

Image not found or type unknown



No 43 - NOVEMBER 2025

Peace in the
digital age

ICIP
✓

SUMARI

Introduction

- Peace in the Digital Age

In depth

- Disinformation, Manipulation, & Hate Speech in Conflicts
- AI Risks for Sustainable Peace
- Bridging Algorithms as Practical Tools for Depolarisation
- Mediation in the Age of Algorithms: Risks and Opportunities for Peace Processes
- Digital Democracy and Cyberactivism in East Africa: The Role of Siasa Place
- Documenting culture before, during and after conflict

Interview

- Stephanie Williams, former Special Adviser to the United Nations Secretary-General for Libya

INTRODUCTION

Peace in the Digital Age

Helena Puig Larrauri

Co-Founder and Strategy Lead of Build Up

In 2022, #BookTok on TikTok boosted literary sales in the US, local commerce via WhatsApp expanded small businesses across Latin America, and the use of Facebook for voter outreach boosted youth and first-time-voter participation in Indian regional elections. That same year, social media amplified ethnic tensions during the Kenyan elections, spread disinformation in the war in Ukraine, and incited violence during protests in Iran. As in previous years, and those that followed, social media shaped culture, politics, and conflict in multiple, often contradictory ways. But 2022 also marked a turning point: global time spent on social platforms began to decline, especially among young people. Could this signal that its influence on society is starting to wane?

Answering that question requires understanding the mechanisms by which social media influences societies and, of special interest to this monograph, impacts conflict. Fundamentally, social media trends are about a consensus held by a specific group. In the peacebuilding field, we worry about how misinformation erodes truth, damages institutions, and leads to violence. But it's essential to dig a little deeper into the causes for this erosion of truth.

Truth in a post-modern sense is not static; it is dialogical, or, in other words, the reflection of a societal consensus that provides stability and the common ground needed for disagreement, until this consensus itself is constructively challenged. Social media has impacted truth and stability by mining societal consensus, and the key characteristic of this mined consensus is not so much misinformation – that is, a signal or, at most, a tactic – as affective polarisation. Affective (or toxic) polarisation is

distinct from issue-based or idea polarisation. It refers to situations in which one believes certain people for who they are, not for what they say. The more powerful social media trends and digital sub-groups are, the more affective polarisation, and the more a broad societal consensus on the truth splinters into sub-group consensus. It's not so much that truth has been eroded; it's more that it has been fragmented.

“ It’s not so much that truth has been eroded; it’s more that it has been fragmented. ”

In this issue, Ahmad Qadi explores in depth how this fragmentation is impacting conflict, looking at how the “triad of disinformation, hate speech, and polarisation” on social media both reflects and actively fuels tensions, deepens divisions, and ultimately sustains conflict. He calls this – and I would agree – “an essential component of modern information warfare”. He describes how Israel has waged this psychological warfare on Palestinians, using AI-generated content and bot farms to amplify anti-Palestinian sentiment and promote Israel’s military actions across Western digital spaces. Sanjana Hattotuwa’s article further explores the impact of these manipulated online sentiments on a specific scenario of importance to building peace: closed-door negotiation rooms and mediation processes. The interview with Stephanie Williams tells a similar story, looking at how online hate speech and foreign interference during the UN peace efforts in Libya between 2019 and 2020 not only deepened divisions on the ground but also endangered women peacebuilders and negotiators (including Williams herself).

I would love to think that the decline in social media use that started in 2022 can also mean a slow return to a less fragmented, less polarised world where conflicts are not as likely to escalate and become intractable; one where we need to worry less about the risks that Qadi and Hattotuwa explore in their articles. However, the rise of AI over the past few years not only continues the trend started by social media use but also complicates it in ways that further intensify conflict.

In her article on AI risks for sustainable peace, Evelyne Tauchnitz explores this from the perspective of relational peace – that is, how AI diminishes the capacity of individuals and their networks (societies) to turn disagreement into constructive deliberation rather than destructive or violent conflict. She shines a light on how AI impacts dignity – through data colonialism and extraction, by mining and responding to our emotional states, through digital surveillance, and by eroding our practical and moral agency – and how it hollows out trust in institutions. In her view, AI systems erode the conditions that enable societies to be resilient to conflict.

“ I would love to think that the decline in social media use can also mean a slow return to a less fragmented, less polarised world where conflicts are not as likely to escalate and become intractable ”

Building on Tauchnitz’s exploration, there are two specific risks of AI to peace worth paying special attention to. First, AI can undermine truth-as-consensus by impoverishing journalism and research. Machines cannot provide the deeply contextual, human-driven nuance and instinct that reporters and researchers offer, such as interviewing witnesses to events. In his article, Wahbi Abdelrahman weaves a beautiful, sorrowful story of how a digital archive is preventing culturcide in Sudan’s war, and how embedding digitisation skills across local institutions is a form of resilience to erasure. Her piece highlights that the digital archive is a socio-technical process – a human-machine collaboration that AI cannot replace. Ironically, this diminishing of human reporting simultaneously threatens the future development of AI itself. Today’s Large Language Models (LLMs) rely on massive amounts of data, and there is some concern (debated) that if AI depletes the quality human-generated content it needs for training – like quality journalism, archiving and research – it may face a data dead end.

Second, AI can weaken deliberative processes. There are some excellent use cases of AI for deliberation. Luke Thorburn's article explores how we can design ranking algorithms (i.e. the computer programs that decide what content we see first or most, in any digital medium) to foster connection—for example, by finding statements where those who otherwise disagree can agree—and how the Alliance for Middle East Peace has used this. Stephanie Williams describes how using AI to summarise positions in the Libya digital dialogues run by the UN helped to put pressure on the political class to acknowledge popular opinion on difficult issues. As with research, these use cases are embedded in socio-technical processes that retain human connection. Where AI is used to represent constituencies without human participants, it erodes the essential human processes of listening and deliberation, fundamental to democracy and peace. Deliberation is central to public debate and legislative processes, enabling groups to challenge biases, solicit information, and forge compromises that lead to more defensible and durable decisions. Over-relying on AI-driven summarisation or modelling fails to capture this vital, messy, lived experience.

“ There is some concern (debated) that if AI depletes the quality human-generated content it needs for training —like quality journalism, archiving and research— it may face a data dead end ”

Neither social media nor AI are inevitable forces. They are business endeavours. What makes social media worse for society is that its business model is premised on capturing attention. For a minute, it looked like the AI business model might be more focused on cloud infrastructure than on advertising/capturing attention, but earlier this year it became clear that this is changing: OpenAI, one of the most prominent players in the AI space, has chosen advertising as its primary business model, and others will follow suit. Here we go again. As long as the design of the internet, digital technologies, and AI is steadfastly focused on maximising attention, it will have a negative impact on conflict— via fragmentation, via polarisation, and through the

erosion of opportunities for our society to deliberately arrive at a consensus that spells out a sustainable, pluriversal peace.

What's more, if we don't apply conflict-sensitive principles to the design and governance of technology, it can—and will—be weaponised for control by powerful actors. Nerima Wako brings to life this risk of capture in her passionate piece on cyberactivism in East Africa—where, she says, the internet is not just a medium, but the movement—and the repressive digital backlash against this new (primarily youthful) democratic activity. Her piece is a cautionary tale about how regulation can mean repression, but also a hopeful and powerful call to defend a free and trusted internet that offers spaces for deliberation when offline spaces are under attack. In her own words, “A digitally peaceful society is one where activists don't need burner phones; where laws protect speech and where we can disagree loudly, passionately, safely.”

People may have had enough of digital spaces that are simultaneously repressive and polarising; maybe the 2022 decline in social media use is a harbinger of things to come. None of these technologies are unstoppable forces of nature. We can stop using them. We can use them differently. We can regulate them. We can redesign them. The six articles in this monograph also point to hope.

Having painted a bleak picture of how digital warfare has been waged on Palestinians, and how this is an example of broader trends in modern information warfare, Ahmed Qadi's article lists a plethora of avenues to turn from division to dialogue. Similarly, Evelyne Tauchnitz's article ends on a hopeful note, arguing that AI can contribute to peace if we align its design principles not only with improving efficiency and robustness, but also with safeguarding human dignity and freedom. To these broader visions of hope, Wahbi Abdelrahman and Luke Thorburn's articles examine concrete, practical examples of how digital technologies can contribute to peace. Stephanie Williams describes how the negative impact of social media on the Libyan peace process sparked a conversation within the United Nations about how digital dialogues could be used to increase transparency in mediation processes.

“ There are as many challenges to peace from the internet, digital technology, and AI as there are opportunities. Addressing them it is a moral and political question that appeals to anyone who is working towards peace in the digital age ”

How do we get peace-supporting, pro-social technology design, such as Thorburn describes, into the mainstream? Here is one possible answer: Sanjana Hattotuwa's article ends with a call for mediators and peacebuilders not just to become more adept at navigating "today's adversarial digital commons," but also to influence how they are designed.

Hattotuwa's conclusion is also the direction I think we should be walking in. There are as many challenges to peace from the internet, digital technology, and AI as there are opportunities – the articles in this issue offer an overview of what we should worry about in a digitally mediated world and how we might turn the tide towards peace. But crucially, this monograph also makes the case that addressing these challenges is not only a technical question reserved for those of us who know how to run social media campaigns to counter online hate or deploy AI-sensemaking for a consultative process. It is a moral and political question that appeals to anyone who is working towards peace in the digital age.

This issue of the "Peace in Progress" e-magazine (Number 43) is a co-edition by ICIP and Build Up. The collaboration stems from the co-organization of Build Peace 2025 in Catalonia, the Build Up's annual conference on technology, innovation, and peacebuilding. The event took place from November 21 to 23 in Santa Coloma de Gramenet (Barcelona).

About the author

Helena Puig Larrauri

Helena is the co-founder of Build Up, a peacebuilding collective dedicated to innovating conflict prevention and addressing polarization in fragile contexts. With over 15 years of experience, she has worked alongside civil society organizations and multilateral institutions to design and implement dialogue, consultation, and conflict analysis programs in divided and conflict-affected environments. Her expertise lies in digital peace process design, digital inclusion in mediation and peacebuilding, and addressing conflict drivers in online and hybrid spaces.

Helena is a Senior Advisor on Digital Technologies and Mediation to the United Nations Mediation Support Unit, a member of the FCDO's Civilian Stabilisation Group roster, and an Ashoka Fellow. She holds a BA in Politics, Philosophy and Economics from Oxford University and a Masters in Public Policy (Economics) from Princeton University.

Photography

Symbolic representation of collaboration between humans and digital technologies.
Author: Rawpixel.com (Shutterstock).

IN DEPTH

Disinformation, Manipulation, & Hate Speech in Conflicts

Ahmad Qadi

Monitoring & Documentation Manager at Zamleh

In today's digital age, conflicts and wars have increasingly spilled over into the online sphere, amplifying the spread of disinformation and hate speech. Social media platforms have become mirrors—and magnifiers—of real-world tensions, serving both as battlegrounds and breeding grounds for digital hostility. While individual users may spontaneously promote content aligned with their political or ideological beliefs, states and non-state actors often engage in coordinated campaigns to deliberately spread false or misleading information with the intention of deceiving, manipulating, or harming. This intentional manipulation of information is referred to as disinformation—the strategic dissemination of falsehoods intended to mislead populations, damage reputations, and sow division.^[1]

Concurrently, hate speech proliferates across these platforms, often incited or legitimized by political elites and power structures. It manifests as communication that expresses or incites hatred against individuals or groups based on intrinsic characteristics such as ethnicity, religion, nationality, or gender.^[2] In times of conflict, hate speech is frequently weaponized to unify populations against a common enemy, justify violence, or intimidate dissenting voices.

At the heart of this digital landscape lies polarization—a condition in which societies are fractured into opposing camps with rigid worldviews. Polarization intensifies as individuals become increasingly entrenched in supporting one side while wholly rejecting or dehumanizing the opposing side. This triad of disinformation, hate speech,

and polarization does not merely reflect existing tensions; it actively fuels them. Together, they form a self-perpetuating cycle that deepens societal divisions, escalates hostility, and sustains conflict.^[3]

“ This triad of disinformation, hate speech, and polarization form a self-perpetuating cycle that deepens societal divisions, escalates hostility, and sustains conflict ”

The Spiral of Polarization in War

Empirical studies across numerous conflict zones reveal a recurring pattern: disinformation and hate speech surge during times of heightened political tension and erupt when conflicts escalate into violence. In such moments, social media becomes an essential tool for competing narratives. Warring parties exploit these platforms to mobilize support, justify military action, discredit opponents, and manipulate public sentiment. Disinformation campaigns often seek to undermine the enemy's morale, instill fear, or destabilize internal cohesion. These digital campaigns are not confined to cyberspace; rather, they have tangible, often deadly, consequences. The United Nations, for instance, concluded that Facebook played a “determining role” in fueling the genocide against the Rohingya in Myanmar. Online hate did not remain virtual—it catalyzed real-world violence.^[4]

Disinformation in conflict settings constitutes a sophisticated ecosystem designed not just to confuse, but to control. It distorts truth, minimizes suffering, erodes empathy, and deepens societal divides. At its core, it is an emotional and psychological weapon—an essential component of modern information warfare. This weaponization of digital spaces is evident in numerous ongoing conflicts.

