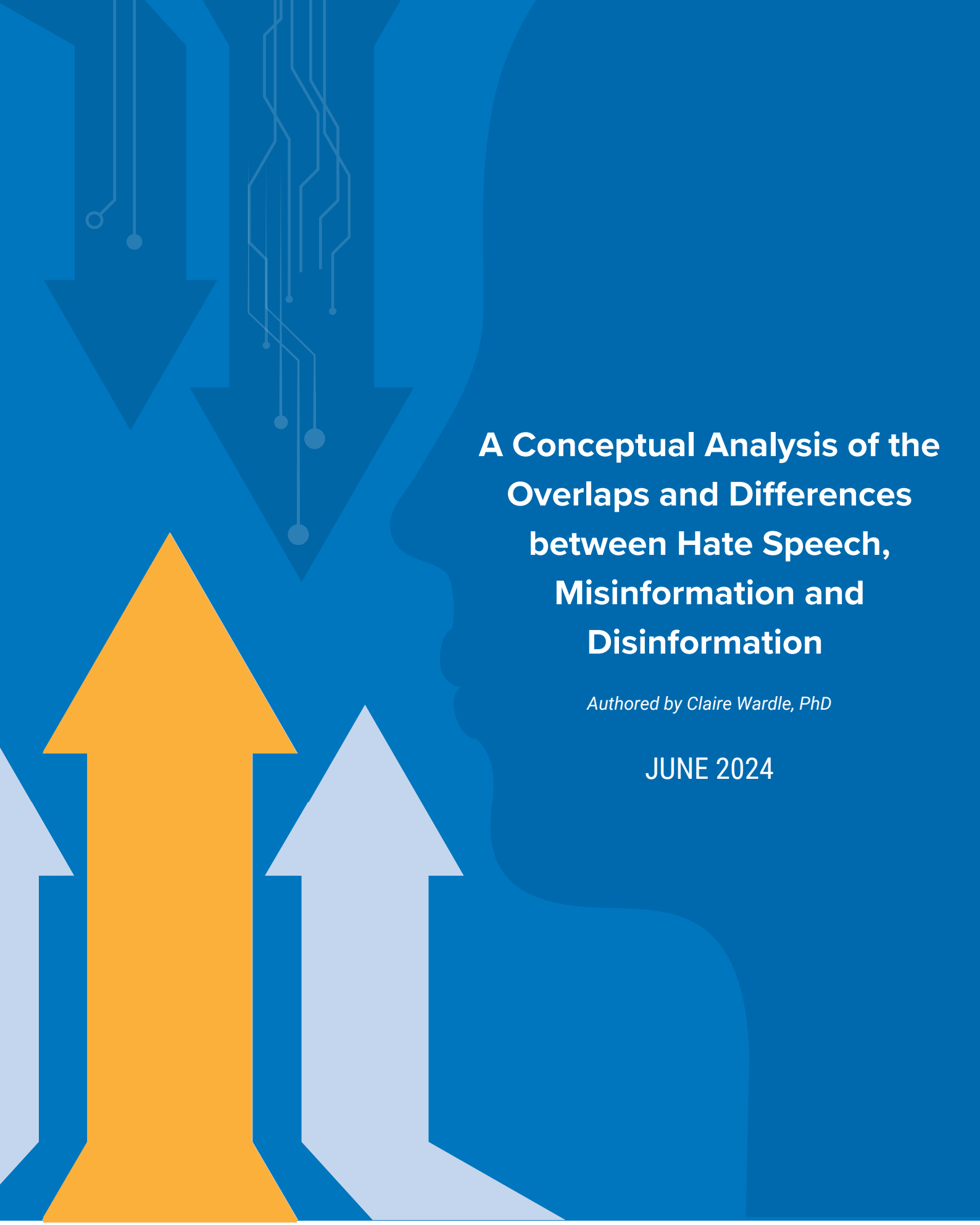

A Conceptual Analysis of the Overlaps and Differences between Hate Speech, Misinformation and Disinformation

Authored by Claire Wardle, PhD

JUNE 2024

United Nations

A Conceptual Analysis of the Overlaps and Differences between Hate Speech, Misinformation and Disinformation

By Claire Wardle, PhD
Brown University

The study was commissioned to inform reflection and policy development. The views expressed here are solely those of the author.

Department of Peace Operations (DPO)
Office of the Special Adviser on the Prevention of Genocide (OSAPG)
United Nations
New York, June 2024

Visit QR Code for downloads and more information

Graphic Design by Isabel Garcia-Sosa

Table of Contents

Executive Summary

This report, commissioned by the UN Department of Peace Operations and the UN Office of the Special Adviser on the Prevention of Genocide, provides a comprehensive conceptual analysis to distinguish between hate speech, misinformation and disinformation. It aims to identify their overlaps and differences, particularly in conflict-affected and high-risk areas. The report reviews relevant international human rights law and international humanitarian law, emphasising the need for responses that respect freedom of expression while addressing harmful speech. It also draws on interviews with a small sample of 15 staff working on either or all of these three areas with the UN and humanitarian organisations.

DEFINITIONAL CHALLENGES

Hate Speech: is defined as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor." [1]

Misinformation: is defined as the unintentional spread of inaccurate information shared in good faith by those unaware that they are passing on falsehoods.[2]

Disinformation: is defined as the intentional spread of inaccurate information, intended to deceive and shared in order to do serious harm. [3]

Hate speech, misinformation and disinformation represent distinct categories of harmful speech, yet there are no universally agreed-upon definitions for these terms. Identifying the "'intent" of the speaker or creator, a key element in distinguishing these forms, adds complexity to their operationalisation and response. Additionally, the intersection of hate speech, misinformation and disinformation can amplify their negative impacts in both online and offline environments. This further complicates efforts to operationalise responses, conduct research and make meaningful comparisons across different contexts.

PARTICULAR CHALLENGES RELATED TO CONFLICT SETTINGS

The report focuses on these issues in the context of "conflict-affected and high-risk areas," defined by the OECD as areas "identified by the presence of armed conflict, widespread violence or other risks of harm to people."[4] According to interviewees, there are a number of characteristics that defined these areas that are critical to consider:

The Volatility of These Areas: In conflict settings, hate speech, misinformation and disinformation, which might be less harmful elsewhere, can quickly escalate and become extremely dangerous.

The Transformation of Conflicts: These types of harmful speech have fundamentally altered the nature of conflict, influencing how battles are fought.

The Limitations of International Law: Many interviewees believe that international human rights law is either inadequate for addressing these forms of speech or that applying the existing framework effectively in daily operations remains a significant challenge.

The Strategic Use of Disinformation: When these types of speech are strategically deployed, it can be combined with real-world tactics to blur the lines between fact and fiction, creating new realities and exacerbating harm.

Undermining the Effectiveness of Organisations: Those pursuing strategic objectives may intentionally undermine the effectiveness of organisations working on peace, security, or humanitarian efforts on the ground.

The Impact on Digital Divide: In regions with low Internet penetration, distrust in media and lower literacy rates, offline spaces can hold significant influence, amplifying the power of those with online access.

TWO KEY TENSIONS

There are two key tensions that run through the literature surrounding these three types of speech and they were mirrored in the interviews with participants. The tensions can be described as the following:

(1) Most speech falling into the three categories discussed in this report is considered protected under international human rights law, which upholds freedom of expression. Responses to this protected speech must be carefully designed and implemented to avoid unintended consequences, such as increased censorship. However, ignoring the potential harms of protected speech can, over time, lead to incitement to discrimination, hostility, or violence. Therefore, the consequences of both action and inaction need to be thoughtfully considered.

(2) There are varying perceptions about the definitions and categories of harmful speech. Some interviewees advocated for an umbrella term to encompass all types of potentially harmful speech, arguing that the strict differentiation between hate speech, misinformation and disinformation leads to siloed efforts and hinders recognising the overall harm these forms of speech cause within communities. Conversely, other respondents emphasised the importance of precise definitions, believing that they are crucial for appropriately and effectively addressing each type of speech, given their distinct historical and legal contexts. They expressed concern that merging these definitions into a single umbrella term could oversimplify and inadequately address the specific harms associated with each type of speech.

THE SIMILARITIES BETWEEN HATE SPEECH, MISINFORMATION AND DISINFORMATION

1. Absence of definitions that can be easily operationalised

While the UN has working definitions, in particular for hate speech, many interviewees drew on different definitions from the literature, some lumped the terms together, some used generalised frameworks such as information pollution,[5] infodemic or information disorder, or harmful information. Misinformation and disinformation were often used interchangeably. While the definitions used by the UN give conceptual clarity and are critical for designing appropriate human rights-based responses, in practice distinguishing between hate speech and misinformation and disinformation can in some contexts be challenging.

2. Hate speech, misinformation and disinformation can have cumulative impacts

As noted frequently in this report, hate speech as well as misinformation and disinformation can have very severe, immediate impacts. However, it's also important to recognise that "low-level" examples of all three types of speech have the ability to cause severe harm over very long periods of time.

3. Some similar tactics for creation, dissemination and amplification tactics for creation, dissemination and amplification

The global nature of the Internet has allowed foreign actors to interfere in conflict settings anonymously, as well as enable highly trusted and technologically savvy diaspora or elite voices to get involved in and often exacerbate local disputes. Whether it's hate speech, misinformation or disinformation, interviewees discussed how they see similar tactics used for creations, dissemination and amplification. That being said, disinformation displays some distinct tactics that attempt to generate new versions of reality, including through the use of forgeries, paid demonstrations, inauthentic digital accounts and/or fake testimonies.

THE DIFFERENCES: HOW HATE SPEECH DIFFERS FROM MISINFORMATION AND DISINFORMATION

1. International law

International law treats certain types of hate speech — incitement to discrimination, hostility and violence and public and direct incitement to genocide — very differently to other types of hate speech, misinformation and disinformation. Certain types of hate speech (reaching the threshold of incitement) are prohibited by international law. There is no such corresponding obligation for misinformation or disinformation as it is not defined explicitly in international law. [6]

2. Identity vs. non-identity characteristics

Hate speech targets people, whether individuals or groups, based on their identity; disinformation can be used as a means of speading hate speech but it may also target individuals and groups based on non-identity characteristics (such as occupation). It may also target institutions, specific facts and values and belief systems.

Hate speech impacts people, and the impact of discriminatory or pejorative speech can be tremendous,

causing death, extreme pain and suffering for the victims. In the most serious cases, it can also lead to the commission of identity-based violence including genocide and related crimes (war crimes, crimes against humanity). In fact, hate speech is recognized as an indicator of risk and potential trigger of such crimes.[7]

3. Hate speech causes different harms

Hate speech causes personal and communal harms in the form of a variety of human rights violations and abuses and in some cases may lead to the commission of genocide and related crimes. Indeed, hate speech and incitement to violence are early warning signs that should be monitored by the UN as a risk factor for such crimes.[8] Disinformation can, by contrast, also cause personal, institutional or societal harms. Disinformation may also harm access to information, and undermine a healthy civic space, particularly when crowding out a plurality of views.

4. Online platforms policies

Many platforms have different policies for hate speech in comparison to misinformation and disinformation and other categories of content.[9] Platform policies on hate speech have emerged over the past 15 years, with help from human rights specialists and academics, as well as learnings from serious mistakes and missteps around certain events (Myanmar in particular). Policies related to misinformation and disinformation have been rolled out since 2016, first in relation to election-based disinformation, and more recently related to health-related misinformation and disinformation surrounding the COVID-19 pandemic.

5. Different prevention and responses needed

Hate speech, misinformation and disinformation intersect with various thematic areas – from conflict prevention and resolution, human rights to digital technologies. Different types of speech may also occur simultaneously in a given context; for example, a UN peacekeeping mission may simultaneously need to respond to hate speech targeting a particular ethnic community during an ongoing election campaign in which misinformation and disinformation abounds, and in an overall context in which the mission is the target of disinformation campaigns concerning the effectiveness of its mandate implementation and questioning the UN's neutrality. The strategic use of disinformation to tarnish a target may require upstream political engagement, compared to misinformation, which may be clarified through awareness-raising or communications campaigns. Therefore, multiple and differentiated responses to each type of speech should be considered, giving priority to the speech with the highest risk of physical harm to civilians, depending on the situations on the ground.

CONCLUSIONS

While these types of speech may initially seem similar, a closer examination of their elements, along with their legal and historical contexts, highlights the need to consider them as distinct phenomena. Various forms of hate speech are governed by different legal and policy frameworks, target different groups, cause different harms, and necessitate different responses. A key factor is the frequent use of these different forms of speech by those intent on causing harm to amplify their impact.

International Human Rights Law and International Humanitarian Law are essential legal frameworks for addressing these issues and guiding appropriate responses. The continuous application of international human rights law, alongside international humanitarian law, is crucial for effectively protecting the right to freedom of opinion and expression, both in peacetime and during conflicts.

