Chapter 5

Navigating regulation, rights and societal resilience


CROCE ROSSA ITALIANA

Chapter 5

Navigating regulation, rights and societal resilience

Contents

Introduction: Information landscape and humanitarian contexts

The spread of false information – whether overt or covert – poses a serious threat to the humanitarian sector, undermining its work and endangering the populations it aims to serve. While not new, digitalization has transformed the scale, speed and complexity of harmful information. This shift is unfolding amid growing distrust in institutions, making its effects even more corrosive. Today's information environment is crowded and layered, with multiple forms of harmful content coexisting and reinforcing one another. Harmful information rarely exists in isolation; it amplifies other risks such as geopolitical tensions and environmental crises. The World Economic Forum's *Global Risks Perception Survey 2024–2025*[1] warns that vulnerabilities linked to online activity are deepening alongside widening societal and political divisions, eroding public trust in information and institutions.

Humanitarian contexts are inherently complex, particularly during disasters, armed conflict and other emergencies. In such situations, the state plays a central role in shaping the information environment – serving as the primary source of official communication, issuing alerts, coordinating messaging and engaging with both domestic and international responders. Timely, accurate and trusted information is essential: it saves lives, fosters trust and supports public order. However, malicious actors often seek to control or distort these narratives. Countering such efforts requires not only credible channels of communication but also organizational and societal resilience.

An important actor navigating this complexity at the national level is the National Red Cross and Red Crescent Society. National Societies hold a unique position as **auxiliaries to public authorities in the humanitarian field** – a status that allows them to support and engage with the state while maintaining independence and adherence to humanitarian principles.

The status of National Red Cross and Red Crescent Societies is not arbitrary; it is formally recognized by states that are signatories to the Geneva Conventions. These states have committed – through both the UN and the International Conferences of the Red Cross and Red Crescent – to uphold and respect the Movement's fundamental principles. These principles – humanity, impartiality, neutrality, independence, voluntary service, unity and universality – are essential to ensuring that National Societies can operate effectively and free from political interference. In practice, this means that National Societies must be able to deliver humanitarian assistance independently, impartially and neutrally, even in complex or politically sensitive environments. Respect for these principles is not only a matter of good practice; it is a legal and moral obligation for states that have endorsed the Geneva Conventions and the commitments made in international humanitarian forums.

Traditionally, this auxiliary role has encompassed disaster preparedness and response, health services, support to vulnerable populations and the promotion of humanitarian principles and international humanitarian law. In practice, it also demands trusted public awareness, communication of information and meaningful community engagement – essential to effective humanitarian action. National Societies contribute in multiple ways: communicating early warnings and public health messages; coordinating with authorities to ensure consistent, timely and localized information; supporting risk

communication and community feedback mechanisms; and increasingly working to counter harmful information and rumours that may undermine humanitarian responses.

Given the complexity of today's information environment, there is a need for greater recognition of – and support for – the independence of National Societies in fulfilling these roles. The 2024 Council of Delegates called on each National Society to strengthen dialogue with public authorities to reinforce their independent action and decision-making, including by anchoring their auxiliary role in domestic law in line with Movement standards and past International Conference resolutions.[2]

5.1 Defining harmful information: A strategic and contextual challenge

Determining what constitutes harmful information is essential for any entity seeking to develop a strategic response aligned with risk management. The impact of harmful information depends on several variables including its scope (scale and severity), duration, magnitude of the incident and the resilience of the affected individual(s), organization(s) and broader context. Harm may result directly from the incident or indirectly, with distinctions often based on the degree of certainty that the information caused the outcome. A single incident can generate multiple forms of harm, classified as direct (primary) or indirect (secondary or tertiary). These may include physical, psychological, societal and deprivational harms. The effects of harmful information are often disruptive and multidimensional, manifesting across several categories simultaneously.

Governments interpret harmful information in diverse ways, shaped by their legal frameworks, political priorities and societal values. Some describe it as deliberately false or misleading content intended to deceive, manipulate or cause harm – whether to individuals, public institutions or national security. Others place greater emphasis on intent, such as its use to influence elections, incite violence or undermine public health measures.

Common definitions distinguish between:

- disinformation: deliberate falsehoods intended to cause harm
- misinformation: false or inaccurate information shared without harmful intent
- malinformation: genuine information used with the intent to cause harm.

In these framings, intent is the key differentiator.

One of the most contested areas in defining harmful information is the boundary between legitimate political dissent and incitement. In some contexts, this line is blurred, raising serious concerns about freedom of expression. Content may be labelled as dissenting, destabilizing or inciting violence, particularly during times of crisis or internal tension. The challenge is ensuring that measures designed to counter harmful information are not used as a pretext for silencing dissent. Clear safeguards are needed

to distinguish between speech that challenges authority and speech that genuinely threatens public order, safety and human dignity.

> **"You know, … I don't think the law has caught up with social media and being able to kind of hold people to account for the information that they've been putting up there. And I think, … you know, we've got a right to kind of free speech. So … there's a conflict there, isn't there? To a certain extent, you know, he can say anything he likes, but it could be false… And then there's the information that people have a right to know, what the truth is. So I think there's a difficult balance to get between the laws and free speech."**
>
> Community member, UK

5.2 The risks of information control in emergencies

While states have a legitimate responsibility to protect territorial integrity and public order, overly broad or opaque security measures can come at the expense of individual rights and community safety. In some contexts, authorities have withheld or delayed the release of critical information, shared incomplete or misleading narratives or used public messaging to shape perceptions during emergencies. In situations of armed conflict, information often becomes a strategic instrument – used to influence, mobilize or obscure.

Measures such as restricting, blocking, filtering, censoring or regulating content are sometimes introduced under the pretext of national security, public order or preservation of culture. Yet, when vaguely defined or applied without transparency, such measures can undermine access to essential information, restrict freedom of expression and compromise the integrity of the information space. Those with disproportionate control over media and platforms can use harmful information to discredit civil society and humanitarian organizations, associating them with malign or unlawful actors or blaming them for crises to justify repressive policies. This manipulation fuels discrimination, human rights abuses and social tensions.

At the core lies a critical tension: addressing harmful information effectively while safeguarding fundamental rights. Vague or overly broad definitions of harmful information risk misuse – suppressing journalism, silencing dissent or curtailing civil society. In some contexts, legislation intended to combat disinformation has been used to detain journalists, restrict reporting or target individuals for expressing political opinions. Such measures can erode public trust and deepen fear and polarization. Effective responses to harmful information must therefore be grounded in legality, necessity and proportionality.

A body of UN human rights instruments, resolutions and guidance underscores the importance of protecting human rights online, including freedom of expression and

access to information. These instruments provide essential guidance in navigating the tensions between addressing harmful information and safeguarding fundamental rights. For example, Article 19 of the Universal Declaration of Human Rights (1948) states that "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers."[3]

One notable multilateral initiative is the Freedom Online Coalition, a group of 42 states committed to advancing internet freedom. The coalition has repeatedly urged governments to refrain from sponsoring disinformation, including campaigns that undermine humanitarian principles or incite violence. It also emphasizes the importance of maintaining a global, free, open, secure and interoperable internet and other digital communications services during times of armed conflict. The coalition has warned that internet shutdowns hinder access to life-saving information for crisis-affected populations, disrupt protection mechanisms and vital services and obstruct the delivery of humanitarian assistance.[4] Importantly, this underscores not only the dangers of disinformation but also the risks posed by the absence of information – a critical concern in emergencies where timely communication can mean the difference between life and death.

Contributor Insight 5.1

Estonia: Upholding freedom while countering harmful information

Estonia ranks among the freest and most digitally advanced societies, with a culture that values openness and democratic resilience. According to Freedom House's ***Freedom on the Net*** report, Estonia has one of the world's most open online environments, with no state-imposed restrictions on expression. The 2025 ***World Press Freedom Index*** ranks Estonia second globally, emphasizing a robust legal and political framework that enables journalists to operate safely and independently.

Estonia distinguishes clearly between free media and hostile propaganda. We do not engage in counter-propaganda but instead enforce EU sanctions to limit harmful disinformation. This regulatory, rather than ideological, approach safeguards facts and democratic integrity. Estonia's response is multifaceted, combining regulation, education and public engagement.

Domestically, harmful disinformation is treated as a national security and societal challenge requiring a whole-of-society response. Estonia promotes media education across all school levels, supports independent journalism, monitors hostile narratives in collaboration with the public sector, NGOs and volunteers, and maintains open communication with both media and citizens. The Government Office coordinates strategic communication across ministries, including psychological defence and crisis messaging. Roles are clearly defined: the Consumer Protection and Technical Regulatory Authority enforces the Digital Services Act, the Information System Authority and CERT-EE manage cyber incidents and the Internal Security Service addresses foreign influence.

Internationally, Estonia promotes both digital rights and media freedom. As the 2025 Chair of the Freedom Online Coalition, we focus on ensuring that emerging technologies, including AI, are governed in a rights-respecting way, while advancing digital inclusion and cross-regional dialogue. As Co-Chair of the Media Freedom Coalition from mid-2023 until mid-2025, Estonia prioritized journalist protection, supporting independent media and raising awareness of disinformation risks. Strong international cooperation underpins this approach, reflecting Estonia's commitment to human rights and resilient, open and secure information ecosystems.

Our lesson is clear: freedom must be protected as a core value, but it must also be paired with proportionate regulation, sanctions enforcement, strong cyber defences, education and societal engagement. Only this balance ensures the protection of freedom of expression in a democratic society, while remaining resilient against information manipulation.

Maarja Kask
Third Secretary, Department for International Organisations and Human Rights
Ministry of Foreign Affairs of Estonia

Silver Küngas
First Secretary, Department for International Organisations and Human Rights
Ministry of Foreign Affairs of Estonia

5.3 Digital ceasefires and harmful information

While traditional ceasefire agreements focus on halting the use of force and physical violence, there is growing recognition that the information space – including the spread of harmful information – can also threaten the sustainability of peace. Harmful information, such as hate speech, incitement to violence and disinformation, have been used to undermine ceasefires, discredit negotiation parties or provoke renewed hostilities. In some recent contexts, mediators and peacebuilders have begun to acknowledge the strategic role of information and have advocated for explicit clauses in ceasefire and peace agreements that explicitly reference the prohibition of hate speech, incitement and disinformation, as well as the importance of access to the internet and accurate information. However, such provisions remain rare, vague or insufficiently monitored.

Incorporating information-related commitments in ceasefire agreements, such as respecting media freedom, refraining from online incitement and establishing joint communication mechanisms, could strengthen trust between parties and help prevent a relapse into violence. This is particularly critical in contexts where digital platforms are actively used to mobilize support, reinforce polarization or spread disinformation.

