Chapter 8

Truth, trust and resilience in the humanitarian sector

Chapter 8

Truth, trust and resilience in the humanitarian sector

Contents

Introduction: The high stakes of the information crisis

This chapter focuses on the shift required for the humanitarian sector to move from treating information as a support function to recognizing it as a core component of resilience. Governments, the International Red Cross and Red Crescent Movement, the broader humanitarian sector and wider society all have a role to play in strengthening resilience to harmful information. It draws lessons from the sector's evolution – from the top-down communication models that once dominated humanitarian action, through the period 2005–2015 when information began to be recognized as a form of aid in itself and participatory approaches such as community engagement and accountability[1] took root. These developments reflected an important commitment to more inclusive, transparent and accountable humanitarian action, yet they emerged at a time when the information ecosystem posed far fewer risks. Today, growing distrust, disillusionment, polarization and the increasingly complex operating environments shaped by harmful information have generated a new kind of crisis – one that strikes at the foundations of humanitarian action: an information crisis.

As UN Secretary-General António Guterres warned in *Our Common Agenda*, "disinformation is an existential threat to humanity,"[2] particularly when it undermines established scientific facts and erodes the social contract.

Over the past decade, the information landscape has undergone seismic change. The humanitarian sector now operates in an environment marked by eroding trust, the rapid spread of harmful information and increasingly polarized public attitudes. Trust – long the bedrock of humanitarian action – is under unprecedented strain, not only from harmful information but also a broader collapse in shared facts and institutional credibility, compounded in 2025 by an abrupt need for a humanitarian 'renewal' or 'reset' triggered by unprecedented deep funding cuts.

Harmful information during emergencies from the COVID-19 pandemic to inter- and intra-state armed conflicts and other humanitarian crises has exposed the vulnerability of both communities and humanitarian organizations. It exploits cognitive shortcuts, such as familiarity bias (the tendency to trust information that is frequently repeated) and availability bias (the tendency to favour emotionally charged or easily recalled narratives). These biases become especially potent in times of crises, when trustworthy and accurate information is most needed but least accessible.

> **"As the saying goes, 'It takes only a moment to spread a rumour, but a lifetime to refute it.' The cost of refuting rumours is completely mismatched with that of spreading them. Moreover, there's nothing we can do about some ordinary people – sometimes they are more willing to believe false information than the truth. This is a harsh reality."**
>
> Community member, China

In digital spaces and connected offline spaces, the boundary between fact and falsehood has become dangerously blurred. The widespread availability of information online and the ease of expressing and amplifying opinions has created a "virtual world of information and misinformation cohabiting side by side … one who will help you and the other who will hurt you".[3] The digital domain itself is far from neutral. As the authors of *LikeWar* observe, the internet has become a contested space – a modern battlefield, where harmful information is wielded as a tool of influence, power and control: "Battle on the internet is continuous, the battlefield is contiguous, and the information that it produces is contagious".[4]

Tactics such as creation of false content, doxing (see Annex I: Glossary, on page 353), smear campaigns and coordinated harassment are increasingly used to silence, discredit and destabilize organizations, while deepening confusion, distrust and division among audiences. Harmful information erodes trust and misrepresents humanitarian action. The consequences are serious: they threaten staff and volunteer safety, community confidence and humanitarian access, challenging the ability to respond effectively to people in need.

New tools and platforms have lowered barriers to participation, but trust and safety in these digital spaces remain fragile, especially where harmful information spreads faster than verification and moderation systems can keep pace. Algorithms and digital platform dynamics amplify polarizing content, pushing users into echo chambers that reinforce harmful narratives and undermine humanitarian principles. The consequences are operational and immediate: reduced acceptance, increased hostility and limited access to vulnerable populations.

The rise of AI adds a new layer of complexity. AI can now generate credible convincing harmful information at scale – in visual, audio and text form, manipulate public opinion and enhance cyberattacks. Once the domain of a few highly resourced states and private actors, AI is now widely accessible, largely driven by rapid private sector innovation. This shift has lowered the threshold for malicious use, enabling the large-scale creation and spread of harmful content. AI technologies are already being used to manipulate public opinion, deepen societal divisions and erode institutional trust. In humanitarian contexts, **this raises important concerns around the protection of affected populations, the integrity of information environments and the responsible use of digital tools in crisis response.**[5] AI also amplifies the potential for targeted exploitation. It can automate the identification of high-value individuals or vulnerable groups, increase the precision of cyberattacks and leverage large datasets to exploit people financially or psychologically. It also enables the production of deepfakes and synthetic narratives that distort public discourse and undermine informed decision-making. These capabilities present urgent challenges for humanitarian organizations, not only for community safety, data governance and digital risk management, but also for the ability to operate within polarized information environments. The responsible use of digital tools and AI – together with appropriate safeguards – is essential to maintaining the accessibility, reliability and trustworthiness of information and supporting principled humanitarian action.

The 'MICE framework' – standing for money, ideology, coercion (or compromise) and ego – has traditionally been used in intelligence and security studies to explain why individuals engage in espionage or other harmful acts. When applied to the information environment, it helps explain why individuals, groups and state actors create and spread harmful information. Today's digital spaces add further motivation including status, revenge and attention-seeking, which can drive both personal and organizational behaviour. For political and state actors, these dynamics often play out at a larger scale:

ideology may advance political or strategic agendas; coercion and compromise can shape the actions of allied groups or populations; ego and status can drive the projection of influence domestically or internationally; and revenge or attention-seeking can be used to destabilize opponents or dominate media narratives. Altogether, these factors show that the spread of harmful information is rarely accidental – it is often guided by clear incentives and pressures, whether at the individual, organizational or state level.

Understanding these motivations is critical for humanitarian actors, as it informs strategies to protect staff, volunteers and communities, maintain trust, counter harmful narratives and mitigate challenges to principled access and response in complex information environments. Practical counter-strategies should be tailored to the specific drivers of harmful information while also **identifying opportunities for engagement and collaboration with states, technology platforms and other stakeholders,** since some measures extend beyond those that humanitarian actors can or should implement. Examples include disrupting financial incentives for criminal actors, providing alternative narratives and engagement pathways to counter ideology or status-driven spread, and establishing accountability mechanisms. Recognizing this spectrum of motivations strengthens the ability of humanitarian, policy and other actors to anticipate and respond effectively to harmful information in complex, high-stakes environments.

Importantly, harmful information rarely exists in isolation. It often amplifies the impact of other crises, intensifying the effects of geopolitical conflicts, climate-related and other emergencies. The World Economic Forum has identified this convergence of risks as a defining feature of today's global risk landscape,[6] posing urgent questions for the humanitarian sector about how to adapt, respond and engage effectively.

In this fragmented landscape, trust in traditional sources of information has sharply declined. Society is increasingly replacing expert knowledge with personal belief, peer experience and influence-driven narratives – a trend described as the 'death of expertise' (Nichols, 2017). People now tend to trust 'people like me' – meaning individuals they perceive as similar to themselves in experience, values or identity – rather than relying on experts or institutions. According to the 2024 Edelman Trust Barometer,[7] trust is becoming increasingly hyper-local, while media and social media continue to rank among the least trusted institutions in a range of countries. Alarmingly, two-thirds of people report difficulty distinguishing reliable news from disinformation. While NGOs are still viewed as ethical leaders, especially among populations with high levels of grievance, even this trust is beginning to erode.

As social theorist Niklas Luhmann noted, **trust reduces complexity and enables action in uncertainty.**[8] In an environment where harmful information spreads rapidly, sustaining **trust, credibility and connection is central to protecting humanitarian space**, ensuring access, and maintaining the effectiveness of principled humanitarian action.

In this context, any humanitarian reset or renewal must prioritize not only explaining the principles of humanity, neutrality, impartiality and independence but also rebuilding both understanding of the impact and effectiveness of humanitarian action and trust in a world where credibility, access and legitimacy are seriously challenged. Understanding the impact of harmful information is essential for **building an evidence base** that informs policy interventions, strengthens operational strategies and guides effective risk strategies and responses. Addressing harmful information is not peripheral – it is central to safeguarding humanitarian space, protecting affected populations and ensuring humanitarian action remains possible.

Contributor Insight 8.1

Information as aid: 20 years on

In 2005, the ***World Disasters Report*** introduced a bold idea: information is as essential in emergencies as food, water or shelter. It popularized the phrase 'information as aid' and asked a provocative question: are humanitarian organizations using information to empower people – or to serve their own interests?

That idea took root. In 2009, it led to the creation of the Communicating with Disaster Affected Communities (CDAC) Network, which bridges media development and humanitarian organizations and champions two-way communication: not just broadcasting messages, but engaging in real dialogue with people affected by crisis.

Today, while 'information as aid' is widely accepted in principle, it is unevenly applied in practice. Too often, accountability initiatives fall back on one-way models where information is delivered and communities passively receive it. Another pitfall is treating information only as data collection – monitoring whether people received a message rather than asking what they need to know. Asking if someone received a red balloon – and whether they liked it – tells you little; you'll never know if someone wanted a yellow one unless you let them steer the conversation. Two-way communication shifts power. It invites participation. It centres people's knowledge and agency.

Today, the power of information is clearest where it is denied. From internet shutdowns to restrictions on or the targeting of journalists, communities face control over what they can access, say and share, as well as being confronted with a flood of overwhelming information platforms and channels. In this environment, digital technologies and AI are double-edged: they can amplify disinformation and bias, but also help filter noise, translate content and support communities in verifying and making sense of information. For CDAC, engaging with these technologies is now central to putting the principle of 'information as aid' into practice.

The crowded and contested informational environment demands that 'information as aid' is understood not as the one-way delivery of neutral facts, but as a relationship between equals. It's not about broadcasting instructions or extracting data, it's about building trust, exchanging knowledge and recognizing affected people as experts in their own lives. When approached this way, information becomes more than aid, it becomes the foundation for solidarity, accountability and shared decision-making.

Ila Schoop Rutten
Information Integrity Lead
CDAC Network

8.1 Why connection matters in humanitarian response

For humanitarian actors, truth and trust are not abstract ideals – they are essential for effective, principled action. Without trust, access can be denied, staff and volunteers face greater risks and affected communities may miss vital support. In today's digital environment, attention is the battleground, and stories – not just facts – shape perception. **Building meaningful connections with communities is now as important as conveying accurate information.** Messages must be culturally relevant, resonant and accessible across multiple languages and formats, while facilitating **two-way engagement** through listening, feedback and dialogue.

In a digital landscape where attention spans are measured in seconds, storytelling must be instant, visual and memorable. Simplicity matters, but so does resonance: narratives must connect to cultural frames and lived realities, allowing audiences to relate to people in need. This shift requires humanitarian organizations to ensure more than one-way broadcasting of information (one to all), to building relationships, facilitating true two-way communication through listening, multilingual engagement, moderation and feedback loops.[9]

Social media has made information-sharing participatory: people no longer just consume stories – they shape and amplify them, often in real time and with an audience. In this environment, responding with the nuance and care that humanitarian principles demand becomes increasingly difficult, yet no less essential. The concept of 'community' in humanitarian response has also expanded and blurred. It no longer refers solely to affected populations, field-based stakeholders and donors. Today, it includes policy-makers, digital influencers, diaspora networks and global publics – often all at once. As a result, the traditional model of 'one message to many', which persists, is likely to land where the boundaries of who is listening – and who is reshaping and amplifying the message – are increasingly unclear. Humanitarian organizations are not only trying to inform, but also to influence, engage and retain the trust of multiple, overlapping audiences, each with different expectations and varying proximity to different crises. Responding effectively means not only knowing what to say, but also understanding *who* is hearing it, *how* they are interpreting it, and *what* they might do with it.

The challenge, then, goes beyond managing harmful content. It is about restoring and maintaining credibility, connection and trust in an increasingly volatile information environment. Yet, most humanitarian organizations lack the structures and resources for continuous, localized dialogue in real time. Those who fail to adapt, whether through hesitation, fragmentation, caution or inertia, risk being left behind. Institutions unable to communicate their story clearly or to respond effectively when harmful narratives take hold lose relevance and public trust.

