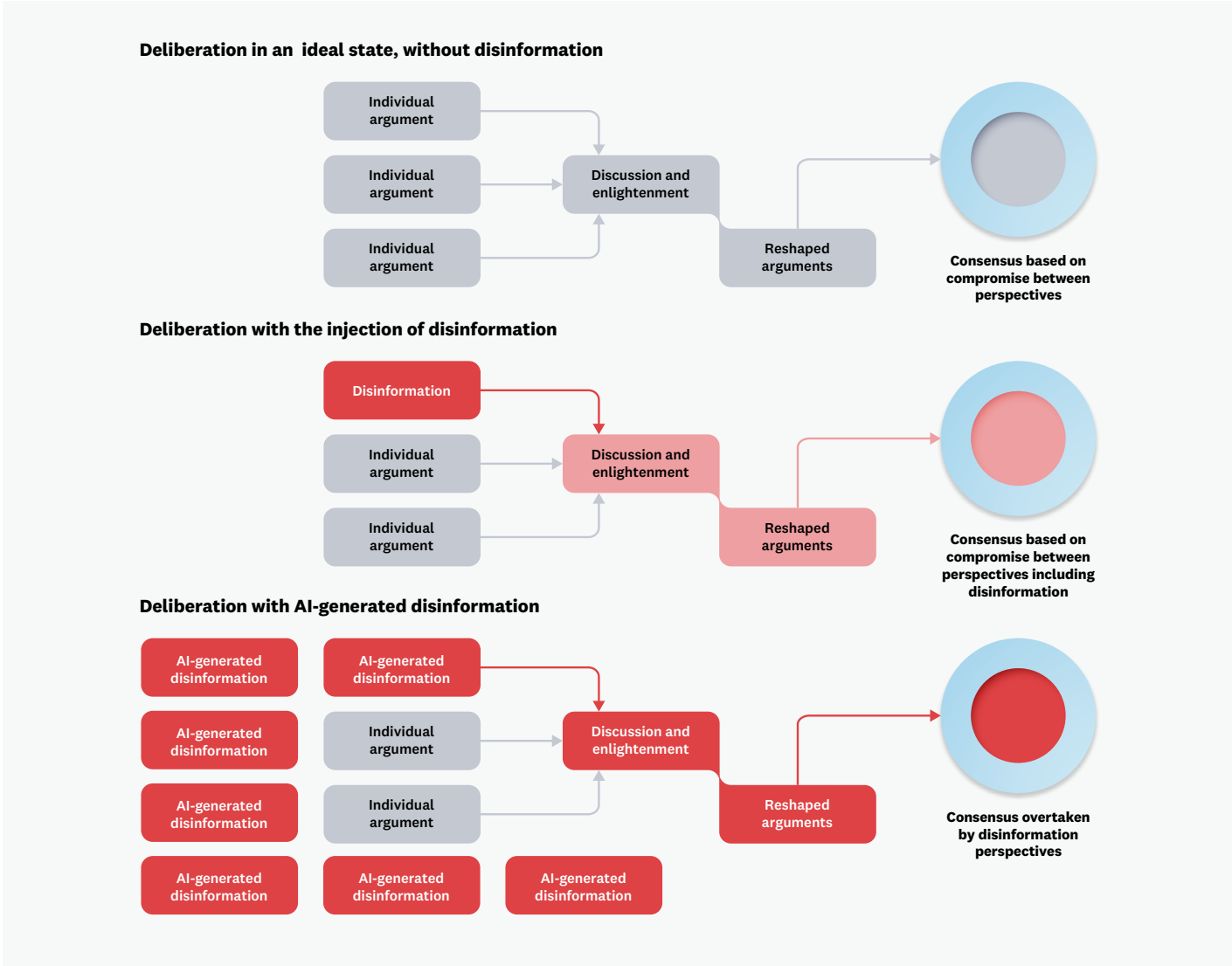
The public launch, in December 2022, of OpenAI's ChatGPT signalled a new era in disinformation. This tool enables a user with no programming skills to create a thousand, or even a million, variations of a disinformation message. This content can then be shared with online bots that spread the messages online. While generative AI companies are attempting to regulate