Contributor Insight 5.2

Humanitarian impacts of digital pollution: The case of Libya

Access to information and the ability to communicate are central to the enjoyment of several human rights, especially in situations of crisis. Just like access to food, access to information is multidimensional – it cannot be satisfied simply by the existence of an information environment.

Internet shutdowns force people into fragmented or unreliable information bubbles, removing them from the ability to verify or even access potentially life-saving information. At the same time, the degradation of content in the information ecosystem can render access to information ineffective or even dangerous.

Unsurprisingly, these two phenomena are increasingly documented in combination, especially in the aftermath of disasters. Local activists have highlighted this interplay during the 2023 floods that ravaged north-eastern Libya, where these risks converged with deadly consequences.

As reported by Libyan journalists and human rights defenders, the mix of internet shutdowns, breaks in connectivity and an unchecked spread of misinformation and disinformation significantly increased the disaster's impact. It increased confusion, eroded credibility and trust in disaster response and delayed life-saving assistance. The result was a slower, less effective humanitarian response but, most tragically, a much higher number of casualties and greater damage to affected communities.[5]

Access Now and other digital rights organizations work to ensure that affected communities have unrestricted access to secure, reliable information, a longstanding challenge in the humanitarian sector where significant progress has been made over time.

However, these gains are now under threat. The normalization of internet shutdowns, combined with the shrinking of funding for local digital rights initiatives and journalism is threatening these gains. Meanwhile, the spread of so-called 'AI-slop' – low quality, AI-generated content – is further polluting the information ecosystem, pushing it beyond a critical threshold. As local initiatives struggle to stay afloat, hostile actors are stepping into the void driven by political or financial motives. In some cases, these actors offer funding under conditions incompatible with humanitarian principles.

As seen in Libya, local journalists, relief organizations and digital rights defenders continue to lead the response, stepping up when their communities need them most. But without sustained donor and tech sector support and stronger protection from the international community, the information environment in crisis settings will continue to worsen and in turn worsen the harmful impacts of disasters.

Giulio Coppi
Senior Humanitarian Officer
Access Now

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

5.4 Sovereignty in cyberspace

Determining whether an influence operation in cyberspace is a breach of sovereignty is a complex issue under international law. It is closely tied to whether the act infringes on a state's territorial integrity or interferes with inherently governmental functions, such as delivering public services, conducting elections or collecting taxes.

Discussions at the UN Open-ended working group on responsible use of information and communications technologies (ICTs) have examined these issues in the context of cyber incidents.[6] Key considerations include whether an incident damages or limits the functionality of infrastructure or related equipment; alters or deletes data; interferes with inherently governmental functions; and whether a state has sought to influence, disrupt or delay democratic processes in another state which could include through propaganda, disinformation or covert actions.[7] Thus, whether such acts amount to a violation of sovereignty depends on their nature and repercussions: Does the incident violate territorial integrity? Does it interfere with or usurp an inherently governmental function?

Researchers such as Pamment emphasize the importance of assessing *foreignness* in influence operations, specifically, whether they have connections to foreign states, citizens or interests aiming to influence public opinion in another state.[8]

5.5 Media as a pillar of societal resilience

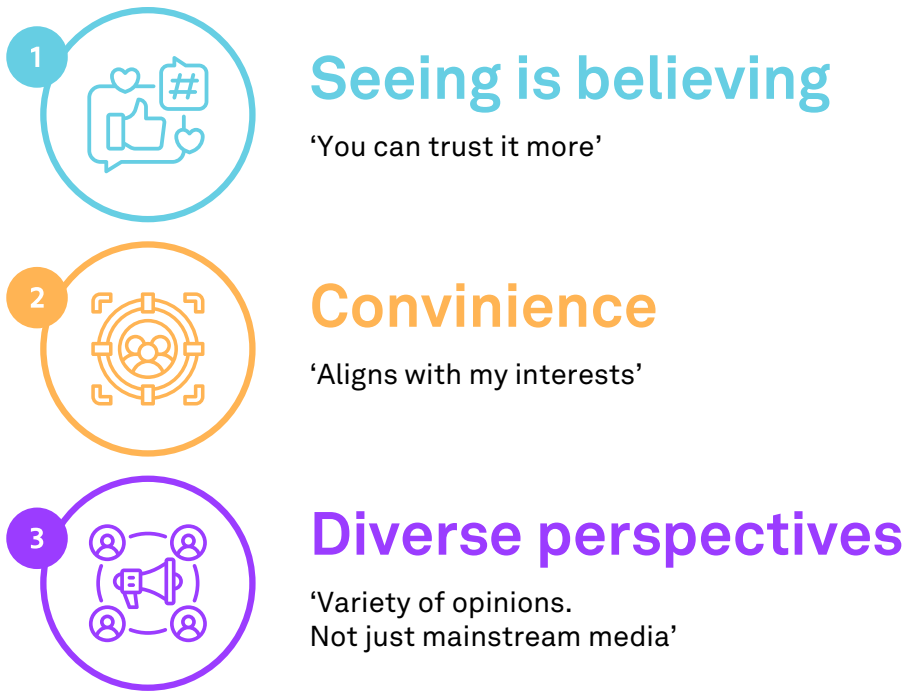
The level of support – or lack of – that states and other stakeholders show for the presence of humanitarian organizations on their territory can significantly influence how they are perceived. When an organization is portrayed as biased, politicized or pursuing its own agenda, its legitimacy is undermined, particularly in contexts where there is limited public knowledge or familiarity with its mandate or work. In such environments, perceptions matter deeply. Humanitarian organizations must be seen as principled, competent and effective. Where public awareness is low and/or where polarization is high, people are more susceptible to harmful information, making them more vulnerable to misconceptions and distrust when forming their opinions.

Researchers Humprecht, Esser and Van Aelst identify five key risk factors that influence a society's vulnerability or resilience to disinformation.[9] These are: (1) high levels of societal polarization; (2) low trust in news media; (3) highly distributed media landscapes (offering more entry points for disinformation); (4) large media markets, where attention-based revenue models incentivize sensational or false content – what the Global Disinformation Index describes as "the loudest voices getting the most attention"; and (5) high levels of social media use, which are consistently correlated with greater susceptibility to disinformation.

According to the Reuters Institute *Digital News Report 2024*, audiences are drawn to video and other content on social and video platforms for three main reasons: the perceived authenticity of unfiltered, user-generated content, which appears more *trustworthy*; a preference for making up one's own mind without editorial framing; and growing mistrust in traditional media.

Fig 5.1

Motivations for using social video



Source: Reuters Institute for the Study of Journalism[10]

Many users, especially younger audiences, tend to trust bystander or first-person footage more than traditional news sources, perceiving it as less filtered, less biased and/or less politically manipulated. Concerns about misinformation often centre less on entirely false content and more on seeing opinions and agendas that they may disagree with, perceived bias or superficial, unsubstantiated journalism.[11]

These dynamics underscore the importance of visibility, transparency and trust-building in humanitarian communication – especially in environments where disinformation can fill gaps left by limited public understanding.

Contributor Insight 5.3

Audiences increasingly rely on social media for news despite concerns about information quality

Social media and video networks have become increasingly central to how people access news in recent years, while other forms of news consumption have dwindled. Data from

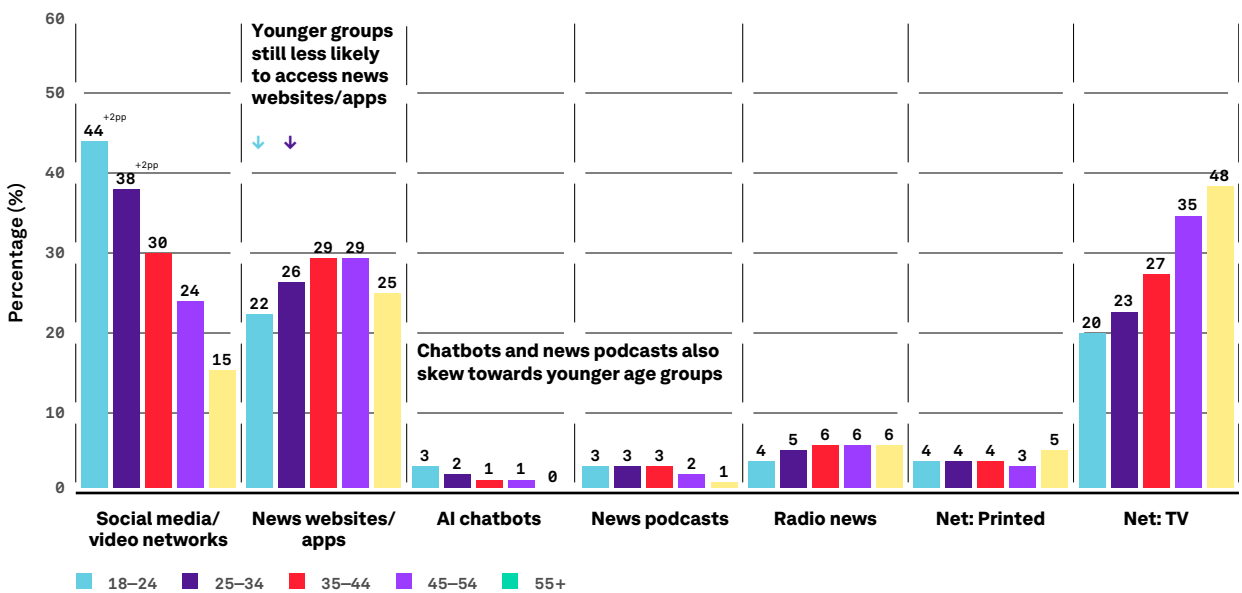
the Reuters Institute *Digital News Report 2025* show that social media overtook television as a source of news in the US for the first time, and in many Latin American, Asian and African countries, social media has been the main source of news for some time.

Generational shifts are driving much of this change. While older adults (55+) continue to rely on traditional media for news, everyone else is now effectively digital-first. Within that online space, younger groups have become less likely to go directly to a news website or app – and more likely to consume news via social media or video networks. Across the 48 countries surveyed, 44% of under-25s and 38% of those aged 25–34 prefer accessing news through platforms, while just a quarter of each group prefer going directly to news websites. Respondents cite convenience and relevance of the news they see through platforms as key reasons – it is often encountered while doing other online activities.