Contributor Insight 8.2

Spotlight: IFRC workshop on AI and harmful information – 'Communities are networks of trust'

An IFRC workshop in July 2025 brought together humanitarian practitioners and technologists to explore the complex intersection of AI, trust and harmful information. Some key insights and recommendations from the discussion were:

Top insights

- **Trust first**: Rather than attempting to counter every piece of misinformation, efforts should prioritize amplifying trusted sources already valued by communities.
- **Timing matters**: Delayed communication – often caused by over-validation processes – can create information vacuums filled by harmful content. Early, transparent engagement is essential.
- **Misinformation is systemic**: It doesn't just affect communities but also confuses decision-makers. Its impact spans every level of the system.
- **Offline trust is foundational**: Digital solutions must be grounded in offline trust-building efforts, such as working with community figures.
- **Silence is not neutral and can be harmful**: The absence of reliable information can itself be a form of harmful information.

Recommendations for humanitarian organizations

- **Integrate community trust indices** into existing assessments and monitoring frameworks.
- **Establish pipelines for 'good information' production**, focusing on quality, frequency and relevance.
- **Re-examine humanitarian risk tolerance** in addressing harmful information**,** acknowledging that inaction or silence can carry its own dangers.
- **Incorporate AI literacy, capacity building and critical thinking** for humanitarian staff and volunteers in training and tools.

This workshop emphasized the need to move from reactive to proactive strategies, embedding human-centred design for trust, transparency and local legitimacy into every layer of AI deployment and information management. By fostering system-wide

alignment and bold experimentation, the humanitarian sector can better navigate the evolving landscape of harmful information.

IFRC, Geneva

8.2 What's ahead? Evolutions and known unknowns

Trust in institutions will likely continue to decline. People increasingly rely on influencers, peer networks and local intermediaries to judge credibility. For humanitarian organizations, credibility will hinge on the ability to engage through these trusted channels, rather than solely on institutional identity. At the same time, gaps in local journalism and media deserts leave communities vulnerable, reducing access to reliable information and amplifying harmful narratives. Strengthening the information ecosystem requires investment in local media capacity, journalism and community-based mechanisms.

Harmful information will increasingly spread in closed, encrypted and algorithmically governed spaces such as WhatsApp, Telegram and social media platforms. These spaces make detection more difficult, but they also offer opportunities for peer-to-peer communication during crises. At the same time, algorithmic moderation and ranking systems can amplify or suppress humanitarian voices, shaping what information people see and highlighting the need for transparency, safeguards and investment by technology platforms, policy-makers and in local media. Further research is still needed to understand the impacts of harmful information on humanitarian operations and on the people they aim to assist and protect.

What is certain is that the information environment will continue to evolve, including the growing use of AI in military operations. It will likely become more volatile, localized and shaped by technology, with AI a defining factor. The rise of AI adds both opportunities and risks. On one hand, generative AI (see Annex I: Glossary, on page 353) is likely to supercharge harmful information – enabling the scaling of the production of convincing deepfakes, synthetic voices and hyper-targeted narratives that can spread faster than verification systems can respond. It also offers tools for translation, fact-checking and multilingual communication when applied responsibly. The development of agentic AI – systems capable of autonomous decision-making, long-term goal pursuit and multi-step task execution – introduces a new phase in the harmful information ecosystem. Unlike current generative models that largely respond to prompts, agentic AI systems can plan, adapt and act strategically in ways that may accelerate the creation, targeting and amplification of harmful narratives. Humanized AI – such as companion bots – risks eroding trust in real human interaction. As AI systems increasingly emulate human traits and communication styles, they risk eroding the intrinsic value of human judgement and interaction. Meanwhile, the deliberate humanization of AI in design – particularly in the form of companion bots and emotionally responsive systems that elicit sentiment, feelings and projections – may further blur boundaries and subtly shape or manipulate human behaviour.

The future of harmful information and cognitive manipulation – meaning increasingly subtle, data-driven techniques that influence how people think, feel or make decisions

without their full awareness – will likely involve highly personalized, AI-enabled tactics that blur the line between persuasion and deception, making stronger societal, technological and regulatory safeguards essential to protect public trust and cognitive autonomy.

In humanitarian crises, the humanization of AI, by mimicking empathy, care or authority, may undermine trust in genuine human interaction at a time when it is most vital. Crisis-affected populations could begin to attribute human-like intentions or credibility to AI systems, projecting emotions and dependencies onto tools that are not capable of moral judgement or accountability. Such systems may also manipulate sentiment in vulnerable populations. For example, a companion bot that 'comforts' individuals could be exploited to spread harmful information, distort perceptions of humanitarian actors or redirect trust away from legitimate support channels. In fragile contexts marked by trauma, displacement or social fracture, this can deepen confusion, exacerbate mistrust and escalate harm. Moreover, the illusion of empathy created by humanized AI risks devaluing authentic human connection in crisis response. **Humanitarian aid relies not only on material assistance but also on empathy, dignity, compassion and trust-building.** If affected communities experience AI-mediated interactions as substitutes for real human engagement, the intrinsic value of human-to-human solidarity may be eroded, weakening the relational foundations of humanitarian action.

On the other hand, AI also brings opportunities: new tools for fact-checking, content provenance and translation, for example, providing multilingual crisis communication. In highly resource-constrained crises, these systems might help bridge gaps – provided they are transparent, well-regulated and carefully integrated into existing humanitarian services. Overall, the challenge lies in ensuring that AI supports rather than replaces human care, reinforcing rather than weakening the trust and solidarity on which humanitarian action depends.

AI holds potential for humanitarian action, but there is a growing risk that cost-driven, unregulated use could harm vulnerable communities. *Building a Responsible Humanitarian Approach: The ICRC's Policy on Artificial Intelligence*[10] provides an overarching framework to guide the organization's exploration and use of AI in ways that align with its humanitarian mission and principles. The SAFE AI project,[11] led by CDAC Network, The Alan Turing Institute and Humanitarian AI Advisory, with support from the UK Foreign, Commonwealth and Development Office, aims to create practical standards, tools and community-driven frameworks to ensure AI is used responsibly and ethically in humanitarian settings. Monitoring the impact of AI on crisis-affected populations will be essential to ensure that its use remains safe, effective and principled.

The state's central role in defining and enforcing the boundaries of legitimate information control and response[12] may be reinforced. Regulation and governance related to harmful information remain fragmented. Some frameworks, like the European Union's Digital Services Act (2024), require major platforms to assess and mitigate systemic risks such as disinformation and to provide transparency on content moderation. Other regulations may restrict freedom of expression or civic space, for example, laws which criminalize vaguely defined 'false information'.

Global norms on responsible behaviour in cyberspace, including discussions at the UN Open-ended working group on information communication technologies (ICTs) and its successor the Global Mechanism on Developments in the Field of ICTs in the Context of International Security and Advancing Responsible State Behaviour in the Use of ICTs, should provide a forum to address harmful information deployed through cyber means.

This would mirror what was achieved with the protection of critical infrastructure and the explicit references to health care during the COVID-19 pandemic.[13]

The UN Security Council Resolution 2730 (2024) on the protection of humanitarian and UN personnel is an important step toward safeguarding humanitarian action – and a foundation that must be built upon. It calls on member states to take appropriate action to address the increasing threat of disinformation campaigns and misinformation that undermine trust in the UN and humanitarian organizations and put humanitarian personnel at risk. Building on this, global cyber governance forums offer a further opportunity to advance calls for greater protections for humanitarian organizations against harmful information deployed through cyber means. The objective is to ensure that digital threats do not compromise access, safety or principled humanitarian action. Voluntary initiatives like the UN Global Principles for Information Integrity seek to establish shared norms for states, platforms and civil society, but their implementation remains voluntary.

In short, the **information environment ahead will be more sophisticated, pervasive,** deeply embedded in daily life and locally fragmented. The key question is whether humanitarian organizations can collaborate and build the resilience to withstand and counter these pressures; and crucially, whether they can move quickly enough to adapt – embedding harmful information analysis, investing in information and media literacy, community engagement and trust-building as a central element of humanitarian action.

Contributor Insight 8.3

The humanitarian imperative to combat misinformation in disasters

Trust is fundamental to effective disaster management. At every stage – preparedness, response, and recovery – communities rely on clear, accurate information to make decisions that reduce risks and support recovery. However, the growing presence of misinformation and disinformation poses a serious humanitarian threat in Australia, compromising access to life-saving information, undermining trust in institutions and exacerbating harm, particularly for already marginalized populations.

False claims about the existence, severity or cause of disasters – such as the widely circulated hashtag #ArsonEmergency narrative during Australia's 2019–20 bushfires which falsely attributed the bushfires to arson rather than climate-related hazards – undermine public confidence and distract from coordinated action. During the COVID-19 pandemic, misinformation competed with and contradicted official health advice in Australia, fuelling vaccine hesitancy and deepening existing health inequities. In many cases, disasters have been opportunistically used to advance unrelated agendas or intensify existing societal divisions, capitalizing on fear and uncertainty.

While regulatory efforts such as the Australian Code of Practice on Disinformation and Misinformation have sought to address this challenge, the scale and speed of misinformation continue to outpace institutional responses. A recently published disinformation playbook outlines practical countermeasures, including:

- **Pre-emption and early detection**: Anticipating false narratives by understanding local tensions and building trusted information-sharing networks.
- **Prebunking (see** Annex I: Glossary, on page 353**) and spread prevention**: Proactively equipping communities to recognize and reject common disinformation themes before disasters strike.
- **Debunking and recovery**: Correcting falsehoods using clear, repeated communication.

The human cost of inaction is clear. Identifying misinformation risks, recognizing that the emblem of the Australian Red Cross helps to reinforce trust when distributing accurate information and addressing misinformation are now core components of day-to-day operations. Tackling misinformation must be treated as a humanitarian priority. This requires equitable, inclusive and culturally informed communication systems. Community-led strategies, trusted messengers and sustained education are critical to protecting lives, strengthening social cohesion and ensuring all communities are supported during times of crisis.

Jenny Gillett
Senior Manager – Policy and Research, External Engagement
Australian Red Cross

Contributor Insight 8.4

The spread of harmful information in situations of conflict and violence

In situations of armed conflict, information can be a lifeline. But when it is distorted, preventing people from accessing life-saving services or leading them to make inadequate decisions about their safety, it can have serious consequences. When information fuels hatred between communities or incites violence, it exacerbates the suffering of people already affected by the harsh realities of conflict.

There is neither an international legal nor agreed-upon definition of harmful information. The ICRC defines it as "information that can potentially cause or contribute to harm, either physically, psychologically, economically or socially".[14] This includes misinformation, disinformation, malinformation, hate speech and other narratives that spread in violation of international humanitarian law even if they do not fall entirely within those categories.

Harmful information increases people's exposure to risks and deepens vulnerabilities during armed conflict. For example, when displaced people are intentionally given misleading information about life-saving services and resources, they may be diverted away from help and toward harm. In situations where hostilities are ongoing, false warnings could lead people to make decisions that put them in harm's way. Speech and narratives that encourage hatred and violence can endanger lives and livelihoods. For example, online

calls for violence against minority groups can trigger acts of violence against individuals and inflict psychological and social harm through harassment, defamation and intimidation. Harmful information also undermines trust in humanitarian organizations' ability to operate, potentially limiting their access to the people they intend to serve.

Addressing harmful information is complex and requires a conflict-sensitive approach. This involves considering the potential harm that the spread of harmful information can have on affected populations. Responses often need to occur at local, national and global levels and may be preventive or reactive. In some cases, humanitarian actors may engage in dialogue with states or non-state actors, while in others or in parallel, they may focus on improving the availability of accurate, timely and reliable information for people. In other cases, humanitarian actors may partner with local and community groups to strengthen preparedness and prevention efforts.

Such responses should be sensitive to existing risks and conflict dynamics and not trigger any unintended harm; this may mean that the focus is not on the veracity of the information or on refuting or correcting it, but rather on the potential for escalation and harm should certain information spread in that context. They should aim to improve the protection of people affected by conflict and violence by both engaging belligerents and other relevant actors, including technology and social media companies, and by mitigating psychological and societal harms. At the same time, it is necessary to strengthen the resilience and agency of people and communities, including diaspora groups, to better navigate information related to risks. This means improving access to knowledge and reliable information and enabling people to identify and understand harmful information and navigate its effects. Equally important is strengthening principled humanitarian action by reinforcing the ability of humanitarian actors to operate and access the people they intend to serve.

