Photo by https://www.vecteezy.com

UNITED NATIONS DEVELOPMENT PROGRAMME

DEVELOPMENT FUTURES SERIES

August 2023

UNDP Global Policy Network Brief

Harnessing the potential of human-in-the-loop artificial intelligence for risk anticipation and violence prevention

by Fabio Oliva and Brian McQuinn[1]

Development organizations and conflict experts struggle to manage the magnitude, complexity and persistent volatility that characterize contemporary crises. Conflicts evolve at such a rapid pace that the amount of data produced by conflict or crisis situations is simply overwhelming. Because of the sheer amount of data and the pace at which they are being produced, human beings are unable to track crisis evolutions and manage effective decision-making processes. Under these radically changing circumstances, artificial intelligence (AI) can help us understand, and even anticipate, the outbreak and evolution of a crisis. Human-in-the-loop (HITL) AI combines the power of machine learning with human intelligence to address complex issues. This brief presents examples from Afghanistan and Ukraine to showcase applications of HILT artificial intelligence in the sphere of conflict resolution, particularly emphasizing risk anticipation and violence prevention.

Introduction

Contemporary peacebuilders and conflict experts are faced with crises of unprecedented magnitude and complexity. These crises are often described as "wicked problems,"[2] or more recently "polycrises."[3] Events on the ground evolve at such a rapid pace that conflict and political analyses are quickly redundant, and planned responses outdated. As trust in institutions and traditional information sources has eroded, the public has increasingly relied on online spaces for news and social interaction, weaving social media apps into the fabric of daily life. More than half of the world's 8 billion people currently use social media.[4]

Benefiting from this societal importance, malevolent and extremist groups have weaponized these online platforms to promote misinformation, manipulate audiences, incite polarization and unrest, and support real-world violence. To date, social media platforms, such as Facebook and Twitter, have been either unwilling or unable to address these fundamental vulnerabilities.[5] Additionally, these companies are often unable to consider the cross-platform impacts of their policies and the behaviours of their users. They lack the resources (and internal incentives) needed to understand the technology's impact in a non-Western context, and thus far, have been unwilling to share the metrics needed to properly evaluate the effects of their policies and interventions.[6]

The amount of data generated during contemporary conflicts is simply overwhelming and even well-resourced organizations are unable to keep up with information that is critical for decision-making. To address these challenges, this paper seeks to map out cutting edge research on the role of artificial intelligence (AI), and specifically human-in-the-loop AI. It also seeks to help development, humanitarian and conflict experts harness social media data to better understand and predict conflict trends while minimizing social media's misuse and weaponization.

Why use human-in-the-loop artificial intelligence for conflict prevention?

This brief draws upon the field of human-in-the-loop AI to harness social media data's heterogeneous and evolving nature to study and predict conflict patterns that could enhance our ability to prevent conflicts and crises.[7] Human-in-the-loop AI is an emerging field of AI research that is uniquely capable of tackling complex and adaptive problems. Traditional AI approaches aim to create systems that eliminate human interaction once completed (e.g. autonomous driving). As a consequence, human input is usually limited to the design stage (e.g. labelling photographs to train AI systems) or testing the system once completed and is normally conceived as a low-skill endeavour, perpetrating precarious employment conditions. However, breakthroughs in the understanding of continually evolving social phenomena such as conflict prevention require a more dynamic approach.

Human input in conflict prevention is critical in ensuring that cultural, contextual and ethical considerations are duly integrated. Conflicts are not solely driven by logical or rational factors. In a conflict context, emotions, social norms and cultural aspects play significant roles. Humans have the ability to empathize with the situation, decipher and make sense of emotions and navigate social complexities, which are essential in conflict prevention efforts. Conflict prevention also requires a deep and nuanced understanding of the local context and its historical background to which human experts can provide insights. Finally, conflict prevention entails making decisions that can significantly impact people's lives and human input ensures that ethical considerations are taken into account during decision-making processes.

Human-in-the-loop AI places human expertise at the centre of systems design and deployment. This approach draws upon human experts' deep domain knowledge, flexibility and creativity while compensating for their limited capacity for processing large information and data sets. As a result, human AI teams achieve a common goal, creating a feedback cycle that simultaneously improves the AI agents' algorithms and domain experts' knowledge. The consequence is a deeper and more purposeful integration of human expertise than is possible in existing supervised machine learning approaches, producing better results than either human-only or AI-only teams, as well as creating a synergistic improvement cycle for the human-AI collaboration.