Fig 5.2

Proportion of each age group that say each are their 'main source' of news, 2025

Aggregate data from 48 countries



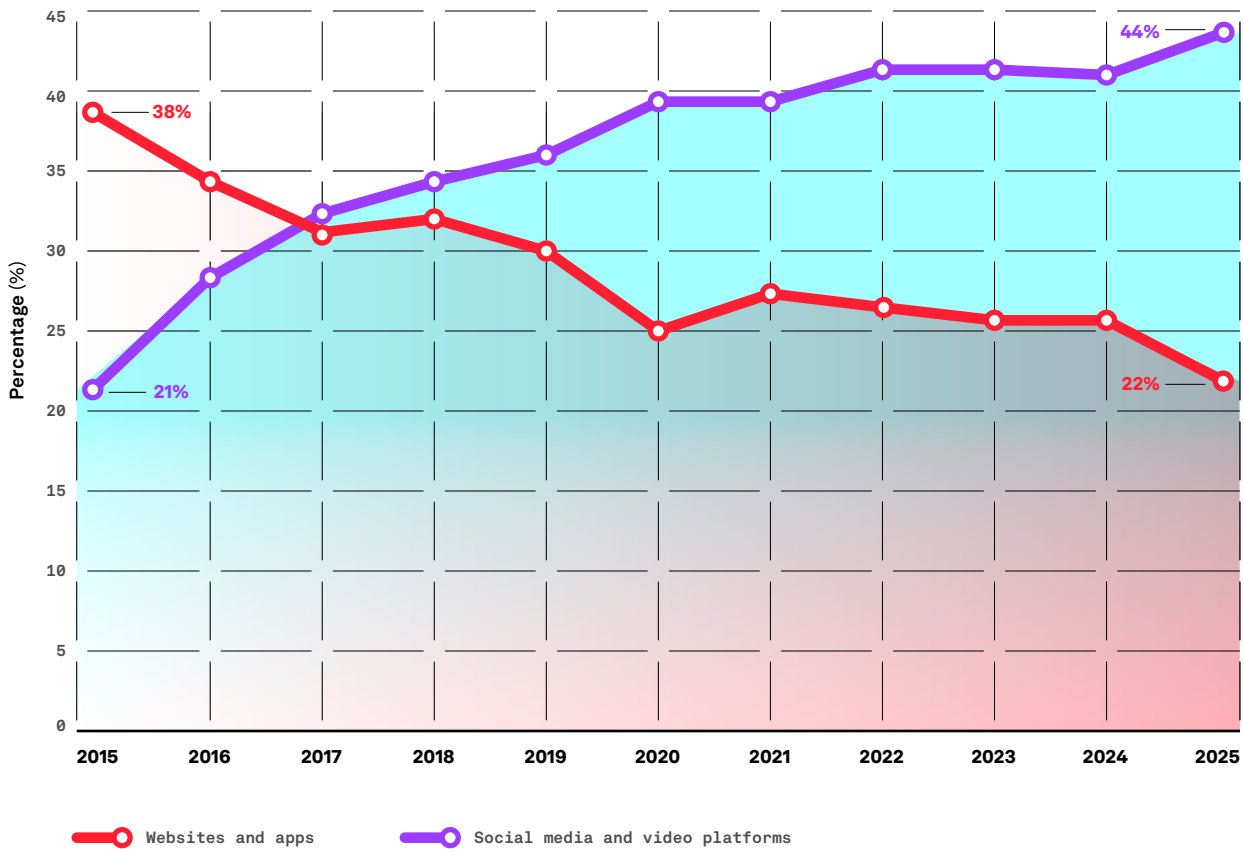
Q **Question:** *You say you've used these sources of news in the last week, which would you say is your main source of news?*

B **Base:** *Respondents who used a source of news in the last week: 18–24 = 9,807, 25–34 = 15,722, 35–44 = 16,354, 45–54 = 15,804, 55+ = 33,449.*

Fig 5.3

Proportion of under-25s who prefer to access news via websites or social networks, 2015–2025

Aggregate data from 48 countries



Q **Question:** *You say you've used these sources of news in the last week, which would you say is your main source of news?*

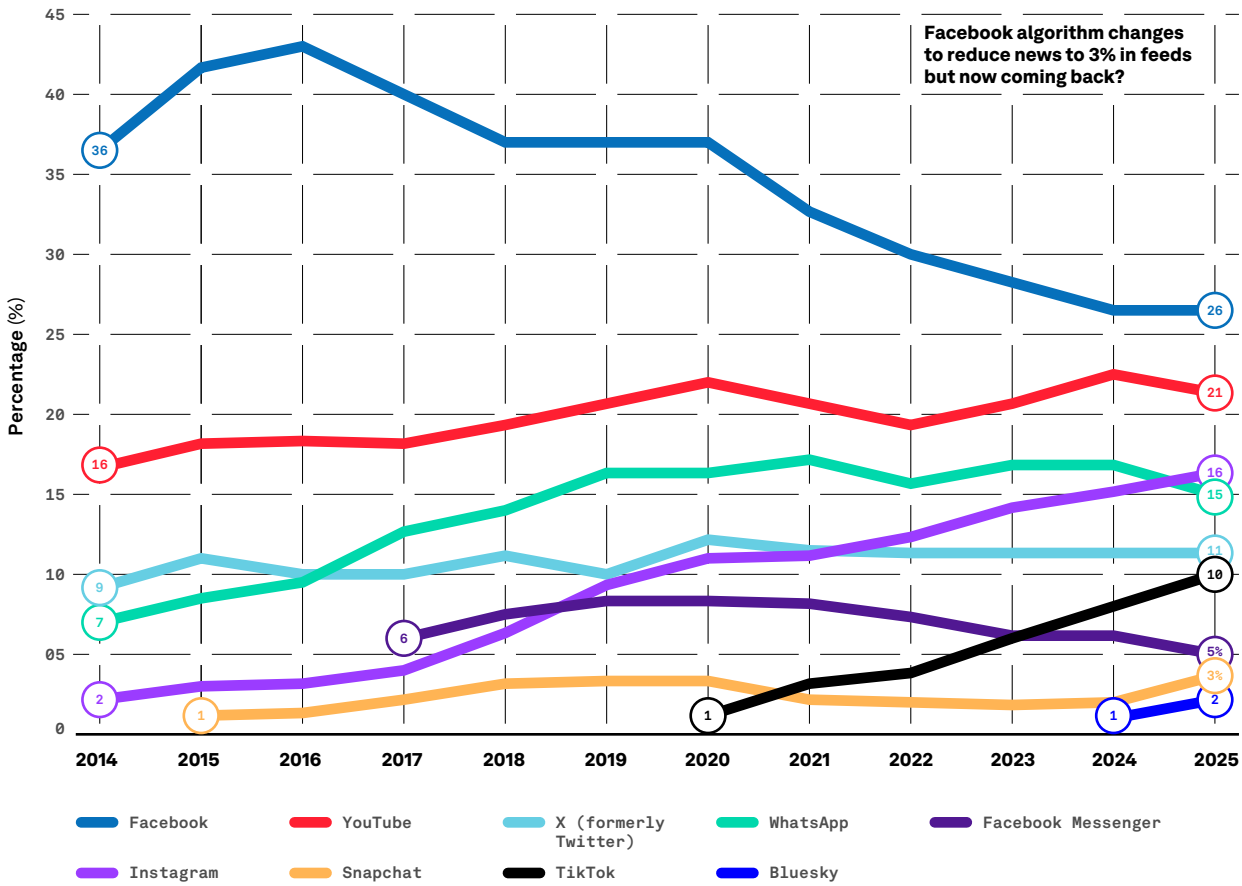
B **Base:** *18–24 = ranging from 5,598 in 2016 to 9,807 in 2025. Note: Number of markets surveyed in 2016 (26), 2017 (36), 2018 (37), 2019 (38), 2020 (40), 2021 (46), 2022 (46), 2023 (47), 2024 (48), 2025 (48)*

The specific platforms used for news have also changed over time. While Facebook remains the main social network for news, its use has declined considerably since 2016. Meanwhile, video-based platforms such as YouTube, Instagram and TikTok have grown in popularity, resulting in more fragmented attention. Across 12 countries tracked since 2014, six networks now have a weekly reach of 10% or more of the population, compared to only two platforms a decade ago.

Fig 5.4

Proportion of respondents who used each as a source of news in the past week, 2014–2025

Selected countries



Q **Question:** *Which, if any, of the following have you used for finding, reading, watching, sharing or discussing news in the last week?*

B **Base:** *Total sample in each country-year in UK, US, Germany, France, Spain, Italy, Denmark, Finland, Australia, Japan (2014–2025), Brazil and Ireland (2015–2025) around 2000. Note: We did not ask about Bluesky in France, Italy, Finland, Denmark, Japan and Canada (2024) and in France, Italy, Denmark and Japan (2025).*

Across these crowded digital platforms, news organizations must compete for attention with a range of other voices, many of whom are often more effective at attracting audiences through partisan commentary, engaging story-telling and relatable personalities often perceived as more authentic.

Our data show that traditional news media particularly struggle to retain audience attention on video platforms, where algorithmic recommendations – rather than, e.g., the suggestions of friends – play a dominant role. While audiences still tend to pay more attention to traditional news media and journalists on Facebook, online creators and personalities now outperform established news media on platforms like TikTok and Snapchat.

Even as audiences are embracing platform-based consumption for news, many express concern about the reliability of the information they find there. Facebook (the most used

platform for news) and TikTok (the fastest growing platform for news) are seen as the biggest problems when it comes to misinformation, with around half of global respondents perceiving each as a major threat. Although trust in news overall has declined in many countries, findings repeatedly show that traditional news media are ***still*** trusted much more than social media or search platforms when it comes to important news.

Across markets, 58% of respondents say they worry about distinguishing true from false news on the internet – a concern that is much higher in Africa and parts of Latin America. When people need to check information that they suspect is false, the most widely cited source is a 'news brand I trust' (38%), ahead of official sources, search engines and fact-checking websites. Even among those who say they would verify information via social media and search engines, it is 'trusted news brands' that people are most likely to be using. In most countries, these trusted sources tend to be the brands (and websites) of news organizations with a reputation for impartial news such as the BBC in the UK, ARD in Germany and NHK in Japan.