Joëlle Rizk
Digital Risks Adviser, Protection Division
International Committee of the Red Cross

8.3 Defining humanitarian resilience in the information age

In the context of harmful information, humanitarian resilience means far more than managing reputational risk. It is the capacity to anticipate, absorb and adapt to information threats – while remaining grounded in humanitarian principles and accountable to affected communities. In an era where digital platforms and algorithmic tools can amplify falsehoods at scale, reactive communication strategies are no longer sufficient.

A truly resilient approach begins with an understanding of harm, not just to organizations, but to communities. Although research remains limited, recent initiatives are shedding light on the damaging effects of harm including at the local level. For example, work by Grand Challenges Canada, Fondation Hirondelle and Internews has highlighted serious consequences: incitement to violence through social media, erosion of social cohesion and deepening mistrust of humanitarian actors.

Research from the University of Melbourne on disinformation during disasters further focuses on how false narratives can fragment communities and inflame tensions, especially in fragile or polarized contexts. Its *City Playbook for Countering Disinformation* underscores the importance of local, context-specific community-led responses that prioritize community trust, civil society engagement and early detection and prevention.[15]

To address these harms, humanitarian organizations must go beyond ad hoc responses. This includes investing in systems that monitor digital threats, assess the impact of interventions and identify structural triggers and vulnerabilities that allow harmful information to spread. **Central to this is community engagement: local actors are not only the first to be affected but often the best positioned to create and sustain effective solutions.**

Building resilience requires understanding the information ecosystem and how people access information and their needs, creating space for feedback, building surge capacity to respond to information crises, supporting digital and information literacy and embedding simulations and anticipatory tools into preparedness planning. It also demands a willingness to confront legitimate criticism with transparency and humility, recognizing that trust is not built through messaging alone, but through sustained, principled and impactful humanitarian action.

The following pages showcase examples from different organizations, illustrating how they are confronting harmful information in practice and the lessons that can help guide the wider humanitarian sector.

Contributor Insight 8.5

Crisis communication preparedness at South Sudan Red Cross

The South Sudan Red Cross has established a Crisis Communication Plan to anticipate and respond to the spread of harmful information during emergencies. This plan sets out clear objectives, key messages, target audiences and communication channels to be used in a crisis. To operationalize the plan, the National Society formed a Crisis Communication Committee comprising the Secretary-General, Communications Manager, Safety and Field Coordinator, Partnerships Coordinator and the Emergency Operations Centre. The Communications Manager conducts regular environmental scanning to detect any mention of the South Sudan Red Cross in public discourse and flags negative narratives for immediate action. The Safety and Field Coordinator monitors the safety and security of staff and volunteers, while the Partnerships Coordinator ensures effective coordination with partners. The Emergency Operations Centre includes departmental managers and

three regional coordinators who facilitate two-way communication between headquarters and branches.

Pascal Ladu
Communications Manager
South Sudan Red Cross

Contributor Insight 8.6

UNHCR's Information Integrity Toolkit

Drawing on extensive testing of activities and responses from pilot projects in Asia and the Americas – and in collaboration with a wide range of stakeholders (including UN agencies, humanitarian partners, civil society organizations and NGOs such as digital rights groups, academia, governments and the private sector) – UNHCR released the Information Integrity Toolkit in April 2025. Designed as a resource for the entire humanitarian sector, the toolkit provides a structured, four-step response framework with practical tools and guidelines to address misinformation, disinformation and hate speech. It is adaptable to different contexts and operational mandates, enabling humanitarian actors to respond quickly and effectively to threats to information integrity.

Gisella Lomax
Senior Advisor, Information Integrity
UN High Commissioner for Refugees

Contributor Insight 8.7

Strengthening information integrity: From principles to practice

Theory of change for information integrity

Emerging normative frameworks on information integrity are grounded in interdisciplinary collaboration and recognize that healthy information environments are essential for human rights, democratic resilience, sustainable development and peace and security. Information integrity also supports critical sectors including business and finance, scientific advancement, technological innovation, public health, education and the creative industries.

Yet, rapid transformations in the global information ecosystem have heightened risks and vulnerabilities. No single actor can address these challenges alone – effective responses require multi-stakeholder collaboration and action. While some stakeholders, like states

and major technology companies, hold the greatest power, resources and responsibilities, others provide vital perspectives and lived experiences that must inform solutions.

Cross-cutting challenges

Strengthening information integrity requires addressing interconnected challenges that create both obstacles and opportunities for building a more resilient global information ecosystem. Key cross-cutting challenges include:

Sectoral silos: Information risks cut across thematic and geographic boundaries, yet responses typically remain compartmentalized in distinct sectors such as elections, public health, climate and conflict. More effective approaches must apply lessons about tackling adversarial behaviour across these domains rather than treating each area in isolation.

Cross-domain information manipulation: Information – and the actors who spread it – move seamlessly between digital and physical spaces. Information risks therefore transcend the artificial boundary between 'online' and 'offline' environments. As trust in digital platforms erodes, particularly amid uncertainty around emerging AI technologies – offline spaces may gain new importance for individuals and communities seeking reliable information. Meanwhile, adversarial actors exploit both domains in tandem to shape public perceptions and influence policy outcomes.

Systematic targeting of information defenders: Researchers, journalists, fact-checkers and civil society activists face strategic harassment and coordinated campaigns designed to undermine their credibility and silence their contributions. These harassment campaigns – frequently gendered and sexualized – create lasting deterrent effects, systematically eroding research capacity precisely when it is most needed.

Research limitations and methodological bias: Data availability varies widely across platforms, pushing researchers to over-rely on single sources and producing skewed assessments that can lead to flawed responses. Geographic and linguistic biases compound these challenges, with most research concentrated in English-language contexts while vast areas of the global information ecosystem remain under-examined.

AI and information integrity: AI is fundamentally transforming how people access information and how adversarial actors generate false content, effectively making societies involuntary participants in a large-scale information experiment with far-reaching consequences. Generative AI tools are proliferating without adequate safeguards, lowering the barriers to producing hate speech and convincing disinformation at scale. In doing so, they undermine every pillar of information integrity – from societal trust to independent media.

The 3R operational framework

Strengthening information integrity requires systematic approaches that move beyond reactive responses to build proactive resilience against evolving information threats. The Research-Risk Assessment-Response (3R) operational model offers a structured framework for organizations to understand and address a range of information risks, and is particularly valuable in resource-constrained settings. Information risks can be understood as actions, conditions or factors that undermine the integrity of information environments and weaken public access to, and understanding of, evidence-based information, informed decision-making, societal trust and cohesion. Socioeconomic and political factors can enable and exacerbate these risks.

Research

Effective interventions begin with rigorous research into information ecosystems, using cost-effective methods including desk reviews, situational analysis and examinations of influence operations or disinformation campaigns. Organizations can leverage existing expertise, partnerships and open-source techniques to identify emerging risks, uncover policy gaps and assess potential solutions. Such research should address critical questions: What risks are present? Who are the drivers? Which tactics make them effective? Which audiences are targeted and with what impacts?

Risk assessment

Risk assessment bridges research and action by setting clear criteria to prioritize responses according to severity, credibility and potential scope of harm. The process assesses factors such as source authenticity, behavioural patterns, narrative content, distribution levels and impacts on target audiences across social, political, operational and human rights domains. Standardized assessment practices classify risks from very low to high, helping determine both the urgency of response and resource allocation.

Response

Response strategies operate across multiple timeframes with four core objectives:

1. **Prevention:** Build long-term resilience to prevent information risks from undermining societal cohesion, human rights, peace and security, and sustainable development.
2. **Protection:** Put in place targeted safeguards in anticipation of high-risk moments.
3. **Mitigation:** Contain risk escalation and reduce impacts in real time.
4. **Recovery:** Restore disrupted capabilities while rebuilding trust and resilience.

Implementation faces significant obstacles, including data limitations, systematic targeting of information defenders and researchers, and persistent research biases. Effective response requires skilled personnel, strong coordination mechanisms and adherence to human rights-based professional standards.

Charlotte Scaddan
Senior Adviser on Information Integrity
UN Global Communications

Contributor Insight 8.8

WHO: Building a resilient emergency system (part 2 of 2)

Common challenges

As funding for crisis responses continues to shrink, risk communication, community engagement and infodemic management (RCCE-IM) is too often considered a 'nice-to-have' element rather than a critical component of emergency preparedness and response. Despite growing awareness and progress, countries and agencies still face persistent challenges including:

- **Resource constraints:** Many health authorities lack stable funding and dedicated specialist capacity for RCCE-IM, hampering both preparedness and response efforts.
- **Fragmented coordination:** Siloed working and absence of well-tested, cross-sector coordination mechanisms, such as simulations or joint emergency planning, can delay unified messaging and confuse or frustrate the public.
- **Gaps in co-creation and localization:** Limited experience in co-designing interventions and messages with communities and insufficient attention to local contexts or language diversity can reduce their relevance and impact.
- **Expertise deficits:** Critical skills in infodemic management, social listening and message testing are often lacking. Without proactive communication, these gaps allow voids in which misinformation and disinformation can thrive.
- **Influencer readiness:** Trusted actors such as health workers and community representatives often lack the motivation, training or resources to fulfil their vital communication roles. Thus, RCCE-IM is too often not implemented.
- **Limited research-to-practice collaboration:** RCCE-IM aims to translate behavioural science, evaluation data and the evidence base into real-world application. However, a lack of training and support in applying RCCE-IM can lead to ad hoc interventions that may be neither effective nor sustainable.

Recommendations for humanitarian organizations

Given these challenges and lessons learned, moving forward requires strategic, systemic action:

- **Embed RCCE-IM in emergency planning:** Make RCCE-IM a statutory and operational priority at all levels of emergency preparedness. Ensure that human resources and regular training are funded to support response.

- **Professionalize and train RCCE-IM teams:** Build multidisciplinary teams with technical expertise in RCCE-IM and behaviour change. Invest in capacity building and mentorship for all practitioners, especially at local levels.
- **Institutionalize coordination:** Create robust, pre-tested systems for multi-agency, multilevel collaboration with clear roles and shared protocols. Allocate resources to test coordination mechanisms through simulations and retrospective reviews of past responses.
- **Forge community partnerships:** Proactively engage civil society organizations, community groups and trusted influencers as co-designers and co-implementers of interventions and messaging.
- **Integrate research and evidence:** Bridge the gap between research and practice by ensuring monitoring and evaluation data form an operational basis for rapid response. Establish partnerships and knowledge-sharing mechanisms that enable continuous learning.
- **Systematize feedback and adaptation:** Embed formal systems for listening, collecting public feedback and adapting RCCE-IM strategies dynamically before, during and after crises.

Nancy Claxton
RCCE-IM Technical Officer
WHO

Leonardo Palumbo
Community Engagement Technical Officer
WHO

Cristiana Salvi
Regional Technical Advisor for Community Resilience and Protection
WHO

Contributor Insight 8.9

WikiRumours: Crowdsourcing verified information for community safety and prosperity (part 1 of 4)

The Sentinel Project employed a combination of community-based monitoring, digital verification and two-way communication platforms to address harmful information in South Sudan and the Democratic Republic of the Congo (DRC). Key tools and strategies included:

- **WikiRumours platform**: An open-source system where users submitted rumours via SMS or voice calls. Trained moderators verified the submitted information using a standard rubric and shared responses back within 24 hours through the same channels.
- **Community ambassadors**: Hundreds of locally trained individuals collected, verified and distributed information. These ambassadors also operated

'peace spaces' where community members could engage in dialogue and fact-check rumours in real time.

- **Preventive messaging**: During high-risk periods (e.g., elections), the team proactively disseminated verified updates to pre-empt misinformation. For example, after a polling station in Bunia was burned, verified updates helped calm the community before rumours spread.
- **Collaborative verification groups**: In the DRC, local authorities such as religious leaders and police joined discussion groups to verify and co-disseminate information. This significantly boosted community trust.
- **Local media partnerships**: Radio programmes, town hall meetings and posters were used to deliver verified information in familiar, trusted formats.

What worked well and why?

- Two-way communication via SMS/voice ensured broad reach, especially in low-connectivity areas.
- Community ownership increased credibility – local people trusted messages verified by people they knew personally.
- Targeted counter-messaging limited the spread of misinformation and addressed rumours at their source.
- Real-time verification improved situational awareness for civilians and humanitarian actors.

What did not work well and why?