Why it is necessary to take advantage of social media data

In many cases, political and military conflicts are now contested as much online as in the real world. For instance, non-state armed groups use social media platforms to recruit globally, while political extremists disseminate disinformation, undermining democracy and transforming warfare and public health communication.

Crucially, the data sets and tools used by social scientists and humanitarians to study social media's impact have not undergone a similar revolution. One consequence of this lag is a limited understanding of the extent to which malicious actors impact communities.[8] Most worrisome is the fact that social media platforms do far less to combat these problems than is commonly assumed.[9] For example, in the months before the Taliban's takeover of Afghanistan, its leaders, spokespersons and news agencies used Twitter to post more than 100,000 tweets to over 2 million followers.[10] More significantly, Twitter only suspended 31 of its 126,000 accounts in the Taliban's supportive ecosystem. Facebook is similarly under intense scrutiny over leaked internal documents questioning their ability and willingness to combat inappropriate content. These reports uncovered "how criminals use the platform for human trafficking and how Instagram affects mental health."[11]

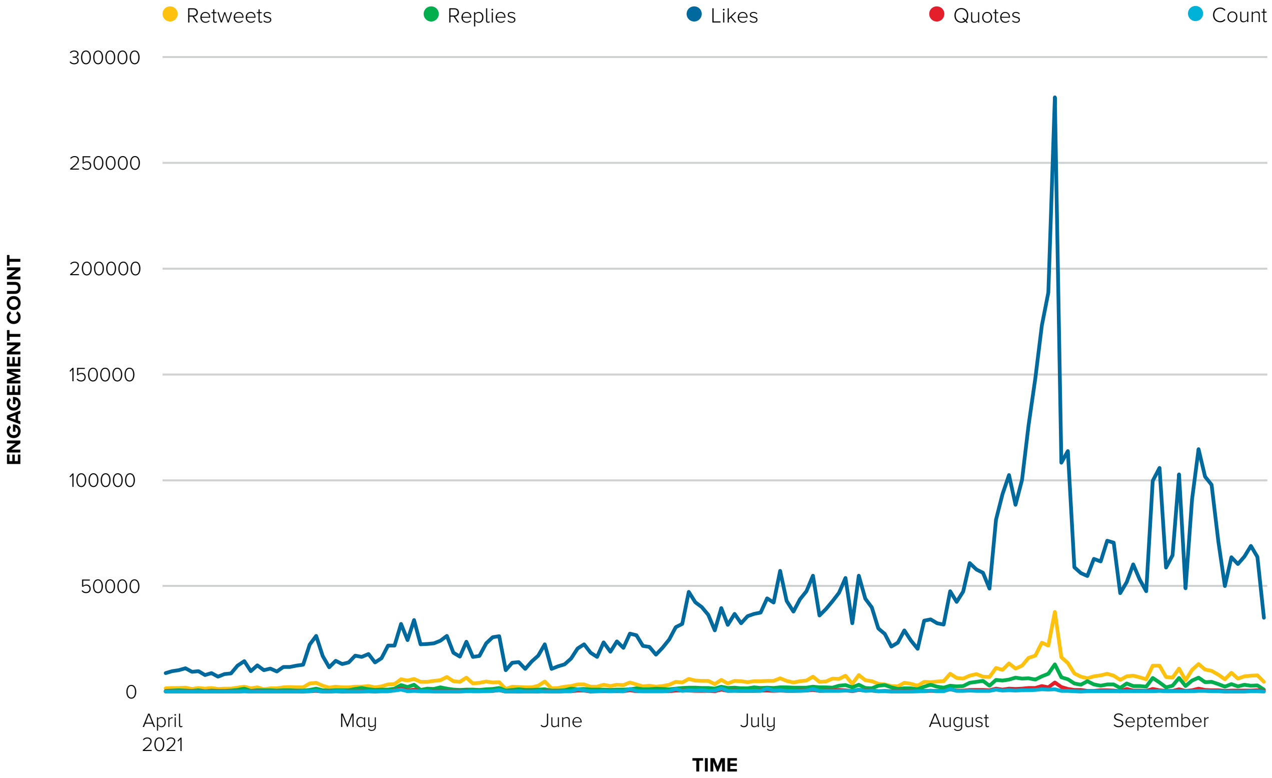
Unpacking emerging field practice

Afghanistan: Social media and the 2021 Taliban takeover

In August 2021, the complete collapse of Afghanistan's army and government defied prevailing analysis.[12] Commentators cited social media as a determining factor in the factor in the rapidity with which the Taliban established themselves as the de facto authorities in Afghanistan.[13] However, there was very limited analysis of how the Taliban used social media in the run-up to the takeover or its impact on the conflict.