Fig 5.5

Top 3 news brands used to check whether something is true

Trusted brands

UK Trusted brands	US Trusted brands	Germany Trusted brands	Japan Trusted brands	Australia Trusted brands
BBC NEWS	CNN	tagesschau	NHK	ABC NEWS
1. BBC News	1. CNN	1. Tagesschau (ARD)	1. NHK	1. ABC News
2. The Guardian	2. Fox News	2. N-tv	2. Yahoo! News	2. BBC News
3. Sky News	3. BBC News	3. ZDF/heute	3. Yomiuri Shimbun	3. 7 News

Q **Question:** *In the previous question you said you would tend to go to a news source you trust to check information. Which one?*

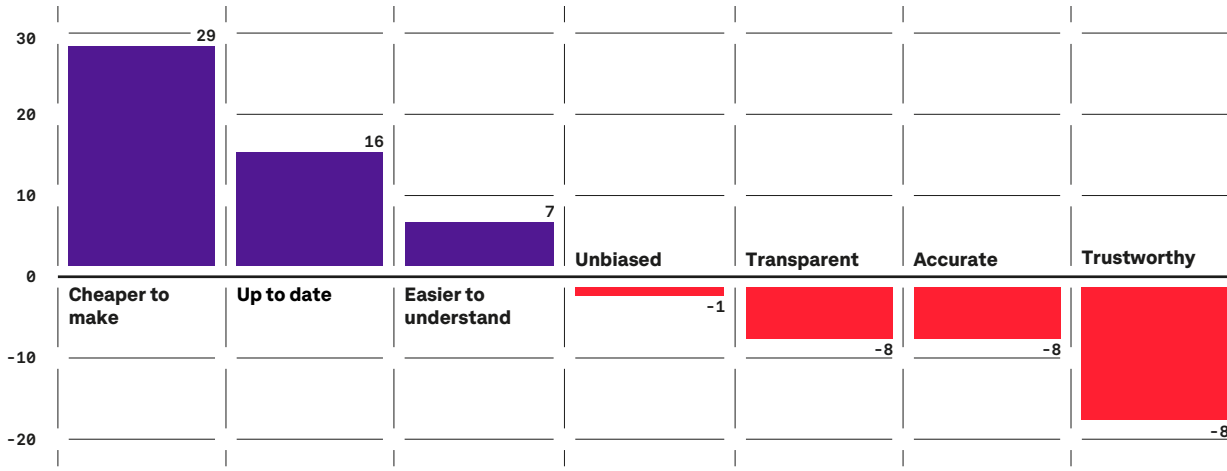
B **Base:** *All those that selected trusted brands in the UK = 867, US = 873, Germany = 811, Japan = 622, Australia = 786 .*

Concerns about (hyper) personalization, polarization and misinformation and disinformation are likely to intensify with AI, as synthetic content floods the internet and more audiences use AI to access information. Among respondents, 7% already use an AI chatbot for news on a weekly basis – rising to 15% among those under-25. The integration of AI into widely used tools like search engines may drastically increase the uptake of these technologies, even if users are not always aware they are interacting with AI. Across countries, survey respondents believe AI will make news cheaper to produce (+29 net difference) and more up to date (+16), but less transparent (-8), less accurate (-8) and less trustworthy (-18). These concerns could increase the value of trusted news brands, including their websites and apps, which may serve as reliable anchors for audiences seeking information on contentious and important stories.

Fig 5.6

Net difference between proportion of respondents who think generative AI will make news more or less of each

Average score across 31 countries



Q **Question:** *In general, do you think that news produced mostly by AI, albeit with some human oversight, is likely to be more or less of each of the following, compared to news produced entirely by a human journalist?*

B **Base:** *Total sample across 31 countries = 54,638.*

The outlook for news organizations remains uncertain. Platform- and algorithm-driven access to news is now the dominant behaviour, especially among younger audiences, significantly weakening the traditional so-called 'post and refer' model, as platforms increasingly try to retain users within their own ecosystem. The rise of generative AI search interfaces is expected to further challenge efforts to draw audiences back to news websites. However, journalism continues to play an important role in the information environment, especially in times of crisis. To remain relevant, publishers will need to rethink how they reach audiences and develop more sustainable business models.

Looking ahead, a central challenge will be how to safeguard trustworthy information and journalism within an increasingly AI-driven, platform-dominated media environment. Beyond financial sustainability, key questions remain for policy-makers, researchers and civil society at large: How can media literacy be strengthened at scale? How should content provenance be communicated – through labelling or other mechanisms? And how can issues such as platform concentration, data licensing and copyright be effectively addressed in the AI era?

Amy Ross
Research Fellow
Reuters Institute for the Study of Journalism, University of Oxford

Nic Newman
Senior Research Associate
Reuters Institute for the Study of Journalism, University of Oxford

A resilient information ecosystem relies not only on strong institutions, but also on the vitality and independence of media actors. Supporting local journalism, fact-checking networks and open media environments is essential to building public trust and effectively countering harmful information. However, several evolving dynamics are reshaping the information landscape in ways that undermine these efforts:

- The rise of 'local media deserts' – areas with little or no access to reliable local news sources – or communities where residents face significantly reduced access to the news of local public discourse.[12] In these environments, people increasingly rely on social media as a primary source of information.

- Shifting media consumption habits are pushing conversations into private, less visible digital spaces – known as 'bounded social media places' – such as private messaging apps and closed groups. These spaces are often viewed as more trustworthy due to their known audiences, controlled visibility and real-time synchronicity by facilitating continuous conversations and personalized content sharing, often outside the influence of public-facing algorithms.[13] However, their opacity makes it harder to monitor or challenge the spread of misinformation.

As highlighted in the World Economic Forum's 2025 *Global Risks Report*, it is becoming increasingly difficult for the public to discern where to turn for trustworthy information. Across 47 countries, only 40% of people say they trust most news, and there are concerns about the risk of misinformation and disinformation over the next two years, especially in high-income countries. It ranks among the top 5 perceived risks in 13 countries including India, Germany and Canada, and in the top 10 in 30 additional countries.[14]

The media plays a dual role in today's information landscape – acting both as a vital conduit for trusted information and, at times, a vector amplifying harmful narratives. This complexity is critical to understand in humanitarian contexts where media coverage can shape public perceptions of crises, humanitarian actors and the legitimacy and/or effectiveness of aid efforts. Not all media are the same and these distinctions matter:

- Independent journalism plays a watchdog role, conducting fact-checking, holding power to account and providing nuanced, evidence-based reporting.

- State-aligned media often reflect government interests, particularly during armed conflict or emergencies, shaping narratives that support national agendas.

- Commercial and tabloid outlets may prioritize sensationalism and speed over accuracy, driven by attention-based revenue models and algorithmic amplification.

- Digital and social media platforms can accelerate the spread of both accurate information and harmful narratives, amplifying reach through network effects, algorithms and user engagement.

Contributor Insight 5.4

Societal resilience: Bridging the information divide by distinguishing reliable from unreliable news sources

Since 2018, NewsGuard has deployed a team of journalists to rate and review the reliability of news sources across the open web, social media and content platforms, using transparent, apolitical journalistic criteria. The need for such labelling of news sources has grown more acute in recent years as the number of unreliable 'news' sources – often presenting themselves as regular news websites while disregarding basic journalistic standards – has surged.

These range from made-for-advertising websites to propaganda outlets spreading harmful false claims, and more recently, AI-generated 'slop' with no human oversight. For example, ahead of Germany's February 2025 snap federal elections, NewsGuard identified a network of 102 AI-generated German-language sites, spreading false claims with apparent authenticity. By August 2025, NewsGuard had documented over 1,270 such AI-generated sites.

Ratings indicate a source's overall trustworthiness and risk of publishing false content. Assessment and review criteria include whether errors are transparently corrected, whether false or egregiously misleading false claims are regularly published, and whether ownership and potential conflicts of interest are disclosed. Each rating is supported by a detailed 'Nutrition Label,' [NewsGuard's term for its ratings profile] explaining the evidence, citing examples of problematic content and including any response from the publisher. NewsGuard's ratings currently cover more than 36,000 online sources across nine countries (Australia, Austria, Canada, France, Germany, Italy, New Zealand, the UK and US), representing more than 95% of online engagement with news in these markets.

Chine Labbe

Senior Vice President, Partnerships, and Managing Editor, Europe and Canada

NewsGuard Technologies

Meanwhile, media outlets themselves are under growing pressure. Trends documented in the *World Press Freedom Index*[15] reveal shrinking press freedoms, alongside rising harassment, censorship and targeted information attacks. Many local and independent outlets face severe economic precarity, leading to newsroom closures and weakened on-the-ground reporting. In parallel, regulatory tools such as licensing requirements or anti-fake news laws have, in some cases, been used to suppress independent journalism, rather than protect the public from disinformation.[16]

In humanitarian settings, the media plays a powerful role in shaping narratives around aid operations. Politicized or misleading coverage can erode public trust in humanitarian actors, hinder access or even place humanitarian personnel, volunteers and affected populations at risk. Despite these challenges, there are also opportunities for constructive engagement. Partnerships with trusted local journalists can enhance the accuracy,

relevance and reach of communication on humanitarian crises. At the same time, media literacy initiatives can help communities distinguish fact from fiction, better understand the risk of harmful information and engage with media content more critically.

Given the diversity of today's media landscape, it is crucial to identify where harmful information is spreading – whether through television, radio, print, digital platforms or social media – and to understand how different communities access and trust information. This requires disaggregated analysis by age, gender, language, geography and other demographic factors, alongside close engagement with community leaders and local stakeholders. As media consumption habits continue to evolve, the need to promote high-quality journalism and empower individuals to navigate an increasingly complex and polarized media environment has never been greater. For the humanitarian sector, this means actively engaging to build understanding of what humanitarian action is and is not, and of the principles that underpin it.

Contributor Insight 5.5

BBC Media Action: A whole-of-society approach is called for

While humanitarian agencies often see the most damaging effects of harmful information during crises, the problem can pre-date the crisis, and sustainable solutions require a whole-of-society approach in which governments, civil society, independent media, technology companies, the private sector and communities each play a role. The humanitarian sector should be more proactive in supporting such approaches, but this demands a step-change. Agencies such as UNHCR, IFRC and others have an essential role in embedding information integrity strategies across their work – not just in communications departments, but as part of their operations, as well as in core humanitarian standards and response plans.

This includes building trusted relationships with local media and community communicators before crises, and embedding these efforts into coordination mechanisms with other societal actors. In practice, this means:

- working with governments to ensure crisis information policies protect rights and reach those most at risk
- partnering with civil society to strengthen community-led responses
- collaborating with and supporting independent media to uphold accuracy and trust
- engaging with technology companies to address harmful content and increase transparency
- demanding adequate legal and regulatory protections from harmful information

— equipping communities with the skills and resources to navigate an increasingly complex information environment.

It is vital that humanitarian actors avoid working in isolation and draw on the existing work and experience of others working on these issues. For example, the media development and democracy-support sectors have been grappling with information integrity issues for many years and have many lessons and best practices (and ongoing initiatives) to share. Linked to this, humanitarian actors should ensure they are leveraging and engaging with global efforts to build normative frameworks in support of information integrity. For example, the UN Global Principles for Information Integrity were launched in 2024, but implementation is a work in progress.