5 Workshop 1, held virtually on 11 July, 2023.
6 Workshop 1, held virtually on 11 July, 2023.
7 "#RwandaIsKilling: la désinformation attise les tensions entre la RDC et le Rwanda," France 24, last accessed on 25 August 2023, https://www.youtube.com/watch?v=ST9iYFpLJos.
8 Ibid. Also see the CongoCheck website, accessible here: https://congocheck.net/.
9 Workshop 1, held virtually on 11 July, 2023.
10 "Mapping Disinformation in Africa," Africa Center for Strategic Studies, 26 April 2022, https://africacenter.org/spotlight/mapping-disinformation-in-africa/.
11 United Nations, *Our Common Agenda Policy Brief 8: Information Integrity on Digital Platforms* (New York: United Nations, 2023). Accessible at: https://www.un.org/sexualviolenceinconflict/wp-content/uploads/2023/06/our-common-agenda-policy-brief-information-integrity-en.pdf.
12 "En RDC, MONUSCO se renforce face à la désinformation," ONU Info, 31 March 2023, https://news.un.org/fr/audio/2023/03/1133812.
13 "Daily Press Briefing by the Office of the Spokesperson for the Secretary-General," United Nations, 26 July 2022, https://press.un.org/en/2022/db220726.doc.htm.

this technology,[14] limiting violent language, for example, and 'tagging' AI generated posts, these efforts are complicated by several factors. First, human beings are quite inventive when it comes to censorship, and often resort to using different words that convey the same meaning and help evade censors.[15] Second, social media and generative AI platforms have not yet put enough effort into first anticipating and then seeking to mitigate conflict risks associated with their tools.

The diagram below illustrates how deliberation,[16] understood in this case as political debates on social media, can be poisoned by disinformation, in particular AI-generated disinformation. As this type of disinformation floods online platforms, it becomes more and more difficult to discern other perspectives. If the disinformation is related to a specific group, for example inciting hate towards its members, it can lead to a poisoned consensus in which members of the conversation begin to believe the hate speech. Recent developments in AI have also increased the capacity of disinformation to be created in a 'tone' that mimics respected public figures.[17] The more convincing and pervasive the disinformation, the more likely it is that it will change the minds of participants in the conversation.

Figure 1: Disinformation Before and After Generative AI[18]




14 See, for example, "Product Safety Standards," Open AI, last accessed on 23 August 2023, https://openai.com/safety-standards.

15 Heng Ji and Kevin Knight, "Creative Language Encoding under Censorship," *Proceedings of the First Workshop on Natural Language Processing for Internet Freedom* (Santa Fe, New Mexico: Association for Computational Linguistics, 2018).

16 This model of deliberation is adapted from: Eleonore Fournier-Tombs, "A New Architecture of the Public Sphere," Thesis, University of Toronto, 2014. Accessible at: https://tspace.library.utoronto.ca/handle/1807/44018.

17 Tshilidzi Marwala, "Deep Learning in Politics," *Artificial Intelligence, Game Theory and Mechanism Design in Politics* ed. Tshilidzi Marwala (Singapore: Palgrave Macmillan, 2023).

18 Note that the ideal state of deliberation does not account for other ways in which AI tools can negatively affect deliberation, for example by increasing polarization.

Additionally, AI-generated disinformation allows for the constant evasion of guardrails by using creative terms, producing disinformation about new events, recaptioning legitimate photos and videos, and producing artificial but increasingly convincing photos and videos. While disinformation evading censorship can be temporarily muted if the new terms are not widely known, this can change quickly, sometimes in a matter of days. New guardrails therefore have to be adopted to catch the new types of disinformation, in a perpetual game of cat-and-mouse.

The diagram below illustrates how a flow of generated AI disinformation, originally flagged by guardrails, is then repeatedly re-written to evade detection. The effects of disinformation may be temporarily muted, or slowed down, as new ways of communicating disinformation are created and shared with social media users. However, after some time, social media users become familiar with the new terminology, and the disinformation campaign can proceed, until new means of detection are installed.