A research team drawing upon conflict expertise, crisis informatics and AI systems investigated the Taliban's heavy investment in influence campaigns on social media.[14] The research began by drawing upon conflict experts to identify influential accounts, images, phrases and other contextual clues to map the social media ecosystem, which was built up from studying the activity of 63 accounts claimed by the Taliban leadership, spokespersons and avowed members. The analysis and mapping were an interactive process, as conflict experts evaluated the results to train the system and update it as new memes and language nuances evolved. At the time, these accounts had more than 2 million followers on Twitter in September 2021. As of May 8, 2022, Taliban content had reached more than 3.3 million verified accounts. Figure 1 outlines the total engagement with Taliban accounts in the five months prior to the takeover.[15] While this example looked at historic data, once an ecosystem is mapped, it can be monitored to identify emerging trends and potential crises in real time.

Figure 1. Total engagement with the Taliban content per week (April 1-September 16, 2021)[16]



Source: Laura Courchesne et al., 2022.

Social media data showed how Twitter became the Taliban's primary social media platform. Twitter's lack of moderation combined with heavy content investments by the Taliban allowed its influence operations to dominate Afghanistan's media landscape and advanced clear narratives during the fighting. The researchers used AI tools to find clear patterns in the group's communication strategies, visual imagery deployed, and the timing and content of social media activity and events on the ground.

The Taliban deployed a consistent repertoire of influence strategies to support their military campaign. In concert with their military operations, the Taliban employed six key influence strategies, each with specific goals, imagery and narratives, to influence international and domestic audiences. The group occasionally promoted plausible – albeit exaggerated or false – assertions. Such disinformation typically involved claims of premature victory or taking undue credit for incidents, all designed to convince Afghans and the international community that the Taliban takeover of Afghanistan was inevitable.

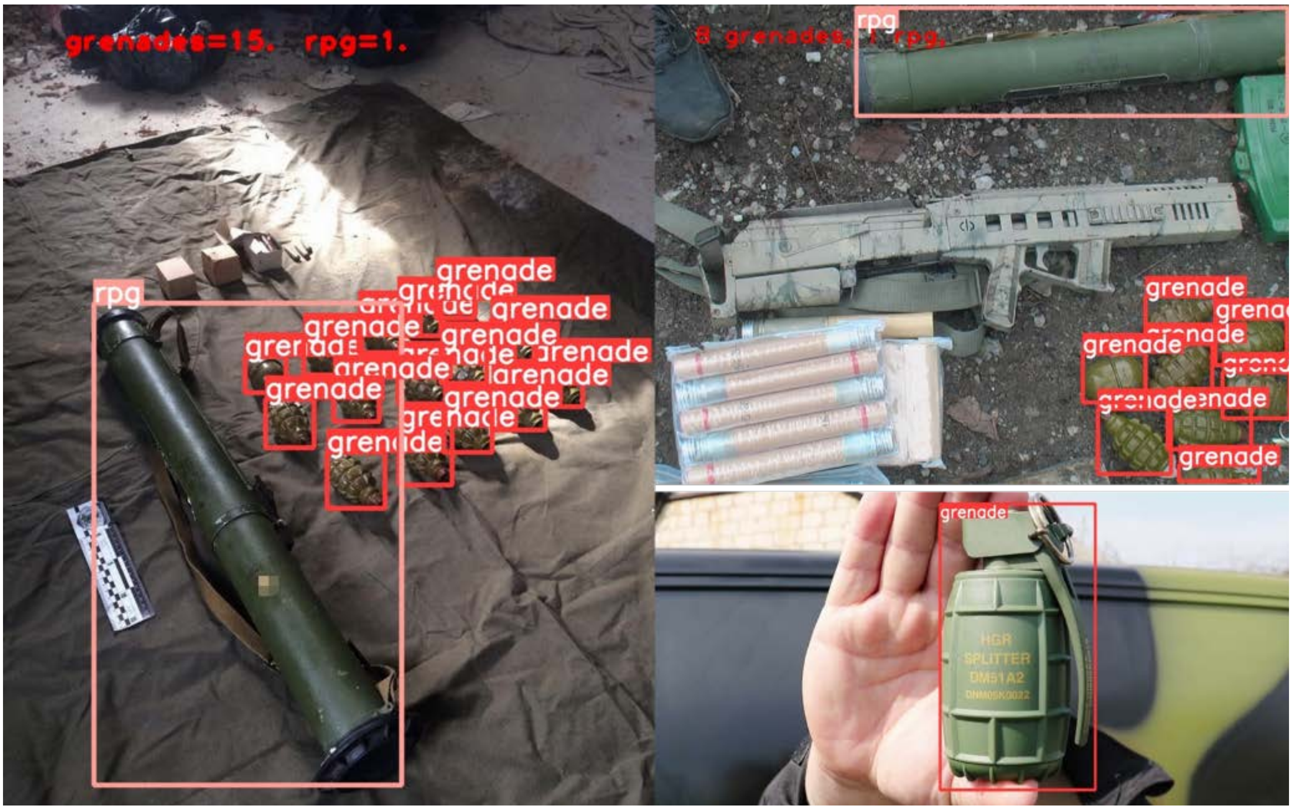
The analysis also provided preliminary evidence that the group's social media efforts coordinated with on-the-ground military operations. The researchers argued that during the summer takeover, the Taliban invested considerable time and organizational resources developing and amplifying a sophisticated online information campaign in real time. The analysis also suggests that the Taliban used Twitter's platform features to amplify their messaging and engage in both domestic and international outreach, including developing specialized hashtags to push specific propaganda narratives in four languages - Dari, English, Pashto, and Urdu - and leveraging mentions to attempt to garner a response from humanitarian organizations, key political players and journalists.[17]