The rapid evolution of AI makes this engagement even more urgent. AI is already transforming the information environment – accelerating the creation and spread of false content, but also offering new tools for verification, translation and rapid information sharing. Decisions being made now about AI governance, design and regulation will profoundly influence how information flows in the next decade. The humanitarian sector must be at the table in these debates, advocating for safeguards that protect vulnerable populations and exploring solutions such as content provenance technologies (e.g., C2PA) to ensure crisis information is identifiable, verifiable and trusted.

Ultimately, harmful information is not a peripheral issue: it is a core risk to effective humanitarian action. Addressing it demands that agencies treat information integrity as part of critical infrastructure – planned for, invested in and maintained with the same seriousness as water, shelter and health systems. By contributing their expertise, partnerships and grounded understanding of at-risk communities to whole-of-society approaches, humanitarian actors can help shape an information ecosystem where truth has a fighting chance – and where communities are empowered to act on it when it matters most.

Alasdair Stuart
Head of Policy
BBC Media Action

Contributor Insight 5.6

When collaboration between media and humanitarians can save lives

In humanitarian crises, communication can be life-saving – or dangerously misleading. And while humanitarian organizations and local media share a commitment to minimizing the impacts of harmful information on vulnerable communities, their relationship is often fraught with tension – too often unaware of how complementary their ambitions truly are.

Humanitarian organizations are guided by neutrality, dignity and 'do no harm'. Local media seek to inform, be objective and hold power to account. These goals aren't oppositional – they're synergistic. But without dialogue, misunderstandings flourish. Journalists may misinterpret operational silence as evasiveness; humanitarian actors may view media

coverage as oversimplified or sensationalist. In the chaos of a crisis, both believe they're helping – yet mistrust and poor collaboration can inadvertently fuel misinformation, leaving affected communities to bear the cost.

When these sectors collaborate, the impact can be powerful. Humanitarian actors bring credible data and balanced contextual understanding; local media offer reach, cultural insight and familiar, trusted voices.

Community-based organizations often navigate this space more naturally. As part of the community, they tend to see local journalists not as outsiders, but as neighbours, former classmates and peers. While this common ground doesn't guarantee a flawless relationship, it does offer a foundation built on mutual lived experience. In contrast, when international agencies arrive, local media is frequently viewed as an unpredictable element, an unacceptable risk to be 'managed', rather than a partner to engage.

At Internews we frequently saw this dynamic and in response we developed the *Information and Risks: A Protection Approach to Information Ecosystems* toolkit. This was designed in collaboration with a global advisory board of protection agencies, the Global Protection Cluster and local media organizations. It provides tools for humanitarian workers, journalists and community organizations to jointly assess how poor information access can exacerbate protection risks. This tool can shine a light on the shared values, approaches and practical ways to collaborate across all disciplines to address harmful information in a crisis situation.

Building meaningful relationships with media requires more than stage-managed community visits. It requires time, mutual respect and an openness to sharing and acknowledging each other's contributions and constraints in a crisis to transform this relationship from transactional to strategic. Some distance must be maintained. Local media must maintain independence to fulfil their accountability role – which they cannot do if they become too entangled in humanitarian operations or funding streams. Humanitarians cannot share all information with the media. But there is a need for spaces of curated neutral ground, where actors in both sectors can engage authentically, without the pressure of a live microphone.

In an era when information can either save lives or spread harm, building trusted partnerships with non-traditional actors isn't optional. It's a necessary part of protecting those most at risk.

Irene Scott
Humanitarian consultant and former Humanitarian Director
Internews Network

5.6 A threat to humanitarian action and to humanity itself

Harmful information undermines trust, deepens social divisions and weakens the ability of institutions – humanitarian or otherwise – to respond effectively in times of crisis. Addressing it requires a whole-of-society response, not just a humanitarian one. It is

intrinsically linked to preserving humanitarian space and the urgent need to prevent the instrumentalization of humanitarian action and actors. Harmful narratives threaten the perceived legitimacy of humanitarian organizations, shaping public perceptions in ways that can directly hinder their ability to access and respond to populations in need. In some cases, this involves misrepresenting an organization's neutrality, intent or operations to incite hostility and disrupt response efforts.

Recognition of this threat is growing. In a significant step, the UN Security Council adopted Resolution 2730[17] in May 2024, spearheaded by the Government of Switzerland and co-sponsored by 98 member states. The resolution explicitly condemns "disinformation, information manipulation, and incitement to violence" against humanitarian and UN personnel. It further raises concern over the increasing use of malicious information and communication technologies including data breaches and information operations that target humanitarian organizations, disrupt their relief operations, undermine trust and threaten the safety and security of their personnel, premises and assets. The resolution encourages member states and the UN system to take appropriate action to address the increasing threat of disinformation campaigns and misinformation that undermine trust in humanitarian organizations, put personnel at risk and hinder humanitarian activities. Importantly, the resolution reaffirms the need for all parties to armed conflict to preserve the ability of humanitarian organizations to act in a manner consistent with the principles of humanity, neutrality, impartiality and independence. This is vital for delivering assistance to persons in need and for ensuring their protection and safety, as well as that of humanitarian personnel.

Contributor Insight 5.7

UK's commitment to protecting humanitarian action from harmful information

Information integrity is critical in humanitarian crises: communities are vulnerable and need to know what they can trust. Access to accurate and timely information for people in the midst of armed conflict can mean the difference between life and death. Conversely, false narratives targeting humanitarian organizations pose a serious threat to perceptions of the neutrality of aid workers and relief operations, damaging community acceptance, restricting humanitarian access and increasing risks for aid workers. And we need truthful reporting of humanitarian crises globally: the absence thereof undermines public support for humanitarian action.

Yet disinformation in the humanitarian space is growing. The UK is tackling this concerning trend directly – through our seat in key multilateral forums, our diplomatic channels and our UK-funded aid. For instance, the UK co-sponsored UN Security Council Resolution 2730, explicitly condemning disinformation and encouraging member states and the UN system to take appropriate action to address the increasing threat of disinformation campaigns and misinformation that undermine trust in UN and humanitarian organizations and put humanitarian personnel at risk. The Political Declaration for the Protection of

Humanitarian Personnel, developed by a Ministerial group led by Australia and including Brazil, Colombia, Indonesia, Japan, Jordan, Sierra Leone, Switzerland and the UK, commits states to take practical action that counters misinformation and combats disinformation, information manipulation and hate speech targeting humanitarian organizations, personnel and activities. This includes efforts to de-politicize humanitarian action, including by building understanding with local authorities and the media, protecting the independence of journalists, raising awareness, and calling out actors that perpetuate disinformation and hate speech and working with technology companies to support these efforts.

As we continue to learn from experience and adapt to evolving threats, the UK is also investing in:

- independent journalism and fact-checking in fragile contexts
- civil society initiatives to counter harmful information and build resilience to it
- accountability mechanisms to promote community feedback and inclusive information ecosystems

We also invest in research, such as a recent project with Grand Challenges Canada, which highlighted the importance of working with communities to build trust, enhance resilience and empower local leadership, and of identifying technological innovations that are scalable and adaptable. These lessons are shaping our humanitarian responses.

The UK remains committed to working with partners to counter harmful information, protect humanitarian personnel and build resilient communities. Together, we can ensure that humanitarian action is guided by truth, trust and integrity.

Laure Beaufils

Director Humanitarian, Food Security and Resilience

Foreign, Commonwealth and Development Office, Government of the United Kingdom of Great Britain and Northern Ireland

The UN General Assembly's resolution on 'Countering disinformation for the promotion and protection of human rights and fundamental freedoms' calls on states to counter all forms of disinformation through policy-based measures, including public education, digital literacy and capacity-building initiatives.[18] As highlighted in Chapter 4, on page 157, the Pandemic Agreement is a landmark international treaty aimed at setting global standards for pandemic prevention, preparedness and response. It underscores the importance of trust, transparency and timely information sharing in effective pandemic communication.[19]

5.7 Red Cross and Red Crescent Appeal to States

Since the Statutory Meetings of the International Red Cross and Red Crescent Movement began in 1867, themes related to information have rarely been addressed.[20] However, 2024 marked a turning point with the adoption of an Appeal to States to take action against harmful information and dehumanizing rhetoric.

Appeal to States:

"We appeal to States to take all appropriate measures to prevent, stop and remedy any abuse, pressure, misinformation, disinformation and dehumanizing rhetoric, through social media or otherwise, that harms the physical, psychological or reputational wellbeing of people in vulnerable situations and the staff and volunteers of the Movement components serving them."

The appeal highlights the increasing impediments faced by principled humanitarian actors in delivering protection and assistance to people in need including the spread of misinformation and disinformation that imperils humanitarian workers and people in their care.[21] While previous resolutions have reflected the Movement's longstanding commitments to provide quality information and combat discrimination, the 2024 appeal reflects a heightened sense of urgency. This shift underscores the importance of the information environment, while reinforcing the need to safeguard unhindered and safe access to people in need to carry out principled humanitarian action.

This resolution (appeal) recognizes that harmful information and dehumanizing rhetoric pose serious risks to people in humanitarian need and humanitarian personnel. Addressing this is thus integral to the protection of people and upholding humanitarian operations.

In disasters, crises and other emergencies, laws, policies and plans that promote the availability and integrity of information – and address the challenges of harmful information – are essential to enable and facilitate effective humanitarian action. Harmful information that obstructs humanitarian response, fuels panic and erodes trust runs counter to these objectives. By integrating measures to counter harmful information, states can uphold their responsibilities while protecting the space for neutral, impartial and independent humanitarian action.

Contributor Insight 5.8

Harmful information and international humanitarian law

While there are no specific provisions on harmful information in international humanitarian law (IHL), this body of international law does regulate the spread of certain forms of information. Specifically, IHL prohibits the encouragement of IHL violations, including war crimes, by states and parties to armed conflict, whether online or offline. It also prohibits "acts or threats of violence, the primary purpose of which is to spread terror among the civilian population".[22] This means that threatening violence is prohibited under IHL if the primary purpose or intent of such activities is to spread terror among the civilian population.

Furthermore, IHL prohibits inciting violence against medical services and humanitarian operations. In particular, spreading disinformation intended to obstruct or frustrate medical and humanitarian work is difficult to reconcile with IHL and may violate the obligation of parties to conflict to respect and protect humanitarian personnel and medical services. Moreover, states must not only abstain from such activities but also protect impartial humanitarian organizations from threats posed by other actors within their jurisdiction or control, including private persons and companies. Propaganda aimed at recruiting children, and publishing images of prisoners of war, are in almost all cases violations of IHL.[23]

While not all harmful information during armed conflict falls within the scope of IHL, for over 160 years, there has been consensus that impartial humanitarian operations and the personnel involved therein must be respected and protected. Additionally, advocating hatred that constitutes incitement to hostility, discrimination or violence, and inciting genocide, may also violate international human rights law or other rules of international law.