In Palestine, disinformation campaigns have contributed to the vilification of victims, the denial of documented atrocities, and the legitimization of targeting journalists and

humanitarian workers. Narratives portraying children and survivors as “crisis actors” or staging their suffering serve to delegitimize real experiences and shield perpetrators from accountability. These strategies not only deny justice to victims but also manipulate global public opinion and shape policy discourse. While some elements may emerge spontaneously, especially during crises, the dominant thrust of such campaigns is deliberate and organized. Through coordinated networks, conflicting parties craft and disseminate content aimed at engineering consent, shifting blame, and constructing binary worldviews that leave little room for complexity, empathy, or dialogue.^[5]

“ Disinformation is an emotional and psychological weapon; in conflict settings it distorts truth, minimizes suffering, erodes empathy, and deepens societal divides ”

While it’s supposed to be true that those forms of content exacerbate during conflict, hate speech and disinformation are often rooted in broader societal narratives—be they nationalistic, religious, or ideological. These narratives, ingrained over time, re-emerge in digital discourse as violent or misleading content. Political actors frequently amplify these stories, framing their own side as morally righteous and aggrieved. While these narratives may simplify complex realities into binary oppositions of good versus evil, social media’s algorithms intensify this reductionist thinking by privileging emotionally charged content and discouraging nuance. This environment is fertile ground for the spread of misinformation and incitement.^[6]

Hate speech and disinformation among other violent content forms are the direct result of toxic polarization which goes beyond normal ideological disagreement and arises when individuals develop deep contempt for those with opposing beliefs while expressing intense loyalty and attachment to their own group’s views. This form of polarization transforms political or ideological differences into identity-based divisions, fostering the perception that the opposing side is not just wrong, but an irreconcilable

enemy. Psychological research identifies three core drivers of this dynamic: dehumanization, dislike, and disagreement. When group members believe that the other side fundamentally dislikes, dehumanizes, or opposes them, polarization intensifies.^[7]

This is exactly what happened in the case of the genocide going on in Gaza. Since the Israeli war against Gaza broke on October 7, 2023, a huge disinformation war has swept digital spaces, with many describing it as unprecedented. Israel launched a comprehensive disinformation and influence campaign designed to justify its military assault on Gaza and delegitimize Palestinian rights. Central to this effort was a \$2 million covert operation funded by Israel's Ministry of Diaspora Affairs, which used AI-generated content and bot farms to manipulate public opinion and promote dehumanizing narratives about Palestinians. This campaign strategically targeted U.S. lawmakers, especially Democrats, through platforms like Facebook, Instagram, and X, with tailored propaganda produced by the Israeli firm STOIC. OpenAI later disrupted some of these efforts, revealing a sophisticated operation aimed at amplifying Islamophobic and anti-Palestinian sentiment across Western digital spaces.

“ The war on Gaza has unequivocally demonstrated how offline conflicts can dramatically intensify online violence and disinformation, and can fuel false and harmful narratives ”

Simultaneously, the Israeli Ministry of Foreign Affairs launched a graphic and emotionally manipulative YouTube ad campaign across Europe and North America, designed to stir support for Israel's military actions. According to 7amleh's analysis, these ads, which included disturbing imagery and appeals framed in child-centered narratives, violated platform standards yet remained active. Additionally, people affiliated with Israel reached out to influencers with offers of payment and “briefing sessions” to distribute pro-Israel messaging on social media, further embedding state-

sponsored narratives into grassroots digital spaces. This orchestrated manipulation highlights the asymmetric digital warfare in which Israel leverages advanced tools and vast resources to dominate discourse and suppress Palestinian voices.^[8]

The war on Gaza has unequivocally demonstrated how offline conflicts can dramatically intensify online violence and disinformation. The existence of an entire Wikipedia page dedicated to tracking the vast amount of disinformation unleashed during the genocide underscores the scale of this phenomenon. It highlights not only how the offline conflict fueled a surge in false and harmful narratives online, but also how much of this disinformation was systematically driven by political actors—beyond the unconscious spread of misinformation by ordinary users.^[9]

At the same time, Israel's opponents—such as Iran—have intensified disinformation efforts, especially during the July 2025 flare-up following Israel's offensive. A coordinated Iranian campaign using over 100 bot accounts on X (formerly Twitter) posted more than 240,000 times, aiming to sway U.S. public opinion and deter potential strikes on Iran's nuclear facilities. The operation promoted Iran's Supreme Leader, spread false claims of Israeli military failures, and portrayed Israel as a terrorist state. Posts used inflammatory imagery and hashtags to maximize virality and undermine the U.S.-Israel alliance. Analysts view the campaign as part of Iran's broader strategy to manipulate global discourse and preempt military action through psychological and informational warfare.^[10]

“ The emotional nature of conflict-related messaging, marked by fear, anger, or grief, further heightens susceptibility to manipulation and confusion ”

Several factors contribute to the spread of disinformation and other polarizing forms of content during conflicts. These include a widespread lack of verification and critical reasoning when assessing received information, as well as a tendency to view content as credible simply because it originates from an in-group source. Disinformation is

often perceived as factual from a specific worldview, particularly when it aligns with deeply rooted stereotypes or long-standing narratives.

Cognitive biases also play a central role: individuals tend to seek information that confirms their existing beliefs, avoid content that challenges those beliefs, and resist updating their views even when confronted with credible counterevidence. The emotional nature of conflict-related messaging—marked by fear, anger, or grief—further heightens susceptibility to manipulation and confusion. People may acknowledge that disinformation exists, yet believe that only others are vulnerable to it, reinforcing blind spots in their own judgment. The echo chamber effect deepens these dynamics by limiting exposure to alternative perspectives, while disinformation cloaked in the language of public, institutional, or scientific authority gains unwarranted legitimacy. In some cases, exposure to accurate information can paradoxically entrench belief in falsehoods, as individuals reinterpret or reject the truth in ways that reinforce their original convictions. Together, these dynamics create a fertile environment for disinformation to thrive and polarization to intensify.^[11]

The Role of Social Media

Social media platforms have played a significant role in escalating conflict by incentivizing divisive and potentially violence-inducing speech, often amplifying content that fosters polarization and mass harassment through algorithms optimized for engagement metrics like shares, comments, and time spent. While platforms have primarily addressed violent conflict through reactive content moderation—responding to specific incidents or outbreaks—this approach fails to address the deeper, systemic design issues that fuel conflict dynamics long before violence erupts. Independent investigations, such as the one regarding Facebook’s role in Myanmar, have shown how platforms can facilitate the incitement of offline violence by fomenting division. Research supports a complex picture: social media correlates with both polarization and increased political knowledge, but the former—especially toxic polarization, where people demonize opposing groups—is a more dangerous prelude to violence than mere policy disagreements.^[12]

“ Social media platforms have played a significant role in escalating conflict by incentivizing violence-inducing speech. Its impact is not confined to the digital realm, it operates within a reinforcing cycle, intensifying polarization and inciting offline violence ”

The NYU Stern report “Fueling the Fire: How Social Media Intensifies U.S. Political Polarization – And What Can Be Done About It” concludes that although social media is not the root cause of political polarization in the United States, it plays a critical role in intensifying affective polarization—deep-seated hostility and contempt between political groups—which in turn erodes democratic norms, weakens trust in institutions, and fuels real-world violence. The report highlights how engagement-based algorithms systematically amplify divisive content, and despite occasional internal measures, platforms have largely failed to regulate themselves.^[13]

The violence offline and online works as a reinforcing cycle, through which any offline incident could fuel online violence and vice versa. A prominent example from Palestine illustrating the role of social media was documented by 7amleh. The organization analyzed the digital buildup to the February 2023 attack on the village of Huwara in the West Bank, where hundreds of Israeli settlers carried out a violent assault that resulted in the killing of one Palestinian, widespread property destruction, including the torching of crops and vehicles, attacks on homes, and the terrorizing of residents. In the months leading up to the attack, 7amleh identified over 15,000 pieces of violent Hebrew-language content on social media platforms that directly targeted the village and its inhabitants. This content served to delegitimize, smear, and dehumanize the local Palestinian population, portraying them in inhumane and threatening terms. Such digital incitement laid the groundwork for real-world violence by normalizing hostility and justifying aggression against the community.^[14]

From Division to Dialogue: Reversing Polarization

A lot can be done to undo prejudice and polarization; one of the hypotheses is simply communication. Contact between groups can have a significant impact between groups or conflicting parties, but only if it's done properly. Sometimes, contact can worsen polarization. For example, following opponents on X might consolidate one's own extreme ideas or prejudice against the others. If the contact is done in a sustainable way, with respectful exchange of ideas among people of the same rank or age, for example, that could reduce polarization. Intervention can be done by governments, Civil Society Organizations, media, social media platforms and other actors to spread the perspectives of others. Listening to stories from the perspective of others could help reduce much of the prejudice and polarization. However, social media has done more harm than good by allowing people to see and engage with similar people or narratives rather than opening them to other perspectives.^[15]

“ Listening to stories from the perspective of others could help reduce much of the prejudice and polarization ”

To counter this, initiatives like Beyond Conflict recommend public awareness campaigns to expose the widespread misperceptions partisans hold about each other and call for holding influential figures accountable when they spread misinformation. Practical strategies to mitigate polarization include fostering intergroup dialogue, encouraging empathy through perspective-taking, and highlighting internal disagreements within political groups to disrupt rigid “us vs. them” narratives. Additionally, promoting kindness on social media can help reduce the normalization of dehumanization, while avoiding the repetition of misinformation and refraining from prejudiced jokes can weaken the social acceptability of divisive rhetoric.^[16]

Other approaches advocate for a shift from reactive moderation to proactive, long-term, and scalable interventions rooted in platform design. The proposal is that platforms move away from engagement-based content ranking in sensitive contexts, limit mass dissemination capabilities, and introduce design features that foster meaningful,

connecting interactions. Platforms should also integrate support for peacebuilding efforts, acknowledging that peace is not just the absence of violence but the presence of social conditions where all communities can flourish. Ultimately, conflict escalates through reinforcing cycles, and breaking this spiral requires disrupting the incentives that reward division and manipulation online.^[17]

Other recommendations call for comprehensive structural reforms, urging social media platforms to transparently redesign their algorithmic systems to reduce the spread of inflammatory content. For example, the NYU Stern report underscores the importance of investing in robust, in-house content moderation teams and fostering deeper collaboration with civil society organizations. It also highlights the critical role of government intervention—advocating for legislation that mandates transparency, empowers regulatory agencies to enforce conduct standards, and supports the development of alternative digital platforms that promote democratic engagement. Ultimately, the report frames unchecked polarization driven by social media as a direct threat to democratic stability, one that demands urgent, coordinated action across sectors.^[18]

Conclusion

Social media platforms and digital tools play a significant and often detrimental role in fueling violence, division, and polarization—both in times of peace and, more acutely, during war. Conflicting parties actively exploit these platforms to amplify their narratives, frequently resorting to disinformation and hate speech as strategic tools to delegitimize the other side. Social media offers a cost-effective and far-reaching means to disseminate such content, making it a powerful instrument in modern information warfare. Crucially, the impact of this content is not confined to the digital realm. It operates within a reinforcing cycle—intensifying polarization, deepening societal fractures, and ultimately inciting offline violence. This dynamic underscore the urgent need for platform accountability and a fundamental rethinking of how digital infrastructures intersect with conflict dynamics.

- [1] Stavros, A., S. Phalen, S. Almakki, M. Nacionales-Tafoya, & R. A. García. 2023. “Disinformation in Conflict Environments in Asia.” Gerald R. Ford School of Public Policy, University of Michigan.
- [2] 7amleh (Arab Center for the Advancement of Social Media) 2022. “A Guide to Combating Online Hate Speech.”
- [3] Polarization Research Lab. 2022. “How Do You Study and Reverse Political Animosity? These Researchers Are Working to Answer That Question.” Charles Koch Foundation.
- [4] Stavros et al. 2023
- [5] Barforoush, S., & S. Plaut. 2024. “Information Disorder in Times of Conflict.” Canadian Museum for Human Rights.
- [6] Cobb, S., S. Kaplan, A. Marc, & G. Milante. 2021. “The Role of Narrative in Managing Conflict and Supporting Peace.” Discussion paper, IFIT, Institute for Integrated Transitions, Barcelona.
- [7] PsicoSmart Editorial Team. 2024. “The Role of Technology in Conflict Mediation: Exploring Digital Tools and Platforms.” Psico-smart Blog..
- [8] Qadi, A. 2025. “From Bot Farms to Censorship: Israel’s Disinformation Warfare against Palestinians.” *Palestine Chronicle*.
- [9] “Misinformation in the Gaza War.” 2025.
- [10] Kahana, A. 2025. “Iran Deployed Bots to Post 240K Times to Block US Strikes on Nuclear Facilities.” NYP.
- [11] Lewandowski, P. 2024. “Psychological Mechanisms of Disinformation and Their Impact on Social Polarization.” *Studia Polityczne* 2: 85-104. Łukasiewicz Research Network - ITECH Institute of Innovation and Technology.
- [12] Stray, J., R. Iyer, & H. Puig Larrauri. 2023. “The Algorithmic Management of Polarization and Violence on Social Media.” Knight First Amendment Institute.

[13] Barrett, P. M., J. Hendrix, & J. G. Sims. 2021. ["Fueling the Fire: How Social Media Intensifies U.S. Political Polarization and What Can Be Done about It."](#) Center for Business and Human Rights, Stern School of Business, New York University.

[14] 7amleh (Arab Center for the Advancement of Social Media). 2023. ["An Analysis of the Israeli Inciteful Speech against the Village of Huwara on Twitter."](#)

[15] Moskalenko, S. 2023. ["What Are the Solutions to Political Polarization?"](#) Greater Good Magazine, Greater Good Science Center, University of California.

[16] PsicoXsmart Editorial Team. 2024. [The Role of Technology in Conflict Mediation: Exploring Digital Tools and Platforms."](#) Psico-smart Blog.

[17] Stray, J., R. Iyer, & H. Puig Larrauri. 2023. ["The Algorithmic Management of Polarization and Violence on Social Media."](#) Knight First Amendment Institute

[18] Barrett, P. M., J. Hendrix, & J. G. Sims. 2021. ["Fueling the Fire: How Social Media Intensifies U.S. Political Polarization and What Can Be Done about It."](#) Center for Business and Human Rights, Stern School of Business, New York University.

About the author

Ahmad Qadi

Ahmad Qadi is the Monitoring & Documentation Manager at 7amleh. He works on content moderation, social media policies, and censorship. Ahmad is a digital rights defender, researcher, and PhD student examining the impact of technology on societies.

Photography

Visual metaphor about disinformation and polarization in the digital environment.
Author: Evan Huang (Shutterstock).

IN DEPTH

AI Risks for Sustainable Peace

Evelyne Tauchnitz

Senior Researcher and Lecturer at the Institute of Social Ethics, University of Lucerne

Peace is often misunderstood as the mere absence of war or overt violence – a state of order and stability. But this narrow conception of what Galtung has called “negative peace”^[1] misses the deeper relational, systemic, and ethical foundations that make peace sustainable over time. In contrast, sustainable peace is not a static achievement or institutional end state, but a dynamic condition that emerges from the quality of relationships, the resilience of social networks, and the presence of human dignity and freedom.