Ukraine: Tracking of weapons via Twitter

On 24 February 2022, Russia invaded Ukraine. The war has devastated the country, killed hundreds of thousands of Ukrainians and Russians and displaced more than 13 million people.[18] Social media, such as Twitter, YouTube, Facebook, have been actively used as "weapons" for online influence and shaping of views and perceptions. However, social media has also been the primary conduit for real-time account of events. The wealth of images and data are an opportunity to study many aspects of the conflict, such as crowd sourcing information on attacks on health-care facilities, placement of mines, and the use of banned weapon systems (e.g. cluster munition).

In order to test the potential of tracking weapons using just social media data, a project examined how human-in-the-loop computer vision systems could be trained to identify types of weapons and the groups using them.[19] Weapon experts of the Small Arms Survey worked with AI and computer vision specialists to build a system that identified weapon types and insignias worn by members of armed actors (Figure 2).

Figure 2. Example of AI systems identifying and counting weapon systems



Source: Centre for Artificial Intelligence, Data, and Conflict, 2023

The weapon experts identified weapon systems that would be particularly challenging to recognize to test whether the low-quality images available on social media and their real-world context could be overcome. The team included group insignias in order to provide analysis not only on the number of systems that were being used but also whether the groups using them could also be identified. The system was able to reliably distinguish and count weapon systems and insignias in the database of pictures. The preliminary research has the potential to identify other objects in humanitarian contexts (e.g. ambulances) and symbols (e.g. Red Cross symbol on a hospital). Again, this was only a test to highlight how much information can be gleaned from publicly available data in a conflict setting where information and data is often difficult to obtain.

Conclusion and next steps

The application of AI technology in conflict and crisis settings requires careful consideration.[20] The most common approach tries to build autonomous AI systems, which eliminate human involvement and expertise. In this brief, we reviewed another strategy, human-in-the-loop AI systems, which we argue holds tremendous potential to responsibly advance United Nations efforts to manage large data sets in conflict situations. In complex and dynamic scenarios, where vast amounts of information are generated from diverse sources, human-in-the-loop AI systems can play a vital role in augmenting the capabilities of UN personnel to identify patterns, trends and potential threats in real time. Moreover, the human-in-the-loop approach ensures that human experts remain actively involved. AI-generated insights are, in fact, reviewed and validated by human experts who possess contextual knowledge of the conflict. This collaborative synergy between human analysts and AI technologies can empower the UN to make more informed decisions, enhance situational awareness and respond effectively to conflicts, ultimately contributing to the maintenance of international peace and security.

This brief primarily has focused on strengthening capabilities among United Nations entities and their personnel because of the universal mandate of the organization and the multi-lateral dimension of its efforts in addressing global crises. The potential risks and negative consequences of artificial intelligence have been recently explored in a UN Security Council debate.[21] In the absence of a new governance framework, AI applications can easily become hostage to individual countries' agendas and other national interest considerations.