- Over-reliance on community ambassadors in the early stages proved insufficient. Residents placed greater trust in authority figures like local chiefs or religious leaders, prompting a shift to include them in verification chains.
- Surveillance of the WikiRumours platform by armed groups posed operational risks in some areas, requiring the adoption of more secure communication methods.
- High resource demands – including logistical costs (especially in South Sudan) and time-consuming verification – challenged scalability without sustained funding.

Unintended consequences and lessons learned

- Early trust-building with community leaders is essential for system adoption and long-term credibility.
- Anticipatory messaging, when deployed quickly, can reduce the risk of rumour-driven panic.

- Localized verification mechanisms, such as peace spaces, promote both information accuracy and community resilience.
- Judicial use of verified data: In Beni, system-generated reports were used in court proceedings, demonstrating the value of verified reports as credible records in conflict settings.
- Family reunification: An unexpected outcome was WikiRumours' role in facilitating the reunification of families separated by conflict and displacement. This is a result of the Sentinel Project's efforts to transmit information across different communities. Subscribers used the WikiRumours platform to request information about their missing family members, while community ambassadors worked to identify and locate them. These efforts successfully reunited 82 children with their families.

Anahi Ayala Iacucci, Nabeel Chudasama, Nabeela Jivraj, Zainah Alsamman
Grand Challenges Canada

Christopher Tuckwood
The Sentinel Project

Contributor Insight 8.10

WikiRumours: Crowdsourcing verified information for community safety and prosperity (part 2 of 4): gaps and support needs

Despite strong results, the Sentinel Project's experience revealed critical gaps in capacity, coordination and specialized support that limited the scale and sustainability of its response to harmful information during the project period.

Identified gaps

- **Technical capacity**: While the WikiRumours system was highly effective offline, the team lacked the capacity to systematically monitor online platforms such as Facebook and WhatsApp, where misinformation also circulated – especially in urban areas.
- **Analytical tools**: The project initially lacked advanced tools for data visualization, rumour-trend mapping and predictive analytics, which could have enabled faster identification of emerging narratives or misinformation spikes.
- **Monitoring and evaluation expertise**: Limitations in designing impact frameworks hindered the ability to capture both behavioural and

perception-level changes across large populations, making it harder to measure long-term shifts in trust or information resilience.

- **Security protocols**: In areas where armed groups surveilled communication channels, the team faced risks without dedicated guidance on secure data handling and risk mitigation in information operations.

- **Sustainable funding**: High operational costs in contexts like South Sudan strained project budgets. The absence of long-term funding commitments made it difficult to retain trained ambassadors or invest in infrastructure (e.g., solar chargers, community radio) beyond initial grant periods.

Opportunities for partnerships, technical support or external guidance

- **Strategic partnerships with tech platforms and digital rights organizations** to enhance online monitoring, detect misinformation patterns and develop secure communication tools tailored for humanitarian contexts.

- **Guidance and toolkits from humanitarian coordination bodies** (e.g., IFRC, CDAC Network) to standardize rumour-verification protocols, establish safety measures for community reporters and integrate responses to harmful information into broader humanitarian coordination systems.

- **Cross-sector alliances with media organizations and trusted local influencers** (e.g., radio stations, religious leaders, teachers) to co-deliver counter-messages and strengthen long-term media literacy.

- **Pooled funding mechanisms or donor consortia to provide multi-year, flexible financing** for community-driven harmful information initiatives, ensuring continuity and enabling scale-up.

To counter harmful information more effectively, system-level strategies must prioritize local engagement, trust-building and cross-sector collaboration. Based on lessons from the Sentinel Project's implementation in South Sudan and the DRC, the following recommendations for organizations are proposed:

Prioritize community-led verification mechanisms

- Empower trusted local actors (e.g., chiefs, faith leaders, teachers, health workers) to act as information stewards, not just recipients.

- Integrate community feedback loops and 'peace spaces' into humanitarian programming as standing structures for rumour tracking, counter-messaging and dialogue.

Develop policy guidance and coordination protocols for harmful information

- Co-develop clear guidance on responding to harmful information in conflict settings, including defined roles, verification standards and escalation pathways.

- Embed coordination within humanitarian cluster systems to avoid siloed responses or purely reactive responses.

Formalize collaboration with local media and radio networks

- Support community radio stations and local journalists as frontline responders to misinformation.
- Establish shared content platforms or templates for rapid, localized dissemination of verified, culturally contextualized information.

Advocate for digital inclusion and responsible platform governance

- Engage social media companies to provide localized misinformation mitigation tools (e.g., flagging features, WhatsApp fact-check bots) in low-bandwidth, multilingual contexts.
- Advocate for AI transparency, stronger content moderation investments and data-sharing frameworks adapted to humanitarian needs in fragile settings.

Promote harmful information literacy across the ecosystem

- Embed media and rumour literacy into school curricula, health outreach and aid distribution channels.
- Train volunteers, government field staff and humanitarian workers – not just technical staff – on identifying, documenting and responding to harmful information.

Invest in anticipatory communication systems

- Deploy proactive messaging protocols before elections, vaccine campaigns or anticipated flashpoints, using SMS, posters and radio to pre-empt likely rumours before they spread.
- Use predictive analysis models, drawing on existing rumour databases such as WikiRumours, to identify potential harmful information hotspots in advance.

Anahi Ayala Iacucci, Nabeel Chudasama, Nabeela Jivraj, Zainah Alsamman
Grand Challenges Canada

Christopher Tuckwood
The Sentinel Project

Contributor Insight 8.11

Centre for Humanitarian Dialogue's disinformation diplomacy: Mediating the digital frontlines of conflict

Formal diplomacy is under pressure. We are witnessing more armed conflicts today than at any time since World War II, alongside a global struggle over truth, trust and technology. Online spaces are increasingly weaponized, fuelling violence, eroding social cohesion, undermining peace efforts and development gains, and hardening the positions of conflict parties – making dialogue and compromise harder to achieve. The rise of AI-powered tools and the rollback of platform moderation are accelerating these risks. Automated online manipulation enables the mass production of synthetic content, bot-driven amplification and coordinated influence operations, while lowering the barriers to 'disinformation-for-hire' services.

Vulnerable communities in conflict zones will continue to bear the brunt of these dynamics, as global diplomacy and international norms struggle to keep pace with rapid technological disruption and geopolitical competition.

Current diplomatic responses remain largely reactive, with an emphasis on counter-disinformation, regulation and digital literacy. While these approaches are essential for addressing the 'disinformation supply chain' (production, distribution, consumption), they are not sufficient on their own. Social media risks must also be addressed at the source: by influencing the behaviour of armed conflict actors not just through sanctions but through broader strategies that reduce the incentives and impacts of information manipulation. Lasting solutions require a comprehensive approach, one that tackles all elements of the chain, from origin to outlet and from platforms to people.

Over the past five years, the Centre for Humanitarian Dialogue, a Swiss foundation with a proven track record in preventing and resolving armed conflicts through discreet diplomacy, has engaged in at least 70% of the world's most violent conflicts,[16] working to reduce suffering, foster dialogue and open pathways to stability and development.

During this period, the centre has expanded its mediation mandate to address the negative impact of digital technologies, including social media, on armed conflicts and peace processes, piloting innovative approaches in more than 15 countries to promote online restraint, including:

- **Facilitating seven social media codes of conduct** in Nigeria (2021), Kosovo (2021), Bosnia and Herzegovina (2022), Thailand (2023) and beyond.
- **Establishing disinformation de-escalation channels** between political actors in the Caucasus.
- **Running mediation workshops** on managing digital harms for local mediators in the Horn of Africa.

- **Building capacity for diplomats and peacebuilders** to address online threats in crises.

- **Creating discreet backchannels** with major social media platforms to flag risks and discuss prevention and mitigation measures.

The Centre for Humanitarian Dialogue's approach blends discreet diplomacy with multi-stakeholder dialogue, engaging conflict actors, civil society, regulators, social media platforms and 'unconventional' conflict stakeholders, such as social media influencers, to identify shared concerns and foster solutions that encourage online restraint and mitigate social media's harmful impact in conflict contexts.

From our work, we have identified at least three key lessons that form the foundation of our 'disinformation diplomacy' – a set of dialogue-based approaches designed to complement, not replace, existing initiatives to address evolving information threats and to strengthen traditional peace and security efforts:

- **Online restraint requires offline dialogue.** Progress becomes possible when conflict parties recognize that disinformation can escalate into risks none can control or are willing to bear. At this stage, conflict actors may agree on 'red lines' and voluntary codes of conduct or norms for responsible behaviour, sometimes even before formal peace talks commence. Establishing formal or informal channels to de-escalate disinformation – similar to cyber diplomacy protocols – remains rare but is increasingly necessary. Such initiatives can also act as confidence-building measures in broader peace processes.

- **Codes of conduct matter but require greater collaboration and accountability.** The Centre for Humanitarian Dialogue has facilitated seven voluntary social media codes of conduct, often linked to elections. While useful in weakly regulated contexts, turning words into action demands institutional accountability, civil society oversight and active collaboration with platforms. Early, sustained engagement with platforms is essential, as their policies and algorithms can inadvertently exacerbate tensions and conflict, positioning them as actors in the conflict and diplomatic space.

- **Local ownership is essential:** The most effective norms are those developed, demanded and defended by those most affected by online harm. Local actors – often best placed to influence community dynamics – must play a central role in shaping and upholding standards for responsible online behaviour. Yet they are too often excluded from formal dialogue and mediation processes.

Today, conflict mediation is no longer solely about persuading parties to lay down physical weapons. It also involves negotiating restraint in the use of digital arsenals and building partnerships with platforms and local actors to prevent, mitigate and resolve armed conflicts. **Cognitive warfare is a race to the bottom.** In this era of fragmentation and multipolarity, the international community must back disinformation diplomacy efforts with both real political will and adequate resources.

Jacobo Quintanilla
Programme Manager, Social Media and Conflict Mediation
Centre for Humanitarian Dialogue

Contributor Insight 8.12

Strengthening information integrity: Towards a preventive and inclusive approach to disinformation

Five key recommendations

Prevent the effects of disinformation by strengthening the professionalism, independence and viability of local media.

- Enhancing these qualities enables local media to respond effectively to community information needs and to create trusted spaces for inclusive dialogue – before rumours and disinformation fill the vacuum and fuel tensions.
- This journalistic approach includes, but precedes, fact-checking, which reacts only after disinformation is already circulating. It also goes beyond prebunking, which focuses narrowly on pre-identified disinformation topics.
- Crucially, support to local media must remain distinct from strategic communication and public diplomacy, which aim to promote the views of funders. Blurring this line risks undermining media credibility and eroding public trust.

Strengthen local-level interaction between media and diverse population segments to understand information needs.

- Digital tools and AI can enhance the media's ability to gauge public sentiment at scale. However, in-person engagement – such as face-to-face meetings, focus groups and field research – remains essential to avoid conclusions distorted by the digital divide or algorithmic bias.

Develop a hybrid (offline and online) inclusive multimedia offer that meets the needs of diverse and specific audiences.

- When selecting formats, broadcast methods, languages, topics and the journalists best suited to address them, it is essential to consider social inequalities and promote inclusivity. This includes taking into account factors such as gender, class, culture, language, religion, urban or rural location, educational background and the digital divide.

Strengthen journalists' and their networks' knowledge of topics that are frequent targets of disinformation.

- Thematic training should cover areas such as conflict dynamics, politics, elections, the environment, gender, the economy, justice and public health. These sessions should be offered as complementary modules – only after

journalists have acquired a solid grounding in the core principles and practices of journalism.

Improve journalists' and audiences' understanding of how digital platforms and social media algorithms work and AI functions.

- Training should address both the opportunities and risks of using these technologies when producing and disseminating public interest media.
- Media outlets should develop clear charters on the use of these technologies – especially for generative AI – and ensure that their application is transparent to audiences.
- Media literacy should be strengthened through accessible and engaging programming that raises awareness of the risks posed by evolving information ecosystems, particularly among marginalized groups.

Sacha Meuter
Head of Research and Policy
Fondation Hirondelle

Contributor Insight 8.13

Rebuilding local information ecosystems: A critical pillar of post-crisis recovery

ICT infrastructure is often overlooked or excluded from post-disaster and post-conflict reconstruction plans. It is frequently excluded from negotiations due to the normalization of its deliberate targeting by political actors or warring parties, often under a veil of limiting transparency and accountability. As a result, millions of people remain in a state of prolonged or permanent disconnection from the internet for months or years after a disaster has struck or a conflict has ended. This not only impacts resilience and recovery but also increases vulnerability to future crises.