## Recommendations

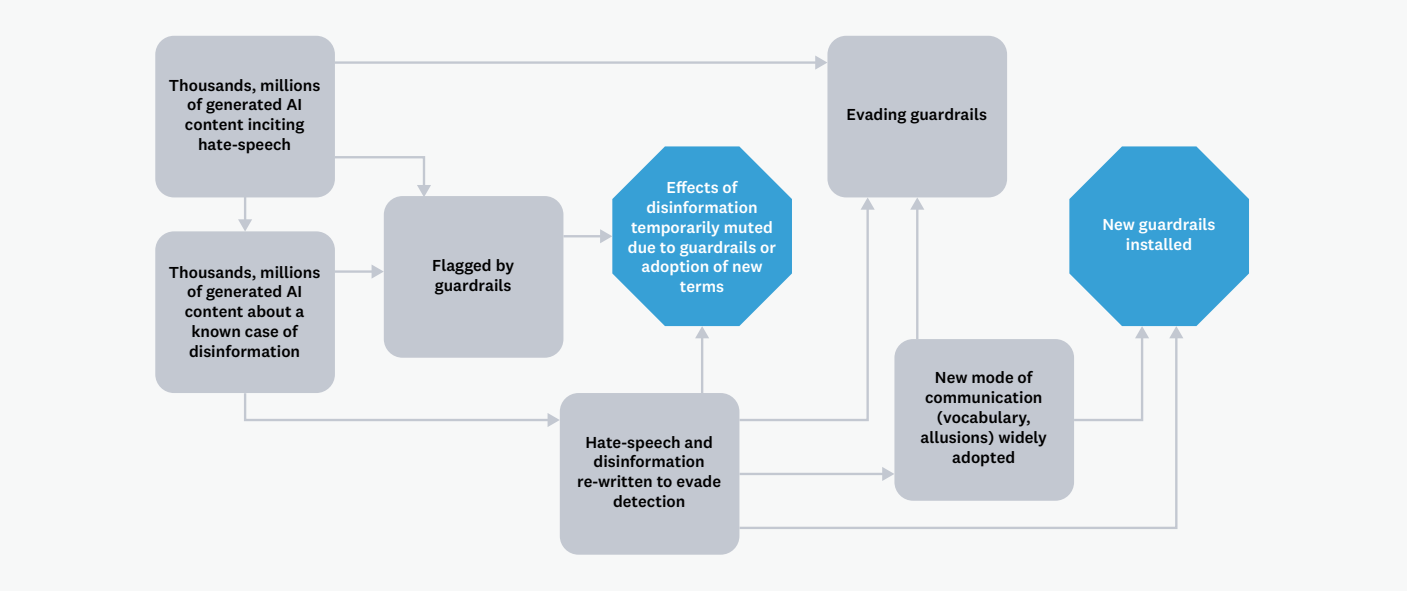
We have identified four recommendations that taken together provide a multi-pronged approach to tackling AI-powered disinformation and reducing its ability to promote conflict in Sub-Saharan Africa.

1. **Disinformation-related efforts should be connected to global, regional, and national initiatives governing AI and digital spaces, especially those working to sustain global peace.**

   There have been many calls recently demanding the global governance of AI, including by the G7,[19] the United Nations,[20] and several large AI companies.[21] In addition, the European Union and Canada have drafted bills to regulate AI; China and the United States both have AI Blueprints to guide policy efforts; and many other countries, such as Thailand, have adopted AI frameworks, including ethical considerations. The United Nations has made the establishment of an international body responsible for global norms and processes a priority, which will be further debated by Member States over the next year. **In this context, addressing AI-generated disinformation and the role it plays to fan conflict and obstruct peacebuilding, peacekeeping, and humanitarian endeavours, should be a priority for the multilateral system.**

2. **Social media platforms should prioritize efforts to contain AI-generated disinformation that can lead to conflict or prevent peace, as explained in the United Nations Secretary-General's June 2023 Policy Brief.**

Figure 2: The Flow of AI-Generated Disinformation




19 Hiroki Habuka, "The Path to Trustworthy AI: G7 Outcomes and Implications for Global AI Governance," Center for Strategic & International Studies, 6 June 2023, https://www.csis.org/analysis/path-trustworthy-ai-g7-outcomes-and-implications-global-ai-governance.

20 "Multistakeholder Advisory Body on Artificial Intelligence," United Nations Office of the United Nations Envoy on Technology, last accessed on 25 August 2023, https://www.un.org/techenvoy/content/artificial-intelligence.

21 Lewis Ho, Joslyn Barnhart, Robert Trager, Yoshua Bengio, Miles Brundage, Allison Carnegie, Rumman Chowdhury, Allan Dafoe, Gillian Hadfield, Margaret Levi, and Duncan Snidal, "International Institutions for Advanced AI," *arXiv preprint arXiv:2307.04699*.