To grasp peace more fully, it is therefore necessary to move beyond abstract ideals or treaty frameworks and attend to how people, groups, and institutions relate to one another. Peace is not a substance or structure, but a positive pattern of interaction – something that is closely tied to human behaviour. It is a phenomenon that happens “in between the nodes”: in the connections, exchanges, and mutual recognitions that bind people together through positive human experiences and exchanges. These linkages can be mapped as networks, where each node – whether a person, group, or institution – carries capacities and vulnerabilities, and each link that connects the different nodes reflects a particular quality of relationship that consists of a unique mix of characteristics such as trust and care, but also conflict or neglect. This networked view of peace shifts the focus from static indicators to the dynamics of connection: how relationships are built, strained, or broken; how power circulates; and how peace and violence coexist on a continuum. Networks can be robust, bridging divides and supporting cooperation – or brittle, marked by fear, exclusion, and fragmentation. Peace is thus not a given, but same as violence, an emergent property of the conditions that

define the characteristics of the system – always at risk of tipping, especially when disrupted by crisis, injustices, or also technological transformations (including AI).

“ At the heart of peaceful networks lie the ethical principles of human dignity and freedom: to handle crises sustainably and transform conflicts though non-violent depends on the levels of resilience ”

At the heart of peaceful networks lie the ethical principles of human dignity and freedom. Dignity is not only intrinsic worth, but a relational quality, sustained when people are recognized in their uniqueness, included, and treated as persons with voice and value. Freedom is not just freedom from coercion, but the ability to participate meaningfully in shaping one's life and relationships. As Hannah Arendt^[2] and Amartya Sen^[3] have argued, freedom is agency-in-context, exercised within structures that either support or suppress it. When dignity and freedom are systematically denied – through surveillance, exclusion, or disinterest – the relational infrastructure of peace begins to unravel. Trust erodes, participation fades, and violence emerges mostly not from hatred, but from systemic breakdown – the “banality of evil,” as Arendt warned, where ordinary actors follow harmful norms in dysfunctional systems. The genocide in Gaza and increased violence by settlers in the Westbank in the aftermath of the Hamas attack on October 7, 2023, provides a tragic example.

While dignity and freedom are the ethical anchors of peace, resilience is its operational backbone – the set of capacities that enable individuals, communities, and systems to withstand disruption and recover or transform without descending into violence. Crises and even conflicts will always happen, but whether violence erupts or, in contrary, individuals, communities, and institutions manage to handle crises sustainably and transform conflicts though non-violent means into states of peaceful co-existence, depends on their levels of resilience. Resilience is not just about survival or bouncing back. It is about navigating crises in ways that preserve core values and relationships,

adapt to changing conditions, and open space for positive change and renewal.

Four interdependent pillars of resilience can be identified:

1. **Resources:** Access to food, shelter, healthcare, knowledge, emotional meaning – all the material and symbolic tools needed to cope with crises.
2. **Social Capital:** Trust, solidarity, and networks of mutual support that enable people to share burdens, access resources, and coordinate responses.
3. **Adaptive Capacity:** The ability to learn, innovate, and adjust strategies when old ones no longer work.
4. **Enabling Environments:** Institutions and policies that ensure fairness, protect human rights, and provide avenues for peaceful change.

Together, these factors determine whether peace can endure under pressure. When resilience is strong, networks flex but do not break; when it is weak, crises can push systems past tipping points where spirals of violence become self-reinforcing (see e.g. recent outbreaks of tribal violence in Syria 2025). These tipping points occur when social norms shift toward fear or aggression, resources become scarce, trust erodes, or opportunistic actors exploit conflicts. As in network theory, a small rupture in one part of the system can cascade – especially if relational ties are already strained. The next section looks at the disruptive force of AI and its potential impact on resilience which shapes the capacity of networks to safeguard peace.

AI as a Risk to Networks Resilience and their Capacity to Safeguard Peace

Artificial intelligence is reshaping how people work, communicate, and are governed – often framed in terms of innovation, efficiency, or optimization. But when seen through the lens of relational peace, AI reveals a deeper risk: it diminishes the resilience of individuals and societies to handle crises and conflicts peacefully.

Understanding AI's impact on peace thus requires moving beyond overt threats like autonomous weapons systems or new cyber threats. Looking at resilience, AI's effects on resources, social capital, adaptive capacity, and enabling environments must be

examined, as these elements together form the relational infrastructure of sustainable peace.

AI and the Disruption of Resources: Dignity Through Livelihood and Emotional Integrity

Resources are the foundation of resilience in any network – not just material goods like food or shelter, but also symbolic, emotional, and informational capacities. AI technologies increasingly control the flow and distribution of these resources, and in doing so, reshape the conditions under which a life in human dignity is affirmed or denied.

“ Through the lens of relational peace, AI reveals a deeper risk: it diminishes the resilience of individuals and societies to handle crises and conflicts peacefully ”

On a structural level, AI-powered tools are increasingly taking over repetitive administrative tasks traditionally performed by back-office clerks – for example, processing invoices, updating payroll records, or verifying electronic forms. While these systems may not fully replace entire roles overnight, they significantly reduce the demand for human labour, leading to fewer entry-level opportunities. Similar trends are visible in other sectors: AI chatbots now manage a large share of customer service inquiries; computer vision systems in warehouses guide autonomous robots in picking, sorting, and packing goods, replacing or reducing manual handling roles; in finance, algorithmic trading systems execute trades and optimize portfolios once handled by teams of junior analysts.

While some new jobs emerge – for example, in AI model training, data labelling, or maintaining and overseeing automated systems – the nature of these opportunities is often narrowly defined. Many are tied to specific functions within the AI development pipeline, such as annotating datasets for machine learning, fine-tuning language

models, monitoring algorithmic outputs for errors, or servicing autonomous machinery. Others involve supervisory roles where humans oversee automated processes, intervening only when the system encounters exceptions or failures. In addition, many of these new job opportunities are often temporary, geographically concentrated, or require advanced skills many displaced workers cannot easily acquire.

The net effect of AI on jobs is therefore likely to be uneven and disruptive for three reasons. First, automation tends to benefit highly skilled or capital-rich actors, while workers in the Global South, precarious labor markets, and routine jobs where many workers are employed face the highest risk of dispossession. Second, the transition periods between job loss and new employment can be long and destabilizing, eroding economic security. Third, even when AI increases efficiency and lowers the cost of goods, this does not automatically translate into better livelihoods. Gains are often distributed unevenly: while some individuals and regions benefit from cheaper products, improved services, and new market opportunities, others experience job loss, wage stagnation, or weakened labor protections. These disparities can widen existing inequalities, both within and between countries, as highly skilled workers and capital owners capture a disproportionate share of the benefits. The result for many is a loss of economic dignity: the ability to support oneself and one's family with purpose and security. These technological disruptions provide a strong argument in favour of an unconditional basic income – a guaranteed, regular cash payment to all citizens, grounded in the right to a life of dignity in which everyone's basic needs are met.

“ AI contributes to what scholars have called “data colonialism”: Vast quantities of data are extracted, and people are rendered legible and exploitable by historical patterns of resource extraction and domination ”

Moreover, algorithmic hiring and performance systems – such as automated CV screening tools or AI productivity tracking software – reduce people to metrics, leading

to discrimination (for instance, rejecting candidates due to gaps in employment) and undervaluing the human dimensions of work like creativity, care, and collaboration.

At the global scale, AI contributes to what scholars have called “data colonialism.”^[4] Vast quantities of data are extracted – often without consent – from individuals, communities, and devices across the world, only to be analysed, monetized, and controlled by a handful of powerful firms and states. This process mirrors historical patterns of resource extraction and domination, with the added twist that the “resource” in question is the relational and behavioural trace of human lives. Dignity is compromised not only because consent is bypassed, but because people are rendered legible and exploitable without reciprocity or recourse.

Even emotional and psychological resources are affected. AI-enhanced platforms – such as TikTok’s *For You* feed, Instagram’s *Explore* page, Facebook’s News Feed, X’s (formerly Twitter’s) timeline algorithm, or YouTube’s autoplay and suggested videos – use machine learning to predict and amplify the content most likely to capture and hold a user’s attention. In social media especially, these mechanisms intensify comparison, competition, and performativity, as users are continually exposed to curated portrayals of others’ achievements, lifestyles, and opinions.

**“ AI reshapes resource flows in ways that may
improve access to information or certain services
yet also undermines the dignity ”**

By mining attention and emotion for profit, these systems erode people’s sense of interior stability and self-worth – essential dimensions of resilience. Surveillance systems further degrade these resources by producing constant low-level anxiety, especially among marginalized populations and civil society actors who already face disproportionate monitoring, whether through predictive policing algorithms, workplace monitoring, automated facial recognition in public spaces, or AI-driven monitoring of online activism.

Taken together, AI reshapes resource flows in ways that may improve access to information or certain services yet also undermines the dignity that comes from having one's labor valued, one's needs met, and one's boundaries respected.

AI and the Erosion of Social Capital: Fracturing the Trust That Sustains Peace

Social capital – the web of trust, norms, and informal relationships that bind people together – is a core pillar of resilience in any community. It enables coordination, mutual aid, and collective action. Where trust and empathy circulate freely, networks tend to bend rather than break under stress. Where they are thin or brittle, crises more easily tip into blame, fear, and fragmentation.

AI systems increasingly intervene in the relational web that underpins peace. Most visibly, social media algorithms, optimized for engagement, prioritize content that elicits outrage, affirmation, or other strong emotions. This shifts the balance of what circulates within and between networks, giving greater visibility to divisive or emotionally charged material while crowding out content that fosters deliberation or nuance. Over time, such algorithmic filtering encourages the formation of polarized echo chambers – tightly knit clusters within broader networks where information flows mainly among like-minded members. These clusters fragment the overall network, weakening bridging social capital – the connections that link people across different backgrounds or perspectives – while reinforcing bonding social capital within homogenous groups.

As in-group identities strengthen, so too does toxic (affective) polarization: a form of division in which individuals not only disagree but develop deep contempt for those with opposing views, alongside intense loyalty to their own group. This “us versus them” mentality narrows the range of perspectives people encounter and undermines the trust, reciprocity, and shared norms that sustain cooperation. While polarization is not new, AI accelerates and amplifies it – often invisibly – shifting the network dynamics that support understanding and mutual recognition.

“ Social media algorithms, optimized for engagement, prioritize content that elicits outrage, affirmation, or other strong emotions. AI accelerates and amplifies polarization ”

Beyond shaping what we see online, AI increasingly shapes how we feel and behave. AI technologies – deployed in social media, retail, education, workplaces, and public administration – often use facial analysis, voice tone detection, and biometric data to infer emotions in real time. Once detected, these emotional states can be used to tailor content or responses: a frustrated customer might receive a calming tone from a chatbot; a student flagged as disengaged by an AI learning platform might be pushed more stimulating material; an employee whose voice is deemed insufficiently “positive” might be prompted to adopt a more upbeat tone in calls.

While often presented as tools to improve service quality or workplace efficiency, the same techniques for reading and influencing emotion can be repurposed in political or social contexts. For instance, political campaigns or activist movements could use emotional AI to identify and target individuals already primed for anger or fear, pushing content that reinforces grievances and strengthens in-group solidarity. Both effects deepen existing divides, making it harder for individuals to engage constructively across differences. When combined with AI-driven surveillance the effects are magnified. Constant awareness of being watched or profiled encourages self-censorship and suppresses dissent. This erodes the authenticity, trust, and openness needed for resilient peace, replacing them with guarded, calculated interactions that weaken the very networks on which social cohesion depends.

In authoritarian regimes such as for example Russia or China, AI-driven surveillance and social credit systems formalize this breakdown of trust. When people fear that any social tie might make them vulnerable to state control or reputational damage, they begin to withdraw from civic and communal life. The social fabric frays not through direct violence, but through relational corrosion – a slow collapse of intermediating institutions and everyday solidarities.

AI Undermining Adaptive Capacity: Locking Societies into Past Harmful Practices

Adaptive capacity is perhaps the most subtle yet essential form of resilience. It is what allows individuals and institutions to navigate novelty, respond to feedback, and change course when old strategies no longer work. Without adaptive capacity, systems become fragile – unable to recover or transform under pressure. In a world shaped by AI, this capacity is increasingly constrained.

Many AI systems are designed to optimize for specific outcomes based on historical data. But in doing so, they often lock in certain assumptions, metrics, or patterns that resist contestation or adaptation. For example, predictive policing systems may perpetuate historical bias based on ethnicity or socioeconomic background. Moreover, individuals affected by AI decisions often lack the capacity to understand, question, or resist those systems. This is especially true for marginalized populations who may be subjected to algorithmic welfare assessments, automated visa denials, or digital surveillance – all without access to explanation, redress, or participation possibilities. This produces not only epistemic injustice but practical disempowerment, frustration and eventual resignation: the denial of one's ability to learn from and respond to the systems that shape one's life.

“ Where people once relied on their own judgment and common sense, they are now often expected to comply with algorithmic protocols ”

Perhaps most importantly, AI's can discourage moral agency and the capacity to make exceptions – both central to adaptive capacity. Where people once relied on their own judgment and common sense, they are now often expected to comply with algorithmic protocols, even when those protocols do not fit the lived reality of the situation. In the criminal justice system, for example, risk assessment algorithms may recommend harsh sentencing for unemployed youth from marginalized neighbourhoods without considering the root causes of criminality, such as unemployment – potentially linked to automation – or the absence of meaningful perspectives for the future. The ability to

make exceptions, to recognize potential for rehabilitation, offer second chances, and weigh factors not captured in the data, is crucial not only for the long-term prospects of individuals' lives, but also for the social cohesion of communities.

Similarly, in social welfare systems, automated eligibility checks may deny support to families in acute need if their circumstances do not match pre-set categories, leaving frontline workers with little discretion to override the decision. In both cases, rigid adherence to algorithmic outputs narrows the space for contextually sensitive moral judgment and pragmatic problem-solving – the very qualities that enable societies to adapt constructively to changing circumstances over time.