In response to the persistent reluctance of the international community to address this issue, civil society is stepping up with coordinated action. For example, 7amleh – the Arab Center for Social Media Advancement, together with the Palestinian Digital Rights Coalition and dozens of international organizations, has launched #ReconnectGaza, a global campaign calling for the rebuilding of Gaza's telecommunications network and the recognition of access to communication as a fundamental human right.[17]

Addressing this challenge requires a holistic approach that includes both short-term emergency connectivity solutions, such as eSIM cards, satellite internet access and mobile communication hubs, to restore basic services, as well as long-term investment in modern telecommunications infrastructure, including fibre optics and renewable energy-powered

networks. Such infrastructure is essential not only for communication but also for the delivery of education, healthcare and economic recovery.

Giulio Coppi
Senior Humanitarian Officer
Access Now

Marwa Fatafta
Middle East and North Africa Policy and Advocacy Director
Access Now

Explainer: Offensive, defensive and integrated responses to harmful information

Defensive responses aim to contain, correct or minimize the damage caused by harmful information. These are typically reactive and focus on protecting an organization's reputation, staff safety and operational access. Key approaches include:

- rapid fact-checking and myth-busting
- activation of crisis communication protocols
- quiet corrections and bilateral engagement with relevant stakeholders
- clarifying harmful information through trusted channels
- strengthening internal alignment on messaging
- ensuring decision-making authority as close to the operational context as possible and escalated only when necessary.

Offensive responses take a proactive approach, aiming to influence the information environment before harmful narratives take hold. These strategies focus on amplifying credible voices, building public trust and pre-empting harmful information by occupying the narrative space early and intentionally. Key approaches include:

- strategic storytelling and values-based campaigns
- partnerships with local influencers and media
- narrative inoculation ('prebunking' tactics)
- community co-creation and distribution of messages
- digital monitoring to anticipate and counter emerging harmful information.

Integrated responses blend defensive and offensive tactics to provide both immediate protection and longer-term influence. These approaches ensure that rapid reaction is

linked to ongoing narrative shaping and community engagement, creating a continuous cycle of protection and trust-building. Key approaches include:

- embedding rapid response capacity within long-term communication strategies
- conducting real-time monitoring to inform both corrections and proactive messaging
- coordinating cross-functional teams to align operational updates with storytelling
- linking incident management to broader reputation and trust-building goals
- sharing lessons from past incidents to strengthen future preparedness.

All three approaches are essential. Defensive responses manage acute incidents. Offensive responses strengthen long-term resilience and trust. Integrated strategies connect the two – ensuring the proactive and reactive shaping of the information ecosystem.

Contributor Insight 8.14

Recognizing the systemic risk

Harmful information is undermining not just communications strategies, but also access, safety and trust. Online narratives – whether grounded in fact, misperception or disinformation – have triggered real-world consequences including operational restrictions, reputational crises, funding freezes and security threats. Yet the humanitarian sector continues to treat harmful information primarily as a communication challenge, rather than a systemic risk. In contexts such as the Sahel, even neutral updates are being reframed as evidence of espionage or political bias, while silence is interpreted as complicity. In Syria and Myanmar, communities turn to platforms like Facebook not only to express anger, but also to request aid.

Proactive monitoring is no longer optional. AI-powered tools, like those employed by Insecurity Insight, have proven essential for detecting sentiment shifts, identifying disinformation surges and flagging rising hostility before it escalates. These tools have captured moments when harmful narratives were not only widespread, but were actively shaping public understanding of humanitarian principles, particularly neutrality.

Four shifts are now needed:

- **Coordinating sector-wide efforts**: Harmful narratives rarely stop at one organization's door. A collective response – including coordinated messaging, data-sharing and joint digital risk assessments – is essential to protect humanitarian space.

- **Investing in digital literacy and local partnerships**: From community influencers to local media, trusted local actors are essential allies in countering misinformation and amplifying accurate information.
- **Reframing digital engagement as essential to humanitarian access**: Social media must be recognized not only as an outreach tool but as a space where acceptance is won or lost. As community-driven content moderation becomes more common, aid actors must actively participate in these spaces, not retreat from them.
- **Developing a humanitarian language for the digital space**: Social media has developed its own language and norms. Traditional humanitarian language that evolved in diplomatic corridors needs to be translated into communication that resonates on social media: clear, sharp and unambiguous. The use of diplomatic phrases is just as out of place on social media as an emoji would be in a UN Security Council resolution.

Christina Wille
Director
Insecurity Insight

Clara de Solages
Researcher
Insecurity Insight

8.4 Trust brokers – the human bridge to credibility

Access to trusted information depends not only on what is communicated – the message – but also on who delivers the message – the messenger. In the Red Cross and Red Crescent Movement, local staff and volunteers will increasingly serve as the 'trust brokers': individuals who play a vital intermediary role in building, maintaining and restoring trust between communities and organizations. Trust brokers help translate institutional intent into locally meaningful terms, bridge gaps in power, knowledge and access and vouch for the credibility of both information and actors. Their effectiveness lies in being perceived as independent, culturally grounded and aligned with community needs. As members of the communities they serve, they are uniquely positioned to navigate sensitive dynamics, defuse tensions and foster dialogue.

In contexts marked by harmful information or low institutional trust, **trust brokers often form the first and most credible line of communication**. Supporting them with timely, accurate messaging, digital access and tools, and clear guidance is essential – not only to counter harmful information but to sustain humanitarian access and reinforce the legitimacy of humanitarian action.

A sustainable presence depends not only on continuity on the ground but on trusted relationships, principled engagement and the ability to adapt to evolving community needs over time.

Humanitarian organizations have a duty of care toward their staff and volunteers to help them cope with the personal and professional impact of harmful information. Exposure to online harassment and harmful information can cause stress, reputational

risk and moral distress. Providing psychosocial support, digital safety training and clear organizational guidance is essential to safeguard their well-being and maintain operational effectiveness.

Who are the influencers in and on the humanitarian sector?

- **Volunteers and frontline staff**: Often the **true bridge** to communities, they are the most trusted voices because they are embedded locally, speak the language(s) and share lived realities.
- **Affected communities**: Displaced people, refugees and disaster survivors often become influencers by sharing lived experiences that reframe global perceptions.
- **Humanitarian organization leadership**: Leadership figures in international organizations and NGOs shape high-level narratives, policy influence and donor agendas. However, they are currently less visible.
- **States** are among the most powerful influencers through funding, legislation, public messaging and control over access. Their endorsement can safeguard humanitarian space, while their rhetoric or laws can just as easily shrink it. States shape narratives, and they can amplify principled humanitarian action and trust or fuel harmful information that delegitimizes organizations and stigmatizes affected communities.
- **Local authorities and officials** play a critical role in shaping perceptions of humanitarian action. Their statements, policies and engagement with communities can either enable or constrain access, influence trust and affect the safety of humanitarian personnel.
- **Community leaders and faith-based actors**: They command local trust and legitimacy and their endorsement can make or break acceptance of humanitarian action.
- **Media and journalists**: They play a major role in shaping public understanding of crises and humanitarian needs.
- **Digital influencers**: Increasingly, popular online voices (YouTubers, TikTokers, diaspora bloggers) can spread humanitarian narratives far faster than official channels.

Does the humanitarian sector 'need' influencers? Yes, for reach and resonance: in a crowded information ecosystem, humanitarian organizations cannot rely only on institutional voices. Influencers – whether local, digital or community-based – help amplify messages, contextualize them and connect them with people's values and concerns. But selectively: not all influencers align with humanitarian principles thus due diligence is required to avoid co-optation, politicization or loss of neutrality. They can provide:

- **Scale**: Influencers can rapidly reach audiences that humanitarian agencies struggle to access.
- **Localization**: Community or local leaders can ensure information is trusted and culturally and contextually relevant.

- **Storytelling**: Personal voices can humanize crises, fostering empathy and solidarity in ways official reports cannot.
- **Countering harmful information**: Influencers can debunk rumours or redirect harmful narratives.

However, there are risks to neutrality and impartiality, of over-reliance and of undermining authenticity or exploiting humanitarian messages for profit. The humanitarian sector *may* need influencers, but it must approach them differently than consumer brands – prioritizing trust, impartiality and long-term community relationships over reach at any cost.

8.5 Humanitarian principles as a compass

The humanitarian principles rely on trust, credibility and clarity of purpose: qualities that can be drowned out in noisy, emotionally charged environments. In a world where **perceived authenticity is the currency of influence**, humanitarian organizations must not only communicate in principled ways but also visibly align their messaging with their actions. The risk is not only that falsehoods outpace facts, but that neutrality, impartiality and independence are misunderstood – or worse, mistrusted – when they fail to align with the prevailing emotional or political narratives of the moment. Humanitarian principles are not only a moral compass but also a vital operational safeguard. The principles serve a dual purpose:

- **aspirational**: reaffirming the humanitarian ideal to alleviate suffering, protect dignity and assist solely on the basis of need
- **practical**: providing a tested framework for maintaining access, navigating contested environments and sustaining trust.

To remain effective, the principles must be actively demonstrated: visible in how principled humanitarian actors engage, how they listen and how they respond. Upholding humanity today means countering dehumanizing narratives, reinforcing dignity through action and communicating with clarity, humility and consistency. Ultimately, principled humanitarian action in the digital age demands more than operational competence. It requires widespread ethical clarity, collective discipline and the courage to resist expedient or reactive narratives. In a world shaped by emotional velocity and harmful information, the **humanitarian principles are more than a moral compass – they are a critical operational safeguard and one of the last defences against the erosion of trust, access and the humanitarian space.**

An emotionally reactive information ecosystem threatens the very conditions on which principled humanitarian action depends – dialogue, trust and space for reasoned engagement. In today's volatile information environment, humanitarian communication is rarely perceived as a neutral act. When handled poorly, it can cross red lines – amplifying harmful narratives, oversimplifying complex realities or becoming co-opted for political or ideological ends. Humanitarian actors face a delicate balance: maintaining transparency while safeguarding operations and security, and upholding neutrality without appearing detached or indifferent. This tension is most pronounced in contexts

where trust has already been eroded. In such environments, even accurate and well-intentioned communication can be met with scepticism, suspicion or outright hostility.

In response, the sector must reaffirm the humanitarian principles not only as a compass, but as a practical framework for navigating contested information spaces. Humanity, neutrality, impartiality and independence are not abstract ideals; they are operational standards, demonstrated through consistent action and credible engagement. In fractured environments, visibly adhering to these principles needs to be reinforced, especially when harmful narratives seek to politicize or delegitimize humanitarian action.

At the same time, the broader information economy is increasingly emotion driven. Surprise, anger and disgust dominate digital platforms, amplified by algorithms and now by AI. Tactical interventions such as digital 'circuit-breakers' and sentiment-based analysis show promise in interrupting the viral spread of emotionally charged harmful information.[18] These measures aim to protect the public sphere by slowing virality, not silencing dissent.

Circuit-breakers

In the digital context, **circuit-breakers**[19] are interventions designed to slow or disrupt the spread of harmful information before it becomes viral. Much like in financial markets, where circuit-breakers halt trading during volatility, these mechanisms temporarily limit the amplification of content that exhibits signs of coordinated manipulation, emotional extremity or rapid spread. This can include platform-triggered slowdowns, content throttling or requiring fact-checking before further distribution. The goal is not to censor, but to create space for verification, reduce emotional escalation and protect public discourse.

Sentiment-based analysis

Sentiment-based analysis uses **natural language processing and machine learning** to detect the emotional tone behind digital content. It categorizes messages as positive, negative or neutral and can further identify specific emotions such as anger, fear or empathy. In humanitarian contexts, this analysis can help organizations understand public mood, track shifts in community perception and anticipate narrative escalation. It supports early warning and communication strategies by highlighting emerging risks, sentiment hotspots or emotional manipulation. It is both what is said – and what is left unsaid – and by whom.

Contributor Insight 8.15

WikiRumours: Crowdsourcing verified information for community safety and prosperity (part 3 of 4) – measuring impact and harm

How was the impact of harmful information assessed?