This policy brief presented two applications of human-in-the-loop AI systems that were deployed to support a sense-making process of online content, including social media posts, news articles and user-generated content produced during the conflict. In the case of Afghanistan, we learned that the 2021 Taliban takeover was accelerated by the group's information operations on social media platforms, which anticipated events and, in some cases, even pre-empted military operations. Several lessons can be drawn from this experience for UN entities to make better use of social media monitoring prior and during rapid onset crises.

Building on the Afghan experience, AI algorithms could be tailored to decipher and track social media narratives attacking humanitarians or disinformation campaigns targeting democratic governments. The Islamic State used social media strategies to affect the morale of local populations and facilitate its seizure of large swaths of territory in Syria and Iraq between 2013 and 2014.[22] In the days prior to the February 2021 military coup in Myanmar, there were speculations and warnings circulating on social media platforms that hinted at an impending takeover by the generals. AI monitoring could have helped monitor those situations, thereby strengthening the UN response and preparedness. In the domain of predictive analytics, the recurrent question is how to interpret and act upon multiple data points. Human-in-the-loop AI introduces the human element as data are being collected, therefore creating the space for preventive decision-making early on. The example of Ukraine illustrated how human-mediated machine learning can analyse large volumes of online content to specific weapon systems and monitor their movements across a large country to anticipate post-conflict community violence or fast-track, area-based programming even as the conflict is still ongoing. If carefully guided by humans, machine learning is poised to advance violence reduction efforts and give hope to ongoing UN global initiatives on disarmament.[23]

More broadly, in conflict situations, real-time understanding of conflict dynamics is most valuable in the identification of at-risk populations and the deployment of humanitarian and development aid where it is most needed. For instance, AI can be employed to analyse satellite imagery to detect changes in infrastructure, movement of troops or displacement of populations. It can enable real-time identification of gendered dimensions of a conflict or identify women's contributions to peace that would otherwise be neglected. There are, however, many risks with these technologies as well, including concerns that these systems would give rise to "surveillance humanitarianism"[24] and an over-reliance on technological solutions that disempower the impacted communities.[25]

Additional research is needed to explore the policy implication of using human-in-the loop AI approaches in so-called "grey zone" tactics and threats by non-state armed groups, cybersecurity actors and mercenaries, which seize on information-contested environments to meddle in electoral processes, political crises and economic activities through unconventional means.

In order to test the feasibility of human-in-the-loop systems within the UN, they could be tested with a small number of UNDP country offices or other UN field operations. The systems would be tailored with UN staff to address the needs of each context. The initial focus could be, for example, to expand prevention efforts in places where tensions are building or identifying disinformation campaigns targeting humanitarian personnel. The goal would be to create a global, state-of-the-art platform to share human-in-the-loop tools, labels, models and algorithms –those created both by UN teams and other innovators. By housing these computational research tools and conflict-related data sets on one platform, this initiative will become the global leader in risk anticipation and violence prevention.

Endnotes

1 Fabio Oliva, PhD, Senior Advisor, UNDP Crisis Bureau, Policy, Knowledge, and Partnerships Team, email: fabio.oliva@undp.org and Brian McQuinn, Associate Professor, University of Regina and Co-Director, Centre for Artificial Intelligence, Data, and Conflict, email: brian.mcquinn@uregina.ca. Acknowledgments: The authors would like to thank Claire Van der Vaeren, Aarathi Krishnan, Calum Handforth, Gonzalo Gomez, Mihaela Stojkoska, Shani Harris and Tanya Pedersen for their review and helpful comments on this brief. The authors are also grateful to the Small Arms Survey.

2 Systems thinking and complexity science use various terms to describe complex problems, including "wicked problems." In this paper, wicked problems are defined as "a class of social system problems which are ill-formulated, where the information is confusing, where there are many clients and decision makers with conflicting values and where the ramifications in the whole system are thoroughly confusing." C. West Churchman, Wicked problems, Management Science, (December 1967), vol. 4, no. 14, B-141-42.

3 In conflict and development literature other terms have also been used, like compound crisis. For original reference for polycrisis, see: Morin, Edgar and Kern, 1999. Homeland earth: A manifesto for the new millennium. Hampton Press (NJ).

4 Chaffey, D., 2022. Global social media statistics research summary 2022. Smart Insights.

5 Washington Post, Only Facebook knows the extent of its misinformation problem. And it's not sharing, even with the White House, 19 August 2021.