AI and the Breakdown of Enabling Environments: Hollowing Out Democratic Institutions

The final pillar of resilience – the enabling environment – includes legal systems, political institutions, and governance frameworks that manage risk, protect rights, and ensure fairness. These institutions play a vital role in structuring relationships, mediating conflict, and upholding the dignity and freedom of all members of a community. Yet AI is developing in ways that increasingly outpace, bypass, or undermine these structures.

The opacity of many AI systems – often protected by intellectual property law or trade secrecy – makes it difficult for affected individuals or even regulators to understand how governance and administrative decisions are made. This undermines accountability and procedural justice. The fact that AI often learns from biased or incomplete data further exacerbates this challenge: when systemic injustices are coded into algorithmic logics, injustice is naturalized and rendered invisible.

“ In many contexts, the use of AI in surveillance, border control, or protest monitoring has actively eroded civil liberties, especially for already vulnerable groups ”

Moreover, the development and deployment of AI are currently dominated by a handful of powerful corporate actors and state agencies. These actors often operate transnationally, with minimal democratic oversight. In many contexts, the use of AI in surveillance, border control, or protest monitoring has actively eroded civil liberties, especially for already vulnerable groups. In China's Xinjiang region, for instance, advanced facial recognition systems have been deployed to monitor and detain Uyghur Muslims – a clear violation of freedom of movement and religious practice. In Greece, EU-funded drones, cameras, and AI-powered systems monitor migrant movements at land borders with Turkey, often in non-transparent ways that raise serious human rights concerns around forced displacement and the criminalization of migration. In the occupied Palestinian territories, Israeli authorities have deployed AI-driven facial recognition systems, such as the “Blue Wolf” and “Red Wolf” programs, to identify and track Palestinians across checkpoints and in public spaces. Elsewhere, during protests in countries such as Russia, India, and Iran, AI-based surveillance tools – including automated identification systems such as facial and gait recognition which analyses an individual's walking patterns – have been used to identify and target demonstrators, severely undermining the rights to freedom of expression and assembly.

The use of algorithmic systems to monitor, profile, or target groups without nuanced context or recourse undermines both human rights and social resilience. As enabling environments falter under these pressures, the networks that once sustained peaceful interaction become increasingly prone to rupture. Trust in democratic institutions diminishes. Recourse to redress becomes elusive. People lose faith that the system can protect them – and, in the absence of peaceful pathways for change, may turn instead to resistance, withdrawal, or violence.

Conclusion: Reweaving Peace in the Age of AI

While not inherently violent, AI systems can erode the very conditions – resources, social capital, adaptive capacity, and enabling environments – that make societies resilient to conflict and crisis. By disrupting access to resources, fragmenting networks, constraining problem-solving capacity, and weakening democratic institutions, AI risks hollowing out the relational fabric that holds communities together. These risks rarely unfold through sudden shocks; they emerge quietly, through disconnection,

disempowerment, and dehumanization. But they are not inevitable.

AI is not a single, deterministic force. Its impacts depend on the purposes it serves, the values embedded in its design, and the contexts in which it is deployed. As the following positive examples show, AI can also be used along the different pillars of resilience to promote peace rather than undermine it:

Equitable access to resources: In contexts where tensions between communities are heightened – such as between herders and farmers competing for scarce land – AI-powered climate forecasting and land-use planning tools can help anticipate droughts, optimize grazing rotations, and reduce disputes over resources. In Kenya, for example, platforms like *Virtual Agronomist* and *PlantVillage* deliver localized farming advice in multiple languages, enabling smallholders to improve yields.

Building social capital across communities: AI-assisted dialogue platforms, such as *Pol.is* which has been used in Taiwan, Canada, Singapore, Philippines, Finland, Spain, and other countries, have shown how technology can bridge divides by mapping areas of agreement across large and diverse groups. In deeply polarized contexts, such tools can strengthen bridging social capital by fostering connections between people who might otherwise never meet across ethnic, religious, or socioeconomic lines.

Enhancing adaptive capacity through inclusive planning: In Brazilian cities like Porto Alegre and Belo Horizonte, AI-enhanced participatory budgeting allows residents – including those from marginalized areas – to propose, debate, and vote on public investment priorities. This not only strengthens local problem-solving capacity but also builds the confidence and skills needed for communities to adapt to future challenges.

Strengthening enabling environments through digital democracy: Civic platforms such as *Decidim* in Barcelona offer transparent, open-source systems for citizens to co-create policies, draft proposals, and hold institutions accountable. Such tools, when managed well and in an inclusive manner, can revitalize trust in governance and enable more direct, participatory forms of democracy.

“ Is not merely to make AI systems safer or more robust and efficient. It is to re-anchor them in a vision of peace: a vision that prioritizes human dignity and freedom, especially for the most vulnerable ”

By aligning AI's use with principles that safeguard human dignity and freedom, we can help transform networks toward trust rather than mistrust, and toward cooperation rather than polarization. In doing so, AI could become not a threat to sustainable peace, but an active contributor to it. The task, then, is not merely to make AI systems safer or more robust and efficient. It is to re-anchor them in a vision of peace: a vision that prioritizes human dignity and freedom, especially for the most vulnerable. This may include ethics-by-design and co-creation approaches, but also to define AI-free spaces that are reserved for humans-only. In any case, two ethical frameworks prove indispensable:

First, the ethics of care reminds us that peace is sustained not through control, but through attentiveness to need, vulnerability, and the context. Care-oriented approaches encourage the design of systems that are inclusive, responsive, and grounded in human relationships. They invite us to value not only outcomes, but the quality of interactions that get us there.

Second, a human rights approach ensures that peace is not left to benevolent intention alone. Rights frameworks embed dignity and freedom into legal norms and institutional practices, protecting individuals from systemic abuse and enabling meaningful participation in shaping technological futures. They provide structural guardrails such as transparency, accountability, and non-discrimination that are essential for preserving peace under conditions of rapid change and crises.

In this sense, peace in the age of AI is not a technical or regulatory problem alone. It is a moral and political project. It requires rethinking not only the tools we build, but the values we encode and the futures we make possible. If we are to preserve peace in an

intelligent machine age, we must begin by defending – and designing for – what makes us human.

[1] Galtung, J. 1969. "Violence, Peace and Peace Research." *Journal of Peace Research* 6 (3): 167–91.

[2] Arendt, H. 2006 [1963]. *Eichmann in Jerusalem: A Report on the Banality of Evil*. Penguin Publishing Group.

[3] Sen, A. 2014 [1999]. "Development as Freedom." A *The Globalization and Development Reader: Perspectives on Development and Global Change*, 525.

[4] See for example Couldry, N., & Mejias, U. A. (2019). *The costs of connection: How data is colonizing human life and appropriating it for capitalism*. In *The costs of connection*. Stanford University Press.

About the author

Evelyne Tauchnitz

Senior Researcher and Lecturer at the Institute of Social Ethics, University of Lucerne, Switzerland. Her expertise lies at the intersection of ethics, technology, AI, peace, and human rights. She serves as Vice-Chair of the Panel on AI & Peacebuilding of the International Panel on the Information Environment (IPIE), Research Associate at the Centre for Technology and Global Affairs (CTGA) at the University of Oxford, and Senior Fellow at the Centre for International Governance Innovation (CIGI). Evelyne is also a Board Member of the Peacenexus Foundation and a former member of the UN Internet Governance Forum's Multistakeholder Advisory Group (MAG). She holds a PhD in International Relations, with a specialisation in Political Science, from the Graduate Institute of International and Development Studies (IHEID) in Geneva.

Photography

Visual representation of the concept of artificial intelligence and digital connection.

Author: Pixel-Shot (Shutterstock).

IN DEPTH

Bridging Algorithms as Practical Tools for Depolarisation

Luke Thorburn

Research fellow at the AI & Democracy Foundation

It is widely recognised that by optimising for “engagement,” social media algorithms tend to disproportionately surface the more extreme, inflammatory voices^[1], raising the emotional temperature of online conversations and contributing to political polarisation^[2]. But could we design ranking algorithms that instead help build common ground? This is the motivation behind an emerging area of research and practice around “bridging algorithms” and “bridging-based ranking.” In this piece, I give a concise introduction to bridging algorithms, orient you to where and how they have been used so far by both civil society organisations and social media platforms, and summarise their current limitations and challenges.

The term bridging-based ranking can describe any method for ranking alternatives – be they social media posts, policy proposals, survey responses, or candidates for elected office – that helps build mutual understanding and trust across divides^[3]. This qualitative goal can be operationalised in many different ways. I will explain the most common approaches, but to help ground this, imagine that you are convening – in whatever community or region you are most familiar with – an online process in which members of groups that have been in conflict with one another submit written proposals for what should happen next. As the facilitator of the process, you need to determine which of the submissions – of which there could be thousands – to make most visible when you report the results back to the community, and you want to do this in a way that is bridging, that is, helps build mutual understanding and trust.

“ Social media algorithms tend to disproportionately surface the more extreme, inflammatory voices. Could we design ranking algorithms that instead help build common ground? ”

The most common approach is to identify proposals that represent some form of common ground or “diverse approval,” that is, those approved by people who usually disagree with each other. If you have some vote data from people in the community about whether or not they support each proposal, then it’s possible to quantify the level of support for each proposal within each group, and then identify those proposals which have a high level of support in multiple groups. Finding such proposals and making them visible has at least two important effects. First, it demonstrates that there is, in fact, common ground —often quite a lot of it —which is not widely accepted in conflict settings. Second, I believe that directing attention to ideas at the intersections of different groups, even if such ideas are not universally supported, helps reduce sorting [4]— that is, the homogeneity of opinion clusters, which is a known risk factor for conflict escalation. Here are some examples:

This approach has been used recently by the Alliance for Middle East Peace (ALLMEP) in a sequence of online dialogues^[5] to help build common ground between Israeli and Palestinian peacebuilders, which resulted in a set of five demands for world leaders and the international community that all had over 90% support on each side.

It has been implemented at scale in Community Notes, a feature on social media platform X (formerly Twitter), where people can propose notes that add important context to tweets that they believe to be misleading, but where these notes are only displayed publicly if an algorithm determines they are likely to be rated as helpful by people on both ends of the political spectrum. This algorithm is open source, and the feature is also being trialled on Meta’s platforms (Facebook, Instagram, Threads), TikTok, and YouTube.

Polis, an open-source tool that facilitates online collective dialogue and has been widely used in civil society and peacebuilding contexts, makes this ranking very accessible. In the standard report from a Polis conversation, scroll down to the “All Statements” section and select to sort the statements by “group-informed consensus.”

“ The term bridging-based ranking can describe any method for ranking alternatives – be they social media posts or policy proposals – that helps build mutual understanding and trust across divides ”

A second approach is to reflect a set of proposals that is representative of the community’s opinions, by ensuring that diverse proposals are included and that there is sufficient context about the level of support for those proposals and where that support is coming from. Such representation and social context are important because in polarised contexts, there are often significant perception gaps^[6] – we tend to think the other side holds more extreme or intolerable views than they actually do. The social context needed to debunk these false perceptions is also usually missing from social media^[7], which both acts as a “prism”^[8] by disproportionately surfacing the most inflammatory perspectives. It only provides raw counts of likes, shares, etc., with no information about the degree to which a post resonates with the population as a whole, or in which sub-communities that feeling is strongest. Because of this, it can be crucial for processes run by civil society to help provide such a social context.

This approach is also implemented in Polis, the tool mentioned above, which provides, in its standard report, both a visualisation of the most prominent opinion clusters and the extent to which each cluster agrees with each statement.

It is also one of the core rationales behind the use of more traditional public opinion polling in peace processes^[9], such as those conducted by Colin Irwin^[10] in Northern

Ireland, Kashmir, the Balkans, Sri Lanka, and elsewhere, which aim — in part — to reduce help people realize that many more people share their views than they realize, that is, to reduce “pluralistic ignorance”.

“ Bridging algorithms are just one of a set of features often present in an emerging ecosystem of online “deliberative tools” that can facilitate deliberation in polarised contexts ”

A third class of approaches uses automated classifiers to identify and uprank proposals that exhibit hallmarks of being written in good faith. Such classifiers are usually more “black box” than the above two approaches, meaning that it is not always possible to know why a given text was scored a certain way by the classifier. Because of this, the use of such classifiers might be less appropriate in processes where transparency and procedural fairness are critical. But they can be helpful in more informal online fora—for example, prompting people to reflect on how their contributions might land with others and empowering them to filter for the kinds of contributions they consider valuable.

For example, Google Jigsaw has built automated classifiers, free to use, that can score text comments in an online forum based on the degree to which they exhibit “bridging attributes” like compassion, curiosity, nuance, or respect.

Challenges and limitations

Each of these approaches has limitations, and the design of bridging algorithms remains an active research area. Three challenges, in particular, deserve attention.

First, ranking items of content is inherently zero-sum, and optimising for anything — be that common ground, representativeness, nuance, or anything else — will mean that you face trade-offs with other goals you have as a facilitator of the process. As an example, consider the goal of common ground. It’s much easier for people to agree on

vague platitudes about creating a more harmonious future than it is for them to decide on, say, precisely worded clauses that could go into a substantive peace agreement. And many peacebuilders will be aware of the diplomatic “move” in which, to allow dialogue to continue despite an irresolvable disagreement, one progressively makes the language of agreement more abstract or procedural. Resulting, for example, not in an agreement about how benefits and burdens will be distributed, but in an agreement merely that they should be distributed “fairly”, or that a particular process will be followed for deliberating on how they will be distributed.

While sometimes the only way forward is possible, such vagueness kicks the can down the road – the “can” being the disagreement which will inevitably have to be resolved – by leaving ambiguous the language – and thus the substance – of what is agreed upon. In this way, the goal of maximising common ground is in tension with the goal of minimising ambiguity, and so ranking by diverse approval can result in relatively vague or ambiguous statements rising to the top. You can get a sense of this in the results from the ALLMEP process mentioned earlier, where some of the most bridging “values” and “visions” for the future have this flavour. I am currently collaborating on an ongoing project that aims to develop a method for identifying statements that are both bridging and unambiguous.