Table 1: Table designed to help identify whether an example of speech should be defined as hate speech, misinformation or disinformation.[10]

<table>
<tr><th colspan="4">ACTORS: Who is involved and impacted?</th></tr>
<tr><td rowspan="2">Sharer of Speech</td><td>Hate Speech</td><td>Disinformation</td><td>Misinformation</td></tr>
<tr><td colspan="3">Private Individuals
Community Leaders: e.g. Local Business Owner, Faith Leader, Civic Representative
Influencers: e.g. Celebrity, Well Known Person
State Actors: e.g. Politician, Official Spokesperson, Military Leader
Non-State Actors: e.g. Armed Group, News/Media organisation, Activist Group</td></tr>
<tr><td>Sharer Intends to Harm</td><td>Yes</td><td>Yes</td><td>No</td></tr>
<tr><td rowspan="3">Target of Speech</td><td colspan="3">An Individual</td></tr>
<tr><td>A group or individuals targeted because of their identity with speech that attacks or uses prerogative or discriminatory language
Certain facts (contemporary and in the past) e.g. Holocaust and genocide denial</td><td colspan="2">Individuals or a group of individuals with a common trait (e.g. occupation) targeted with false or distorted information. This does not need to be hateful or discriminatory.
State actor or organisation
Non-state actor or organisation
Certain facts (contemporary and in the past) e.g. Holocaust and genocide denial
A value or ideal (e.g., democracy, science)</td></tr>
<tr><td>Example: A well-known journalist, activist, politician targeted because of their identity (e.g. race, gender, religion) with speech that is discriminatory</td><td colspan="2">Example: A well-known journalist, activist, politicians targeted because of his/her occupation with mis- and disinformation</td></tr>
<tr><td>Audience of Speech</td><td colspan="3">Limited Audience
Would reach a defined community, either in person or online
Large, multi-community audience with the likelihood of repetition over multiple days or weeks
National or Global Audience (via news or exceptionally large online audience)</td></tr>
</table>

<table>
<tr><th colspan="3">CONTENT: What has been created, with intent to harm?</th></tr>
<tr><td rowspan="3">Content</td><td>Hate Speech</td><td>Disinformation</td></tr>
<tr><td colspan="2">• Denial and distortion of some historical events (for example the Holocaust or other genocide demonstrated by international court of law)
• Content designed to emphasise in-group/out-group differences
• Harmful content created using AI (e.g., deep fakes)</td></tr>
<tr><td>• Explicit calls to violence based on identity, including genocide
• Explicit calls to discriminate based on identity
• Content designed to demonise and/or dehumanise based on identity
• Explicit recommendation to take action that would cause someone harm based on identity
• Use of dog-whistles and coded language related to identity
• Use of identity-based slurs
• Using words or phrases designed to evade content moderation</td><td>• Content designed to deceive or evade (e.g., sharing false claims; creating false accounts; deceptive editing)
• Content designed to mislead (e.g., cherry picking statistics, editing quotes, out of context images)
• Content designed to undermine trust in institutions and official processes (e.g., conspiracies)</td></tr>
<tr><th colspan="3">DISTRIBUTION: What platforms and tactics are being used to encourage the distribution of disinformation or hate speech?</th></tr>
<tr><td rowspan="2">Platforms</td><td>Hate Speech</td><td>Disinformation</td></tr>
<tr><td colspan="2">• Offline mechanisms: speech, pamphlets, posters, peer-to-peer conversations Broadcast mechanisms: radio, television, newspapers
• Closed digital spaces: encrypted messaging apps, online groups and communities
• Public digital spaces: video sharing apps, social networks, online advertising
• Disinformation: forgeries, paid demonstrations, fake testimonies, use of inauthentic accounts
• Other: academic conferences and journals and various types of art, graffiti, memes, songs</td></tr>
</table>

<table>
<tr><th colspan="4">HARM: What damage could be caused?</th></tr>
<tr><th>Harms</th><th>Hate Speech</th><th>Disinformation</th><th>Misinformation</th></tr>
<tr><td rowspan="4">Harms</td><td colspan="3">Impacts on both physical (loss of life or injury, including sexual violence) and mental health
Self-protective or forced withdrawal of certain groups of people from the public square (offline or online) (e.g., female politicians, journalists, voters)
Serious reputational damage to the target of the speech, which can lead to barriers to accessing rights and services, restrictive or discriminatory measures, acting as a trigger for additional disinformation and/or hate speech
Increased societal polarisation and climate of fear
Heightened hostility and hatred against the target of speech
Loss of morale and self-belief amongst target group</td></tr>
<tr><td colspan="2">Demonisation, dehumanisation, marginalisation and discrimination against the target of the speech</td><td>Decline of trust in an institution or value (e.g., belief in science, democratic values, freedom of expression)</td></tr>
<tr><td colspan="2">Hate crimes or violence against targets of hate speech</td><td>Undermining belief in peace and cooperation and increasing justification for conflicts and barriers to integration, possibility of triggering violence and conflict</td></tr>
<tr><td colspan="2">Self-protective or forced displacement; segregation of communities and creation of identity-based enclaves vulnerable to violence</td><td>Resistance to public policy measures (e.g., climate, public health) leading to different material impacts (financial sector/economy, employment, environmental).</td></tr>
<tr><td colspan="2">Genocide, crimes against humanity or war crimes</td><td>Elections impacted (including votes suppressed, distortion of policy debates, results not considered legitimate).</td></tr>
</table>

The challenge for professionals working in affected fields is to develop operational responses that draw on the expertise of specialists in each area. They must create workflows and processes that acknowledge the distinctions, including the specific policy and human rights law frameworks, while also recognising that harmful speech is often intentionally designed to evade clear definitional boundaries.

While this report focuses on conflict-affected and high-risk areas, the discussion, findings and recommendations would also be relevant to crisis, fragile and non-conflict settings.This report sets the stage for a subsequent policy paper focused on specific response strategies, aiming to provide actionable insights for policymakers and practitioners engaged in mitigating the harmful effects of hate speech and misinformation in conflict-affected and high-risk environments.

Introduction

PURPOSE OF THIS REPORT

This report was commissioned by the UN Department of Peace Operations and the UN Office of the Special Adviser on the Prevention of Genocide to explore three different notions of speech[11] and information — hate speech, misinformation, and disinformation. How are they different? Where do they overlap? And most critically, how can they be identified and responded to appropriately? The report is based on a comprehensive literature review and informed by interviews with a small sample of 15 staff working on either or all of these three areas with the UN and humanitarian organisations. They work at headquarters as well as in conflict settings.

While there are some mentions of potential responses as part of the discussion, the primary purpose of this report is to provide a conceptual analysis for understanding the similarities and differences between hate speech, misinformation and disinformation. A policy paper to follow will focus specifically on responses.

In addition, the report focuses on these issues in the context of "conflict-affected and high-risk areas," defined by the organisation for Economic Cooperation and Development (OECD) as areas "identified by the presence of armed conflict, widespread violence or other risks of harm to people."[12] The discussion, findings and recommendations would also be relevant to crisis, fragile and non-conflict settings.

DEFINING THE TERMS

There are no universally agreed-upon definitions of hate speech, misinformation and disinformation and there are additional terms that are also used to describe these or similar phenomena. These include: dangerous speech; information manipulation; influence operations; and computational propaganda. This makes it difficult to operationalise, research or make comparisons across contexts. Recent UN initiatives have proposed operational definitions for the three key terms and these will provide the basis for the discussions that follow.[13] It should nevertheless be noted that these three terms do not capture all informational harms.[14]

Within the UN, however, there is broad agreement on the way that misinformation and disinformation are defined, although the vagueness of these definitions means there are significant issues related to the way different legal and institutional environments globally consider these types of speech. In addition, the fact that the definitions for misinformation and disinformation rely so heavily on identifying "intent", which can be exceptionally hard to operationalise, causes additional challenges even when there is agreement on the wording of the definitions.

MISINFORMATION

is defined as the unintentional spread of inaccurate information shared in good faith by those unaware that they are passing on falsehoods.[15]

DISINFORMATION

is defined as the intentional spread of inaccurate information, intended to deceive and shared in order to do serious harm.[16]

Similarly, while there is no international legal definition of hate speech the UN System has a working definition set out in the UN Strategy and Plan of Action.[17]

HATE SPEECH

is defined as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor."

Hate speech, misinformation and disinformation have a negative impact on the information environment, this is particularly the case in conflict-affected and high-risk areas. In too many situations there are longstanding and engrained social, cultural, political, ethnic, religious or other divisions and rivalries fueling hate speech. In some instances, such hate speech intersects with rumors, conspiracy theories and myths, exacerbated by targeted disinformation created and funded by private or state actors. The toxic information flows between online and offline settings, is often supercharged by influential individuals, both in-country and part of the diaspora, and by politicians. This can then be amplified by the media, a mixture of state-controlled outlets, bloggers, and independent voices on radio and television, as well as digital platforms.

In certain contexts, the impact of such narratives could be described as hazardous. The UN Secretary-General has noted that:[18]

> **HATE SPEECH**
> *HAS BEEN A PRECURSOR TO*
> **ATROCITY CRIMES**
> *INCLUDING*
> **GENOCIDE,**
> *FROM RWANDA TO BOSNIA TO CAMBODIA.*
>
> UN Secretary-General, 2019[18]

The same has been highlighted in academia, as Sandrine Tiller and co-authors argue, hate speech misinformation and disinformation "can lead to death, injury, imprisonment, discrimination or displacement. [They] can, directly or indirectly, fuel vicious cycles of violence and further entrench already protracted conflicts."[19] It should also be noted that the growing weaponisation of information aimed at humanitarian workers, peacekeepers, and journalists in the field has led to increased harassment, threats, and fatalities among these individuals. Consequently, these organisations face heightened challenges in effectively carrying out their mandates while maintaining their credibility.[20] Furthermore, civilians in conflict zones who are making life-or-death decisions regarding their safety may also be harmed by misinformation and disinformation, for example related to safe areas, evacuation routes, or access to medical care or humanitarian aid.

These threats are being taken seriously at the highest level of the UN, such as with UN Security Council resolution S/RES/2686 (2023) which "condemns misinformation, disinformation and incitement to violence against United Nations peacekeeping operations intended to negatively affect their safety or their ability to implement their mandates." Resolutions have also been adopted by the General Assembly, e.g., A/RES/73/328 (2019), A/RES/75/309 (2021) , A/RES/77/318 (2023) and the Human Rights Council, e.g., A/HRC/RES/49/9 (2022), A/HRC/RES/49/21 (2022).

The UN System has been working on the harms caused by hate speech[21] and false information for decades, particularly their impact in conflict settings.[22] The lessons from the Rwandan genocide against the Tutsi and the use of *Radio-Télévision des Milles Collines* and community information networks to spread hate and incite violence is an example of this.[23] More recently, the launch by the UN Secretary-General of the UN Strategy and Plan of Action on Hate Speech highlighted the increased focus and importance of tackling hate speech with the onset of social media. Over the past years, there has also been increased focus on misinformation and disinformation, a consequent response to the global role of new technologies in facilitating coordinated inauthentic behavior[24] to reshape people's perceptions, opinions and understanding of what is true.

The UN's response varies by entity, depending on mandate. For some agencies or departments, these types of speech have been combined, with some staff having MDH (representing their focus on Misinformation, Disinformation, and Hate Speech) in their job titles, while in other settings the work has been quite distinct, with some people, taskforces and workstreams focusing solely on hate speech, and others on misinformation and disinformation. This is caused in large part by the different contexts in which different UN entities work.

In recent months, the new term "information integrity" has also been utilised in the UN System and Member State circles. It is not designed to replace existing language, but to help bring attention to the need to proactively strengthen information environments — including the need to empower individuals and communities to communicate, be informed and make decisions, as a way of mitigating the potential harms of hate speech, misinformation and disinformation.[25] The UN Secretary-General's Policy Brief on Information Integrity on Digital Platforms,[26] calls on the international community to strengthen information integrity, described as information that is accurate, consistent and reliable. In addition, the UN Office of the Special Adviser on the Prevention of Genocide (OSAPG) has developed guidelines for technology and social media companies and Member States on tackling online hate speech, based on the UN Strategy and Plan of Action on Hate Speech and three years of dialogue with technologycompanies, the UN System and civil society.[27] UNESCO has further developed Guidelines for the Governance of Digital Platforms around content that can be permissibly restricted under international human rights law and standards.[28] The Governments of Canada and The Netherlands launched a Global Declaration on Information Integrity Online in the margins of the UN General Assembly in September 2023, endorsed by 30 Member States as of October 2023.[29]

This report attempts to provide a conceptual analysis of where these types of speech overlap and where they differ, with a view to identifying operational implications for how these different phenomena can be identified, mitigated, and addressed.

PARALLEL TENSIONS

There are two fundamental tensions when considering these three types of speech. Other than the very worst forms of hate speech (incitement to discrimination, hostility or violence or incitement to genocide), the majority of the speech that falls into these three categories discussed in this report qualifies as protected speech, in line with international human rights law that protects freedom of expression (See Part 2 of this report for an expanded discussion of the categories of hate speech that are restricted under international law.) Attempts to respond to this type of protected speech have the potential of causing certain consequences, specifically increased censorship, if not carefully designed and implemented – hence, the consequences of action need to be considered. Nevertheless, failing to address the potential harms of protected speech may lead, particularly over time, to incitement to discrimination, hostility or violence. Hate speech, misinformation and disinformation as well as some of the state responses to deal with it so far have had a negative impact on the humanitarian space, protection of civilians, human rights, and peace and security work, including the prevention of genocide. This tension is a recurring theme throughout the report.

The second tension involves different perceptions about how to think about definitions and categories. Some interview respondents were passionate about a need for an umbrella term to describe different types of potentially "harmful speech" and to stop thinking about hate speech "vs." misinformation and disinformation, as they expressed that the focus on definitions was causing "siloed working" and consequently preventing a recognition of the harms that all these types of speech exerted within the communities they were serving.[30] They saw the definitions as less useful because the speech they perceived as causing harm in their communities rarely fit neatly into categories. While there certainly might be a local manifestation of these information harms that do not fit neatly, there is still a critical need to understand and take into account the separate terms, in particular as it relates to the various international human rights law standards for the terms, as well as the impact on the targets of hate speech, misinformation and disinformation, when considering the response.