6 The mission regrets that Facebook is unable to provide country-specific data about the spread of hate speech on its platform, which is imperative to assess the adequacy of its response. United Nations Human Rights Council Report of the independent international fact-finding mission on Myanmar, A/HRC/39/64, 12 September 2018.

7 Zanzotto, F.M., 2019. Human-in-the-loop artificial intelligence. Journal of Artificial Intelligence Research, 64, pp. 243-252.

8 Bastug, M.F., Douai, A., & Akca, D., 2020. Exploring the 'demand side' of online radicalization: Evidence from the Canadian context. Studies in Conflict & Terrorism. 43:7; 616-637.

9 Kang, C., 2021. Facebook whistle-blower urges lawmakers to regulate the company. New York Times.

10 Courchesne, L., Rasikh, B., McQuinn, B., Buntain, C. 2022. Powered by Twitter? The Taliban's takeover of Afghanistan. A joint report by the Centre for Artificial Intelligence, Data, and Conflict, The Empirical Studies of Conflict Project at Princeton University and the New Jersey Institute of Technology.

11 Bouché, V., 2018. Survivor Insights. Thorn.

12 Reuters, Afghan army collapse 'took us all by surprise,' U.S. defense secretary says, 28 September 2021

13 CSNB, 'Intelligence failure of the highest order' - How Afghanistan fell to the Taliban so quickly, 16 August 2021

14 Courchesne, L. Rasikh, B., McQuinn, B., and Buntain, C., 2022. Powered by Twitter? The Taliban's Takeover of Afghanistan. Centre for Artificial Intelligence, Data, and Conflict

15 Figure 1 is reproduced from Courchesne, L. Rasikh, B., McQuinn, B., and Buntain, C. 2022. Powered by Twitter? The Taliban's takeover of Afghanistan. Centre for Artificial Intelligence, Data, and Conflict

16 The group used Twitter to generate four times more domestic engagement than the content of 18 mainstream Afghan news organizations combined. See the following report for further analysis: Courchesne, L. Rasikh, B., McQuinn, B., and Buntain, C., 2022. Powered by Twitter? The Taliban's Takeover of Afghanistan. Centre for Artificial Intelligence, Data, and Conflict.

17 These references included links to human rights reports on the situation in Afghanistan. See the following report for additional examples: Courchesne, L. Rasikh, B., McQuinn, B., and Buntain, C., 2022. Powered by Twitter? The Taliban's Takeover of Afghanistan. Centre for Artificial Intelligence, Data, and Conflict.

18 About 8m refugees and over 5m internally displaced. UNHCR, May 2023 https://bit.ly/42DxA30

19 The project was led by Professor Abdul Bais and Muhib Ullah, publication forthcoming.

20 Beduschi, A., 2022. Harnessing the potential of artificial intelligence for humanitarian action: Opportunities and risks. International Review of the Red Cross, vol. 104, no. 919: pp. 1149–1169. For the implication of AI stems and autonomous weapon systems see: Schuller, A.L., 2017. At the crossroads of control: The intersection of artificial intelligence in autonomous weapon systems with international humanitarian law. Harvard National Security Journal, 8, p. 379.

21 Artificial intelligence: Opportunities and risks for international peace and security – UN Security Council, 9381st meeting, 19 July 2023 (https://media.un.org/en/asset/k1j/k1ji81po8p)

22 Macnair, L. & Frank, R., 2018. The mediums and the messages: Exploring the language of Islamic State media through sentiment analysis, Critical Studies on Terrorism, 11:3, 438-457.

23 The preliminary work and consultations that are shaping the upcoming New Agenda for Peace include a call for expanding research and institutional partnerships to better understand the linkage between disarmament and development. Besides being a key element of the 2030 Agenda under SDG 16, the reduction of all forms of violence is also part of the renewed commitments towards operationalising conflict prevention to be included in the New Agenda for Peace.

24 Weitzberg, K., Cheesman, M., Martin, A., and Schoemaker, E, 2021. Between surveillance and recognition: Rethinking digital identity in aid, Big Data & Society, Vol. 8, No. 1.

25 Duffield, M., 2016. The resilience of the ruins: Towards a critique of digital humanitarianism, Resilience, Vol. 4, No. 3.

The views expressed in this publication are those of the author(s) and do not necessarily represent those of the United Nations, including UNDP or the UN Member States.