“ To allow dialogue to continue, one progressively makes the language of agreement more abstract or procedural; the goal of maximising common ground is in tension with the goal of minimising ambiguity ”

Second, just because you can use an algorithm to recognise common ground or nuance, that doesn’t guarantee there will be much of it to find. The most successful applications of these methods have been in contexts where substantial effort is put into creating conditions in which people engage in good faith and contribute ideas they believe have the potential to resonate broadly. A widely acknowledged limitation of

Community Notes,^[11] for example, is that only a relatively small percentage – under 10% – of all the notes written are found to be bridging and are displayed publicly. Its effectiveness as a check on falsehoods in the public sphere is thus significantly limited by the thoughtfulness of the note contributors. Still, it is challenging to build such a culture at the scale of a modern social media platform. In contrast, when bridging algorithms are used in processes with additional onboarding and trust-building, with time for deliberation, and in communities small enough for people to have a shared sense of responsibility – for example because they live in the same place and recognise that they have to build some form of everyday life together – they will likely be more successful because there will be more bridging ideas to be found.

Third, it is essential to acknowledge that the accessibility of bridging-based ranking as a tool for facilitators and peacebuilders is not yet where it should be. Perhaps the most accessible implementation is the open-source tool Polis. If you are comfortable using a spreadsheet, it is also straightforward to use bridging-based ranking to curate the results of processes run on Remesh – the tool used by ALLMEP – or of votes collected using any standard survey tool. Bridging algorithms are just one of a set of features often present in an emerging ecosystem of online “deliberative tools” that can facilitate deliberation in polarised contexts. Financial support is available to access these tools if needed^[12].

“ The impact of bridging algorithms is likely to come from their strategic use by peacebuilding and civil society organisations in polarised contexts ”

Where to from here

Bridging algorithms remain an active area of research^[13]. In the social media context, a large academic study recently tested several bridging algorithms with roughly 6,000 participants for 4 months across multiple platforms (Facebook, X, and Reddit), and we

will know soon whether it was successful in reducing affective polarisation. A smaller recent study^[14] testing a form of bridging algorithm on X has already shown that this is possible. But while much of the focus has been on social media, the impact of bridging algorithms is – at least in the short term – likely to come from their strategic use by peacebuilding and civil society organisations in polarised contexts at a more local or regional scale.

*If you are interested in using these methods but are not sure where to begin, please feel free to reach out to the author or to the Deliberative Tech hub of the [Council on Technology and Social Cohesion](#), an emerging community of practice for civil society organisations using these tools.

[1] Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao & Anca D. Dragan. March 2025. “Engagement, user satisfaction, and the amplification of divisive content on social media”. In PNAS Nexus, Volume 4, Issue 3.

[2] Tiziano Piccardi, Martin Saveski, Chenyan Jia, Jeffrey T. Hancock, Jeanne L. Tsai, & Michael Bernstein. 2024. “[Social Media Algorithms Can Shape Affective Polarization via Exposure to Antidemocratic Attitudes and Partisan Animosity](#)”. Preprint.

[3] Aviv Ovadya & Luke Thorburn. 2023. “[Bridging Systems: Open problems for countering destructive divisiveness across ranking, recommenders, and governance](#)”, ArXiv.

[4] Luke Thorburn, Maria Polukarov & Carmine Ventre. 2024. “Societal Sorting as a Systemic Risk of Recommenders”. In 18th ACM Conference on Recommender Systems (RecSys ’24), New York: Association for Computing Machinery, 2024, p. 951–956.

[5] Andrew Konya, Luke Thorburn, Wasim Almasri, Oded Adomi Leshem, Ariel Procaccia, Lisa Schirch, & Michiel Bakker. 2025. “[Using collective dialogues and AI to find common ground between Israeli and Palestinian peacebuilders](#)”. In The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’25), New York: Association for Computing Machinery, 2025, p. 312–333.

[6] Jonathan Stray. 2025. “[Why Do They Think We’re Extreme?](#)”. In *Better Conflict Bulletin*.

- [7] E. Glen Weyl, Luke Thorburn, Emillie de Keulenaar, Jacob Mchangama, Divya Siddarth & Audrey Tang. 2025. “Prosocial Media.” ArXiv.
- [8] Chris Bail. 2021. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Prince University Press.
- [9] Colin Irwin. 2025. “The People’s Peace Third AI and UN Edition. Public Opinion, Public Diplomacy and World Peace”. *Peace Polls Publication*.
- [10] See the [Peace Polls](#) site.
- [11] Madison Czopek. 2023. “Why Community Notes mostly fails to combat misinformation”. In *Poynter*.
- [12] Specifically, the AI & Democracy Foundation makes deliberative tools freely available to try under its [Deliberative Tools Access Program](#). If this is of interest, please do feel free to fill out the linked form.
- [13] Madon Revel, Smitha Milli, Tyler Lu, Jamelle Watson-Daniels & Max Nivkel. 2025. “[Representative Ranking for Deliberation in the Public Sphere](#)”, ArXiv.
- [14] Tiziano Piccardi, Martin Saaveski, Chenyan Jia, Jeffrey T. Hancock, Jeanne L. Tsai & Michael Bernstein. 2024. “[Social Media Algorithms Can Shape Affective Polarization via Exposure to Antidemocratic Attitudes and Partisan Animosity](#)”, ArXiv.

About the author

Luke Thorburn

Research fellow at the AI & Democracy Foundation and a final year PhD student in computer science at King’s College London. His research currently focuses on how social media recommender systems can be designed to mitigate conflict risks, an idea sometimes called “bridging-based ranking”. He coauthors the Understanding

Recommenders project of the Center for Human-Compatible AI at UC Berkeley, is a member of the GETTING-Plurality Research Network in the Allen Lab for Democracy Renovation at Harvard University, and a research affiliate of the Machine Intelligence and Normative Theory Lab at Australian National University. Previously, Luke has worked with or advised Ofcom, the newDemocracy Foundation, and Meta. His background is in probability and statistics.

Photography

Symbolic representation of algorithms and their role in polarization in the digital sphere. Autor: Achira22 (Shutterstock).

IN DEPTH

Mediation in the Age of Algorithms: Risks and Opportunities for Peace Processes

Sanjana Hattotuwa

PhD in Politics and Social Media (University of Otago)

“Online mediation has great potential to assist in resolving disputes. In the Asia Pacific region, it has a particular place in enabling access to justice for large populations who have little access to dispute resolution by other means.”^[1] This observation, from two decades ago, remains profoundly relevant. My argument then was that a Western-centric, technology-based conflict-resolution paradigm was inadequate for Asian contexts. Long before the ubiquity of social media and smartphones, our focus was on developing online conflict-resolution architectures that leveraged nascent technologies such as mobile telephony and community internet radio.

A prime example of this was in 2002, as an integral part of the official ceasefire negotiations in Sri Lanka. Here, a specially adapted commercial-off-the-shelf (COTS) software, Groove Virtual Office, was used in an unprecedented manner to support active mediation processes anchored to a one-text process^[2]. This platform, which I co-architected and led the development of tools to backstop the mediations, enabled encrypted, asynchronous, and multilingual communication, creating searchable repositories of documentation, decision-support tools, and multi-stakeholder position mapping. Crucially, it was designed to integrate input not just from the primary negotiating parties, but also from the diaspora and those involved in Track 2 and Track 3 of the peace processes,^[3] demonstrating a long-standing practice of using ICTs for complex mediation far from the Global North.

In 2025, two decades after the One-Text process in Sri Lanka, the role and relevance of technology in peacebuilding is far better established. However, the intervening years reveal a fundamental challenge for mediators. The very platforms that offer unprecedented opportunities for dialogue also serve as vectors for spoiler dynamics that can derail fragile peace talks. This dilemma is acute because these new vectors of information production lie entirely beyond the remit of the Chatham House rules and the sandboxing required to incubate trust. Mediators are now bombarded with information, and their carefully managed processes face a constant threat from online campaigns of mis- and disinformation designed to inflame hate, spread incendiary falsehoods, and erode public confidence in a potential agreement. This 'emotional contagion' effect, where online sentiment directly impacts how people feel offline, means that the security of a closed-door negotiation room is no longer sufficient to protect the integrity of the process. What, if any, frameworks can reconcile technology's capacity to both ruin and rebuild trust within the specific context of a mediated negotiation?

“ The very platforms that offer unprecedented opportunities for dialogue also serve as vectors for spoiler dynamics that can derail fragile peace talks ”

Immediately evident is the dark(er) side of technology and social media, which poses a direct and increasing threat not just to social cohesion writ large but to the very mechanics of peace mediation. Algorithmic amplification of inflammatory content on social media can harden the positions of negotiating parties and their constituencies, making compromise, which is the cornerstone of mediation, politically untenable. The spread of disinformation can be weaponised to undermine the credibility of mediators, derail specific talks, or violate the confidential 'sandboxing' essential for fragile negotiations. For a mediator, this creates an unprecedented challenge: key aspects of the conflict environment are being shaped in real time by forces entirely beyond the remit of the Chatham House rules or established codes of conduct. Spoiler dynamics

are no longer confined to physical acts of violence but manifest as viral campaigns of hate and falsehood that can unravel delicate progress achieved at the negotiating table.

Conversely, and less often highlighted, is the transformative potential of technology in mediation is most profound when it moves beyond elite processes and empowers grassroots communities. As I envisioned in 2006, the future of Online Dispute Resolution (ODR) in the Global South lay not in replicating PC-based systems, but in leveraging technologies already in people's hands. The explosive growth of mobile telephony presented an opportunity to design ODR systems with a 'human face'. Concrete applications for community-based mediation included: using SMS to send vernacular notifications of settlements to disputants; enabling in-field mediators to gather audio and video testimonies via their mobile phones; creating expert systems that could provide mediators with real-time options for resolving common disputes, such as those over land; and using mobile video-conferencing to connect parties in remote areas with Alternatives Dispute Resolution (ADR) centres. This approach takes mediation to the paddy fields, the post office, and the village chieftain's residence, transforming it from a centralised, inaccessible process into a pervasive, user-friendly, and culturally resonant service for nonviolent dispute resolution.

“ Spoiler dynamics are no longer confined to physical acts of violence but manifest as viral campaigns of hate and falsehood that can unravel delicate progress achieved at the negotiating table ”

This dualism reflects the work by Prof Miriyam Aouragh, who likened social media to Damocles' sword, and argued “that those who are empowered by taking the seat under the sword do so haunted by the constant threat of being hurt by the same sword, because slaughter could come at the slightest disruption”^[4]. As Prof Admire Mare avers, Aouragh's central thesis is that “social media is an open-ended technology without closure, which allows state and non-state actors to harness it for good and

nefarious purposes. In other words, though it is rare to find it acknowledged in policy debates or even in academic literature from the Global North, social media is simultaneously good and bad, helpful and harmful, conciliatory and divisive, peaceful and violent.

This dual nature of digital tools manifests directly in mediation processes themselves. WhatsApp, for instance, has revolutionised how mediators engage with conflict parties – the UN Office of the Special Envoy for Yemen uses WhatsApp groups to maintain real-time communication with negotiators[5], whilst simultaneously grappling with how the same platform spreads disinformation about the peace process. In Libya, UNSMIL established “Trusted Partner” relationships with Facebook to remove harmful content targeting members of the Libyan Political Dialogue Forum, particularly women participants, whilst using the same platform to promote peace narratives[6]. This paradox – where the same app, technology, platform, product or tool that enables inclusive consultation can weaponise spoiler dynamics – requires mediators to develop sophisticated digital strategies that leverage opportunities whilst mitigating risks.

Beyond traditional web-based platforms, innovative digital mediation approaches have emerged across diverse contexts. Build Up’s 2021 WhatsApp consultations reached 93 Yemeni women across 11 governorates, creating dialogue spaces where physical meetings were impossible[7]. The Civil Society Support Room for Syria established an interactive website that enables Syrian civil society actors to submit input directly to the UN Special Envoy, bridging the gap between Track 1 negotiations and grassroots perspectives[8]. These initiatives reveal how digital tools, when designed with specific mediation objectives, can overcome traditional barriers of geography, gender, security, and access.

“ Mediators need to develop sophisticated digital strategies that leverage opportunities of technologies whilst mitigating their risks ”

The evolution of online dialogue extends beyond formal peace processes into emergent digital societies. Writing on the future of Online Dispute Resolution (ODR)^[9], I pointed to the complex social and commercial transactions occurring in virtual worlds like ‘Second Life’. These environments, with their own economies, property rights, and social norms, were already generating novel disputes that spilt over into the real world, including the first murder induced by a conflict over a virtual artefact. This raised a crucial question for the future of mediation: do we need ODR systems explicitly designed for disputes that arise and exist entirely within virtual domains? The ‘metaverse’ of today is grappling with these same challenges of governance, harm, and resolution. This demonstrates that the need for innovative mediation frameworks, across varied technology products and platform surfaces, is constantly expanding, requiring us to bridge not only the physical world but also the increasingly complex, fluid, and dynamic interplay between online, virtual, and real-world domains where communities form and conflicts arise.

Today’s mediators increasingly recognise that effective digital mediation requires meeting parties where they communicate – whether through encrypted messaging apps, social media platforms, or mobile-first interfaces – rather than expecting universal access to traditional web platforms. And yet, fundamental platform reforms are essential to support this verdant potential. Algorithms must prioritise constructive dialogue over engagement metrics, with circuit-breakers that slow the viral spread of incendiary and false commentaries during heightened, sudden-onset crisis periods or in contexts defined by democratic deficits or democratic backsliding^[10]. Resource allocation must achieve parity between Global North and South operations, hiring local language speakers and cultural experts while investing in AI systems designed for non-English and non-Western contexts.

Community-led and community-level interventions require sustained investment and capacity building. Traditional peacebuilders need training in adopting and adapting digital platforms in grounded, gendered, sustainable, accessible, and equitable ways. At the same time, the constellation of domestic policy, regulatory, media, and legal frameworks around information integrity must also be strengthened. In fact, I would argue that digital media literacy should be made a public education priority, from

children to adults. Women's groups and youth organisations that use social media for peace require dedicated support and resources, recognising their roles as both vulnerable populations and innovative peace agents.