It was also noted multiple times that different mediums of speech were used to exacerbate harm. For example, one WhatsApp message that suggests a journalist is taking bribes, is made more harmful when it is accompanied by a doctored photo of the journalist holding a brown envelope, along with an offensive slur connected to the journalist's ethnicity. The harm is magnified when multiple actors are paid to send out that same image on multiple platforms, using the same hashtag, and applying coordinated tactics to "brigade"[31] the Facebook page of the newspaper for whom the journalist works. While some examples could be described as disinformation, and other examples specifically as hate speech, in other examples the two interact, for example, when hate speech is exacerbated by disinformation. This tactic amplifies the extent or reach of the speech and increases the likelihood of risk of harm. These are two key elements of the Rabat Plan of Action, which applies a threshold test based on six parameters to assess whether speech rises to the level of prohibited speech.[32]

Other respondents viewed the definitions as

critical to responding appropriately and effectively to the different types of speech, citing the very different historical and legal contexts. They were worried about collapsing these definitions and using an umbrella term as a catch all for any types of speech that might cause harm. A message that was repeated often by these respondents was the fundamental need for definitions in terms of differentiating between protected and restricted speech and developing nuanced and tailored responses to different types of harms, as well as their interactions. Another key differentiation is that hate speech is speech targeting persons or groups on the basis of identity, whereas misinformation or disinformation does not require this component. Mixing the terms therefore also risks blurring this important distinction.

These two positions are not direct contradictions. For those working on the ground, particularly in conflict settings, there needs to be as much collaboration across agencies and experts as possible. The focus has to be on understanding the ways in which harms are being caused by speech, or in which forms of speech inter-relate and influence each other, in order to mitigate those harms in real-time. At the same time, there needs to be a shared recognition of the clear definitional and legal lines that need to be relied upon for responding appropriately to different types of speech.

BOUNDARIES

In the context of this discussion on definitions, it's important to note that while the report is focused on these three types of speech, there are other categories of speech or tactics that cause harm, and/or exacerbate hate speech, misinformation and disinformation. An example would be malinformation,[33] speech that is based on genuine information, but shared deliberately with the intention of causing harm, such as the deliberate leaking of private information relating to an individual or organisation. Focusing too much on false or misleading information often means this type of strategic manipulation or handling of accurate information is ignored.

There is also a risk that legitimate criticism of an organisation, such as the UN, might be conflated with misinformation or disinformation or even hate speech (even though hate speech usually targets individuals and/or groups based on identity). Repeated attacks can gradually erode trust in the protections offered by the UN and ultimately harm those the UN is mandated to protect, especially in conflict settings. Therefore, when the UN analyses misinformation or disinformation, it should distinguish between genuine criticism and falsehoods.

Criticism of an organisation, however upsetting for those who work for the organisation, cannot be confused with hate speech, misinformation or disinformation against a person or community. Repeating actual examples of when a humanitarian organisation took actions that harmed a community is not disinformation. Sometimes repeating these examples is a genuine effort to hold an organisation accountable and to encourage change. Or, depending on the actor, their pattern of behavior and their likely motivations (e.g., to inflame anti-UN sentiment for political or other purposes), it could be described as malinformation — the strategic use of genuine information designed to undermine confidence in the organisation. Therefore, the UN needs to be aware of manipulation, including patterns of artificial amplification, negative sentiments about its work, and whether levels of trust with the communities they serve are decreasing, and how

this might be impacting protection of civilian activities or the implementation of other mandate areas. However, it needs to be categorised very differently from hate speech, misinformation or disinformation.

WIDER CONTEXT

Hate speech, misinformation and disinformation cannot be divorced from the context in which they are created, consumed or shared. Speech exists within a wider ecosystem, one that has a history that impacts the present, and one where different actors within the ecosystem have motives for creating, disseminating or making sense of it. It is important to understand the wider political, societal and human rights contexts in which such narratives are being utilised. And like any ecosystem, an information ecosystem is fragile and vulnerable when pressure is applied. Attempts to impact speech have consequences. Even those who have positive intentions about wanting to mitigate harms from certain types of speech might cause unintended consequences by creating justification for additional censorship.

It should also be noted that identifying and responding to hate speech, misinformation and disinformation, particularly within the case of conflict settings, is more nuanced. Those who are trying to cause harm take advantage of the chaos, the declining levels of trust in traditional gatekeepers, caused in part by deliberate campaigns designed to undermine trust in journalists, and the overwhelming fast-paced nature of digital networks (often amplified via low-technology environments, which leads people to rely on heuristics while falling victim to confirmation bias). Remembering the disconnect between the precise language of legal frameworks and the messiness in the online and offline information landscape is very important when discussing these issues and when developing operational responses.

However, definitions are important to inform appropriate responses by UN entities whose mandates are engaged. Responses to such narratives must be consistent with the UN Charter and international human rights law. Having a clear focus on definitions and categorisation of various types of speech within an information ecosystem helps to guide the appropriate range of responses through the messiness. This is particularly the case for responses where the UN engages with or supports responses, whether by states, civil society or private sector, including technology and social media companies, to ensure that they adhere to international human rights standards and principles, and ensure important considerations such as “do no harm”.

Attention, particularly within the context of the UN, is quite rightly focused on speech that could lead to imminent harm, because of core UN priorities and prevention of genocide and related crimes, as well as displacement, exclusion and unequal access to resources. However, there needs to be more consideration of the long-term impacts of content that may not rise to the level of incitement to discrimination, hostility or violence and imminence of harm according to the Rabat threshold test and does not pose an immediate threat. Examples of this might include misogynistic, racist, religiously motivated or xenophobic “memes” or coded language that are too often dismissed without recognising the cumulative impact of this type of speech on the victims and communities. A number of UN entities are working to tackle the root causes of hate speech and disinformation but too often the

focus on these types of speech tends to be on monitoring, understanding impact and working on regulation. There is a need for increased investment and support to help UN entities understand the root causes and sources, so as to recognise the nature of the risks they may be facing and to help plan potential responses designed to tackle those root causes. [34]

It is important to note the significant impacts hates speech, misinformation and disinformation are having on UN operations and priorities, including peacekeeping missions. Missions themselves are being targeted with significant consequences (most recently, the peacekeeping mission in Mali, MINUSMA), including speech targeted at types of staff, (for example: UN mission staff in the Central African Republic in 2020 being falsely accused of trafficking weapons; human rights staff or others undertaking sensitive work; national staff on the basis of their identity or work; and defamation campaigns targeting peacekeepers and specific individuals such as spokespeople),[35] as well as hate speech targeting the most vulnerable communities, including ethnic minorities, women and girls, as well as journalists or political and human rights advocates.

The strategic use of disinformation is designed to sow distrust, which is vital for maintaining peace and security and undermines the very foundation of any peace mandate. By leveraging genuine and legitimate grievances against peacekeeping, the animosity instigated in some mission areas has resulted in a contraction of consent from host state communities. In certain instances, this has resulted in demonstrations or "blockades" of UN convoys, restricting freedom of movement and negatively impacting mandate implementation, further putting at risk the civilians that some UN peacekeeping operations are mandated to protect. Indeed, disinformation has targeted and created mistrust and confusion around central peacekeeping tasks, such as the provision of support to ongoing peace processes, confidence-building or the protection of civilians.[36] However, this does not fit within the definition of hate speech, although in many contexts the same actors will use hate speech to also target vulnerable groups, particularly ethnic or religious minorities, undermining social cohesion. [37] Again, this provides an important reminder that the distinctions between hate speech, misinformation and disinformation are critical. Understanding the content, context and impact, as well as power structures and interests that the speech serves, is fundamental as the response required is very likely different.

Many UN entities, whether in mission or non-mission settings, often have protection mandates: protection of civilians (in UN Peacekeeping Operations); protection of other groups such as refugees and Internally Displaced Persons (IDPs) (UNHCR); and/or to protect and promote human rights for all (OHCHR). In both settings the UN would be mandated to address hate speech, including through human rights monitoring and reporting, individual protection of civilians under direct threat, as well as the use of good offices, engagement with and capacity building for authorities and community leaders particularly to foster civic space, and effective dialogue and facilitation. This is set out in the UN Strategy and Plan of Action on Hate Speech. The UN Office of the Special Adviser for Prevention of Genocide is the UN focal point for the Strategy and provides support to such national-level initiatives, including the development of context-specific action plans. The UN Department of Peace Operations (DPO) also provides support and guidance to peacekeeping missions on misinformation, disinformation, malinformation and hate speech through policy and guidance, trainings and monitoring and analysis expertise and tools.

Part 1: Reflections on the Contemporary Information Environment

ONLINE AND OFFLINE SPACES

A key consideration in initiatives to address hate speech, misinformation and disinformation is to examine the ways that information travels across digital platforms, gets passed into offline settings, and then moves back to digital spaces again. In many operational contexts the dynamic between online and offline spaces will differ, impacted by questions of inequality, connectivity, digital inclusion, and device access. This is particularly true in conflict settings. Understanding these considerations in each operational context, (for example does potential harmful information move primarily from offline spaces to online, or from online to offline?) is a priority for planning potential responses.

The mechanisms that have always been used to disseminate ideas (one-to-one conversations, small group discussions, public speeches, pamphlets, posters and newspapers, radio and television broadcasts, as well as various forms of art) continue to be potential vehicles for hate speech, misinformation and disinformation, and now, those mechanisms have been supercharged by a range of digital technologies that make the creation, distribution and amplification of content both low cost and frictionless, enabling the immediate distribution and amplification across vast geographic distances.

Digital technologies have also fundamentally changed who people can and do trust.

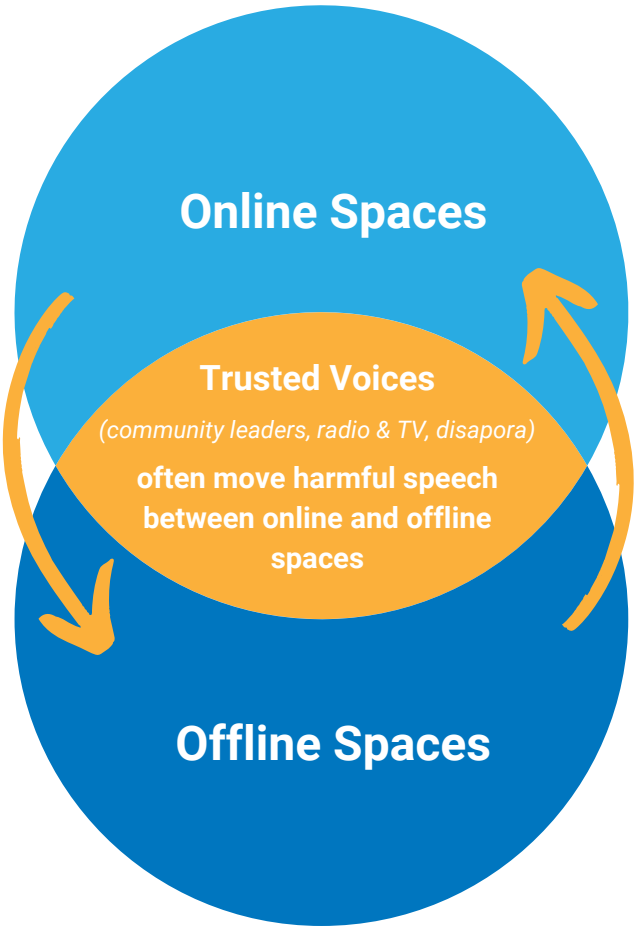

Figure 1: Diagram explaining the impact of some actors in moving speech from offline to online spaces and back again.

While people could always make choices about which "sources" (both personal and institutional) to trust, the choice was somewhat limited (and continues to be in countries where authorities exercise full control of the media and public debate within the borders they control). The Internet has transformed the information ecosystem, giving voice to people and communities, particularly those living in repressive regimes, who had few or no means to have their views heard, let alone amplified. However, despite the explosion of content available online, the Internet, through algorithmic filtering and the manipulation of online communities by actors attempting to build a following, can too often encourage the consumption of certain types of content that reinforces existing world views and encourages connection with others who share those views. Information, both online and offline, can be created and targeted by those deploying tactics of manipulation, who also know that the information will be filtered through existing world views, histories, and cognitive biases.

As underlined in the UN Strategy and Plan of Action on Hate Speech,[38] the focus should always be on the targeted victims, as well as the context in which speech is being disseminated — the historical, political, and social context as well as the strength and independence of the media sector, socio-economic levels, and levels of education and economic stability.

Digital technologies have enabled malicious behaviors that are impossible in offline settings, for example:

⚠ Actors can abuse someone or something anonymously, and hundreds of times per day, almost for free, from anywhere in the world.

⚠ Actors can use editing software to alter a photo to show something very different from the original.

⚠ Actors can track down someone's family via personal information available online, and share private details online to encourage offline violence.

⚠ Deployment of computational techniques including networked bots,[39] hashtag squatting and generative AI[40] can supercharge all of the above.

However, what happens online does not just stay there; information may have started offline and then been digitised, or these additional "opportunities" offered by digital technologies can move offline, meaning that even in settings with low digital penetration, only one person with access to the Internet or smartphone can exert influence. In mission contexts, there is frequently inequality of access to online technology, giving elites and the diaspora with such access, as well as protection from reprisals or other risks, outsized influence over the narratives' spread offline.

In addition to the free-flowing transfer of information between offline and online spaces, there are techniques (social media bots, deepfake technology, coordinated disinformation campaigns, to name a few) used to manipulate the information ecosystem to maximise amplification opportunities. The advertising business model that powers many social media platforms has resulted in platforms designed to maximise engagement. Actors attempting to manipulate the information ecosystem understand these design features intimately, and have perfected tactics to take advantage, and essentially weaponise those features.

In addition to these design features, those trying to manipulate understand how interconnected these platforms are, and the ease and speed with which information can travel across communities within platforms and then crossover into others. Such actors also understand the ways in which rumors can be seeded in small communities (both online and offline) with the intention that they will move quickly via trusted messengers and will then be amplified by politicians or mainstream media.[41]

This dynamic is a fundamental characteristic of the contemporary information landscape. This reality, originally described as "trading up the chain,"[42] and adapted by the author as "the Trumpet of Amplification"[43] provides a tool for determining the best moments to intervene and mitigate hate speech or misinformation and disinformation. Not all disinformation or hate speech is disseminated and amplified in this way, but deliberate campaigns are often seeded using these methods.