Philippe Stoll
Lead on Movement Initiative on Harmful Information
International Committee of the Red Cross

Contributor Insight 5.9

Disaster law and harmful information

Legal and policy frameworks for disaster risk management (DRM) provide the foundation for preventing and mitigating risks, preparing for crises and ensuring effective response and recovery in societies when disasters strike.[24] To remain effective, these frameworks need to also address emerging challenges that undermine their objectives – such as the spread of harmful information.

Misinformation, disinformation and rumours can heighten vulnerability and disrupt every stage of the DRM continuum. Harmful narratives may undermine prevention and mitigation initiatives and distort public understanding of hazards. False narratives may discourage communities from taking preparedness measures or erode trust in early warning systems.[25] During response, they may delay life-saving action, undermine coordination among responders or incite hostility toward humanitarian actors.[26] In recovery, harmful information can fuel stigma, exacerbate inequalities and obstruct community resilience.

These growing challenges demand greater attention within legal and institutional arrangements for DRM. Addressing information risks requires legal and policy frameworks that: (1) recognize information integrity as a core component of DRM; (2) protect individuals by balancing rights with measures to mitigate harmful falsehoods; and (3) enhance coordination, accountability and public trust through mechanisms for information verification, community engagement and transparent communication.[27]

Safeguarding information integrity is critical to ensuring that DRM efforts truly protect and empower communities. States should therefore take harmful information into account as a potential barrier to effective DRM and in their efforts to strengthen legal preparedness for disasters.

Isabelle Granger
Global Lead, Disaster Law and Auxiliary Role
IFRC

5.7.1

The evolving place of information in humanitarian standards

The 1994 Code of Conduct for the International Red Cross and Red Crescent Movement and NGOs in Disaster Relief makes little reference to information, except in relation to the availability of data "pertinent to the implementation of effective disaster response." Building on this, the 1997 Humanitarian Charter highlighted the "transparency of information and decision-making" as essential to effective and accountable humanitarian action. Subsequent frameworks have placed growing emphasis on information as central to dignity, participation and protection.

The Humanitarian Charter and Sphere Standards (1997–present) identify access to information as a core element of participation; the Core Humanitarian Standard on Quality and Accountability (2014) includes explicit commitments to information-sharing and communication with affected people; and both the ICRC Rules on Personal Data Protection and IFRC Policy on the Protection of Personal Data frame personal data and information management as integral to protection, dignity and legal compliance. In parallel, the CDAC Network Guidelines advanced the principle of 'communication as aid,' focusing on community engagement and misinformation management, while the International Civil Society Centre's Solidarity Action Network has more recently turned attention to the impact of disinformation campaigns on civil society.

Contributor Insight 5.10

Personal data protection as a defence against harmful information

Personal data has become a highly exploitable commodity in disinformation campaigns, used to create, target and disseminate harmful content through AI-generated media such as deepfakes (see Annex I: Glossary, on page 353) or via large-scale unauthorized data harvesting on social platforms. Where specific regulation surrounding targeted disinformation or AI is absent, existing data protection laws – now in place in over 100 countries – can provide a foundational layer of defence. While limited in scope, these laws require consent, transparency and restrictions on data sharing, and offer individuals enforcement mechanisms.

As the humanitarian sector increasingly relies on the collection of personal data and digital tools to deliver assistance, organizations must take proactive measures to protect staff and beneficiary data. This includes minimizing data collection, enforcing adequate security measures, carefully selecting digital service providers and managing the risks of publicly shared content.

Disinformation campaigns targeting humanitarian operations, such as allegations of fraud or misuse of funds, may also trigger government or donor demands for greater data disclosure in the name of accountability. In such cases, organizations need counter-disinformation strategies that increase transparency without compromising humanitarian principles or individual privacy protections.

Manuela Cardoso
Consultant, Legal Department and Data Protection Office
IFRC, Geneva

Contributor Insight 5.11

Humanitarian standards in a changing information ecosystem

Government policies and regulations on digital platforms, artificial intelligence and data governance are not keeping pace with the rate of technological change and are fragmented across jurisdictions. In the US, for example, Section 230 of the Communications Act of 1934 provides broad immunity to service providers for content generated by users of their platforms. For humanitarian organizations, this creates a heightened responsibility to safeguard communities from harmful information while navigating a shifting and uneven regulatory environment.

The Sphere Minimum Standards, along with others in the Humanitarian Standards Partnership (HSP) portfolio, are built through consensus and grounded in international law, rights, evidence and practical experience. They distil vast knowledge into accessible guidance for frontline practitioners and others. Yet, in an era defined by fast-moving digital innovation and disinformation, clear gaps remain. New Connectivity As Aid standards are coming soon, but the HSP portfolio still lacks sufficient guidance on data stewardship, intellectual property and the use of AI. Addressing these areas is urgent, as harmful information increasingly undermines trust, accountability and the protection of crisis-affected people.

Sphere, as host of the HSP, recognizes the need for inclusive and consultative development of new standards to meet these gaps. Future standards should highlight how harmful information, accelerated by technology, disproportionately affects vulnerable populations and provide practical guidance to mitigate these risks. Strategic partnerships will be essential. Sphere encourages standards developers to collaborate with organizations committed to meaningful community engagement, such as People First Impact Method (P-FIM), to ensure that community voices shape new standards from the outset. In a complex and fast-changing information ecosystem, humanitarian self-governance through widely owned and versatile standards will save lives. This is one way the sector can and must work together to build trust, uphold truth and strengthen resilience in the face of harmful information.

Tristan Hale
Director of Operations
Sphere Standards

Taken together, these instruments underscore that information is not peripheral to humanitarian action but fundamental to its effectiveness, accountability and legitimacy. Looking ahead, this recognition should also encompass harmful information, ensuring that standards explicitly address how it can undermine safety, dignity, humanitarian access and action response.

5.7.2 The tech sector: Platforms, power and responsibility

The explosive growth of online content, driven by new technologies and user-generated platforms, has made harmful information more prevalent and harder to detect, moderate and remove. What began as spaces for personal connection have evolved into powerful tools for influence, distortion and manipulation, shaped by the speed, scale and reach of digital information. Social media platforms are designed to reward engagement, not accuracy. What goes viral is not necessarily what is true. Built around attention-based algorithms, these platforms fuel contests over narrative and influence with real-world consequences. These platforms are designed for maximum engagement, where every like, share and comment release a hit of dopamine, reinforcing addictive use and deepening echo chambers.[28]

The rise of AI has further concentrated power in the hands of a small number of companies that develop and control the systems including algorithms shaping what billions of people see, search for and share. Their influence extends beyond access – these systems increasingly affect how public opinion is formed, how issues are framed and which narratives gain prominence. Generative AI has drastically lowered the barriers

to creating and distributing false or misleading content, across text, video, audio and images. Distinguishing AI-generated from human-made content is increasingly difficult. While some material is simply the result of error or AI hallucination, other content is deliberately produced by threat actors – state actors, activist groups or individuals – aiming to mislead, influence or disrupt. This rapid, low-cost production and deployment of synthetic content presents significant challenges for information integrity.

5.8 UN action on AI and information integrity

The UN General Assembly's March 2024 Resolution on AI systems[29] highlights the growing risks to **information integrity**, access to information and human rights. It warns that the improper or malicious design, development, deployment and use of AI systems could undermine the Sustainable Development Goals (SDGs), deepen digital divides and reinforce structural inequalities and biases. The resolution raises concerns about the potential accidents and the risk of compounded threats from malicious actors. It encourages member states and stakeholders to develop effective and interoperable tools, reliable content authentication and provenance mechanisms – such as watermarking or labelling. These measures would enable users to identify information manipulation, determine the origins of and distinguish between authentic and AI-generated or manipulated digital content. It further encourages efforts to strengthen media and information literacy to strengthen societal resilience.[30]

The June 2023 UN policy brief on Information Integrity on Digital Platforms was the result of consultations developed under the auspices of the Secretary-General as part of the Our Common Agenda and Summit of the Future initiatives. This seeks to provide a concerted global response to information threats in the information environment. The principles aim to guide member states, digital platforms and other stakeholders in fostering a more inclusive, transparent and safe digital space, particularly freedom of opinion, expression and access to information.[31] In September 2025, the UN released the first in a new Issue Brief series entitled From Principles to Practice: Strengthening Information Integrity.

The UN's 2024 *Governing AI for Humanity* report[32] warns that large parts of the world are excluded from international AI governance conversations, with only seven countries party to seven prominent non-UN AI initiatives and 118 countries not party to any.[33]

5.9 Content moderation and the power of platforms

Content moderation has become one of the most powerful and contested functions exercised by global technology platforms. Companies behind social media and messaging services play a central role in shaping the information ecosystem, determining what

is amplified, removed or monetized. The model of social media platforms – functioning as intermediaries allowing user-generated content to spread at scale – has effectively shifted much of the responsibility for identifying and managing harmful information onto the private sector. This places content moderation and governance in the hands of technology companies rather than public institutions. Content moderation is carried out under each platform's own terms of services and definitions, typically through a combination of human moderators and automated systems. Platforms may block or remove content either in response to legal requests or based on internal assessments. However, definitions of harm vary and enforcement is uneven.

This power is exercised with limited transparency, especially regarding how content is moderated and curated. Many governments have been reluctant to regulate legal but harmful information citing their obligations to uphold freedom of expression. At the same time, global technology platforms often comply with national content laws, even when those laws conflict with international human rights standards. This results in an uneven and fragmented landscape, where users' access to content depends largely on their location and the legal environment of that jurisdiction. While some platforms publish transparency reports outlining how often governments request content removal, reporting practices vary widely and often lack the detail necessary for meaningful public accountability.

While content regulation can be lawful and necessary, especially to prevent harm, some states use it to control narratives, limit transparency and suppress dissent. This raises a deeper set of questions: Who decides what constitutes harmful information? Who is accountable? And how can we safeguard fundamental rights in a fragmented and global information space?

Under international law, for example, Article 19 of the International Covenant on Civil and Political Rights, freedom of expression may only be restricted when three conditions are met: the restriction must be provided by law (clearly defined in legislation), necessary (to achieve a legitimate aim) and proportionate (not excessive or overly broad). Yet many current censorship practices fail to meet these criteria, raising serious concerns about compliance with international standards.