“ Algorithms must prioritise constructive dialogue over engagement metrics, with circuit-breakers that slow the viral spread of incendiary and false commentaries ”

Social media's impact on peacebuilding in Global Majority contexts defies simplistic, binary categorisations of it as either beneficial or harmful^[11]. As Prof Mare argues, the platforms represent “open-ended technology without closure” that requires a sophisticated understanding of local contexts, power dynamics, and sociotechnical relationships^[12]. My own doctoral research in 2021 established how, in Sri Lanka, social media “simultaneously contributed to authoritarian entrenchment as well as resistance to democratic erosion”, how “different user motivations on Facebook and Twitter simultaneously supported prosocial and violent frames during moments featuring significant offline unrest”, and how the likes of Facebook, and Twitter “simultaneously amplified hate as well as produced prosocial, nonviolent and conciliatory content that called for civility, upheld democratic institutions and celebrated diversity”.

The path forward for mediation in the digital age requires a fundamental rethinking of the online environment in which peace processes unfold. Platform business models that profit from a “war of stories” directly undermine the mediator's core task of de-escalation by amplifying divisive narratives and accelerating epistemic decay^[13]. The solution lies not in tweaking harmful platforms, but in designing bespoke digital mediation architectures grounded in genuine partnership with local civil society. This ensures that those with situated experience in conflict transformation are central to creating tools that are not just technically functional but contextually grounded. For a mediator, the recognition that code is not neutral is paramount; it reflects inherent

biases that can either support the delicate balance of a negotiation or sabotage it entirely.

For the contemporary mediator, the persuasive, projected, and perceived authority of sophisticated AI tools must be met with profound professional scepticism. Their role is not to supplant the mediator's craft, but to augment their capacity. As demonstrated by the One-Text platform during the Sri Lankan peace process decades ago, the true value of technology lies, aside from its endogenous development, in its ability to help mediators manage vast amounts of information, facilitate structured dialogue across distances, and model complex scenarios to enhance human judgment. We must constantly question the biases in AI's data sources and flawed analyses, remembering that no algorithm can replicate the essential, human-led tasks of building rapport, demonstrating empathy, and making critical, nuanced judgements at the negotiating table.

“ We must question the biases in AI's data sources, remembering that no algorithm can replicate the essential, human-led tasks of building rapport, demonstrating empathy, and making critical, nuanced judgments at the negotiating table ”

Ultimately, transforming digital platforms from arenas of conflict into practical tools for mediation demands that we centre the wisdom of those who negotiate peace daily in the world's most fractured places. It is local communities and peace practitioners, including those from First Nations, Adivasi, and indigenous communities, who understand the granularities of trust-building and narrative management essential to any successful mediation. Their insights, experience, and knowledge must inform the design of a new generation of digital mediation environments. This requires abandoning Silicon Valley's extractive logics, rapacious platforms, and mercenary endeavours, and instead building online spaces architected to support dialogue,

confidentiality, and consensus-building, which are the very pillars of the mediation process that current algorithms are incentivised to destroy.

To wit, the modern mediator's task is no longer confined to a physical negotiating table, but extends to online content, commentary, and currents that feed conflict. Success, therefore, will not be found in only skilfully navigating today's adversarial digital commons, but in actively reshaping them. The craft of mediation, with its emphasis on de-escalation, confidentiality, and measured dialogue, must now inform the code's very logic. This is the next frontier challenge: not adapting peace processes to the whims of the algorithm but bending the algorithm to the enduring principles of just peace.

[1] Hattotuwa, S., & M. C. Tyler. 2005. *An Asian Perspective on Online Mediation*. *Asian Journal on Mediation* 1 (1): 1-24. U of Melbourne Legal Studies Research Paper No. 158.

[2] ICT4Peace Foundation. 2013. *The Janus Effect: Social Media in Peace Mediation*. Zürich: ICT4Peace Foundation.

[3] Women's Peace and Humanitarian Fund. Definitions of peace process, track 1 and track 2, and implementation of a peace agreement.

[4] Mare, A. 2024. *Social Media, Conflict, and Peacebuilding in Southern Africa: A Primer*. Kujenga Amani, Social Science Research Council. December 17, 2024.

[5] UNITAR. 2021. *'WhatsApp Diplomacy': The Future of Multilateralism in a Post-COVID-19 World?* United Nations Institute for Training and Research. May 19, 2021.

[6] Sustaining Peace Select. 2023. *Platform Engagement*.

[7] Build Up. 2023. *Feminist Approaches to Online Consultations and What They Reveal*. Blog by Build Up. May 18, 2023.

[8] Office of the Special Envoy of the Secretary-General for Syria (OSE-Syria). 2016. *Civil Society Support Room*. September 26, 2025.

[9] Hattotuwa, S. 2006. *The Future of ODR: One Brief Glimpse*. ICT for Peacebuilding (blog). February 22, 2006.

[10] Bunse, S. 2021. *Social Media: A Tool for Peace or Conflict?* SIPRI Commentary. August 20, 2021.

[11] Reuss, A., i S. Stetter (eds.). 2023. *Social Media and Peacebuilding: How Digital Spaces Shape Conflict and Peace*. Palgrave Mcmillan.

[12] Ibid

[13] Hattotuwa, S. 2024. *A 'War of Stories': Humanitarianism in the Disinformation Age*. ICT for Peacebuilding (blog). December 23, 2024.

About the author

Sanjana Hattotuwa

He did his doctoral research on the intersection of social media, political communication, propaganda, and information disorders in Sri Lanka, as well as how New Zealand's Christchurch massacre in March 2019 was represented on Twitter. Specialising in and advising on social media communications strategy, digital security for journalists and human rights defenders, social media activism, online advocacy and grounded, mixed-methods social media research, Hattotuwa's experience in studying, negotiating, and developing policies against information disorders spans two-decades, and work in South Asia, Southeast Asia, North Africa, the United States, Europe and the Balkans.

He founded in 2006 and till June 2020 curated the award-winning Groundviews, Sri Lanka's first civic media website. From 2021 to 2024 he was the Director of Research at The Disinformation Project, based in New Zealand. He remains a Special Advisor at the ICT4Peace Foundation, based in Switzerland.

Photography

Symbolic representation of mediation in a digital and algorithmic context. Author:
Anbu-Creations (Shutterstock).

IN DEPTH

Digital Democracy and Cyberactivism in East Africa: The Role of Siasa Place

Nerima Wako

Executive Director of Siasa Place

In 2015, armed with nothing but passion, borrowed WIFI from a city café that double up into a bar in the evening in the middle of downtown, and a hunch that Twitter might matter someday, we launched Siasa Place, a youth-led civic-tech organization based in Nairobi, Kenya. I say we, a group of young ambitious people who somehow connected on democracy, we met at various functions around the city, and we were drawn to each other by our sense of general knowledge on country and current affairs and this passion to change the way things were. I had just completed my studies abroad, so when I moved back home after being away for 7 years, I was happy to find this new community, one that I connected with intrinsically and immediately. We were patriotic, and we didn't have money, just a sense of urgency. The civic space was shrinking, the air thick with disillusionment, and young people were hungry for a way to speak, to be heard, to organize. Ten years later, that bet on digital democracy has put us on the frontlines of East Africa's cyberactivity movement.

Back then, it felt like shouting into the void. We ran a weekly conversation called #SiasaWednesday, where we debated budget allocations, corruption scandals, leadership and youth engagement in governance. We had a specific hour that we would meet virtually and sometimes it felt like we were talking to ourselves. Every week without fail, for up to 3 years, we met at the same time, discussing various issues and mobilizing others to join us. And slowly, people joined. Threads grew longer. Followers multiplied. Can you imagine we would type conversations, sometimes stay on a subject for as long as three hours. This was before spaces, this was when we would have

threads of conversations, sometimes you have no idea who was reading or questions would come in halfway, while some questions came in days later after someone coincidentally stumbled on the conversation. So, back then, you would have to type a response. These days, we really take X spaces for granted, where thousands can join live discussions, speaking, listening, from the simple unmuting of a button and challenging power in real time. Or even TikTok live videos, where people can comment in real-time.

“ The internet hasn’t just become the medium, it is the movement. Hashtags are petitions; influencers are political analysts; memes are protest signs ”

Today, memes are protest signs, comical but with strong political stances and messaging. A clever and different political language that young people love to interact with. Hashtags are petitions. Influencers are political analysts. The internet didn’t just become the medium; it became the movement. But even as our voices amplified, so did the backlash.

The Double-Edged Sword of Digital Empowerment

Over the past decade, digital platforms have grown into the primary arena for civic engagement across East Africa. From Nairobi to Kampala, Dar es Salaam to Kigali, young people have built powerful communities online, pushing for transparency, accountability, and justice.

Siasa Place exists to strengthen that movement. We develop institutional frameworks, offer civic education, and train the next generation of leaders on how to use tech tools for political engagement. Initiatives like “Politics, Tech, and Rights” address the nuances of digital labor, content moderation, and gender-based violence in online spaces. We also support platforms such as zKE that collect opinions from mainly young people on bills, asking for suggestions or recommendations to bills in a more

streamlined and effective way, basically gathering public opinion which is part of our constitution when it comes to public participation. The way feedback is submitted will change greatly, especially as younger generations continue to get accustomed and cultured to utilizing digital platforms.

But the very tools we use for organizing are now being weaponized against us. In Kenya, Parliament has recently been working on passing a controversial social media bill that would grant authorities access to citizens' devices. Under the guise of regulation, it paves the way for surveillance and repression. In Uganda and Tanzania, entire platforms have been shut down during elections. Ethiopia experienced a full internet blackout in 2023. Digital democracy is not just uneven across the region—it's under siege.

“ Digital platforms have grown into the primary arena for civic engagement across East Africa. Young people have built powerful communities online, pushing for transparency, accountability, and justice ”

What Digital Democracy Looks Like in East Africa

To speak of digital democracy in East Africa is to acknowledge contradiction. Connectivity is uneven. Access is costly. Kenya has high speed internet, connectivity is struggling even though we have a growing reach, it is still difficult to access the internet in very rural areas. When we speak in terms of affordability, data bundles are still costly to the average person. Even owning a smart phone is expensive. What ends up happening is that if a household can afford to get a smart phone device, it often belongs to the male of the household. That also translates to users being male dominated, most online users will be more male than female. Online violence is already high, but tech assisted violence against women is even greater. When you are already a minority online, then the chances of you being harassed online are high – then many reduce their engagement on issues – especially political matters that tend to be quite

emotive.

There is also the fact that Kenya is one of the most engaged countries in the world, top 5 actually when it comes to how many hours citizens spend online. However, there is a contradiction: in countries where mainstream media is dominated by politically affiliated owners, the internet often becomes the only unfiltered space for dissent. In fact, we have seen a growth of alternative media, from podcasters to YouTubers. However, even here, we can't rely solely on the online world.

“ Digital democracy means creating alternative systems of trust and accountability, because too often institutions fail us ”

Digital democracy means pairing Facebook live sessions and tweets with town halls. It means understanding that just because something is legal doesn't mean it's just. It means creating alternative systems of trust and accountability, because too often institutions fail us. Who would have thought that we would witness impersonation of militaries, such as Iran, for example, who posted that they were attacking Israel as defense on X platform. And in a matter of hours, it had over 100,000 views. It has made mysterious institutions that often are protected by veils of authority almost feel reachable and sometimes more human facing. The law can be bent to serve repression; the internet must not be.

Stories of Resistance and Risk

From Siasa Place we've supported campaigns demanding the release of political prisoners in countries like Mozambique. We've worked with Africiviste to highlight repression in Senegal. Locally, we've seen how digital tools allow young people to respond fast mobilizing against injustices like police brutality or public graft. We have had conversations with similar organizations like Yiaga in Nigeria, comparing our elections and getting an understanding firsthand on processes, speaking to each other and collaborating and building reports that can be useful for future recommendations.

But we've also seen the risks. Government-sponsored trolls spread disinformation to discredit activists. More and more, this technique has become ever so visible and aggressive. Dismissing activists as being commercial and Western-sponsored to bring about anarchy and to destroy the countries that we live in. A contradiction of sorts, because the same western countries or eastern funds most of our government budgets, including essential pillars such as police service, and health. Governments receive the most amount of support figuratively.

“ Digital tools allow young people to respond fast mobilizing against police brutality or public graft. But we've also seen the risks of internet: disinformation, surveillance and intimidation ”

There has been a global rise in State agencies monitoring phones and infiltrating WhatsApp groups. But this is a great fear for us because it is literally the difference between life and death. People have been abducted and even murdered because of a post that they put. This has become normalized that even abductions no longer seem unlawful. It has become a weekly event and even elected members that serve in Assemblies are being abducted, not just activists, but ordinary citizens who choose to be vocal.

In early June, a brave teacher called Alfred Ojwang had put up a post about corruption in the police sector. He was picked up by the police and driven over 300km from his home (passing several police stations) for questioning. He was taken to the Central Police Station in downtown Nairobi. A few hours later, he died in custody. The police tried to frame it as though he committed suicide in the cell. But due to pressure from the public and the fact that activists refused to leave his side, the autopsy concluded that Alfred was tortured and beaten to death. Alfred was tracked on his device, and this was because of a post that he made that “tarnished” someone's record. His story haunts us because it reminds us that we are not only fighting narratives; we are fighting for our lives.

Still, what keeps us going is belief. Belief that another Kenya, another East Africa, is possible. One where voices are not silenced but amplified. One where justice is not just spoken about in encrypted chats but practiced.

The Boniface Mwangi and Agather Atuhaire Case: A Flashpoint

In May 2025, human rights activists Boniface Mwangi (Kenya) and Agather Atuhaire (Uganda) travelled to Dar es Salaam (Tanzania) to support opposition leader Tundu Lissu during a court appearance. They were abducted by Tanzanian security officers, blindfolded, stripped, assaulted, and dumped at their respective borders. The brutality was chilling. There was clear cross-border coordination and now we worry that we are not just being tracked in our countries but in our community, in the region.

What came after was equally significant: the digital roar. Activists were forced to come together cross-border and amplify – across East Africa they rallied, flooding timelines with calls for justice. Siasa Place, among others, issued a 72-hour ultimatum to regional bodies, demanding action. Hashtags trended between nations. Partnering with local and international media, which picked up the story. For once, silence was not an option.

“ Cyberactivism has demonstrated that borders are often insignificant in the face of repression. With a viral video and a coordinated outrage we can make noise, and noise can lead to justice ”

This case showed both the peril and the power of digital activism. Had there been no Twitter thread, no viral video, no coordinated outrage, perhaps their story would have ended in silence. But we made noise. And noise can lead to justice; we are still on that path to justice.