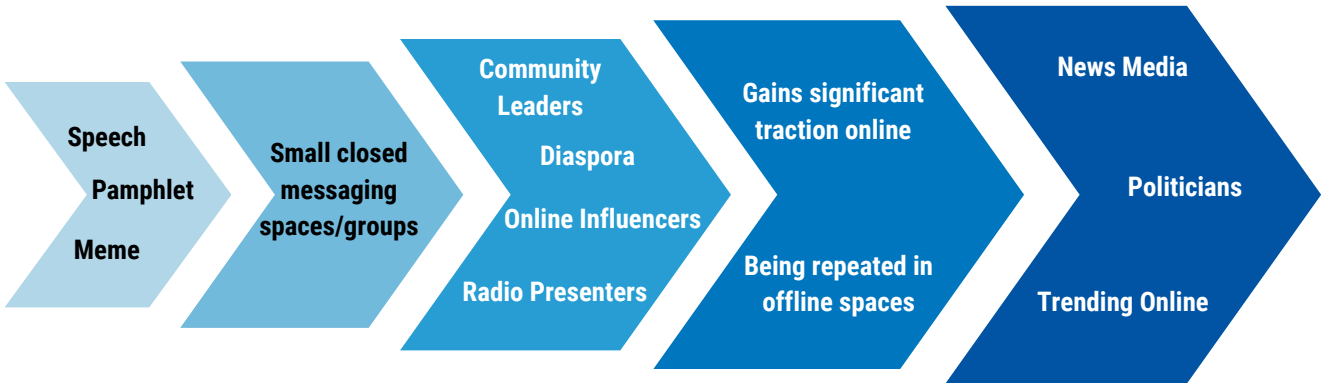

Figure 2: Adaptation of the Trumpet of Amplification to emphasise the ways information can move through the information ecosystem, flowing between offline and online spaces.

USE OF FRAMEWORKS TO MAKE SENSE OF THE INFORMATION ENVIRONMENT

The definitional muddiness has led a number of scholars and policymakers to develop and use alternative frames as a way of helping people understand the current challenges with the information environment. The most common frames are: information warfare,[44] ecological frame, and information needs frame.

INFORMATION WARFARE

Those who use the information warfare frame tend to be those working closely on issues related to foreign interference and information operations.

As the term propaganda became less able to capture the ways in which information systems, platforms and technologies were being weaponized in the service of ongoing conflict, and interference in foreign elections, terms including Coordinated Inauthentic Behavior,[45] Foreign Information Manipulation and Interference[46] and Foreign Malign Influence[47] are used to explain the ways in which information has become a weapon used by countries against one another.

ECOLOGICAL FRAME

Another frame is an ecological frame, namely a

pollution frame. Notably the work of the United Nations Development Programme (UNDP) work in[48] this area has focused on this Information Pollution frame. Comparing the spread of certain types of speech (in particular misinformation) to a toxin, pollutant or virus allows researchers to consider the response in the same way public health professionals think about slowing down the spread of a virus; for example, building "speech bumps" and friction into the design of platforms would be equivalent to social distancing.[49]

Relatedly, Whitney Phillips and Ryan Milner also write about the need to frame the problem through an ecological lens: "As it is impossible to eradicate our environments of all toxins or pathogens, the focus according to this paradigm should be on three factors: understanding the proportion of exposure to low- and high-integrity information, people's receptivity to finding misinformation credible, and a risk analysis to hone in on the toxins least likely to be abated."[50]

As the work of UNESCO stresses, "the rise of online misinformation and hate speech has shown that we must massively upscale efforts to teach people of all ages to think critically and click wisely in online spaces, and to understand the algorithms and processes that underpin them. In other words, empower people with media and information literacy." [51]

INFORMATION NEEDS FRAME

While much of the work over the past decade has focused on defining and identifying misinformation and disinformation in online spaces, more recently, there has been a focus on emphasising what can be done to strengthen and improve the entire information ecosystem. The idea is not new. Around 2009-2011, the work of Internews on "Mapping Information Ecosystems to support Community Resilience" received a great deal of attention.[52] Designed to help humanitarian organisations map information ecosystems during emergency situations, the assessment tool has been used in a number of contexts to identify where information needs are not being met. It provides a very useful starting point for mapping information ecosystems and identifying factors that make them more vulnerable to speech that has the potential to cause harm.

UNDP's work on its Information Pollution Mapping Programme builds on these ideas. It was designed in response to its country offices needing to "better understand what is motivating disinformation and how it is being produced and disseminated across different socio-political contexts." The country maps include media monitoring, data collection from sources tracking information pollution, online listening, review of online and offline influencers, analysis of levels of trust in public information sources, and review of initiatives and organisations working to monitor and counter misinformation and disinformation.[53] In addition, UNESCO's Social Media 4 Peace project has been piloted to better understand the impact of harmful content in conflict-prone environments and to pilot concrete initiatives.[54]

Additionally, UNDP has been working on preventing and addressing hate speech by understanding its drivers through monitoring and analysing online hate speech narratives, along with other areas of digital harms more broadly, including violent extremist narratives, through open-source research process. For example, through partnership with SecDev, UNDP has conducted Digital Ecosystem Mapping—a methodology involving the automated collection and analysis of publicly available online content from accessible social media platforms, segmenting the content by actor, audience, content, and theme, then analysing it over time — across Central Asia and Southeast Asia and is developing a methodological toolkit for measuring the impact of the digital harms to inform policies and programming around violence prevention.

As noted above, and as part of the implementation of the UN Strategy and Plan of Action on Hate Speech, in July 2023, the Office of the Special Adviser on the Prevention of Genocide launched a policy paper on "Countering and Addressing Online Hate Speech: a guide for policy makers and practitioners." The guide and its recommendations build on three years of engagement by the Office, bringing together technology and social media companies, UN Working Group on Hate Speech, Special Procedure mandate holders and civil society working on tackling online hate speech. The policy sets out recommendations addressed to Member States, technology and social media companies, within the UN System and for civil society under areas of:

(1) Ensuring Respect for Human Rights and the Rule of Law when Countering Online Hate Speech. Apply these Standards to Content Moderation, Content Curation and Regulation;
(2) Enhancing Transparency of Content Moderation, Content Curation and Regulation;
(3) Promote Positive Narratives to Counter Online Hate Speech, and Foster User Engagement and Empowerment;
(4) Ensure Accountability, Strengthen Judicial Mechanisms and Enhance Independent Oversight Mechanisms and;
(5) Advance Community-Based Voices and Formulate Context-Sensitive and Knowledge Based Policymaking and Good Practice to Protect and Empower Groups and Populations in Vulnerable Situations to Counter Online Hate Speech. [55]

The UN Secretary-General's June 2023 Policy Brief on Information Integrity on Digital Platforms has a similar focus on strengthening the information ecosystem. [56] The Brief calls for more evidence-based information from trustworthy sources, delivered via technology or media systems less dependent on algorithms designed for emotion and sensationalism. It focuses on improving the quality of information made available to people, using messengers that are trusted by the community to deliver the message, and limiting barriers to accessing information (for example digital divides, language, or cultural barriers).

58%

OF **UN PEACEKEEPERS**
SAID **MIS**INFORMATION
AND **DIS**INFORMATION
SEVERELY OR CRITICALLY IMPACTED
THE WORK OF THE MISSION

UN DPO 2023 INTERNAL SURVEY TO PEACEKEEPING MISSIONS [57]

Part 2: Framing Provided by International Human Rights Law and International Humanitarian Law

International human rights law and international humanitarian law are both fundamental legal frameworks when considering these issues and provide guidance for how to respond to these problems. The application of international human rights law at all times alongside international humanitarian law is vital for the effective protection of the right to freedom of opinion and expression, in both peacetime and during conflicts.

During armed conflict, international human rights law continues to apply alongside international humanitarian law, which is triggered only at the onset of armed conflict and is concerned primarily with the conduct of military operations and the protection of certain classes of persons in international and non-international conflicts. As such, international humanitarian law covers freedom of expression and access to information issues "only tenuously and non-systematically."[58] Human rights law, principles, and standards provide clarity and protection where international humanitarian law is silent, absent or unclear.

However, it needs to be acknowledged that the relationship between international humanitarian law and international human rights law is complex and multifaceted. For example, Human Rights Committee General Comment 29 on States of Emergency sets out the circumstances and limitations in which State parties can derogate from international human rights standards when declaring states of emergency, which can be applicable in situations of conflict.

Secondly, Human Rights Committee General Comment 34, addresses Freedom of Expression and discusses the circumstances in which this qualified right can be limited on grounds of national security and public order. [59]

INTERNATIONAL HUMAN RIGHTS LAW

The right to freedom of opinion and expression is a fundamental human right set out in the Universal Declaration on Human Rights, Article 19 and the International Covenant on Civil and Political Rights (ICCPR) Article 19. The right to hold opinions without interference is an absolute right from which no derogation is permitted as acknowledged by the Human Rights Committee. The Human Rights Committee has further emphasised that the freedoms of opinion and expression form a basis for the enjoyment of a wide range of other human rights, including the rights to freedom of assembly and association among others.[60] Finally, it enshrines both the right of people to impart information, but also their right to seek and receive information.

Resolutions by the General Assembly[61] and the Human Rights Council[62] acknowledge the importance of respect for freedom of opinion and expression in tackling hate speech and disinformation and recognise the positive impact of the right in addressing these phenomena. In addition, the UN Strategy and Plan of Action on Hate Speech has as one of its fundamental principles the promotion of more speech, not less, and the best means for addressing hate speech. Similar recommendations are outlined in the Rabat Plan of Action.[63]

Therefore, any measures to address hate speech, misinformation or disinformation need to emphasise the rights to freedom of opinion and expression and consider measures to address phenomena in line with this and related rights and the guidance provided by the international human rights law framework. [64]

Under international human rights law only the most serious forms of hate speech amounting to incitement speech are prohibited, in particular: (a) "direct and public incitement to commit genocide "[65] (b) "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence."[66] The Rabat Plan of Action and its six-part test (considering the context, speaker, intent, content, extent and likelihood of harm)[67] provides a framework for assessing whether speech reaches the threshold of incitement to hostility, discrimination or violence.[68] In addition, Article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination imposes a duty upon States to criminalise "all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin." Incitement requires a triangular relationship between the hate speaker, an audience, and the target group.

Article 20(1) of the ICCPR also includes specific prohibition of any propaganda for war. It should be noted that loose definitions of "propaganda for war" make this harder to operationalise. The prohibition is interpreted by the Human Rights Committee as extending to propaganda threatening or resulting in an act of aggression or breach of peace contrary to the Charter of the United Nations; it does not prohibit advocacy of the sovereign right of self-defense or the right of peoples to self-determination and independence in accordance with the Charter of the United Nations.[69]

In addition, certain forms of speech may be prohibited under international human rights law, even if they do not reach the above-mentioned threshold of incitement, in specific circumstances. Namely, under Article 19 of the ICCPR, certain types of expression may be restricted if such restrictions meet strict conditions. Such limitations need to: (a) be provided by law; (b) pursue a legitimate aim, such as being necessary for the respect of the rights or reputations of others or for the protection of national security or of public order (ordre public) or of public health or morals (c) be necessary in a democratic society and proportionate (the "three-part test").[70] Restrictions may therefore be imposed for example to protect individuals from hate speech based on their protected characteristics (or identity factors) in order to ensure their rights to equality and non-discrimination, but as long as the conditions of the three-part test are met. However, the Human Rights Committee has stressed that such restrictions may not "put in jeopardy the right [to freedom of expression] itself" – setting a high threshold for restrictions. Of note, the Human Rights Committee has further confirmed that restriction under Article 20(2) also need to meet the 3-part test as set out in Article 19.

In relation to disinformation, international human rights law protects the right to express information and ideas of "all kinds,", so falsity on its own is not a reason to limit expression. Article 19(2) of the ICCPR provides that: "Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any media of his choice." As noted by the UN Secretary-General: "State efforts to address the impacts of disinformation should avoid approaches that impose an undue burden on the freedom of

expression or are susceptible to politicised implementation. Not all inaccurate information is harmful, and only some harms – such as those that in fact implicate public health, electoral processes or national security – may warrant State intervention. Even when there is a legitimate public interest purpose, the risks inherent in the regulation of expression require a carefully tailored approach that complies with the requirements of legality, necessity and proportionality under human rights law."

INTERNATIONAL HUMANITARIAN LAW

As noted above, public and direct incitement to commit genocide is prohibited in the Genocide Convention;[71] the use of propaganda, misinformation and disinformation during armed conflict is not specifically prohibited under international humanitarian law, nor international human rights law.

International humanitarian law prohibits the encouragement of international humanitarian law violations (including war crimes), online or offline[72] and it prohibits "acts or threats of violence the primary purpose of which is to spread terror among the civilian population" (Article 51(2) Additional Protocol I and Article 13(2) of Additional Protocol II).

This may include attacks against media professionals for the sole purpose of intimidating them into silence. [73]

As the UN Secretary-General has noted, the impact of information operations during and relating to armed conflicts is of particular concern, although the phenomenon is not new.[74]

These harmful consequences have prompted calls for states to:

> **Take measures to protect the human rights of individuals within their jurisdiction from violation by information operations or activities carried out by other States and non-state actors.**
>
> UN Secretary-General, 2019[74]

Growing evidence of the use of this type of speech being used to exacerbate conflicts has led to calls for international humanitarian law to specifically deal with two issues.[75] First, the intersection of international humanitarian law and computational propaganda,[76] particularly in an age of Artificial Intelligence (AI). As Henning Lahmann argues in Protecting the Global Information Space in Times of Armed Conflict, the legal implications of information activities in the context of armed conflict against the background of the digital transformation have so far received only scarce attention. [77]

Second, the responsibilities of states that inject, spread or sponsor propaganda, disinformation or incitement to discrimination, hostility or violence from across borders in conflicts to which they are not a party. As the Special Rapporteur for freedom of expression and opinion argues, "Neither international human rights law nor international humanitarian law appear to have a clear answer to [this] thorny question." [78] Of note, the international human rights standards outlined in the previous section continue to apply in times of armed conflict and therefore need to be considered simultaneously.