- Baseline and follow-up surveys measured self-reported changes in access to reliable information, trust in sources and perceived safety.
- Focus group discussions provided context-rich insights into how misinformation influenced decisions, such as whether to flee, return home or accept aid.
- Monitoring of offline and SMS-based channels tracked rumour types, frequency, response time and the effectiveness of efforts to counter misinformation.

Reach: Over 27,000 direct subscribers across South Sudan and DRC received verified updates, with indirect reach exceeding 2 million people through community sharing, radio and posters.

Behavioural change: Surveys found that 85% of users believed the platform (WikiRumours) helped prevent rumours. Communities reported exercising more cautious behaviour – verifying claims before reacting or sharing information.

Service uptake: After verified messages were shared during Ebola/COVID-19 outbreaks, communities showed greater willingness to engage with health services, reflecting improved trust in both humanitarian responders and the information they provided.

Offline feedback: Community ambassadors regularly gathered anecdotal feedback in 'peace spaces' and town hall meetings, which informed strategy adaptations.

Methods used

- **Baseline and follow-up surveys**: These measured changes in self-reported access to accurate information, trust in sources and perceptions of safety.
- **Focus group discussions**: These provided qualitative insights into how misinformation shaped community behaviour and attitudes toward humanitarian aid and services.
- **Rumour report analytics**: The WikiRumours system automatically logged rumours submitted via SMS and voice calls, enabling the team to assess frequency, response time and rumour resolution rates.

- **Community feedback loops**: Regular in-person 'peace spaces' and dialogue sessions gathered real-time insights from residents and local stakeholders.
- **Primarily offline monitoring**: The project relied on SMS short codes, toll-free voice lines and community ambassador reports. These offline channels were most relevant in low-connectivity, rural environments.
- **Limited online monitoring**: In areas like Bunia and Beni with some digital access, project staff informally tracked social media trends, especially during elections. However, systematic online monitoring was not a core component due to infrastructure limitations.

Changes tracked

- **Behaviour**: Community members increasingly verified information before acting. For example, after receiving counter-messaging about a false attack, people chose not to flee, avoiding unnecessary displacement.
- **Trust**: 85% of users surveyed believed that WikiRumours helped prevent the spread of false information. Humanitarian actors also reported improved coordination with communities that received verified updates.
- **Access to services**: In areas where rumours initially discouraged vaccine uptake or use of health services, sharing corrected information led people to attend and engage more with both humanitarian and health services.

Data gaps or challenges in measurement

- **Limited online monitoring**: Low internet penetration meant the project focused on SMS, radio and word-of-mouth channels rather than systematic tracking of online platforms.
- **Pandemic restrictions and logistical barriers**: In rural or volatile areas, these factors reduced the frequency of data collection.
- **Attribution challenges**: Without larger-scale studies, it was difficult to isolate the specific impact of misinformation from other conflict drivers such as insecurity or displacement.

Anahi Ayala Iacucci, Nabeel Chudasama, Nabeela Jivraj, Zainah Alsamman
Grand Challenges Canada

Christopher Tuckwood
The Sentinel Project

Contributor Insight 8.16

WikiRumours: Crowdsourcing verified information for community safety and prosperity (part 4 of 4) – policy and framework gaps

1. **Lack of formal misinformation, disinformation and hate speech policy**: While operational tools exist, there is no dedicated organizational policy on harmful information that includes a standard template and operating procedures across regions and scenarios. Such a policy would make this work easier and more standardized.

2. **Absence of a universal safeguarding policy for digital engagement**: With the growing use of SMS and digital rumour platforms, there are no specific protocols addressing data protection, digital surveillance risks or safe engagement in contested online spaces.

3. **No escalation framework**: There is no formal process for escalating misinformation that threatens aid operations or community safety to local authorities, humanitarian clusters or social media platforms.

4. **No application of a standardized impact measurement framework**: Although monitoring tools are in use, Grand Challenges Canada's impact measurement framework (developed as part of our misinformation, disinformation and hate speech scoping study and designed to track changes in behaviour, trust and rumour prevalence over time) was not yet developed during the project period.

Suggested frameworks and tools that would be beneficial

1. A comprehensive harmful information response protocol integrating rumour tracking, rapid verification, community engagement and staff safety measures.

2. Cross-sector toolkit for humanitarian responders with templates for counter-messaging, local risk assessments and training curricula on harmful information.

3. Safeguarding and risk mitigation guidance for digital rumour collection and verification in high-risk areas, including protocols for staff and volunteer protection.

4 A harmonized harmful information impact measurement framework aligned with broader humanitarian indicators to evaluate changes in trust, access and behavioural response.

Anahi Ayala Iacucci, Nabeel Chudasama, Nabeela Jivraj, Zainah Alsamman,
Grand Challenges Canada

Christopher Tuckwood
The Sentinel Project

Contributor Insight 8.17

International Red Cross and Red Crescent Movement Initiative on Harmful Information

To address the negative impact of harmful information on trust and acceptance in the Red Cross and Red Crescent Movement, the ICRC, IFRC and Swiss Red Cross launched the Movement Initiative on Harmful Information. The initiative's overarching objective is to strengthen the Movement's ability to address harmful information while leveraging its unique strengths, values and global network to safeguard the space for principled humanitarian action. It aims to:

- **Build capacities** of all Movement components to address harmful information through a multidisciplinary approach.
- **Establish a coordination mechanism** for crisis management, enabling early detection, in-depth analysis and collective responses to harmful information incidents.
- **Contribute to global knowledge, humanitarian diplomacy and advocacy** efforts.

Its governance structure includes a Steering Committee, Permanent Secretariat and Coordination Group overseeing four interconnected workstreams: crisis management, training, thematic, and external engagement. The initiative also undertakes to engage in humanitarian diplomacy activities to influence global decision-making in this field. As of June 2025, the Swiss Red Cross is hosting the Movement Initiative on Harmful Information Hub.

Swiss Red Cross

Contributor Insight 8.18

The importance of research and evidence-based practices to counter harmful information and support trust in the humanitarian sector

In an environment where trust is fragile and the consequences of misinformation and disinformation can be immediate and severe, research and evidence-based practices are essential for generating actionable insights that reinforce credibility, resilience and accountability.

Robust research helps humanitarian organizations design interventions that are responsive to local realities and grounded in verified information – reducing the risk of perpetuating inaccuracies and often resulting in more innovative and cost-effective practices. It provides the foundation for understanding the sources, patterns and impacts of harmful information and allows humanitarian actors to move beyond reactive responses and instead develop proactive strategies informed by data, context and community insights.

Evidence-driven communication also fosters transparency and trust with affected populations. When communities see that humanitarian organizations are guided by reliable data and open about both what is known and what remains uncertain, trust grows. This trust is essential not only for countering harmful narratives but also for ensuring that humanitarian responses are accepted, relevant and effective.

As a distributed network of research, academic and scientific entities and initiatives within the Movement, the Red Cross Red Crescent Research Consortium (RC3) supports continued commitment to research and the integration of evidence-driven practices and policies. It aims to ensure the Movement remains trusted and responsive in meeting the needs of affected communities worldwide.

In a sector defined by urgency, constraints and complexity, research is sometimes viewed as a luxury. But in today's volatile information environment, it is a necessity. Evidence-based practice is not just a matter of accuracy – it is a matter of ethics, impact, trust and, ultimately, saving lives. Looking ahead, evidence must become more than a retrospective tool – it must serve as a compass for anticipating challenges, adapting to evolving contexts, informing policy and mitigating the effects of harmful information. This requires investing in predictive research, real-time data analysis and community-based monitoring systems that can detect emerging narratives and inform timely, context-specific responses.

Red Cross Red Crescent Research Consortium (RC3)

8.6 Recommendations for resilience

Building resilience against harmful information and safeguarding principled humanitarian action requires more than isolated interventions. It demands a coordinated, values-driven roadmap embedded across humanitarian diplomacy, community engagement and accountability, safer access initiatives, programme design, preparedness, risk management and communication strategies. The following eight pillars offer actionable steps across short-, medium- and longer-term horizons.

8.6.1 Trust as a strategic asset

Trust is central to humanitarian action – supporting access, operational effectiveness, delivery and legitimacy. Harmful information seldom creates mistrust in humanitarian action on its own: it amplifies existing tensions, inconsistencies and perceived shortcomings. Trust is not static or binary. It exists on a spectrum shaped by cultural experience, power dynamics and exposure to harmful narratives. It evolves along a continuum: **tell me → show me → prove it → keep proving it.** At each stage of this continuum, trust can be weakened or strengthened, but never assumed; in humanitarian crises, it must be continually earned and safeguarded against the risks and corrosive effects of harmful information.

The messenger matters: volunteers, local staff and local leaders often serve as trust brokers. Supporting them to share accurate, timely and accessible information builds credibility from the ground up. Engagement and community information networks are central to resilience and to reduce the vacuum in which harmful information thrives. Leadership accountability and principled consistency across humanitarian organizations and operations are non-negotiable. People assess institutions not only by what they say, but by what they do and whether the two align.

Two dimensions of trust are critical:

- **Operational trust,** grounded in presence and (human) proximity in interactions with affected communities, authorities, armed actors, media and peers. This trust must be personal and embodied: every staff member and volunteer must carry the organization's humanitarian integrity in their behaviour, grounded in principles, standards and professionalism.
- **Institutional trust**, built through principled behaviour, ethical conduct, accountability and regulatory compliance so that stakeholders believe the organization's stewardship, competence, effectiveness and values.

Action points

- **Short term***:* Strengthen internal messaging and training that emphasize trust-building behaviours and responsibilities of individuals as ambassadors of principled humanitarian action. Map trusted community and authority figures (e.g., chiefs, religious leaders, elders) and include them in credibility chains; formalize partnerships to co-verify and disseminate information.

- **Medium term***:* Monitor and analyse community perceptions of trust linked to the operational environment (not only reputation) and impact of harmful information. (The IFRC's Community Trust Index could be extended to measure harmful information.)

- **Long term***:* Integrate trust-based performance metrics that capture both relational and technical dimensions of humanitarian action. Institutionalize multi-layered trust networks as part of preparedness and accountability systems.

8.6.2 'Right-touch' compliance in a digital age

Accountability in the information space is essential to sustaining trust and legitimacy in humanitarian action. Humanitarian organizations must apply the same standards of transparency, responsibility and protection to their communication as in their operations – verifying information, mitigating against harmful content and addressing unintended impacts. Feedback mechanisms should enable communities to question and influence how information about them is used.

Humanitarian actors should promote norms and accountability across the information ecosystem – engaging with stakeholders from technology, media and states to uphold humanitarian principles and protect people in need from harm.

While strong compliance systems support credibility and accountability, excessive bureaucracy and overly rigid procedures can erode trust, add unnecessary burden and weaken the human proximity that is central to humanitarian action. A 'right-touch' (i.e., striking the right balance) approach balances safeguards with flexibility, ensuring compliance reinforces – not replaces – principled and ethical judgement and humanitarian integrity.

States and other donors can support **right-touch compliance** by promoting due diligence that reinforces ethical judgement and humanitarian integrity, rather than imposing overly rigid procedures that slow responses or erode trust. Flexible frameworks ensure accountability while enabling principled, timely decision-making in complex operational and information environments.

Action point

- **Continuous**: Support compliance approaches that balance accountability with flexibility, enabling humanitarian actors to exercise principled and anticipatory decision-making.

8.6.3 Informational and digital literacy and capacities

Effective responses to harmful information require confidence and competence in navigating the information ecosystem and digital space, supported by strong internal capacities, technology access, strategic partnerships and shared standards. This shifts the focus from countering individual messages to understanding broader dynamics of how information is created, shared and trusted within communities, while emphasizing local media ecosystems, social trust and inclusive access to reliable information.

Staff and volunteers should be equipped to navigate digital environments responsibly, recognize emerging threats and engage constructively. Collaborative partnerships with technology actors, civil society and media organizations help promote safe and principled digital practices and advocate for accountability where harm occurs. Community-based digital literacy should be viewed as a protection strategy – empowering people to assess, challenge and contextualize information. Coordination across the Movement and with external actors is essential to share insights, align standards and amplify principled voices, as no single actor can address this challenge alone.

Co-creation with communities ensures they are active partners, not merely recipients, in shaping credible, context-appropriate responses. This approach considers five pillars: the **message** (content and framing), the **medium** (channels of transmission), the **audience**, the **actors** creating and circulating information, and the **impact** on people and systems.