The UN Guiding Principles on Business and Human Rights and related standards provide important guidance for technology companies. These principles call for heightened human rights due diligence and risk management to prevent and address adverse impacts, including from the spread of harmful information. This is particularly critical in contexts affected by humanitarian crises, where such information can exacerbate vulnerabilities and fuel harm. Companies are expected to invest in and implement robust measures to mitigate potential negative impacts of their activities on people's safety and dignity. This includes ensuring that their policies, procedures and practices are consistent with international human rights standards and take into account international humanitarian law.

Complementing this, the UNESCO framework for regulating digital platforms is grounded in the recognition that information is a public good. As such, it calls for government action to protect and support the integrity, accessibility and equitable availability of information ecosystems.[34]

Meta (formerly Facebook), for example, does not explicitly list humanitarian organizations as targets in its reporting on inauthentic behaviour. Yet, its findings on narrative manipulation and inauthentic behaviour in contexts of armed conflict have clear

implications for humanitarian actors, whose operations may be disrupted or delegitimized by such tactics.[35]

Microsoft highlights that not all threats posed by AI originate within the systems themselves; many stem from the broader information ecosystems.[36] Two prominent categories stand out:

- **Impersonation:** AI-generated deepfakes (audio, video or images – see Annex I: Glossary, on page 353) increasingly enable impersonation of individuals, with serious risks including fraud, blackmail, coercion, defamation and information warfare.

- **Content production**: AI tools can be misused to produce harmful synthetic content at scale, including disinformation, spam, non-consensual intimate imagery and grooming scripts. These threats are often amplified versions of pre-existing problems, now rendered faster and more pervasive by generative AI.

Microsoft has also called on states to adopt clear norms and restrictions to curb harmful foreign influence operations, particularly those targeting crisis and emergency contexts, humanitarian and emergency response organizations, elections, marginalized communities and protected groups such as ethnic minorities and LGBTQ+ populations. It advocates for limits on the use of certain tools and techniques including synthetic media (e.g., deepfakes) and emphasizes that social media data from foreign citizens should not be exploited for influence operations.[37]

5.10 From self-regulation to state oversight: The evolving governance of online content

Despite their language of *community*, digital and social media platforms are, above all, businesses. Their primary responsibility is to shareholders, not the public good – so the metrics that matter most are financial. This profit-driven logic shapes how platforms approach both problems and solutions. With an engineering-first mindset, they often frame even complex societal challenges as design problems solvable through technical fixes. Ultimately, platform terms of service – written by private companies – administer communities of unprecedented scale.[38] Yet fundamental dilemmas remain: Should these companies restrict the flow of information? What content should be restricted, how and by whose standards?

The US legal framework, particularly Section 230 of the Communications Decency Act, has played a central role in shaping this landscape.[39] It shields US platforms from liability for user-generated content while also protecting them when they moderate content in "good faith." This approach effectively left much of the internet to self-regulate, with content moderation emerging as a corporate practice – not out of legal obligation, but as a strategy to pre-empt stronger regulation. In contrast, enforcement around intellectual

property – governed by the more stringent 1998 Digital Millennium Copyright Act has drawn much clearer lines.[40]

The European Union's Digital Services Act, which came fully into force in 2024, establishes a comprehensive legal framework to regulate online services and address the spread of illegal and harmful content, including misinformation and disinformation, while safeguarding fundamental rights such as freedom of expression. Non-compliance can result in significant penalties, including fines of up to 6% of a company's global annual turnover. The Digital Services Act seeks to strike a balance between mitigating systemic societal risks associated with online content and protecting freedom of expression and information.[41] Key provisions relevant to harmful information include:

1. **Systemic risk management**: 'Very large online platforms' (with more than 45 million active users in the EU) and 'very large online search engines' are required to identify, analyse and mitigate systemic risks stemming from the design, functioning and use of their services. This includes risks posed by disinformation that could threaten democratic processes, public health or public security.

2. **Transparency and accountability**: Platforms must disclose their content moderation policies, advertising practices and key parametres of algorithmic decision-making and provide users with accessible complaint-handling and appeal mechanisms.

3. **Code of practice**: In February 2025, the strengthened EU Code of Practice on Disinformation was formally integrated into the Digital Services Act framework. Measures include demonetizing disinformation, strengthening fact-checking and improving access to reliable and authoritative information.

4. **Enhanced user-reporting mechanisms**: Online platforms must provide accessible and effective systems for users to report illegal content.

5. **Crisis-response mechanism**: In exceptional circumstances, such as threats to public security or public health, the European Commission may require platforms to take targeted measures to address the rapid dissemination of harmful information during crisis periods.[42]

These developments build on earlier EU initiatives, including the European Commission's Code of Practice on Disinformation and the Strengthened Code of Practice on Disinformation.[43]

5.11 Framing a response: Supply and demand solutions to disinformation

Professor Anya Schiffrin of Columbia University developed a 'taxonomy of solutions' – an analytical framework that categorizes various policy responses to online misinformation and disinformation into two broad approaches: supply side and demand side.

Contributor Insight 5.12

Understanding the universe of fixes for online misinformation and disinformation

Since 2016, scholars, policy-makers and practitioners have all been worried about the spread of misinformation and disinformation online. The COVID-19 pandemic and associated spread of vaccine resistance has made the problem even more urgent in the public health and humanitarian spheres. To better understand policy proposals, I developed an analytical framework and was the first to create a taxonomy of solutions which distinguishes between 'supply-side' and 'demand-side' proposals,[44] many of which have since been implemented around the world.

Demand-side solutions emphasize the role of the consumer, while **supply-side solutions** emphasize the supply of information, looking more to its producers, suppliers and purveyors. In the taxonomy, supply-side solutions fall into two subcategories: 1) suppressing poor-quality, dangerous or illegal information; and 2) creating and/or promoting high-quality information either online or by supporting journalism.

Demand-side solutions include efforts to teach media literacy, journalist efforts to engage audiences and verification efforts such as labelling and fact-checking. These solutions all emphasize audience demand for information and the role of individual choice. Media literacy attempts to build discernment skills among audiences so they can identify which sources to trust. Similarly, solutions involving community participation include efforts by journalists to build trust by seeking to bolster engagement with reliable, relevant material. Solutions involving verification mechanisms, such as fact-checking and labelling, provide a means of establishing what is objectively correct or 'true.' A recent synthesis of hundreds of academic studies carried out by the International Panel on the Information Environment[45] found that demand-side interventions such as media and information literacy, labelling and publishing corrections appeared effective in more than 10% of studies. However, these solutions are expensive and difficult to scale.[46] Table 5.1, on page 199 shows the range and classifications of supply-side and demand-side solutions.

Supply-side solutions

While the US has largely followed the demand-side path, many other governments, including in Europe, have put less emphasis on consumer decisions and focused more on the supply of information available to consumers. Germany's 2019 Network Enforcement Act (NetzDG) attempts to hold social media platforms responsible for combating online speech deemed illegal under domestic law. The European Digital Services Act 2022 harmonizes different national laws to address illegal content. Both aim to suppress false or potentially harmful misinformation or disinformation and are supply-side solutions. So is the use of AI to screen and filter information and legal action – defamation suits – against purveyors of falsehoods. Efforts to promote good-quality journalism, provide accurate information through platforms such as Google and YouTube, to support public broadcasters or fund local news are all attempts to boost the *supply* of high-quality information.

There are, of course, overlaps between supply-side and demand-side solutions. Middleware tools serve both sides by shaping what is available (supply) while helping audiences make informed choices (demand). These include NewsGuard's rating systems on the credibility

of online sources and tools to counter misinformation, and the Journalism Trust Initiative which establishes indicators for the trustworthiness of journalism. Similarly, the international fact-checking movement protects the accuracy of the supply of information but also requires active consumer involvement.

Table 5.1

Taxonomy of solutions – supply side versus demand side

Demand-side solutions	Supply-side solutions	
	Suppression of poor-quality information	**Promotion of a healthy information ecosystem**
Media literacy training Assisting audiences to better distinguish between what is true and false	**Controlling information flows** Includes content suppression, downranking of content, removal of bots and de-platforming to restrict what information is shown	**Increasing quality information** Attempts by YouTube to increase quality of content and by Google to highlight accurate information
Community participation Includes journalists' efforts to establish trust in and engagement with high-quality information, as well as to develop citizen journalism	**AI and content moderation** Using AI to distinguish between true, false and illegal information and supplementing with human content moderators where required	**Advancing AI** Using AI to promote good-quality information, while recognizing current limitations
Fact-checking Includes labelling and browser extensions that audiences can use	**Regulation and laws** Hate speech laws existing in many parts of the world, such as NetzDG and the Digital Services Act	**Support for high-quality journalism** Policies such as funding the BBC, giving subscription vouchers, tax credits and subsidies, and relying on innovation funds and philanthropic or donor support
Raising awareness Reporting on the platforms for a stronger understanding of the effects of misinformation and disinformation	**Defamation lawsuits** Demonstrating repercussions for actively spreading lies, such as the 2023 lawsuit U.S. Dominion v. Fox News Network about voting machines	**Healthy journalism business model** Engaging with private and public sector adverts to ensure a healthy supply of advertising revenue can fund journalism

Source: Schiffrin, 2017

Dr Anya Schiffrin
Senior Lecturer in Practice
School of International and Public Affairs, Columbia University

5.12 Civic trust and societal resilience

> **"So what I would want to suggest is: they should start with the ground level in schools. Educating people in schools, telling them the quality source of information. Once that is done, those children will keep that and they will grow with that kind of mentality. So the education part should continue using the Ministry of Education in schools that should know the correct source of information. Our community meetings like the church and other political gatherings should also be emphasized. But now the only problem is because of these issues, political issues, people want to gain support, they want to use whatever means. But what I would suggest is let the children be taught the correct things from the Ministry of Education first."**

Community member, Zambia

Some countries have adopted whole-of-society approaches to build societal resilience – to inoculate – against information threats. These include education programmes, public tracking of foreign disinformation, election protection measures, transparency for political advertising and campaign activities. In addition, some governments use legal measures, such as laws for the removal of illegal content.

Contributor Insight 5.13

Why Finland is a forerunner in media literacy

For many years, Finland has been regarded as a global leader in developing and promoting media literacy. The small northern European country has attracted international attention and frequent requests to share experiences behind its strong reputation. Some of the key elements of the Finnish approach to media literacy include:

- **Long traditions.** Media education has been part of Finland's democracy and education system for decades. References date back to the 1950s, with many organizations and practitioners accumulating extensive experience. This long-term perspective has ensured continuity in the national approach.
- **Policies and strategies**. Finland was an early adopter of national policies to promote media literacy. It has a dedicated national authority for media education with a statutory mandate to advance media literacy and strengthen resilience as part of comprehensive security. Media education is integrated into curricula across all levels of education, ensuring that children and young people develop these skills throughout their educational path.