AI GOVERNANCE

The speed at which AI, particularly generative AI, has emerged for mass use and consumption has created a regulatory vacuum around the specific harms that will be caused by this new technology. There has been a flurry of advisory activity and multistakeholder convenings, including an Executive Order by the US President,[79] and an AI Safety Summit hosted by the UK. [80] In December 2023, the UN Secretary-General's AI Advisory Body published its Interim Report on Governing AI for Humanity calling for governance of AI to be developed in accordance with the UN charter, human rights and international law. [81]

FIVE GUIDING PRINCIPLES IN THE UN SECRETARY-GENERAL'S REPORT STATE:

1. AI should be governed inclusively, by and for the benefit of all.

2. AI must be governed in the public interest.

3. AI governance should be built in step with data governance and the promotion of data commons.

4. AI must be universal, networked and rooted in adaptive multistakeholder collaboration.

5. AI governance should be anchored in the UN Charter, International Human Rights Law, and other agreed international commitments such as the Sustainable Development Goals.

Part 3: Recognising the Specific Context of Conflict Settings

Information ecosystems in conflict-affected and high-risk areas are exceptionally vulnerable to manipulation, due to fragile institutions, limited access to reliable information restrictions on the press, and technological vulnerabilities. The amount of hate speech or disinformation is often significantly elevated, while social cohesion and trust in authorities is often eroded, leading to the sharing of rumors[82] and misinformation by different actors. Critically, the impact and subsequent harms caused by this type of speech are exacerbated in these contexts, when tension and division are already high, and those consuming content are desperate for any information to make decisions to keep themselves and their families safe.

While platform policies are designed to apply equally across the globe, this is not the case in practice. Platforms have invested heavily in content moderation in major languages (and the failures in these major languages are well known),[83] but have failed to put in place necessary safeguards in other languages. This was highlighted most clearly during the serious human rights violations and violence committed against the Rohingya in Myanmar in 2017, violence which may amount to genocide or related crimes, when Facebook did not have the ability to adequately moderate content in Burmese and other local ethnic languages[84] (the same issues exist today in Myanmar, particularly on Telegram).[85]

It should be noted that in August 2022, in response to recommendations by the Facebook Oversight Board, Meta published its Crisis Policy Protocol to "codify [its] containment policy response to crisis.[86]" While this attempt by one platform to provide more resources around crises is a step in the right direction, it needs to be acknowledged that the moderation of content is less likely to be actioned in places where it is most needed.[87] In addition, the reality of potentially harmful speech being created and shared in offline spaces and in traditional information environments needs to be remembered.

As the Special Rapporteur on Freedom of Expression and Opinion's Report on Disinformation and Freedom of Opinion and Expression During Armed Conflicts writes:[88]

> **Disinformation, propaganda and hate speech are not peculiar to armed conflicts. They are used also at other times and spread in an amorphous way across the various phases and cycles of tensions and unrest that precede or follow armed conflicts. The underlying causes of conflict, namely, historic grievances, systemic inequalities, discrimination, intercommunal and ethnic rivalry, political tensions, and poor governance, provide a perfect breeding ground for them. The dynamics of division, polarisation and dehumanisation that characterise violence and conflicts sustain and are sustained by such manipulation of information.**
>
> Special Rapporteur on Freedom of Expression and Opinion, 2022[88]

While information manipulation in conflict settings has a very long history, the impact of new digital tools and techniques powered by computational mechanisms is beginning to be felt.

The Special Rapporteur continues: "There is growing evidence that information is being manipulated to trigger, aggravate, and sustain violence over prolonged periods, increasing the fog of war with contradictory and false news and fostering a climate of distrust. The dynamics of armed conflict and disinformation work in a complex interplay with other grievances to exacerbate human suffering, feed hatred and target vulnerable groups."[89]

In addition, the disproportionate impact on minorities and marginalised groups in relation to intersectionality has been emphasised repeatedly by the UN Secretary-General, particularly through the Strategy and Plan of Action on Hate Speech;[90] by the Special Adviser on the Prevention of Genocide,[91] and by many other UN entities[92] as well as special procedures mandate holders. In this context, the gravity of hate speech may increase when somebody is targeted for multiple reasons. For example, an under-privileged minority, disabled woman may be targeted because of her race, sex and disability status simultaneously.[93]

As noted by the UN Secretary-General:[94]

> **Hate speech incites violence, undermines diversity and social cohesion" and "threatens the common values and principles that bind us together. … It promotes racism, xenophobia and misogyny; it dehumanises individuals and communities; and it has a serious impact on our efforts to promote peace and security, human rights, and sustainable development.**
>
> UN Secretary-General, 2022[94]

(1) The tinderbox nature of conflict-affected and high-risk areas is critical to understand. Hate speech, misinformation and disinformation that might not be harmful in other settings can quickly become exceptionally harmful in these settings. As one interviewee stated: "When people have nothing to lose, when they have lost their family, their business and community to a conflict, they're very quick to turn to violence when provoked." This in turn increases the Rabat threshold factor of likelihood and imminence of harm.[95]

(2) These types of speech, particularly speech that is created and disseminated via digital technologies, has fundamentally changed the ways conflicts are being fought. As International Committee of the Red Cross (ICRC) notes: "[Misinformation, Disinformation and Hate Speech] also illustrates the presence of new, unconventional, anonymous and/or non-local actors in conflict and humanitarian ecosystems. These may include various segments of society that do not fall into any traditional category of conflict actor, such as technology companies, the private sector, or digital influencers. Each can play different roles in propagating information, raising questions as to their respective responsibilities to prevent, mitigate and respond to [Misinformation, Disinformation and Hate Speech]."[96] This element corresponds to the Rabat threshold factor of the speaker and their status in society.

(3) Many of those interviewed shared a belief that international human rights law is insufficient to address these types of speech, or if it is sufficient, the challenge remains the application of this framework in day-to-day work. There appears to be a need for increased understanding of the application of the relevant provisions of international human rights law regulating freedom of expression to understand further the scope of relevant responses.

(4) When deployed in pursuit of a strategic objective by those creating disinformation, harmful online speech can be combined with tactics in the physical world, blending fact and fiction, and creating a new reality. For instance, the mobilisation of the population by civil society organisations with little organic buy-in can be streamed online or televised to amplify the impression of grassroots support,[97] or fake or coerced testimonies and other forms of false "evidence" of wrongdoing circulated on social media.[98] Going beyond speech, hate speech and disinformation can use multidimensional tactics to create harm.

(5) In some cases, a key tactic is to limit the effectiveness of organisations working on peace, security, or humanitarian affairs on the ground. This can lead to these organisations being the subject of disinformation campaigns, staff affected by harassment and defamation, and their operations subjected to movement restrictions or other operational obstructions. It also needs to be noted that a number of high-profile scandals involving international organisations working in conflict settings provides a fertile ground for those trying to undermine trust in these organisations to misrepresent narratives that are based on "kernels of truth." Finally, some of these conflict settings also have a long history of problematic colonial rules, enabling the use of powerful narratives that draw on these deep and painful histories to shape the way contemporary conflicts and international actors are being perceived.

(6) Lower Internet penetration, low levels of trust in media, and lower literacy rates in many of these settings can give offline spaces significant influence and amplify the power of those with online access. Rumors can quickly spread from pamphlets to speeches by community leaders, local radio broadcasts, comments by politicians, and ultimately reach national media. Even in countries with limited digital access, those who are online (both domestically and in the diaspora) can still influence offline communities. Platforms like X (formerly Twitter), though not widely used by the general population, can shape news coverage, which is then further disseminated through traditional media and private communication channels such as WhatsApp or Telegram groups.[99]

Part 4: Examining the Similarities and Differences between Hate Speech, Misinformation and Disinformation

ILLUSTRATIVE GUIDE

This section will outline the similarities between the different types of speech, but then underline why they need to be considered as separate and distinct entities.

In the following table, the different elements of investigating hate speech or misinformation and disinformation are listed. This is meant to be an illustrative guide. Not all categories are mutually exclusive because those trying to cause harm take advantage of the overlaps that can exist between these categories.

Actors: Who is involved and impacted? (Sharer of speech, Intent of sharer, Target of speech, Audience of speech)

Content: What has been created?

Distribution: What mechanisms are being used to encourage the distribution of misinformation and disinformation or hate speech? (Platforms and Tactics)

Harm: What damage could be caused?

Table 1: Table designed to help identify whether an example of speech should be defined as hate speech, disinformation or misinformation.[100]

ACTORS: Who is involved and impacted?			
	Hate Speech	*Disinformation*	*Misinformation*
Sharer of Speech	• Private Individuals • Community Leaders: *e.g. Local Business Owner, Faith Leader, Civic Representative* • Influencers: *e.g. Celebrity, Well Known Person* • State Actors: *e.g. Politician, Official Spokesperson, Military Leader* • Non-State Actors: *e.g. Armed Group, News/Media organisation, Activist Group*		
Sharer Intends to Harm	Yes	Yes	No

<table>
<tr><td rowspan="3">Target of Speech</td><td colspan="2">An Individual</td></tr>
<tr><td>A group or individuals targeted because of their identity with speech that attacks or uses prerogative or discriminatory language</td><td>Individuals or a group of individuals with a common trait (e.g. occupation) targeted with false or distorted information. This does not need to be hateful or discriminatory
State actor or organisation
Non-state actor or organisation
Certain facts (contemporary and in the past) e.g. Holocaust and genocide denial
A value or ideal (e.g., democracy, science)</td></tr>
<tr><td>Example: A well-known journalist, activist, politician targeted because of their identity (e.g. race, gender, religion) with speech that is discriminatory</td><td>Example: A well-known journalist, activist, politicians targeted because of his/her occupation with mis- and disinformation</td></tr>
<tr><td>Audience of Speech</td><td colspan="2">Limited Audience
Would reach a defined community, either in person or online
Large, multi-community audience with the likelihood of repetition over multiple days or weeks
National or Global Audience (via news or exceptionally large online audience)</td></tr>
<tr><td colspan="3">CONTENT: What has been created, with intent to harm?</td></tr>
<tr><td rowspan="2">Content</td><td>Hate Speech</td><td>Disinformation</td></tr>
<tr><td colspan="2">Denial and distortion of some historical events (for example the Holocaust or other genocide demonstrated by international court of law)
Content designed to emphasise in-group/out-group differences
Harmful content created using AI (e.g., deep fakes)
Explicit calls to violence based on identity, including genocide
Explicit calls to discriminate based on identity
Content designed to demonise and/or dehumanise based on identity</td></tr>
</table>

Content	• Explicit recommendation to take action that would cause someone harm based on identity • Use of Dog whistles and coded language related to identity • Use of identity-based slurs • Using words or phrases designed too evade content moderation	• Content designed to deceive or evade *(e.g., sharing false claims; creating false accounts; deceptive editing)* • Content designed to mislead *(e.g., cherry picking statistics, editing quotes, out of context images)* • Content designed to undermine trust in institutions and official processes *(e.g., conspiracies)*
DISTRIBUTION: What platforms and tactics are being used to encourage the distribution of disinformation or hate speech?		
Platforms	*Hate Speech*	*Disinformation*
	• **Offline mechanisms**: speech, pamphlets, posters, peer-to-peer conversations Broadcast mechanisms: radio, television, newspapers • **Closed digital spaces:** encrypted messaging apps, online groups and communities • **Public digital spaces**: video sharing apps, social networks, online advertising • **Disinformation:** forgeries, paid demonstrations, fake testimonies, use of inauthentic accounts • **Other:** academic conferences and journals and various types of art, graffiti, memes, songs	
HARM: What Damage Could Be Caused?		
Harms	*Hate Speech*	*Disinformation* / *Misinformation*
	• Impacts on both physical (loss of life or injury, including sexual violence) and mental health • Self-protective or forced withdrawal of certain groups of people from the public square (offline or online) (e.g., female politicians, journalists, voters) • Serious reputational damage to the target of the speech, which can lead to barriers to accessing rights and services, restrictive or discriminatory measures, acting as a trigger for additional disinformation and/or hate speech • Increased societal polarisation and climate of fear • Heightened hostility and hatred against the target of speech • Loss of morale and self-belief amongst target group	

Harms	Demonisation, dehumanisation, marginalisation and discrimination against the target of the speech	Decline of trust in an institution or value (e.g., belief in science, democratic values, freedom of expression)
	Hate crimes or violence against targets of hate speech	Undermining belief in peace and cooperation and increasing justification for conflicts and barriers to integration, possibility of triggering violence and conflict
	Self-protective or forced displacement; segregation of communities and creation of identity-based enclaves vulnerable to violence	Resistance to public policy measures (e.g., climate, public health) leading to different material impacts (financial sector/economy, employment, environmental).
	Genocide, crimes against humanity or war crimes	Elections impacted (including votes suppressed, distortion of policy debates, results not considered legitimate).

THE SIMILARITIES BETWEEN MISINFORMATION AND DISINFORMATION AND HATE SPEECH

This section describes the similarities between the different types of speech.