Generally, there has been a certain reticence among social media platforms to meaningfully engage in content monitoring, a problem that is now compounded by generative AI. This has led to many failures to appropriately address conflict-fuelling disinformation, as evidenced, for example, by the case against Facebook by victims of the Rohingya genocide.[22] Poorly trained – and paid – workers, many of whom are vulnerable,[23] including in Kenya[24] and neighbouring countries, have not only been tasked with monitoring traumatic content shared on social media, but also enabling generative AI companies to avoid creating hateful or otherwise harmful text and images in the first place. An improved approach to content monitoring and preventing hate speech and disinformation could involve several components: 1) More investment in AI tools that counter disinformation; 2) Protection and training for workers responsible for monitoring content; 3) Increased investment in multilingual moderation, especially in conflict-prone or fragile settings; 4) More careful consideration of disinformation being injected in the training data for these tools, and 5) More transparent data sharing with researchers and peace and security actors.

3. **Governmental organizations, civil society, and private sector companies should commit further funding to fact-checking initiatives, run by both journalists and platforms, and be more transparent about their road-maps for fact checking.**

Fact-checking initiatives managed by journalists and community moderators, such as CongoCheck, should be supported, and their findings displayed on social media to warn users against harmful content and disinformation. Efforts are also needed to ensure the information they produce is provided in local languages, ensuring no community is left behind.

4. **National governments and other bodies in Sub-Saharan Africa should develop digital literacy programmes to help people identify disinformation more easily.**

While citizens wait for action from regulators and from AI companies, progress can still be achieved through digital literacy programmes. Many initiatives that use digital literacy to reduce online polarization have had promising effects, at least in localized, pilot form. There remain doubts as to whether an ability to identify disinformation reduces the propensity to share such content online. However, there has been consensus that digital and media literacy could help diffuse political tensions and reduce violent extremism. This is a strategy that several peacekeeping operations, such as MONUSCO, have recently adopted to educate people and remain proactive in the face of disinformation risks.

22 Amnesty International, "Myanmar: Facebook's Systems Promoted Violence against Rohingya; Meta Owes Reparations," 29 September 2022, https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/.

23 Phil Jones, "Refugees help power machine learning advances at Microsoft, Facebook, and Amazon," Rest of World, 22 September 2021, https://restofworld.org/2021/refugees-machine-learning-big-tech/.

24 Sarasvati NT, "Kenyan Workers Call for Probe into Disturbing Work Conditions in AI Content Moderation for OpenAI," Medianama, 13 July 2023, https://www.medianama.com/2023/07/223-kenyan-workers-call-for-investigation-into-exploitation-by-openai/.

**About this policy brief:** This policy brief is the first output of a research partnership between Interpeace and UNU-CPR, which explores the way AI-driven disinformation contributes to conflict. A first internal workshop discussing these issues with Interpeace country and regional teams in Sub-Saharan African was conducted on July 11, 2023. The findings from the workshop have informed this brief.

**Author bios:** Dr Eleonore Fournier-Tombs is the Head of Anticipatory Action and Innovation at UNU-CPR where she focuses on the development of methodological tools and policy recommendations related to AI and data at the United Nations. She is also an Adjunct Professor at the University of Ottawa Faculty of Law and a recurring lecturer on new technologies and cybersecurity at McGill University and Université de Montréal.

Dr Rebecca Brubaker is the Director of Policy, Learning, and Advisory Services at Interpeace, an international peacebuilding organization. She leads Interpeace work on 'AI for Peace.' Prior to joining Interpeace, Brubaker worked for the United Nations in New York, Tokyo, Istanbul, and Geneva.

Dr Eduardo Albrecht is a Senior Fellow (Non-Resident) at UNU-CPR. He is currently an Associate Professor in the Department of Social Sciences at Mercy College in New York and was previously Associate Professor of Anthropology and International Studies at Pukyong National University in South Korea. He was also a Visiting Fellow at the European Institute for Asian Studies in Brussels and the International Peace Institute in New York.

**Acknowledgements:** This paper benefited greatly from the many ideas, suggestions, and general peer reviews provided by Naomi Miyashita and Joseph Bullock of the United Nations Department of Peace Operations.



**ISBN:** 978-92-808-6606-3

**Citation:** Eleonore Fournier-Tombs, Rebecca Brubaker, and Eduardo Albrecht, "Artificial Intelligence-Powered Disinformation and Conflict," UNU-CPR Policy Brief (New York: United Nations University, 2023).