In a post-truth era, staying principled is both an ethical and operational necessity, and digital literacy and access are core enablers of trust and risk management.

Information and digital literacy and access are now core enablers of trust and risk management. Limited literacy, especially at decision-making levels, undermines the ability to anticipate and respond effectively to harmful information. Without the skills, support and infrastructure to navigate today's complex digital information environment, even the most principled strategies risk being reactive rather than anticipatory.

Action points

- **Short term***:* Integrate digital literacy, access and harmful information considerations into programme planning tools, risk matrices and community engagement strategies.
- **Medium term***:* Establish cross-functional teams to design anticipatory strategies and pre-emptive messaging, and to embed digital monitoring skills.
- **Long term***:* Promote sector-wide dialogue to align standards, share insights and amplify trust-based approaches to communication and engagement, backed by sustained investment in information and digital literacy and inclusive access.

8.6.4 Embed risk management in core systems

Risk assessments should address not only physical and reputational risks but also trust-related vulnerabilities, including perception gaps, harmful narratives and community backlash. Humanitarian diplomacy must reaffirm the relevance of humanity, neutrality, impartiality and independence, especially in contested environments.

Actions points

- **Short term***:* Integrate trust and harmful information management and monitoring into risk frameworks and preparedness planning. Invest in systematic audience research, combining digital tools with in-person

engagement to ensure diverse perspectives shape humanitarian engagement and communication strategies.

- **Medium term***:* Establish cross-functional teams to design anticipatory strategies to address potential information risks before they escalate and coordinate principled, pre-emptive messaging.

- **Long term***:* Ensure sector and system-wide dialogue to align standards, share insights and amplify trust-based approaches to humanitarian risk frameworks.

8.6.5 Anticipation and integrated response strategies

A proactive response begins with anticipation and foresight. Humanitarian actors must invest in tools and practices that help them understand and anticipate harmful information, not just react to it. This includes scenario planning, risk mapping and early warning systems to help identify potential sources, narratives and impacts on access, trust and safety. Understanding triggers and enablers – such as grievances, power dynamics and moments of societal stress – supports more targeted preparedness and response. Frameworks like the ABCDE[20] approach can help map actors, messages, distribution mechanisms and effects, though they require analytical capacity and timely data access.

Anticipatory strategies must be community led or guided, locally relevant and adaptive to evolving digital threats.

Resilience demands more than reaction: it requires offensive, defensive and integrated proactive strategies connected in a cycle of protection and influence.

Actions points

- **Short term – defensive:** Rapid fact-checking, crisis communication, quiet corrections and escalation protocols. Develop standard operating procedures for safe use of rumour-tracking and verification platforms. Deploy two-way communication systems (SMS, voice, radio call-ins) for immediate rumour verification. Set up a sector-wide crisis communication taskforce on harmful information and rapid information-sharing protocols on harmful information in risk assessments and cluster mechanisms.

- **Medium term – offensive**: Prebunking, strategic storytelling, values-based campaigns, partnerships with local influencers and community co-creation. Train staff and volunteers on digital security and establish incident reporting protocols for platform misuse. Embed anticipatory messaging protocols into crisis communication plans to pre-empt harmful narratives.

- **Long term – integrated**: Link rapid response to narrative shaping, embed real-time monitoring and coordinate cross-functional teams so defensive and offensive approaches reinforce each other. Embed digital safeguarding and secure communication standards into organizational policies and donor frameworks. Scale and systematize real-time, two-way verification platforms as core infrastructure for community resilience.

8.6.6 Information diplomacy and norm setting

Addressing harmful information requires more than technical fixes – it also demands dialogue, restraint and shared norms. Just as ceasefire agreements limit the use of physical weapons, **harmful information diplomacy** seeks to create voluntary guardrails around the instrumentalization of information. This complements existing humanitarian and peacebuilding efforts by preventing escalation, fostering trust and creating space for principled humanitarian action.

Action points

- **Short term**: Facilitate offline dialogue mechanisms where conflict parties acknowledge the risks of harmful information and agree to red lines or voluntary codes of conduct. Use these as confidence-building measures in peace and mediation processes. Build awareness that ICT networks form part of critical civilian infrastructure and are protected under international humanitarian law. Advocacy efforts include awareness of connectivity gaps that disproportionately affect marginalized groups, including women, displaced persons and people with disabilities, who rely on communication networks for safety, services and participation.

- **Medium term**: Support the development and monitoring of voluntary codes of conduct for digital behaviour in weakly regulated contexts backed by civil society oversight and early engagement with platforms. Advocate for states to recognize ICT restoration in reconstruction plans and include digital access in humanitarian negotiations. Integrate digital safety and harmful information safeguards into all emergency connectivity responses.

- **Long term**: Strengthen local ownership of digital norms by ensuring communities most affected by harmful information are central to shaping, demanding and defending standards for responsible digital and offline behaviour.

8.6.7 Policy and governance for information resilience

Policy gaps remain a barrier to systematic action. While tools exist, most organizations lack dedicated harmful information policies and safeguarding standards for digital engagement. Escalation frameworks for harmful information, protocols for volunteer and staff protection, and standardized understanding of harms and impacts are urgently needed.

Action points

- **Short term:** Develop organizational harmful information policies with clear procedures for prevention, escalation and response. Identify and document impacts and harms through a standard taxonomy or framework of harms.

- **Medium term***:* Establish safeguarding and risk mitigation standards for digital engagement, including data protection and safe use of SMS and rumour-tracking platforms. Encourage media outlets to adopt transparent charters on AI use, ensuring audiences understand how content is generated or assisted by technology.

— **Long term***:* Adopt and institutionalize a harmonized harmful information impact measurement framework aligned with broader humanitarian indicators.

8.6.8

Research, evidence and partnerships

Harmful information is borderless and adaptive; tackling it requires evidence, innovation and collaboration. At present, most humanitarian actors document incidents only anecdotally or as part of broader communication or access challenges. This leaves significant gaps in evidence: the human, social and operational impacts and harms[21] of harmful information remain under-measured compared to physical damage to lives, infrastructure or livelihoods. Without this evidence base, policy responses risk being reactive, fragmented or misaligned with humanitarian principles.

Greater investment is therefore needed in research, tools and partnerships that strengthen resilience at scale. Evidence turns anecdote into accountability. A robust understanding of the impacts of different forms of harmful information enables more effective advocacy with states, regulators and platforms, and helps ensure that humanitarian concerns are embedded in emerging governance frameworks. It also strengthens internal accountability by ensuring responses are data driven, anticipatory and principled.

Action points

— **Short term***:* Map existing incident reporting systems across the sector and align them with harmful information categories (e.g., emblem misuse). Establish rapid reporting channels that capture not only what content spreads, but how it affects safety, access and community trust. Develop and share a cross-sector toolkit for humanitarian responders, including templates for counter-messaging, local risk assessments and training curricula on harmful information. Document and share evidence of its impacts and unintended positive outcomes. Provide thematic training for journalists on issues frequently targeted by harmful information, complementing core journalism standards. Ensure systematic engagement with local media and journalism.

— **Medium term***:* Develop shared metrics to quantify the impact of harmful information on humanitarian outcomes (e.g., delays in aid delivery, reduced health service uptake, safety incidents). Pilot integration of harmful information indicators into needs assessments, early warning systems and programme evaluations.
Invest in research and evidence on harmful information as a humanitarian risk, including its impact on trust, behaviour and access. Identify and pilot sustainable funding models for collaboration on monitoring, verification and moderation systems, including cross-sector partnerships.

— **Long term***:* Institutionalize a global evidence base for harmful information, feeding into humanitarian diplomacy and policy advocacy. This could include an inter-agency repository of cases, impact studies and lessons learned to inform regulation, funding support and norms on information integrity. Build multi-stakeholder partnerships with states, platforms, media actors and community influencers to ensure approaches are locally

grounded but globally coordinated. Develop measurement frameworks that capture co-benefits (e.g., peacebuilding) alongside humanitarian outcomes.

8.7 Prerequisites and cross-cutting enablers

Building resilience to harmful information depends not only on trust, compliance, policy and partnerships but also on a set of enabling conditions that determine whether recommendations translate into practice. Three enablers stand out:

8.7.1 Crisis communication preparedness

Effective responses to harmful information depend on preparedness before a crisis hits with established crisis communication structures with clear roles, escalation protocols and links to staff and volunteers. Regular environmental scanning and decision-making frameworks (when to engage publicly, prioritizing internal communication) ensure rapid, coordinated responses that balance transparency with risk management.

In today's complex information environment, communication is not merely a support function – it is a critical enabler of principled humanitarian action. Moving from reactive to strategic communication enables actors to shape the information environment through transparency, inclusion and dialogue – reinforcing trust and community resilience. Strategic and context-sensitive communication helps to safeguard humanitarian space, sustain access and build trust with communities. When grounded in sound analysis and principled practice, it can prevent the escalation of tensions and reduce harm through effective message framing, audience engagement and dialogue. The aim is to promote alternative narratives, enhance community resilience to harmful information and foster information and media literacy. This goes beyond just correcting falsehoods and aims to change behaviours.

This requires an understanding of local drivers and triggers of harm, as well as investing in trusted, locally anchored engagement. Volunteers, including digital volunteers, can serve as early responders in the information ecosystem to detect and respond to emerging narratives, offering scalable, community-based interventions. This requires solid engagement, internal communication and support. Mapping influencers and narrative dynamics can build understanding of who shapes opinion in specific contexts and how that influence can be used constructively in support of humanitarian response.

Timely, inclusive and principled communication is essential. While information is not water or shelter, it often determines how – and whether – those needs are met and the basis on which people make decisions.

8.7.2 Standardized tools and frameworks

Fragmented responses increase vulnerability. Information Resilience or Information Integrity Toolkits (e.g., Movement Safer Access Framework (being updated), UNHCR

Information Integrity Toolkit (2025),[22] ICRC Framework[23] (2025), IFRC Organizational Capacity Assessment and Certification[24]) demonstrate how structured resources can provide a common approach to prevention, detection, escalation and response. Standardized, adaptable tools which could be socialized cross-sector and with digital rights groups, academia and the private sector to help humanitarians act faster and more consistently across contexts.

8.7.3 Resourcing and professionalizing RCCE

Risk communication and community engagement (RCCE) is not just an add-on service for responses to disasters and emergencies – it enables and drives community-centred and evidence-informed responses by placing the needs, feedback and realities of communities at the forefront.

In times of crisis, RCCE has proven to be not just a tool but a lifeline to protect the health and well-being of communities. Placing the needs, feedback and realities of communities at the forefront of responses strengthens accountability, improves public confidence in response efforts and minimizes the impact of risks on the lives, livelihoods and well-being of those affected. Failing to integrate RCCE and wider community protection technical areas from the outset of a response contributes to miscommunication, mistrust, poor compliance with public health guidance – all of which ultimately lead to a slower and less effective response.

RCCE capacities must be embedded and invested in as a foundational element of disaster and emergency preparedness, so that the systems, capacities and expertise can be leveraged when crisis strikes. This includes:

- Trained, multidisciplinary teams capable of collecting, analysing and applying social science evidence and feedback in real time.
- Proactive community partnerships that build trust, foster co-creation of solutions and ensure marginalized voices are heard.
- Robust feedback loops so that community concerns and insights are systematically captured, analysed and visibly acted upon.

By institutionalizing these elements, responses move from being reactive to systematic, evidence-driven and locally anchored, ultimately strengthening both the effectiveness and legitimacy of emergency response.

Without these enablers, the eight roadmap pillars risk being applied unevenly or only after harm is done. With them, humanitarian actors can move from fragmented, reactive responses to system-wide resilience: prepared, equipped and embedded in communities before, during and after crises.

8.7.4 Harnessing influence for humanitarian resilience

Resilience to harmful information depends on enabling trusted voices to carry principled narratives. Volunteers and frontline staff remain the strongest enablers, embedded in communities and building trust locally. Community leaders, youth networks, etc. amplify this trust and diaspora groups extend influence across borders. Independent media and

journalists provide credibility through accurate reporting, and digital influencers and activists offer reach into fragmented digital spaces, though with risks for neutrality. Above all, technology platforms act as 'meta-influencers', shaping which voices are amplified or suppressed. Advocacy for transparency and rights-respecting governance are vital.