1. Absence of definitions that can be easily operationalised

While the UN has working definitions, in particular for hate speech, many interviewees drew on different definitions from the literature, some lumped the terms together, some used generalised frameworks such as information pollution, infodemic or information disorder, or harmful information. Misinformation and disinformation were often used interchangeably. In interviews with people working in conflict settings, there were often references to rumors, which were used differently from misinformation. As one interviewee explained, they used the term "rumor" to describe the type of falsehoods that are so ingrained in a society that it is no longer possible to investigate the "creator" of the original piece of disinformation.[101]

While the definitions used by the UN give conceptual clarity and are critical for designing appropriate human rights-based responses, in practice distinguishing between hate speech, misinformation and disinformation can in some contexts be challenging.

Take an example of someone posting a falsehood about the personal life of a person belonging to a particular identity group. Without any further investigation of the motive of the person, it can be difficult to define that example as hate speech, disinformation, or misinformation.

However, it's very important that those attempting to mitigate the potential harms understand how to operationalise these definitions, even if it's time consuming.

They need to investigate whether:

- The person who posted the content created the content.
- If not, who was the original creator?
- What was their motive?
- What was their intended and actual harm?
- And if the post continues to be disseminated by more people, do they understand and share that original motive and do they intend to cause further harm, and if so, who is the target?
- And is the individual post part of a wider campaign on multiple platforms and offline spaces?

It is important that patterns of tactics, techniques and procedures that may have been used to spread false information are considered and analysed. It is likewise important to understand whether the speech was directed at the person because of their religious identity and if there is a broader context and/or risk of hate speech against this religious group. Understanding these elements is crucial when thinking about a response and deciding whether additional information is needed.

2. Hate speech, misinformation and disinformation can have cumulative impacts

As has been noted frequently in this report, hate speech as well as misinformation and disinformation can have very severe, immediate impacts. However, it's also important to recognise that "low-level" examples of all three types of speech have the ability to cause severe harm over very long periods of time. With hate speech, years of demonising and dehumanising speech can create the conditions under which genocide, and other related crimes, are more likely to occur. Similarly with disinformation, a drip, drip, drip of conspiratorial thinking that, for example, connects international agencies with secret plans to depopulate nations[102] can undermine trust in the same institutions tasked with protecting civilians in those same locations. And years of harassment against female politicians and journalists can result in women removing themselves from public office or ceasing to remain a public figure.

As Susan Benesch argues in her work on "dangerous speech," she writes: "[I]t isn't the case that speech is either dangerous or not dangerous at all. Rather, more or less dangerous speech can be imagined along a spectrum, or like dominoes in which each piece affects its neighbor. As people come to accept a moderately dangerous message, they also become a bit more likely to accept an even more dangerous one. In this way, normal social barriers to violence erode as increasingly dangerous speech begins to saturate the social environment."[103]

3. Some similar tactics for creation, dissemination, and amplification

Many interviewees discussed the ways in which hate speech and disinformation have always been a challenge in their work, even before the advent of digital technologies; that peer-to-peer rumors, posters, speeches by leaders, and radio could do enormous damage. But there was a recognition that the introduction of digital technologies has taken these dynamics and supercharged them in different ways. For example, digital manipulation of images and videos has allowed free, high-quality production

tools for creating harmful content. The global nature of the Internet has allowed foreign actors to interfere in conflict settings anonymously, as well as enable highly trusted and technologically savvy diaspora or elite voices to get involved in and often exacerbate local disputes.

That being said, disinformation displays some distinct tactics that attempt to generate new versions of reality, including through the use of forgeries, paid demonstrations, inauthentic digital accounts and/or fake testimonies.

Interviewees discussed the challenges of differentiating hate speech from disinformation, as the same actors, tools, distribution, and amplification techniques were often used for each type of speech (see Table 1).

The practices honed in the pre-digital era have now been amplified, and the harms that were occurring offline are now reaching many more people in online settings, which in turn risks inspiring copycat behavior and generating additional risk of harm. But while the technologies, tactics and techniques are similar for the creation, dissemination and amplification of disinformation and hate speech, it's critical that the motives of the actors, and the harms caused are carefully considered because of the different response frameworks available.

THE DIFFERENCES: HOW HATE SPEECH DIFFERS FROM MISINFORMATION AND DISINFORMATION

While the previous section outlines the similarities between the different types of speech, this section underlines why it is so important to consider the ways the various types of speech are different.

1. International law

International law treats certain types of hate speech — incitement to discrimination, hostility and violence — very differently to other types of hate speech and misinformation and disinformation.

Certain types of hate speech (reaching the threshold of incitement) are prohibited by international law. There is no such corresponding obligation for misinformation or disinformation as it is not defined explicitly in international law.[104] Only disinformation or misinformation that also falls within the requirements of incitement, or that falls within the scope of article 19(3) of ICCPR can be prohibited. This is a critical distinction when considering the different types of speech and planning responses.

There are a number of regional and national laws related to misinformation and disinformation that have been passed over the past seven years, but the lack of a clear, operational definition of these types of speech has led to some criticism that, in these contexts, legislation overreached and prohibited protected speech and the laws had a chilling effect with real impact, including arbitrary detention of those alleged to spread misinformation or disinformation.

2. Identity vs non-identity characteristics

Hate speech targets people, whether individuals or groups, based on their identity; disinformation can be used as a tactic for spreading hate, but can also target individuals and groups based on non-identity characteristics (such as occupation). It can also targets institutions, specific facts and values and belief systems.

Hate speech impacts people, and the impact of discriminatory or pejorative speech can be tremendous, causing death, extreme pain and suffering for the victims. In the most serious cases it can also lead to the commission of

identity-based violence including genocide andrelated crimes. In fact, hate speech is recognised as an indicator of risk and potential trigger of such crimes.[105] In this sense, for those UN entities responsible for protecting civilians, human rights or other protection mandates, hate speech constitutes a threat and risk of human rights violations/abuses, including in the most serious cases, risk of genocide and related crimes to which missions have a duty to intervene. In essence, it triggers the UN's responsibility to prevent and respond to harm.

In comparison, disinformation can also target states, governments, inter-governmental and non-state actors and institutions, beyond legitimate criticism or opposition. Tactics are frequently designed to damage reputations and ability to carry out work by specific organisations. As well as "entities," disinformation can be targeted at "value systems" or "concepts," such as democratic systems and processes or the belief in freedom of religion, climate change or public health emergencies such as the COVID-19 pandemic or the Ebola epidemic. In this sense, disinformation becomes an operational and reputational risk for targeted UN entities, distinct from their respective mandates. In essence, it inhibits the UN's ability to prevent and respond to harm.

At times, disinformation is used to target individuals or communities, but if the false information is specifically aimed at discriminating against a person or a group, it should also be considered hate speech, and responded to through international human rights law, if it meets the threshold of incitement. For example, increasingly we're seeing how disinformation and hate speech are being used to target and harass women working in journalism. Sometimes this is labeled misogyny[106] and other times "gendered disinformation."[107] In these cases, while it might be tempting to frame the speech in the context of disinformation, it should be considered using international human rights law as it relates to prohibited speech, as the target is being discriminated against on the basis of their identity, in this case their gender.[108]

3. Hate speech causes different harms

As discussed above, hate speech targets people and groups, whereas disinformation can also target governments, institutions, organisations, specific facts, and value systems. Both can have very serious consequences and there is of course overlap, but the nature of the harm is different. Hate speech causes personal and communal harms in the form of a variety of human rights violations and abuses, and in some cases may lead to the commission of genocide and related crimes.[109]

Indeed, hate speech and incitement to violence are early warning signs that should be monitored by the UN as a risk factor for such crimes.

Disinformation can, by contrast, cause personal, institutional, or societal harms. For example, it can result in the undermining of an election result, the destabilisation of UN Peacekeeping Operations, or increasing distrust around the use of vaccines. In addition, disinformation on a number of different topics significantly impacts the ability of people to access accurate information in order for them to make informed decisions. Disinformation may also harm access to information, and thus the right to freedom of opinion and expression. In doing so, it may undermine the civic space, particularly when crowding out a plurality of views.

4. Online platform policies

In the digital age and particularly in the "Global North," multinational online platforms regulate a

wide range of potentially harmful content, including hate speech, misinformation, disinformation, calls to violence,[110] coordinated inauthentic behavior, bullying,[111] and invasions of privacy.[112] Notably, many platforms have different policies for hate speech in comparison to misinformation and disinformation, and other categories of content.[113] Platform policies on hate speech have emerged over the past 15 years, with help from human rights specialists, academics, learnings from serious mistakes and missteps around certain events (Myanmar in particular). Policies related to misinformation and disinformation have been rolled out since 2016.

Though platforms target the same kinds of speech, they take different approaches. Irrespective of the policies as written, researchers, human rights investigators and journalists continue to demonstrate how the platforms fail to enforce their own policies. This is particularly the case in countries with minority languages, or in countries less likely to regulate technology platforms.[114]

Across all platforms there is a preference to focus on behaviors rather than content. This is particularly the case with Meta, which has a large team looking at "coordinated inauthentic behavior"[115] — a term that is broader than disinformation. For misinformation, it shies away from explicitly defining the term, or attempting to "articulate a comprehensive list of what is prohibited."[116] Meta also created an independent Oversight Board in 2020, tasked with advising the company on content moderation decisions globally.[117]

X (previously known as Twitter) and YouTube use the term misinformation, not disinformation, which is somewhat strange considering both platforms talk about wanting to tackle content that is deliberately designed to be false or misleading with the intention of causing harm. X has a crisis misinformation policy,[118] but rather than addressing inaccurate content, it focuses on elevating reliable information during armed conflict, public health emergencies, and large-scale natural disasters. It also has a manipulated and synthetic media policy,[119] designed to tackle deepfakes and other forms of generative AI.

YouTube has a misinformation policy, and three additional policies related to elections, COVID-19 and vaccines.[120] It has an alternative definition of misinformation which is expanded to include disinformation by referring to: "Certain types of misleading or deceptive content with serious risk of egregious harm are not allowed on YouTube. This includes certain types of misinformation that can cause real-world harm, like promoting harmful remedies or treatments, certain types of technically manipulated content, or content interfering with democratic processes."[121]

TikTok has an Integrity and Authenticity section in its Community Guidelines that includes specific guidance around: misinformation, civic and election integrity, synthetic and manipulated media, fake engagement, unoriginal content, spam and deceptive account behaviors. But for the misinformation section, for example, the advice is disturbingly vague, stating: "Content is ineligible for the [For You Feed] if it contains general conspiracy theories or unverified information related to emergencies. To be cautious, content that warrants fact-checking is also temporarily ineligible for the 'For You Feed' while it is undergoing review."[122]

Platforms' definitions of hate speech also take different tracks. Meta's is a model of specificity, outlining a three-tier system, giving over a dozen examples of "dehumanising speech" alone: comparing people to "[i]nsects (including but not

limited to: cockroaches, locusts)," "[f]ilth (including but not limited to: dirt, grime)," or "[f]eces."[123] YouTube casts a much wider net — prohibiting "dehumanising . . . by . . . comparing [people] to animals, insects, pests, disease or any other on-human entity"[124] — as does X, which forbids simply "dehumanisation."[125] TikTok includes a section on Hate Speech and Hateful Behaviors in its Community Guidelines, stating it does not allow hateful behavior, speech or the promotion of hateful ideologies based on "caste, ethnicity, national origin, race, religion, tribe, immigration status. Gender, gender identity, sex, sexual orientation, disability or serious disease."[126]

WhatsApp has Terms of Service that outlines prohibited activities that it they would take action against if reported. These include "sharing content (in the status, profile photos, or messages) that's illegal, obscene, defamatory, threatening, intimidating, harassing, hateful, racially or ethnically offensive, or instigates or encourages conduct that would be illegal."[127] Telegram, which also has an encrypted messaging functionality, also has "channels" which allow users to broadcast to millions of followers. Telegram has a much less hands-on approach, focusing on illegal content such as "child abuse," "violence," and "illegal drugs." It also discusses users being able to flag fake accounts, examples of copyright infringements or spam. But there's no mention of hate speech, misinformation or disinformation.[128]

Platform policies are currently undergoing a period of change, motivated by Elon Musk's purchase of Twitter, his disbanding of the Trust and Safety Council[129] and leaving the European Union (EU) Code of Practice on Disinformation.[130] He also rolled back certain policies, including ones around COVID-19 and election misinformation, which has opened the door to other platforms loosening their policies.[131]

The recent impact of the EU's Digital Services Act does provide some hope that platforms will be required to address their actions related to potentially harmful speech.[132]

5. Different prevention and responses needed

The complexity of the information environment dictates that those working on the front lines of monitoring, countering and responding to hate speech, misinformation and disinformation need to be given the tools and support to differentiate between different types of speech, including coordination and information-sharing, and adequate human, technological and financial resources for UN actors monitoring these different types of speech. As underlined regularly throughout this report, speech that is circulating cannot be defined simply through a content analysis. The speech and the power structures or interests its dissemination serves, must be investigated. Who created it? What was their motive? Who or what is the target? What harm are they trying to cause?

The investigation helps to identify the response. If it's a rumor targeting a particular ethnic or religious group, the response will likely involve building trust between community or faith and traditional leaders through dialogue and other activities, and where relevant helping them to craft effective debunks or counternarratives. If it's a coordinated inauthentic campaign, funded and created by a state actor, the response could involve engagement with political leaders, leveraging Member State political and diplomatic tools or flagging inauthentic behavior to media outlets, including offline and online platforms. If the speech involves any type of incitement to violence or discrimination based on identity, the response might be a legal one, and it will require evidence that can be used in a court of law.

For all three, there may be baseline preventive and response actions that may be implemented.