The Role of Regional Solidarity

Cyberactivism has demonstrated that borders are often insignificant in the face of repression. Whether in Nairobi, Kampala, or Dar es Salaam, the playbook is familiar: surveillance, disinformation, intimidation.

That's why regional solidarity matters. We don't just need allies, we need systems. Systems that are constantly adapting because they often find ways to try and shut down platforms or the internet itself. We need protocols and mutual protection mechanisms. When Mwangi and Atuhairi were taken, we didn't wait for governments to act. Concerned citizens and civil society filled the vacuum.

Building Resilience: The Work of Siasa Place

Siasa Place's work now spans training, research, and advocacy. We advocate for rights that are already provided for in our constitution. Gig workers deserve contracts, protections, and dignity. We're mapping out the tech ecosystem to identify where rights are being violated and where interventions are possible. Through TrustLab, we train grassroots organisations on digital security. We teach the use of VPNs, two-factor authentication, phishing awareness, and fact-checking strategies. We help communities build digital hygiene the way past generations built unions or savings groups. We're also investing in storytelling. Because if we don't tell our stories, someone else will—and they will distort them.

We work with community-based organizations, mentoring them to institutionalize and also become leaders in their own communities and learning about budget cycles and the importance of social accountability. The constitution grants citizens the right to participate in that process. They need to learn how to hold local leaders accountable especially utilizing already provided spaces and platforms to do so.

**“ A digitally peaceful society is one where
activists don't need burner phones; where laws
protect speech and where we can disagree loudly,
passionately, safely ”**

We work with young leaders who have an interest in pursuing political leadership. Partnering them with political parties and also encouraging them to be members of parties and play an active role in politics. Our main work is educational awareness, because when these processes become known, young people are often ready to take them up.

What Peace Looks Like in a Digital Age

Peace is no longer just about ceasefires or election cycles. It's about data protection. It's about the right to post without fear. It's about whether your activism gets you a retweet or a jail sentence.

A digitally peaceful society is one where activists don't need burner phones. Where laws protect speech, not criminalize it. Where we can disagree loudly, passionately, safely, and it's not just the responsibility of governments... platforms, donors, international allies, they all have a role to play. Because if one voice is silenced online, we all lose something.

This isn't just a moment. It's a movement. We are not coding apps; we are coding resistance. Every tweet is a vote. Every view is a vigil. Every WhatsApp message, every livestream, every meme, everything matters.

We know the risks. But we also know our power. As we build digital peace in East Africa, we do so knowing that silence was never an option. The internet didn't save us. We saved each other. And we're not done yet.

About the author

Nerima Wako

Nerima is the Executive Director of Siasa Place, a youth organization established in 2015, dedicated to collaboratively create an environment that enables youth of Kenya to directly engage with the political mainstream in a meaningful way. Siasa Place

educates youth on electoral processes, constitution, government institutions and functions. It specializes on civic-tech, digital political campaigns and advocacy.

Nerima attained her Bachelor of Arts in Journalism and Sociology from Jacksonville State University 2010 and her Master's in Public Administration in 2012. She has held notable positions including: Vice Chairperson of Youth Coordinating Committee; Member of the Kenya Eminent Peace Panel (2022); Member of Council for Responsible Social Media; Member of the UNDP Youth Sounding Board or Member of the Youth Advisory Committee to the Kingdom of Netherlands (2022-2023).

Photography

Moment of intervention during a gathering organized by Siasa Place. Author: Siasa Place (organization's Facebook page).

IN DEPTH

Documenting culture before, during and after conflict

Wahbi Abdalrahman

Coordinator of the Sudan Memory Project

Merowe locality nudges up against the Nile River in Sudan's Northern State, close to the hydroelectric dam and more than 300km north of the capital city, Khartoum. When Wahbi Abdalrahman and his team from Nile Valley University (NVU) arrive here with their Epson and Canon scanners, their laptop and digital cameras, the outside temperature is often pushing towards forty degrees Celsius. The area has changed in the last couple of years since the Rapid Support Forces (RSF) were repelled from Merowe soon after the start of Sudan's civil war in April 2023. The traditional Wednesday market is still bustling, but many of the goods being sold are now Egyptian, rather than Sudanese, and customers include huge numbers of displaced people who have escaped the conflict that rages elsewhere in the country.

Against this challenging backdrop, Wahbi's team sets up its travelling digitization unit. Since April 2023, funded by two six-month Cultural Emergency Response grants secured with Professor Marilyn Deegan, they have scanned more than 60,000 pages of documents in Sudan's Northern and River Nile States. Many of these documents are from family collections. Still, they constitute and record important events in Sudan's history that were either always absent from official narratives during Omar al-Bashir's dictatorship or are now being erased by the ongoing civil war.

Using their digital camera and audio-recording equipment, Wahbi's team has also captured video footage documenting accounts of abandoned markets, forgotten historical figures, and traditional practices in the region. He always travels with a

notebook. In this he records any stories or extra details told to him by the collection owners about the items being digitized because, over the years, his team has come to appreciate that each belonging is more than an image, it has a story. Preserving these stories and historical traces will be crucial for understanding and remembering this period.

“ Sudan has one of the richest, most diverse heritages in the world with 19 major ethnic groups. In just over a decade, Sudan Memory have digitised more than 400,000 items ”

Sudan Memory digitization project

Wahbi's work contributes to the broader Sudan Memory (SM) project initiated in 2013 to digitise cultural heritage at risk of decay and destruction in Sudan from climate change or possible future conflicts, the scale of which has far exceeded what was imagined at that time. NVU is one of several Sudanese and international partners contributing to the project, that also includes Sudan's National Records Office (NRO), The Sudanese Society for Archiving Knowledge (SUDAAK), the Women's Museum of Darfur, the Sudan Radio and Television Corporation (SRTC), University of Khartoum, University of Durham, and King's College London, amongst others.

Sudan has one of the richest, most diverse heritages in the world with 19 major ethnic groups that speak over one hundred languages and dialects. Its archaeological heritage reaches back several millennia, and the country has more than 200 pyramids, as well as being rich in funerary goods and remains, wall paintings, and artefacts, about which most of the world beyond Sudan is ignorant. In just over a decade—and despite a coup, regime change, and conflicts—SM partners have digitized more than 400,000 of these Sudanese cultural heritage items.

During the project, the partners uncovered unexpected riches, such as the original copy of a telegram sent by Zubair Pasha from Cairo in 1883, hidden inside a manuscript in a village north of Berber, and original letters handwritten by the Mahdi, all of which Wahbi and his team carefully digitised. The growing digital collection now comprises films, photographs, manuscripts, museum objects, audio files, oral histories, and even a 3D, interactive, historic model of Suakin Island, as well as all associated Arabic and English metadata. What all of these items attest to is a rich and resilient culture that is stronger for its long history of diversity and co-existence.

“ To erase any record of the diversity and cultural co-existence is a tried and tested weapon of war worldwide ”

Conflict and culturcide

Attempting to erase any record of such cultural co-existence is a tried and tested weapon of war worldwide. In recent years, we have witnessed the Taliban dynamite the Bamiyan Buddha statues, the Islamic State destroy the temples at Palmyra, and the deliberate burning of the Sarajevo library, an event that Bosnian theatre director Gradimir Gojer described as, “a triumph of barbarism and the death of the cohabitation of Muslims, Orthodox, Catholics and Jews that had existed for centuries in Bosnia-Herzegovina.” As spaces for “shared memories and identities” when such monuments, archives, museums, and libraries are erased in conflicts, they leave behind only what SOAS professor Dina Matar describes as memory “gaps” (Matar, 2023, cited in Khaled et al., 2023).

Since April 2023, we have seen these kinds of gaps emerge across Sudan, threatening to puncture its memory. In Khartoum, as elsewhere, RSF forces looted, damaged, or destroyed thousands of precious artefacts at the National Museum that documented Sudan’s long history and culture and the many civilisations that occupied this region over the centuries. UNESCO has warned of a “threat to culture”. At the same time, Ikhlas

Abdel Latif Ahmed, director of museums at Sudan's National Corporation for Antiquities and Museums (NCAM)—one of SM's founding partners—stated more emphatically that RSF forces “destroyed our identity, and our history” (Copnall, 2025).

“ The Sudan Memory collection is now all that remains of large parts of Sudanese culture; so many physical artefacts, collections, and buildings have been destroyed since 2023 ”

Reports and anecdotes from SM partners and acquaintances still in Sudan describe the destruction of the Sudan Radio and Television Corporation buildings, previously one of the most extensive film archives in Africa with radio, video and film recordings that date back to the 1940s and provide a unique historical resource for Sudan. We understand that partners in other parts of the country have certainly lost collections, and as the full extent of the conflict is revealed, we anticipate news of many more memory gaps.

In South Darfur state, in the town of Nyala, the collection of over 4000 items, which was collected and curated for the Darfur Women's Museum to tell the many different stories of the region, has been rescued and placed in a secret location. However, under current circumstances, it is not possible to display these items. As Kate Ashley, SM Consultant Project Manager, noted, this collection, “really demonstrated how just personal, everyday collections and objects can be put together and understood in this way that really sort of documents the social history of a place and has so much meaning.” Currently, only the digital collection that SM spent months photographing in 2021—along with a beautiful interview with the museum's founder, Fatima Mohamed Al Hassan, sadly now deceased—is all that is currently accessible, and only through the SM website.

Archival sanctuary

We know that the Sudan Memory collection is now all that remains of large parts of Sudanese culture since so many physical artefacts, collections, and buildings have been destroyed since 2023. The project has, of course, taken on a new urgency against this backdrop, both in terms of ensuring the long-term survival of the existing digital artefacts and facilitating the digitization of other heritage records that are threatened with loss.

That Wahbi and his NVU team can continue their work with very limited assistance from outside Sudan is a testament to the SM project's decentralised approach to digitization. In the early days of SM, project leaders Dr Badreldin Elhag Musa (SUDAAK) and Professor Marilyn Deegan (King's College London) envisaged conducting all digitization activities at a central hub in Khartoum's Africa City for Technology. In 2017, however, at the first project meeting after securing grants from the British Council and Aliph Foundation for £800K to initiate SM, Sudanese partners made it clear that they instead wanted scanning activities to take place on location, in decentralised hubs. Intense negotiations ensued, but the funders agreed to support this.

**“ Memory gaps must be addressed so that people
can rebuild their culture and their stories with
dignity and hope ”**

One of the many positive outcomes from this decision is that digitisation skills have become embedded in these partner institutions. Individuals trained by the SM—such as Wahbi—were empowered to train their colleagues and develop their own digitization teams, enabling work to continue even now.

Partner institutions, like NVU, also kept the equipment provided during the project and made decisions about which materials they would digitise. Of course, like all institutions, entrenched bureaucracy and power politics influenced which materials were selected for digitization. In some cases, items that the SM team deemed valuable, but the holding institution considered too controversial—or reputationally sensitive—were not selected for digitization. On the other hand, Sudanese partners, like

Wahbi, who have an intimate understanding of how culturally sensitive certain topics are for Sudanese society, were able to make careful assessments about whether private collections—like those held by families in the Merowe locality—should, in fact, be made public.

This approach to selection and appraisal was essential in building trust with Sudanese partners who were understandably suspicious about being exploited by the kind of extractive digitization projects experienced elsewhere on the African continent. In these cases, Africans had committed their labour and cultural heritage resources to digitization projects funded by institutions in the Global North only to find they had no access to, or control over, the digital artefacts they had produced (Breckenridge, 2014; Pickover, 2014; Rassool, 2018; Chamelot, Hiribarren, and Rodet, 2020).

In a concerted effort to mitigate these concerns about neocolonialism, the SM project is based on a series of agreements and memoranda of understanding that ensure partner institutions maintain ultimate control over how the digitised items are reused, as well as keeping digital copies of the items. So, while archival-quality digital copies of thousands of SM items are stored in the King's College London research data repository (KORDS), permission to reuse any of the digital items first requires contacting the original rights owner. Tragically, the ongoing conflict means many of these rights owners are now missing, some presumed dead.

“ In the immediate post-conflict period, the digital collection will be an important resource for remembering and revitalising Sudan’s shared, diverse identity and its rich heritage ”

In other cases, where Sudanese partners did not give permission for digitized copies of materials to leave the country, these have also been destroyed along with the tangible versions, so that nothing remains other than a gap. Yet, as Ikram Madani, Head of the Natural History Museum, University of Khartoum, notes, “If our physical collections have

been destroyed as a result of the war, then at least we will have the digital records of items to rebuild our collections from.”

Post-conflict

The SM project’s intention has always been to return the digital repository to a Sudanese institution and to keep another version as a copy somewhere outside the country. In the immediate post-conflict period, the SM digital collection will be an important resource for remembering and revitalising Sudan’s shared, diverse identity and its rich heritage, important building blocks for negotiating a future based on cultural coexistence for which there are, undoubtedly, clear precedents. In a recent oral history project conducted by Sara El-Nager, where she interviewed participants about their roles and reflections on the SM project, many noted the positive personal impact the project had on them. They spoke effusively about travelling to other, less familiar parts of Sudan, where they met people from different walks of life, which showed them the cultural diversity of Sudan but also what many had in common.

How these resources are deployed will be determined by Sudanese communities both within the country and its growing diaspora. For Asia Mahmoud, an SM Local Coordinator and Collection Researcher, this digitization project is about so much more than its content: “A lot of people actually faced this shock of losing everything [during the war], and it became personal to everyone. Projects like SM not only preserve our history, culture, and heritage, it will actually be like a solace for everyone; it feels like we’re here, we still exist.”

Making sure that the digital resource is available post-conflict depends, however, on ensuring that SM in its entirety receives sanctuary somewhere that safeguards and preserves against the threats of corruption, degradation, and obsolescence in the long term. The urgency of this is incomparable with the need to alleviate the suffering currently being inflicted on Sudan’s people and to secure an end to the conflict. Our wish is, however, that by guaranteeing the long-term future of SM, it can address the memory gaps so that people can reconstruct their culture and histories with dignity and hope, and in ways that safeguard peace agreements.

Bibliographic references:

Breckenridge, Keith. 2014. "The Politics of the Parallel Archive: Digital Imperialism and the Future of Record-Keeping in the Age of Digital Reproduction." *Journal of Southern African Studies* 40 (3): 499–519.