Ultimately, it is **trust, not reach, that turns influence into resilience,** safeguarding humanitarian space and ensuring that communities can act on accurate information.

The IFRC's Community Trust Index offers an evidence-based framework to measure, track and enhance trust between humanitarian organizations and the communities they serve. It assesses community perceptions across two dimensions: competence (technical skills, effectiveness, relevance) and values (integrity, transparency, participation) while also identifying enablers and barriers to trust in areas such as early warning systems, climate resilience, migration and public health. Other organizations and governments could adopt or adapt the index to generate structured, standardized data that complements real-time community feedback. This combination provides a reliable baseline and trend analysis to inform strategic decision-making.

Looking ahead, the Community Trust Index could be further strengthened by integrating targeted focus on harmful information. Doing so would allow organizations to: track the spread and impact of rumours systematically, identify community segments most affected, and bridge qualitative feedback with quantitative insights. By evolving in this way, the index can help anticipate challenges and co-create tailored solutions with communities. Ultimately, the Community Trust Index is then more than a measurement tool – it is a call to action. By diagnosing trust gaps and empowering communities as partners, it equips organizations to rebuild and sustain trust well beyond crises.

8.7.5 Systemic support needs

Expertise and experiences highlight that addressing harmful information cannot rely on isolated projects or short-term fixes. Sustained resilience requires system-level investment, coordination and safeguards. Several gaps stand out:

- **Technical and analytical tools**: Humanitarian actors need the ability to monitor both online and offline spaces, supported by data visualization, narrative and rumour trend mapping and predictive analytics to anticipate harmful narratives before they spread.
- **Monitoring and evaluation frameworks**: Impact must be measured not only in outputs (e.g., messages delivered) but also in behavioural shifts, perception change and trust dynamics at community level.
- **Security protocols:** Dedicated standards are needed for secure data handling and staff safety in contexts where armed groups or political actors monitor or exploit information systems.
- **Sustainable financing**: Rumour-tracking and verification systems are resource-intensive. Without multi-year, flexible donor support, community capacity and infrastructure cannot be maintained or scaled. There is real potential for cross-sector collaboration to benefit from economies of scale.

- **Policy and coordination guidance**: Clear roles, standards and escalation pathways are required within humanitarian coordination mechanisms to avoid fragmented or reactive responses.
- **Integration of local media:** Community radio stations and local journalists should be actively engaged as frontline responders to harmful information, co-producing and distributing trusted, localized counter-messages.
- **Digital inclusion and platform accountability**: Platforms must adapt their tools to humanitarian realities, providing low-bandwidth, multilingual and locally contextualized solutions while ensuring transparency and safeguards.
- **Ecosystem-wide literacy**: Media and rumour literacy should be extended beyond technical staff to volunteers, health workers, teachers and community members, embedding resilience across whole populations.
- **Anticipatory systems**: Predictive modelling and pre-bunking strategies – delivered via SMS, posters or radio ahead of flashpoints such as elections or vaccination drives – can reduce the space for harmful rumours to take hold.

Together, these support needs point to a critical shift: from isolated interventions to systemic resilience, where humanitarian actors, governments, donors, platforms and communities co-invest in shared infrastructure for trustworthy information.

Conclusion: Together, we can uphold and reclaim space for humanity

Trust is not assured; it is built, reinforced and renewed. In the face of harmful information, it remains the most powerful safeguard for humanitarian space. By investing in trust, embedding right-touch compliance, addressing policy gaps and advancing evidence-based partnerships, the humanitarian sector can shift from reactive counter-narratives toward systemic resilience.

Harmful information cannot be addressed piecemeal. Building resilience requires trust at the centre, compliance that enables rather than obstructs, integrated risk management, stronger policies, deeper evidence and partnerships – and above all, proactive strategies that connect defensive and offensive efforts. Silence and delay carry their own dangers; early, transparent and trusted communication and engagement is the most powerful safeguard for humanitarian space. Governments need to act to preserve this humanitarian space.

The humanitarian sector cannot afford to cede the information space. While the speed, scale and sophistication of harmful information poses significant challenges, disengagement is not a viable option. To maintain access, credibility and principled impact, humanitarian actors must engage deliberately grounded in the humanitarian principles,

supported by collaboration and partnerships, and informed by a clear understanding of the social dynamics shaping today's contested narratives.

Reclaiming narrative space requires more than correcting harmful information. It calls for a reframing of communication as connection, rooted in listening, empathy, proximity, humility and consistency. It also requires organizations to understand how harmful information spreads and why people believe or accept it, addressing not just the content of falsehoods but the emotions, fears and identities that give them power.

In times of crisis, harmful information thrives by reducing complex realities into simplistic explanations and easily identifiable enemies. Principled humanitarian action must not only speak truth, it must also understand fear, identity and belonging. In this fragmented, emotionally charged landscape, reclaiming narrative space is not about controlling the story. It is about restoring trust, rebuilding connection and reasserting the relevance of humanitarian principles and action in the eyes of affected communities and the broader public. Governments have an important role in this regard to reinforce the importance of principled humanitarian action and preserving the space for humanitarian organizations to operate.

The Movement's Resolution on Tolerance offers a valuable foundation to reinvigorate efforts against harmful information – reminding us that respect, diversity and non-discrimination are not abstract ideals but practical tools to reduce polarization, counter dehumanizing narratives and preserve humanitarian space. Tolerance online also has its limits: "what is often framed as a fight over *speech* is actually a fight over *reach*" (DiResta[25]) – the algorithmic amplification that determines which voices are elevated, repeated and made unavoidable. The humanitarian sector must advocate for changes that reduce the reach of hate speech and malicious content that imperil humanitarian action and endanger staff, volunteers and crisis-affected populations.

The scale of the challenge may feel overwhelming, but it need not paralyse us. By acting collectively – rooted in principles, grounded in trust and united across communities, states, platforms and humanitarian organizations – the sector can move forward with courage and clarity.

The intersection of resilience, principled communication and institutional credibility is now a critical arena for action. Humanitarian actors must shift from a defensive posture to proactive, adaptive and systematic approaches that prepare for and mitigate the impact of harmful information, particularly where it threatens access, undermines trust or endangers lives. Delivering this requires a cross-functional approach in collaboration with communities. **The right to information is contested, but in humanitarian crises, the need to know elevates information to a core element of response.**

Together, we can uphold and reclaim space for humanity.

Endnotes

1 IFRC. *Community Engagement and Accountability Toolkit*. (2020) https://communityengagementhub.org/resource/cea-toolkit

2 UN. *Our Common Agenda: Report of the Secretary-General*, UN Doc. A/75/982 (2021) p.27 www.un.org/en/content/common-agenda-report/

3 Levitin, DJ. *A Field Guide To Lies: Critical Thinking in the Information Age*. (2016) He highlights that we need to use just *some* of the time we saved in information acquisition to perform proper information verification, p.253

4 Singer, PW. and Brooking, ET. *LikeWar: The Weaponization of Social Media*. (2018) p.261

5 Kreps, S. *Democratizing Harm: Artificial Intelligence in the Hands of Non-state Actors*. Foreign Policy at Brookings, (2021) pp.2–5 www.brookings.edu/articles/democratizing-harm-artificial-intelligence-in-the-hands-of-non-state-actors

6 World Economic Forum. *Global Risks Report 2025*. (2025) www.weforum.org/publications/global-risks-report-2025/digest. The report identifies the convergence of threats – including geopolitical, environmental, societal and technological risks – as a defining feature of today's global risk landscape.

7 Edelman. *2024 Edelman Trust Barometer: Global Report*. (2024) The report notes that media became the least trusted institution in 2020, while social media has been the least trusted source of news and information since 2016. www.edelman.com/trust-barometer

8 Luhmann, L. *Trust and Power.* (John Wiley and Sons, 1979)

9 Singer, PW. and Brooking, ET. *LikeWar: The Weaponization of Social Media*. (2018) p.186

10 ICRC. *Building a Responsible Humanitarian Approach: The ICRC's Policy on Artificial Intelligence*. (2024) www.icrc.org/en/publication/building-responsible-humanitarian-approach-icrcs-policy-artificial-intelligence

11 CDAC Network, The Alan Turing Institute and Humanitarian AI Advisory. *SAFE AI: Standards and Assurance Framework for Ethical Artificial Intelligence in Humanitarian Action*. (2025) www.cdacnetwork.org/safe-ai

12 Kaspersen, A. and Cavelty, MD. Digitalisation of Conflict, GESDA Science Breakthrough Radar. p.3

13 UN. *Open-ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security: Final Substantive Report*. UN Doc. A/75/816 (2021) https://digitallibrary.un.org/record/3906060

14 ICRC. *Addressing Harmful Information in Conflict Settings: A Response Framework for Humanitarian Organizations* (2024) https://shop.icrc.org/addressing-harmful-information-in-conflict-settings-a-response-framework-for-humanitarian-organizations-pdf-en.html

15 University of Melbourne. *City Playbook for Countering Disinformation*. Melbourne Centre for Cities and Centre for Artificial Intelligence and Digital Ethics. (2022) www.unimelb.edu.au/cities/research/projects/current-projects/disinformation-in-the-city/disinformation-in-the-city-response-playbook

16 Defined as 'wars' or 'limited wars' by the Heidelberg Institute for International Conflict Research.

17 7amleh. '#ReconnectGaza: A global campaign to restore connectivity in Gaza.' News. Association for Progressive Communications. 26 February 2025. www.apc.org/en/news/reconnectgaza-global-campaign-restore-connectivity-gaza; York, JC. 'Connectivity is a Lifeline, Not a Luxury: Telecom Blackouts in Gaza Threaten Lives and Digital Rights.' Electronic Frontier Foundation. 16 June 2025. www.eff.org/deeplinks/2025/06/connectivity-lifeline-not-luxury-telecom-blackouts-gaza-threaten-lives-and-digital

18 Johns Hopkins University, Imperial College London and Georgia Institute of Technology. *Countering Disinformation: Improving the Alliance's Digital Resilience*. NATO Review. (2021) ch.1 https://archives.nato.int/nato-review-countering-disinformation-improving-the-alliance-s-digital-resilience. The article discusses AI-based sentiment analysis, emotional predictors, and automated 'circuit-breakers' to slow the virality of harmful content while protecting freedom of expression.

19 Ibid

20 A – Actor, B – Behaviour, C – Content, D – Degree, E – Effect. Pamment, J. *The EU's Role in Fighting Disinformation: Crafting A Disinformation Framework*. (2020)

21 ICRC. *Addressing Harmful Information in Conflict Settings: A Response Framework for Humanitarian Organizations*. (2025). The ICRC determines that decisions about if, when and how to respond to harmful information are based on the harm potential, spread potential of information and associated risk indicators, p.12; Lindsey, C. and Glasser, G. *Report of Second Expert Meeting on the Development of a Harms Methodology*. CyberPeace Institute. (2024) pp.5–6 defines harm as: "an impairment or disruption of an entity's capacity or ability to function and exist as it otherwise would have in its usual context". This definition identifies four levels of harm affecting individuals, organizations, societies and international peace and security. See also `UN High-Level Advisory Body on Artificial Intelligence. *Governing AI for Humanity: Final Report*. (2024) cited in Paris Peace Forum. *Forging Global Cooperation on AI Risks: Cyber Policy as a Governance Blueprint*. (2025). This highlights that a strong emphasis on risks and harms should be at the centre of AI governance "that focuses on who is at risk and accountable, and not just what is at risk".

22 UNHCR. *Information Integrity Toolkit*. (2025) www.unhcr.org/handbooks/informationintegrity

23 ICRC. *Addressing Harmful Information in Conflict Settings: A Response Framework for Humanitarian Organizations*. (2024) www.icrc.org/en/publication/addressing-harmful-information-conflict-settings-response-framework-humanitarian

24 The OCAC is a comprehensive process that helps National Societies review their capacity and performance. It enables them to identify strengths and weaknesses, focus efforts to become strong and sustainable service providers, and measure themselves against the minimum standards expected of modern humanitarian and development organizations. IFRC. Organizational Capacity Assessment and Certification. (2019) https://data.ifrc.org/en/ocac. See also Policies and key commitments. www.ifrc.org/our-promise/trust-and-accountability/policies-and-key-commitments

25 DiResta, R. *Invisible Rulers: The People Who Turn Lies into Reality.* (2024) p.318