Hate speech, misinformation and disinformation intersect with various thematic areas – from conflict prevention and resolution, to human rights to digital technologies. Different types of speech may also occur simultaneously in a given context; for example, a UN Peacekeeping mission may simultaneously need to respond to hate speech targeting a particular ethnic community during an ongoing election campaign in which misinformation and disinformation abounds, and in an overall context in which the mission is the target of disinformation campaigns concerning the effectiveness of its mandate implementation and questioning the UN's neutrality. Therefore, multiple responses to each type of speech should be considered, giving priority to the speech with the highest risk of physical harm to civilians, depending on the situations on the ground.

There are responses designed to pre-empt or counter the root causes of hate speech, but when disinformation is deliberately used to exacerbate the impact of hate speech, those tactics should be taken into account and will shape the response. For example, a response to hate speech using religious hatred and intolerance might be to build a closer relationship with faith leaders. If some of that hate speech is relying on an outright falsehood organised through an automated network, it would warrant reporting the suspected networked behavior to the social media platforms for removal. It would also be advisable to brief the faith leaders about the existence of these networks so they could talk to their communities about the deliberate attempt to manipulate opinion.

In conflict settings, the focus is on mitigating the root causes of the divisions that are being inflamed by the potentially harmful speech, through peacebuilding activities, supporting civic space and media freedom as well as digital and information literacy. Staff from across the UN System are also implementing awareness-raising initiatives with different actors and engaging with communities to understand what harmful speech is and the impact it is having on the country or region. Furthermore, there is an ongoing need to push for increased transparency across the information ecosystem, specifically regulation focused on transparency and accountability for all players.

In addition to this pre-emptive work, when hate speech, misinformation and disinformation circulates, various officers on the ground take action. Human rights officers, political affairs officers, public information officers, protection of civilians advisors, and peace and development advisers make crucial decisions. They determine when to debunk false information and when to engage with those disseminating it to encourage removal or correction. They assess whether to report to offline and online platforms to seek removal, demotion, or demonetisation according to moderation policies. They also decide when to advocate for other response measures, including legal or diplomatic action.

Recognising when responses might overlap requires expert knowledge and experience of monitoring and responding to different types of speech in conflict settings, particularly as speech can sometimes act as an early warning signal that violence is becoming more likely or more severe. There are many instances when hate speech is being exacerbated by disinformation; therefore, understanding the most appropriate response in these cases is critical for anyone working in these contexts. Due to the overlapping nature of these kinds of speech, there is certainly a need for coordination across UN entities.

Conclusions

This paper has provided a conceptual analysis on the similarities and differences between hate speech, misinformation and disinformation, with a particular emphasis on conflict-affected and high-risk areas.

Examples of hate speech, misinformation or disinformation rarely fit neatly into the categories outlined in reports like this and rarely align with the precise language set out in legal frameworks. Potentially harmful speech is messy, particularly the local manifestations of those harms, and it is often tempting to collapse these three categories into a wider umbrella category with an emphasis on the harm rather than a thorough examination of the content or actor that created or shared that content.

While acknowledging this reality, this report firmly argues that it is critical that the three types of speech outlined here be treated as distinct types. While there are a number of ways that these types of speech appear similar, a deeper dive into the different elements of speech, alongside their legal and historical contexts, emphasises how we have to think about them as different phenomena. Certain types of speech may have different legal and policy frameworks, have different targets, cause different harms, and require different responses.

Disinformation and hate speech require intent on the part of the entity that creates and shares the content. Misinformation requires no such intent. Certain types of hate speech (that reach the threshold of incitement) are prohibited by international law, whereas no such international legal framework exists for misinformation or disinformation. The target of misinformation and disinformation can be individuals as well as organisations, whereas hate speech is directly at a person or group based on their identity.

An example mentioned above provides the most compelling explanation of why terminology matters. Using the term "gendered disinformation" rather than misogyny fundamentally changes the way people think about the speech. Framing it as misogyny forces a recognition that someone is being targeted with hate speech on the basis of their identity, and if there is evidence that it might lead to incitement to discrimination, hostility or violence, there may be legal remedy. If framed as disinformation, where there is no available international legal response, there might be less consideration of those options.

By blurring these distinctions, we make it considerably more difficult for those working in conflict-affected and high-risk areas to understand how to hold accountable those who create and share these types of speech. It also complicates the design and implementation of appropriate preventative strategies, whether that is building education curricular to roll out with community leaders or the news media, so there is awareness of the legal repercussions of hate speech that crosses the threshold to incitement to discrimination, hostility or violence or whether its advocating platforms implement stronger content moderation for hate speech over misinformation.

As the potential harms of generative AI become clearer, these foundational concepts are even more important to ground discussions.

Generative AI will make easier the creation of hate speech and disinformation, and it will superpower the targeting of that content. The result will be more inadvertent sharing of false and misleading sharing of disinformation or misinformation. Rather than focusing on the new, shiny technology harm, we should understand that AI will reinforce the tactics, techniques and strategies that are already being deployed in conflict settings.

The challenge for those working closely on these different types of speech is to find a way to operationalise a response that leverages the expertise of specialists in hate speech, misinformation and disinformation. This report, in particular Table 1, is designed to help those working in conflict settings, as well as policymakers and researchers to understand these critical differences.

The characteristics and complexity of the contemporary information environment dictates that those working on the front lines of monitoring, countering and responding to hate speech, misinformation and disinformation require adequate human, technological and financial resources to differentiate between, prevent and respond to, different types of speech. A critical aspect of this work includes workflows and processes that encourages coordination and information-sharing.

This report provides a shared language and framework to help ground conversations across the UN and partner organisations around issues related to these three types of speech. This report provides a starting point for building workflows and processes that acknowledge the differences between these three types of speech, while recognising that speech designed to cause harm is deliberately crafted to avoid fitting neatly into definitional categories. It also sets the stage for a subsequent policy paper focused on specific response strategies, aiming to provide actionable insights for policymakers and practitioners engaged in mitigating the harmful effects of hate speech, misinformation and disinformation in conflict-affected and high-risk environments.

End Notes

[1] UN (2019) UN Plan of Action and Strategy on Hate Speech. https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/
Action_plan_on_hate_speech_EN.pdf

[2] United Nations. June 2023. Our Common Agenda Policy Brief 8: Information Integrity on Digital Platforms, p.5
https://www.un.org/sites/un2.un.org/files/ourcommon-agenda-policy-brief-information-integrity-en.pdf

[3] Report of the Secretary General. December 2021. Countering disinformation for the promotion and protection of human rights and fundamental freedoms. https://documents-dds-ny.un.org/doc/UNDOC/GEN/N21/416/87/PDF/N2141687.pdf

[4] OECD (2016), OECD Due Diligence Guidance for Responsible Supply Chains of Minerals from Conflict-Affected and High-Risk Areas: Third Edition, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264252479-en

[5] Information pollution – "Information pollution refers to false, misleading and manipulated online and offline content, which is created, produced and disseminated intentionally or unintentionally, and which has the potential to cause societal or physical harm. An overabundance of information and a high incidence of low-quality information within an ecosystem reduce our ability to find and trust information. Information pollution can be categorized as disinformation, misinformation or malinformation" (UNODC)
https://www.unodc.org/e4j/en/cybercrime/module-14/key-issues/information-warfare--disinformation-and-electoral-fraud.html

[6] Although if an example of disinformation falls within the scope of incitement as defined by Article 20/Article 4 the International Convention on the Elimination of All Forms of Racial Discrimination, it would be prohibited.
https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg_no=IV-2&chapter=4&clang=_en

[7] UN Framework of Analysis for Atrocity Crimes. https://www.ohchr.org/en/documents/tools-and-resources/framework-analysis-atrocity-crimes

[8] UN Framework of Analysis for Atrocity Crimes. A/70/741–S/2016/71

[9] Although different, all adhere to the UN Guiding Principles on Business and Human Rights. "The UN Framework [...] addresses the human rights responsibilities of businesses. Business enterprises have the responsibility to respect human rights wherever they operate and whatever their size or industry. This responsibility means companies must know their actual or potential impacts, prevent and mitigate abuses, and address adverse impacts with which they are involved."

[10] Table is for illustrative purpose and does not provide a defined list. Some of the categories may also appear in hate speech, misinformation and disinformation. Each incident needs to be individually evaluated based on the various criteria including inter alia, the Rabat threshold test.

[11] In this report speech as a term is used to encompass all forms of communication, including text, visual, graphic and moving images.

[12] OECD (2016), OECD Due Diligence Guidance for Responsible Supply Chains of Minerals from Conflict-Affected and High-Risk Areas: Third Edition, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264252479-en

[13] It should be noted that the UN Strategy and Plan of Action on Hate Speech provides a UN-wide agreed upon operational definition of hate speech. There has not been the same agreed upon alignment around mis- and disinformation.

[14] For example, speech that can be used to promote violent extremism. This type of speech is different to hate speech and disinformation, but violent extremist groups use elicit tactics that exploit and capitalize on the different forms of speech to maximize harms. https://www.undp.org/prevent-violent-extremism/preventing-violentextremism-report-series

[15] United Nations. June 2023. Our Common Agenda Policy Brief 8: Information Integrity on Digital Platforms, p.5
https://www.un.org/sites/un2.un.org/files/ourcommon-agenda-policy-brief-information-integrity-en.pdf

[16] Report of the Secretary General. December 2021. Countering disinformation for the promotion and protection of human rights and fundamental freedoms. https://documents-dds-ny.un.org/doc/UNDOC/GEN/N21/416/87/PDF/N2141687.pdf

[17] UN (2019) UN Plan of Action and Strategy on Hate Speech. https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf

[18] Secretary-General's remarks at launch of UN Strategy and Plan of Action, June 2019:https://www.un.org/sg/en/content/sg/statement/2019-06-18/secretary-generalsremarks-the-launch-of-the-united-nations-strategy-and-plan-of-action-hate-speech-delivered

[19] Sandrine Tiller et al. (2021) The Fog of War…and Information https://blogs.icrc.org/law-and-policy/2021/03/30/fog-of-war-and-information/

[20] The increased targeting of journalists is outlined in UNESCO's 2021/2022 report, Journalism is a public good: world trends in freedom of expression and media development. https://unesdoc.unesco.org/ark:/48223/pf0000379826. The increased targeting of peacekeepers is described by Albert Trithart (Dec 2022), Disinformation Is a Growing Threat for UN Peacekeepers https://theglobalobservatory.org/2022/12/disinformation-a-growing-threat-for-un-peacekeepers/. James Blake outlines the security provisions NGOs need to t ake to protect their staff in, Disinformation and security risk management for NGOs, (Jul 31, 2020) https://www.gisf.ngo/ blogs/disinformation-and-security-risk-management-for-ngos/

[21] https://www.un.org/en/hate-speech/united-nations-and-hate-speech/role-of-the-un

[22] UNESCO. The Impact of UN Peacekeeping Radio Stations. https://www.unesco.org/en/days/world-radio/un-peacekeeping

[23] Shannon Fyfe. Tracking Hate Speech Acts as Incitement to Genocide in International Criminal Law. Leiden Journal of International Law. 2017;30(2):523-548. doi:10.1017/S0922156516000753

[24] What is Coordinated Inauthentic Behavior. Media Manipulation Casebook, Harvard Kennedy School. https://mediamanipulation.org/definitions/coordinated- inauthentic-behavior

[25] See as an example UNDP Oslo Governance Centre's resource on Information Integrity https://www.undp.org/policy-centre/oslo/information-integrity

[26] Our Common Agenda Policy Brief No. 8: Information Integrity on Digital Platforms, June 2023

[27] Countering and Addressing Online Hate Speech: A Guide for policy makers and practitioners https://www.un.org/en/genocideprevention/documents/publications- and-resources/Countering_Online_Hate_Speech_Guide_policy_makers_practitioners_July_2023.pdf

[28] UNESCO, Internet for Trust: Guidelines for the Governance of Digital Platforms. https://www.unesco.org/en/Internet-trust?hub=71542

[29] https://www.government.nl/latest/news/2023/09/20/canada-and-the-netherlands-launch-the-global-declaration-on-information-integrity-online

[30] There are number of examples of excessively punitive legal systems stifling freedom of expression while leaving vulnerable groups unprotected. For instance, the UNESCO Project Social Media 4 Peace proved that in the target countries of the project, the lack of incorporation of specific segments of the society to be protected by law often leaves women, LGBTIQ+ people, and religious and ethnic minorities legally unprotected. Regulations and public policies are focused on punishing perpetrators without provisions for protecting and defending victims of hate speech.

[31] Brigading refers to "all coordinated abusive engagement behaviour online. This engagement can come in the form of retweets, comments, quote retweets, email campaigns and more.". Tony Blair Institute, What is Brigading. (March 2021). https://www.institute.global/insights/tech-and-digitalisation/social-media-futures-what- brigading.

[32] UN Human Rights Council, "Report of the United Nations High Commissioner on Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred," (Rabat Plan of Action), UN Doc. A/HRC/22/17/Add.4, 11 January 2013. The parameters of the test are; context of the statement, speaker's position or status, intent to incite audience against target group, content and form of the statement, extent of the dissemination, and likelihood of harm, including imminence.

[33] First defined by Wardle and Derakhshan in Information Disorder: Toward an Interdisciplinary framework for research and policymaking. Council of Europe (2017).

[34] Guidance exists for example in the Detailed Guidance for Field Presences on the UN Strategy and Plan of Action on Hate Speech: https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf

[35] Albert Trithart, (2022) Disinformation Against UN Peacekeeping Missions. Institute for Peace. p.3 and Panel of Experts report, S/2020/662, 8 July 2020, pp. 13 - 14

[36] United Nations, Department of Peace Operations, "Protection of Civilians Newsletter: Mis/Disinformation and the Protection of Civilians," 2022, Fourth Issue Brief

[37] Ibid

[38] https://www.un.org/en/genocideprevention/documents/Guidance%20on%20COVID-19%20related%20Hate%20Speech.pdf

[39] Networked bots – "A group of compromised computers running software under external" Also known as botnets. (UNECOSOC)

[40] Generative AI – "Generative AI is a subfield of artificial intelligence (AI) and machine learning (ML) that involves the creation of original data or content, including images, video, text, code and 3D renderings"

[41] UNESCO, (2023) Platform problems and regulatory solutions: findings from a comprehensive review of existing studies and investigations, https:// unesdoc.unesco.org/ark:/48223/pf0000385813

[42] Holiday, R. (2012) Trust Me, I'm Lying: Confessions of a Media Manipulator, New York: Portfolio/Penguin.