Chamelot, Fabienne, Vincent Hiribarren, & Marie Rodet. 2020. "Archives, the Digital Turn, and Governance in Africa." *History in Africa* 47: 101–18.

Copnall, James. 2025. "From Prized Artworks to Bullet Shells: How War Devastated Sudan's Museums." *BBC News*, April 26.

Khaled, Mai, Heba Saleh, Lucy Rodgers, Alexandra Heal, & Dan Clark. 2023. "How the Loss of Entire Families Is Ravaging the Social Fabric of Gaza." *Financial Times*, December 13, 2023.

Pickover, Michele. 2014. "Patrimony, Power and Politics: Selecting, Constructing and Preserving Digital Heritage Content in South Africa and Africa." Paper presented at *IFLA WLIC 2014 – Libraries, Citizens, Societies: Confluence for Knowledge*, Lyon, France, August 16–22.

Rassool, Ciraj. 2018. "Digitisation and the Government of Collections." Paper presented at *Postcolonial Digital Connections*, Halle, Germany, May 16–17.

About the authors

Wahbi Abdalrahman

Director of the Nile Valley Centre of Documentation and Ethnographic Studies at Nile Valley University. He has held key administrative roles, including Deputy Dean of Libraries and leading the Sudan Memory Project team in River Nile State. His career spans decades of leadership in documentation, digitisation, and library sciences.

Sara El-Nager

Sudanese British journalist who supported the Sudan Memory project with digitisation activities while living in Khartoum. Since leaving Sudan in 2023, she has developed a written history of the Sudan Memory project based on oral histories with key participants.

Marilyn Deegan

Emeritus Professor of Digital Humanities at King's College London and the Project Leader for Sudan Memory. She has over thirty years of experience leading digitisation projects and has a background as a medievalist, specialising in Anglo-Saxon medical texts.

Laura Gibson

Senior Lecturer in the Department of Digital Humanities at King's College London. Her research on decolonisation and digitisation is informed by several years working in South African museums, including as Collections and Digitisation Manager at the Luthuli Museum national legacy project.

Photography

Man reviewing archival documents by hand, surrounded by folders and papers, during the process of digitizing historical materials. Author: Wahbi Abdalrahman (Sudan Memory).

INTERVIEW

Stephanie Williams, former Special Adviser to the United Nations Secretary-General for Libya

Rita Costa

Arts & Creative Peace Lead, Build Up

Stephanie Williams served as a senior official in the United Nations between 2018 and 2022. Her last post was as Special Adviser to the United Nations Secretary-General for Libya.

In this conversation, she reflects on how disinformation and digital manipulation have influenced one of the most intricate international mediation processes in recent years. Drawing on her experience leading the UN peace mission in Libya, Williams explains how online hate speech, disinformation, and foreign interference not only exacerbated divisions on the ground but also directly endangered women negotiators and peacebuilders.

She also describes how the UN mission responded to digital violence through direct collaboration with social media platforms and by organising online dialogues, an experience she recounts in her book *Libya Since Qaddafi: Chaos and the Search for Peace*.

How would you describe the political and social context in Libya when you first arrived as a UN mediator?

The country emerged from 42 years of authoritarian rule in 2011, when Muammar Qaddafi was ousted following a popular uprising. During the four decades of his capricious and violent rule, the government tightly controlled information. There was absolutely no freedom of expression, speech, or assembly. Following Qaddafi's

overthrow, the population went from zero freedom of expression to complete, untethered, unfettered freedom. It was, of course, a remarkable transformation as Libyans finally found their voices: many Libyan media outlets emerged—radio stations, newspapers, television channels.

During the Arab Spring, of which Libya was a part and parcel, the uprisings were to a great extent organised online. I was on the other side of the Arab world, in the tiny island nation of Bahrain, where social media also energised the short-lived uprising there. In general, during that tumultuous period in the region, the internet became a massive engine of mobilisation. The same happened in Libya, and the trend continued after Qaddafi's overthrow, mainly for the better, but in some cases for the worse, as anyone could create and edit news or shape public opinion.

How did social media change the conflict dynamics in Libya?

Social media became a double-edged sword as the Libyan conflict evolved. The platform that thrived in Libya—and continues to thrive—is Facebook. There are more Facebook accounts in Libya than there are people. After Qaddafi's ouster, things did not go as envisioned: Libya began to fracture and then descended into armed conflict. There are many reasons for that devolution, but one of them was the use of social media to divide the population and to spread incitement to violence through the use of hate speech.

After 2011, Libya went through two civil wars: one in 2014 and another between 2019 and 2020, the latter being the one I witnessed. In this most recent conflict, it often felt as though the war was unfolding as much online as on the physical battlefield, with polarising rhetoric, hate speech, and narratives of "otherness." The dehumanisation of the "enemy" made it extremely difficult for us, as the United Nations mission, to promote dialogue.

From our position, we strongly opposed hate speech and sought to collaborate with various media outlets. We brought together some of the leading influencers to encourage them to moderate their tone, with mixed results. Still, it was impossible to get media figures from opposing sides into the same room because the fighting on the ground was too intense.

It was during this period, when I served as the United Nations' lead mediator, that digital violence became most dangerous.

“ Social media became a double-edged sword as the Libyan conflict evolved. It was used to divide the population and to spread incitement to violence with hate speech ”

What other challenges did you encounter regarding disinformation?

A significant challenge was external interference. Most observers are aware that countries deployed mercenaries and supplied weapons in blatant violation of the UN Arms Embargo. Still, certain foreign capitals also employed electronic armies and coordinated disinformation campaigns. For instance, when General Haftar launched his surprise attack on Tripoli in April 2019, there was a significant (dis)information component attached to it to make his attack appear a fait accompli, even though it faced significant resistance.

After the cessation of hostilities and the ceasefire, the latter of which I mediated in October 2020, we moved to political talks under the umbrella of the internationally blessed Berlin Process. It was during the political negotiations that we saw the most immediate and personal digital harm caused by disinformation and hate speech. Women participants, in particular, became the primary targets of online abuse, character assassination, and intimidation.

In fact, violence against women in Libya had already manifested itself earlier.

Before 2019, several prominent Libyan activists, politicians, and parliamentarians had already been targeted. A preeminent female human rights activist, Salwa Bugaighis, was killed in 2014 in Benghazi, at the time of the parliamentary elections. Legislator Fariha Berkawi was also killed in Derna in the summer of 2014. In July 2019, during the height of that civil war, Seham Sergiwa, a parliamentarian from Benghazi, was forcibly

disappeared. She had been outspoken in her criticism of the use of military force; she was very brave. One night, after expressing what I would call mild criticism of General Haftar's forces on television, masked men raided her home in Benghazi and abducted her. She has never been seen again.

So, there was already this climate of fear. And then, when we began the political talks in November 2020, we gathered all of the Libyan participants –17 of whom were women— outside of the country, because conditions inside the country were still too fragile. Naturally, the female participants were concerned about their safety.

How did the attacks against women take place in the digital sphere?

Almost immediately, the women delegates started to be targeted on Facebook. It began with the creation of fake Facebook pages meant not only to publicly shame them for their political participation, but also to intimidate them and their families.

One of the steps we took to counter this was to create a trusted relationship with Facebook. By then, it was well known that disinformation had played a role in fuelling conflicts, for example, in Myanmar, so Facebook was sensitive to this phenomenon. Libya was not a massive market for the company – only 7 million people – but Facebook was aware of the negative implications of its inaction in other instances, so when we flagged these false accounts, they took them down within 24 hours. That enabled us to keep the women in the room.

“ Women became the primary targets of online abuse, character assassination, and intimidation through false accounts in social media ”

However, while we were conducting the political mediation in November 2020, Hanan al-Barassi, another female activist in Benghazi, was gunned down in broad daylight for speaking out on social media, Haftar's sons. Needless to say, it was a hazardous climate for women. We ended up losing one woman from the political dialogue—she withdrew. She was from Eastern Libya; I guess you can draw your own conclusions as to her

reasons. It wasn't my position to question her. You could see that all these women were under terrible pressure. Direct digital violence created fear among the participants in the political dialogue.

Was there also coordinated disinformation targeting the UN mediation itself?

During the first round of our political negotiations, a large-scale disinformation campaign took place, the origins and impact of which we did not understand at the time. We saw a lot of activity on social media but couldn't tell where it was coming from.

It wasn't until a month after that round of talks that we realised what had happened. A study by Stanford University's Internet Observatory revealed that the Internet Research Agency, linked to the late Russian Wagner Group leader Yevgeny Prigozhin, had coordinated with electronic armies from the Middle East region to carry out a massive disinformation campaign against the UN mediation in Libya.

Prigozhin personally harassed me. As I describe in my book, the worst incident occurred in January 2022, when I visited Moscow. The Russian warlord – or someone impersonating him – appeared at my hotel with a group of cohorts and he followed me into the elevator. We didn't realise what was going on until the next day, when photos were published online as part of a false story claiming I had struck a deal with Prigozhin. I was still in Moscow when the story broke, and it was genuinely frightening for me and my team, given what we knew of his brutality. The UN mission in Libya had already suffered violent attacks; three staff members had been killed in Benghazi in 2019 under still-murky circumstances.

The Stanford study noted that what was happening in Libya was part of a broader pattern also seen in Syria and Sudan. The report prompted Facebook to take down hundreds of pages linked to the operation.

Which technologies contributed most to promoting digital violence?

I think it's now well established that disinformation campaigns, the use of algorithms, and the manipulation of social media were factors in Brexit; they were also a factor in the U.S. elections in 2016. So, it shouldn't be any surprise that essentially some of the

same or similar forces were applying the same methodology in the Libyan context.

Russia had a strong interest in Libya and was fully backing one side of the conflict – they had already supplied Haftar with weapons and mercenaries. The Russians were interested in producing an outcome that would not jeopardise their standing or interests on the ground in Libya and more broadly on the African continent. It wasn't a big surprise, but the revelation reinforced our commitment and approach—or at least my approach—that the only way to combat this was to be completely transparent about the mediation process and the end goal to reassure the Libyans.

“ The only way to combat disinformation was to be completely transparent about the mediation process and the end goal to reassure the Libyans ”

You can see in the algorithms that people go down these rabbit holes and believe all kinds of conspiracy theories. It's very hard to reach people whose primary source of information is social media and who reside in an echo chamber. It becomes an identity trait at some point: you create a new identity, or an added identity, complementary to your offline identity. Phones, computers, and any technology that enabled social media became weapons used by either side in the conflict.

How did your commitment to transparency translate into your mediation strategy?

The best disinfectant is sunlight, right? You just have to shine a light on what you're doing, because if you're not open and transparent and on the offensive rather than the defensive, you're not really responding. So, we needed to become much more transparent and grew increasingly committed to what I'd call radical transparency.

It was quite a difficult conversation within the United Nations context, because — and I respect this — there is a great deal of caution. But I felt that the social media environment in Libya, including the foreign interference, was so aggressive that if we

weren't out there telling our story, others would – and it wouldn't be the right one.

In this context, we organised digital dialogues – mainly with young people – to counter false narratives circulating on social media. We held five sessions over five months and conducted surveys to understand young people's views on key issues, including elections, economic reform, militia disarmament, and human rights. This helped us counter widespread claims like, "You know, elections can't be held in Libya – that's just a Muslim Brotherhood thing."

Through these exchanges, we quickly found that the vast majority of Libyans wanted elections, a change in the status quo, greater transparency in oil revenue management, accountability for human rights violations, and transitional justice. And all of this happened live, in Arabic, and outside the reach of the "Big Brothers" watching online.

These dialogues became hugely popular and even trended on social media. We used humour to connect with young audiences and show that their voices could be heard.

How did you expand that transparency during the political dialogue?

In February 2021, during our second political dialogue, I decided to broadcast most of it live. It was the post-pandemic period, and when the dialogues took place, people were still suffering from the aftereffects of COVID-19 – including its psychological impact – which made it even more critical to ensure maximum transparency in the room, allowing the Libyan public to witness what was happening, whether it was the voting process or the interviews with candidates for the unity government.

This experience confirmed my belief that any political or peace process must be inclusive. Of course, the preference of many countries – and of Libya's political elite – is to keep negotiations behind closed doors. You know, five or three men meet in a smoke-filled room, and an agreement is reached, right? But I believed, and still do, that in places like Libya, the more inclusive peace is, the more likely it is to take root – or at least to have a real chance of doing so. Conversely, the more it becomes a simple division of spoils or a power-sharing deal, the more likely it is to slide back into violence.

“ Any peace process must be inclusive. We need greater inclusion of women, youth, and civil society. The more inclusive peace is, the more likely it is to take root ”

We need greater inclusion of women, youth, and civil society, as well as representatives of vulnerable groups – whether ethnic communities or tribes.

In this context of digital violence, is there room for hope?

My feeling is one of concern rather than optimism. During the Arab Spring, for example, social media was a major driver of mobilisation, but once a technology begins to be used positively, other forces soon arrive to sabotage and manipulate it.

Traditional institutions – whether democratically elected governments or the United Nations – are always at a disadvantage, trying to keep up with rapidly evolving technology. We are now witnessing the use of fake news even by prominent world leaders, such as President Trump, and disinformation has become normalised.

My question is: if the British voters had known how the Brexit campaign was being manipulated online, would they have made the same decisions? Democratically elected governments and responsible institutions tend to act cautiously. They will always be vulnerable because malicious actors will stay one step ahead, and they are willing to use the same tools without ethics. For them, morality doesn't exist; the common good doesn't exist.

We need to raise public awareness about how and why people are being manipulated online.

About the author

Rita Costa

Rita Costa is Build Up's Arts & Creative Peace Lead. Since 2020 she has collaborated with Build Up facilitating participatory processes of social media analysis in Western Africa and Southern Europe, co-creating an international artist residency, developing frameworks to understand and address digital harm across the Foreign, Commonwealth & Development Office (UK Government), and organizing the annual Build Peace conference. She has a background in International Law, Political Science and choreography.

Photography

Acting Special Representative of the Secretary-General Stephanie Williams briefs the press during the fourth round of the 5+5 Libyan Joint Military Commission, Palais des Nations. 21 October 2020. Author: Violaine Martin (UN).