[43] Wardle, C. (2018) 5 Lessons for Reporting in an Age of Disinformation. First Draft

[44] Information warfare – "A term used to describe the collection, distribution, modification, disruption, interference with, corruption, and degradation of information in order to gain some advantage over an adversary (Marlatt, 2008; Prier, 2017)".

[45] Coordinated Inauthentic Behavior, Meta. https://about.fb.com/news/tag/coordinated-inauthentic-behavior/

[46] European Commission. (2022) 1st EEAS Report on Foreign Information Manipulation and Interference Threats. https://www.eeas.europa.eu/eeas/1st-eeas-report- foreign-information-manipulation-and-interference-threats_en

[47] Patrik Szicherle, Fighting Foreign Malign Influence in Democratic States. Center for Democracy and Resilience, GlobSec.

[48] UNDP. Information Pollution Mapping Programme https://www.undp.org/policy-centre/oslo/information-pollution-mapping-programme

[49] David Scales et al. (2021) The Covid-19 Infodemic — Applying the Epidemiologic Model to Counter Misinformation. New England Journal of Medicine. https:// www.nejm.org/doi/full/10.1056/NEJMp2103798

[50] Talking Information Literacy. An Interview with Whitney Philips and Ryan Milner: Thinking Ecologically about Our Polluted Information Networks. https:// projectinfolit.org/smart-talk-interviews/polluted-information-networks/

[51] UNESCO (2023) Global Media and Information Literacy Week 2023, Media and information literacy in digital spaces: a collective global agenda. https:// unesdoc.unesco.org/ark:/48223/pf0000387226

[52] Internews. Information Ecosystem Mapping Framework, v2 Mapping Information Ecosystems to Support Resilience Framework .pdf (internews.org) https:// internews.org/areas-of-expertise/humanitarian/assessments/humanitarian-information-ecosystem-assessments

[53] UNDP. Information Pollution Mapping Programme. https://www.undp.org/policy-centre/oslo/information-pollution-mapping-programme

[54] UNESCO (2023) Social Media 4 Peace. https://articles.unesco.org/en/articles/social-media-4-peace

[55] United Nations, July 2023: Countering and Addressing Online Hate Speech: a guide for policy makers and practitioners: https://www.un.org/en/ genocideprevention/documents/publications-and-resources/Countering_Online_Hate_Speech_Guide_policy_makers_practitioners_July_2023.pdf

[56] United Nations. June 2023. Our Common Agenda Policy Brief 8: Information Integrity on Digital Platforms, https://www.un.org/sites/un2.un.org/files/our-common- agenda-policy-brief-information-integrity-en.pdf

[57] UN DPO. (2023). Second Mis/disinformation Survey. Addressing Mis/Disinformation Unit. Te sample size of this perception survey was 261, including 35 respondents from MINUSCA, 51 respondents from MINUSMA, 67 respondents from MONUSCO, and 44 respondents from UNMISS

[58] Henning Lahmann. January 2022. Protecting the global information space in times of armed conflict. ICRC.

[59] An overall summary is found in paras 42 and 43 of Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion andexpression, August 2022. Disinformation and freedom of opinion and expression during armed conflicts. A /77/288 https://daccess-ods.un.org/tmp/ 7787461.28082275.html

[60] HRC General Comment 34: https://documents.un.org/doc/undoc/gen/g11/453/31/pdf/g1145331.pdf?token=Dfv7HU8YZc6VKuEkLw&fe=true

[61] A/RES/77/318

[62] A/HRC/RES/49/21

[63] Detailed Guidance on the UN Strategy and Plan of Action on Hate Speech (https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf)

[64] Convention on the Prevention and Punishment of the Crime of Genocide, article 3(c)

[65] Article 20 (2) of the International Covenant on Civil and Political Rights.

[66] A/HRC/22/17/Add.4, appendix, para. 29; the Rabat threshold test is available in 32 languages online at https://www.ohchr.org/en/freedom-of-expression

[67] https://www.ohchr.org/en/freedom-of-expression

[68] Human Rights Committee General Comment No. 11, para. 2.

[69] Art. 19 (3) of the International Covenant on Civil and Political Rights

[70] Detailed Guidance on the UN Strategy and Plan of Action on Hate Speech (https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf); refer also the Human Rights Committee General Comment 34 (on article 19).

[71] Convention on the Prevention and Punishment of the Crime of Genocide, article 3(c)

[72] ICRC: https://www.icrc.org/en/document/general-misinformation-disinformation-and-hate-speech-questions-and-answers#:~:text=Does%20spreading%20disinformation%20violate%20international humanitarian law,crimes)%2C%20online%20or%20offline.

[73] British Red Cross. Field Guide. Media Professionals and Armed Conflict: Protection and Responsibilities under International Humanitarian Law. 2014.

[74] Report of the Secretary General. Countering disinformation for the promotion and protection of human rights and fundamental freedoms. A_77_287

[75] Robin Geiss and Henning Lahmann, 2021. Protecting Societies: Anchoring A New Protection Dimension In International Law In Times Of Increased Cyber Threats. Geneva Academy.

[76] Computational propaganda has been defined as 'the use of algorithms, automation, and human curation to purposefully distribute misleading information over social media network. EU Parliament, 2018 https://www.europarl.europa.eu/RegData/etudes/ATAG/2018/628284/EPRS_ATA(2018)628284_EN.pdf.

[77] Henning Lahmann. January 2022. Protecting the global information space in times of armed conflict. ICRC.

[78] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, August 2022. Disinformation and freedom of opinion and expression during armed conflicts. A/77/288 , p.14

[79] US White House. (Oct 2023) Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/ briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[80] UK Safety Summit 2023. https://www.gov.uk/government/topical-events/ai-safety-summit-2023

[81] UN Secretary General's AI Advisory Board (Dec 2023) Governing AI for Humanity. https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf

[82] Rumor: "a specific proposition for belief, passed along from person to person, usually by word of mouth, without secure standards of evidence being present" (Allport and Postman, A Psychology of Rumor, 1947)

[83] Sarah Roberts. (2019) Behind the Screen. Content moderation in the shadows of social media. New Haven. Yale University Press and Tarleton Gillespie. (2021) Custodians of the Internet: Platforms, Content Moderation and the Hidden Decisions that Shape Social Media. New Haven: Yale University Press.

[84] BSR (October 2018) Human Rights Assessment: Facebook in Myanmar. https://about.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria_final.pdf

[85] OHCHR. Myanmar: Social media companies must stand up to junta's online terror campaign, say UN experts. 13 March 2023. https://www.ohchr.org/en/press- releases/2023/03/myanmar-social-media-companies-must-stand-juntas-online-terror-campaign-say

[86] Meta. January 25, 2023. Crisis Policy Protocol. Meta Transparency Center. https://transparency.fb.com/en-gb/policies/improving/crisis-policy-protocol/

[87] Article 19 (June 2022) Content Moderation and Freedom of Expression: Bridging the Gap between Social Media and Local Civil Society (with UNESCO and the EU) https://www.article19.org/wp-content/uploads/2022/06/Summary-report-social-media-for-peace.pdf

[88] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, August 2022. Disinformation and freedom of opinion and expression during armed conflicts. A/77/288 , p.9

[89] Ibid, p.6

[90] https://www.un.org/en/hate-speech/un-strategy-and-plan-of-action-on-hate-speech

[91] https://www.un.org/en/genocideprevention/documents/22-00041_OSAPG_PolicyPaper_Final.pdf

[92] https://www.ohchr.org/en/press-releases/2023/06/un-human-rights-chief-hate-speech-has-no-place-our-world

[93] Dhanaraj Thakur and Devan L. Hankerson. (2021). Facts and their Discontents: A Research Agenda for Disinformation, Race and Gender. Center for Democracy and Technology

[94] UN Secretary General (2022) Message on the International Day for Countering Hate Speech. https://unsos.unmissions.org/secretary-generals-message- international-day-countering-hate-speech

[95] https://www.ohchr.org/en/freedom-of-expression

[96] ICRC. Harmful Information: Misinformation, Disinformation and Hate Speech In Armed Conflict And Other Situations Of Violence. https://shop.icrc.org/download/ ebook?sku=4556/002-ebook

[97] See for example Microsoft Threat Analysis Center report: "Russia's influence networks in Shael activated after coups," https://blogs.microsoft.com/wp-content/ uploads/prod/sites/5/2023/09/Sahel-Gabon-Coup-Playbook-PDF.pdf

[98] HoronTV: "Au Mali, autopsie d'une attaque informationelle contre la Minusma", 24 May 2023. https://horontv.ml/au-mali-autopsie-dune-attaque-informationnelle- contre-la-minusma/

[99] In one interview, the respondent talked at length about the ways that Twitter had been targeted by disinformation actors during the referendum in Colombia in 2020 to distort mainstream news coverage.

[100] This table is for illustrative purpose and does not provide a defined list. Some of the categories may also appear in hate speech, misinformation and disinformation. Each incident needs to be individually evaluated based on the various criteria including inter alia, the Rabat threshold test.

[101] This builds on the classic 1947 definition of rumor "a specific proposition for belief, passed along from person to person, usually by word of mouth, without secure standards of evidence being present" (Allport and Postman, A Psychology of Rumor, 1947)

[102] Nadesan, M. (2022). Crises Narratives Defining the COVID-19 Pandemic: Expert Uncertainties and Conspiratorial Sensemaking. American Behavioral Scientist, 0(0). https://doi.org/10.1177/00027642221085893

[103] Dangerous Speech Project. April 2021. Dangerous Speech: A Practical Guide. https://dangerousspeech.org/guide/

[104] Although if an example of disinformation falls within the scope of incitement as defined by Article 20/Article 4 the International Convention on the Elimination of All Forms of Racial Discrimination, it would be prohibited. https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg_no=IV-2&chapter=4&clang=_en

[105] UN Framework of Analysis for Atrocity Crimes. https://www.ohchr.org/en/documents/tools-and-resources/framework-analysis-atrocity-crimes

[106] Lucina Di Meco (2023) Monetizing Misogyny. Gendered Disinformation and the Undermining of Women's Rights and Democracy Globally. #ShePersisted. https://she-persisted.org/wp-content/uploads/2023/02/ShePersisted_MonetizingMisogyny.pdf

[107] Julie Posetti et al. (2021) The Chilling: global trends in online violence against women journalists; research discussion paper UNESCO

[108] UNESCO. (2023) How to combat hate speech and gendered disinformation online? https://www.unesco.org/en/articles/how-combat-hate-speech-and-gendered- disinformation-online-unesco-provides-some-ideas

[109] UN Framework of Analysis for Atrocity Crimes. A/70/741–S/2016/71

[110] https://transparency.fb.com/policies/community-standards/bullying-harassment/

[111] https://transparency.fb.com/policies/community-standards/violence-incitement/

[112] https://help.twitter.com/en/rules-and-policies/abusive-behavior

[113] Although different, all adhere to the UN Guiding Principles on Business and Human Rights. "The UN Framework [...] addresses the human rights responsibilities of businesses. Business enterprises have the responsibility to respect human rights wherever they operate and whatever their size or industry. This responsibility means companies must know their actual or potential impacts, prevent and mitigate abuses, and address adverse impacts with which they are involved."

[114] Tarleton Gillespie. (2021) Custodians of the Internet: Platforms, Content Moderation and the Hidden Decisions that Shape Social Media. New Haven: Yale University Press.

[115] https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/

[116] https://transparency.fb.com/policies/community-standards/misinformation

[117] https://www.oversightboard.com/

[118] https://blog.twitter.com/en_us/topics/company/2022/introducing-our-crisis-misinformation-policy

[119] https://help.twitter.com/en/rules-and-policies/manipulated-media

[120] https://support.google.com/youtube/topic/10833358?hl=en&ref_topic=2803176&sjid=17986620533551083705-NA

[121] https://support.google.com/youtube/answer/10834785?hl=en

[122] https://www.tiktok.com/community-guidelines/en/integrity-authenticity/

[123] https://transparency.fb.com/policies/community-standards/hate-speech/

[124] https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=2803176#zippy=%2Cother-types-of-content-that-violates-this-policy

[125] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

[126] https://www.tiktok.com/community-guidelines/en/safety-civility/

[127] Staying Safe on WhatsApp. WhatsApp. https://faq.whatsapp.com/1313491802751163

[128] EUDisinfoLab. December 2022. Disinformation on Telegram: Research and Content Moderation Policies. https://www.disinfo.eu/wp-content/uploads/ 2022/12/20221220_TD_Telegram.pdf

[129] Cat Zakrzewski et al. (Dec 12, 2022) Twitter dissolves Trust and Safety Council, The Washington Post, https://www.washingtonpost.com/technology/2022/12/12/ musk-twitter-harass-yoel-roth/

[130] Gillard, F. (27 May 2023) Twitter pulls out of voluntary EU disinformation code, BBC, https://www.bbc.com/news/world-europe-65733969

[131] Fischer, S. (June 6, 2923) Big technology rolls back misinformation measures ahead of 2024, Axios, https://www.axios.com/2023/06/06/big-tech-misinformation- policies-2024-election

[132] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-128 act_en#:~:text=Digital%20Services%20Act%20(DSA)%20overview&text=Its%20main%20goal%20is%20to,and%20open%20online%20platform%20environment.