- **Cross-sectoral approach**. Media education in Finland is diverse and widely implemented. Schools, public authorities, cultural institutions, NGOs and private sector actors all contribute to the promotion and development of media education.

- **Trust and stability**. Finland consistently ranks first in both the Media Literacy Index and the World Happiness Report. These results reflect a society with high levels of trust in authorities and the media. Citizens' strong critical media literacy skills and digital competences foster inclusion, strengthen stability and safeguard democracy.

Julia Alajärvi
Senior Adviser
National Audiovisual Institute of Finland

International research[47] highlights several systemic challenges facing civil society actors, which hamper their ability to assess information threats and develop effective responses. While a range of initiatives are emerging, many remain isolated, pointing to an urgent need for stronger coordination, knowledge-sharing and the importance of skill diffusion to prevent duplication and build collective capacity. A critical constraint is the limited access to social media data, which limits civil society's ability to assess the scale of disinformation and to evaluate the effectiveness of different interventions. Furthermore, engagement with technology platforms remains uneven: many organizations, particularly in the Global South, struggle to have their concerns heard and are frequently overlooked by platform decision-makers.

The OECD identifies five key determinants of public trust in government[48] that may provide some interesting insights for civil society in building institutional trust:

- responsiveness and reliability in delivering services and anticipating needs (reflecting competence)

- perceptions of integrity, openness and fairness (reflecting public values).

Academic literature further distinguishes between: trust in competence – the ability to deliver on expectations – and trust in intentions – the perception that actions are taken in good faith. These two dimensions are interdependent and influence how people evaluate institutions. Trust is ultimately built on both performance and principle – the perceived capability of institutions and the values guiding their actions. Transparency, fairness and accountability serve as mutually reinforcing attributes that strengthen both pillars and are essential for building societal resilience in the face of harmful information.

Concluding remarks: A collective responsibility for preserving principled humanitarian action

For the humanitarian sector, strengthening cross-sector engagement is essential – not only to ensure the flow of trustworthy information but also to identify and respond to harmful information. Disruption alone is not enough; responding effectively requires collective action, similar to the collaborative efforts developed for global health security during the COVID-19 pandemic.

Harmful information is borderless, adaptive and often directed at the people most at risk. Algorithmic systems amplify its spread, while global disparities in data governance, AI capacity and political and societal polarization intensify its reach. Effective responses require greater algorithmic transparency on how content is sorted, ranked, amplified and targeted, combined with approaches that protect privacy, dignity and authenticity.

Resilience must be built at every level – individual, institutional and societal – drawing on community trust, local knowledge, behavioural science, digital and information literacy. Evidence shows that simple interventions, such as digital prompts,[49] can reduce the sharing of false content, underscoring the value of pairing regulation with context-specific solutions. Governments, platforms, civil society and researchers must work together to test and scale what works. This underscores the importance of pairing regulation with behavioural science and practical interventions that are context specific. Governments, platforms, civil society and researchers must work together to test what works, for whom and in which context.[50]

Humanitarian crises – from pandemics to armed conflict – create fertile ground for harmful information. Information related to the COVID-19 pandemic exposed both the speed of harmful information spread and the cost of delayed action. Meeting the challenges of identifying, preventing and mitigating harmful information requires a systems approach and urgent, cross-sectorial action. Fragmentation between humanitarian actors, digital rights advocates and the technology communities must be bridged. This includes creating space for emerging questions, fostering diverse and underrepresented perspectives, and prioritizing principles and accountability. Strong, forward-looking partnerships are essential, not only to shape governance frameworks, but to influence the design and deployment of tools used to navigate this rapidly evolving landscape.

Ultimately, states must act. As underscored in the Movement's Appeal to States, addressing harmful information is now central to principled humanitarian action. Trusted information may not be water, food or shelter but it is imperative to accessing all three, as well as to ensuring safety, dignity and autonomy.

Asks, aims and recommendations

Asks

Balance the protection of humanitarian space, the safety and dignity of affected populations and the integrity of humanitarian operations with freedom of expression by building rights-respecting, transparent, accountable and resilient information ecosystems.

Aims

Ensure laws and policies protect principled humanitarian action and safeguard trusted and reliable information.

Hold platforms accountable for measures that protect principled humanitarian action and implement transparent, rights-respecting crisis protocols that safeguard affected populations and humanitarian personnel.

Embed harmful information analysis into humanitarian operations to safeguard trust, access and principled humanitarian action.

Engage affected communities in information strategies and work with digital platforms to ensure timely, accurate and safe information flow.

Build trusted information ecosystems through digital and media literacy, local dialogue, initiatives that counter polarization and ongoing monitoring to evaluate and adapt interventions.

Recommendations

States and policy-makers

- Adopt rights-respecting regulations to counter harmful information while safeguarding humanitarian action and trusted information.
- Coordinate with humanitarian actors to ensure laws and regulations protect and respect principled humanitarian action.
- Integrate risks from harmful information that could affect principled humanitarian action into negotiations, frameworks and operational planning.
- Respect and support the independence of National Societies in their auxiliary role, ensuring they can operate without interference.

Regulators and technology companies

- In humanitarian crises, report transparently on harmful content removal, moderation practices and algorithmic adjustments to ensure trust, accountability and prevent harm to people in need and humanitarian personnel and volunteers.
- Collaborate with principled humanitarian actors to design localized mitigation tools such as fact-check bots, verified information hubs and multilingual content.
- Strengthen enforcement against coordinated harmful information campaigns targeting principled humanitarian organizations and the safety and dignity of affected communities, humanitarian personnel and volunteers.
- Develop and apply rights-respecting crisis-response protocols in collaboration with humanitarian actors and share relevant data with humanitarian organizations and research hubs to support trusted, principled action.

Humanitarian actors

- Embed trusted information as essential for principled humanitarian action and for combating harmful information into humanitarian standards and operational frameworks.
- Share evidence of harmful information trends with regulators and platforms in ways that respect data protection rules and support principled humanitarian action to mitigate harmful information, protect humanitarian action and safeguard affected populations, humanitarian personnel and volunteers.
- Partner with local journalists, fact-checkers and trusted content creators to amplify accurate, contextualized and life-saving information for populations in need.
- Strengthen community engagement by integrating media literacy, information resilience and feedback mechanisms into humanitarian programming, ensuring communities can access, understand and act on reliable information.

Communities and local leaders

- Promote community-based media and independent journalism to provide timely, accurate and accessible information.
- Lead digital and media literacy initiatives to build critical thinking and resilience to harmful information.
- Act as trusted intermediaries by disseminating verified information through local channels such as radio, schools and faith institutions.
- Facilitate dialogue to counter polarization, address stigma and prevent harmful information from escalating into violence.

Endnotes

1 World Economic Forum. *Global Risks Report* 2025 (2025) p.34–35 https://reports.weforum.org/docs/WEF_Global_Risks_Report_2025.pdf

2 Council of Delegates of the International Red Cross and Red Crescent Movement, Resolution 5: Call for respect and support for principled humanitarian action (2024). https://rcrcconference.org/app/uploads/2024/10/CoD24_R5-Res-NIIHA-EN.pdf

3 UN. Universal Declaration of Human Rights (adopted by UN General Assembly Resolution 217 A (III)). 10 December 1948. www.un.org/en/about-us/universal-declaration-of-human-rights

4 "Global, free, open, secure, and interoperable access to the Internet is essential during armed conflict, allowing people to request, access, or deliver aid and obtain or share vital information and early warnings. This access also enables humanitarian actors to coordinate operations and distribute information and aid and is equally crucial for post-conflict restoration and peacebuilding." Freedom Online Coalition. Joint Statement on Protecting Human Rights Online and Preventing Internet Shutdowns in Times of Armed Conflict, adopted June 2025. https://freedomonlinecoalition.com/joint-statement-on-protecting-human-rights-online-and-preventing-internet-shutdowns-in-times-of-armed-conflict

5 Access Now. 'Libya floods: people need reliable internet now.' Press release. 22 September 2023. www.accessnow.org/press-release/libya-floods-internet

6 The UN Open-Ended Working Group on Security of and in the Use of Information and Communications Technologies (OEWG), established by the UN General Assembly (A/RES/75/240), is mandated to advance responsible state behaviour in cyberspace and strengthen the security and stability of the ICT environment.

7 A number of states have issued position papers on the application of international law in cyberspace, many of which are available through the UN Office for Disarmament Affairs (UNODA) repository and the UN OEWG documentation pages. https://meetings.unoda.org/open-ended-working-group-on-information-and-communication-technologies-2021

8 Pamment, J. Countering Information Influence Activities: *The State of the Art*. Riga: NATO Strategic Communications Centre of Excellence, (2018).

9 Humprecht, E., Esser, F. and Van Aelst, P. Resilience to Online Disinformation: A Framework for Cross-National Comparative Research, *The International Journal of Press/Politics*, 2020:25(3). See also Humprecht. E. 'Why resilience to online disinformation varies between countries.' *Democratic Audit* (2020) www.democraticaudit.com/2020/03/24/why-resilience-to-online-disinformation-varies-between-countries

10 Reuters Institute for the Study of Journalism. *Reuters Institute Digital News Report 2024* (2024) p.16 https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2024-06/RISJ_DNR_2024_Digital_v10%20lr.pdf

11 Newman, N. *Overview and Key Findings of the 2024 Digital News Report*. Reuters Institute for the Study of Journalism (2024) https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024/dnr-executive-summary

12 Torre, L., Ramos, G., Noronha, M. et al. Sourcing Local Information in News Deserts. *Media* 2024:5(3), 1228-1243 www.mdpi.com/2673-5172/5/3/78

13 Malhotra, P. "What You Post in the Group Stays in the Group": Examining the Affordances of Bounded Social Media Places. *Social Media + Society* 2024:10(3) DOI:10.1177/20563051241285777

14 World Economic Forum. *Global Risks Report* 2025 (2025), p.35. https://reports.weforum.org/docs/WEF_Global_Risks_Report_2025.pdf

16 See, for example, Lim, G. and Bradshaw, S. *Chilling Legislation: Tracking the Impact of 'Fake News' Laws on Press Freedom Internationally*. Center for International Media Assistance. (2023) www.cima.ned.org/publication/chilling-legislation

17 UN Security Council, Resolution 2730 (2024) on the protection of humanitarian and UN personnel https://docs.un.org/en/S/RES/2730(2024)

18 UN General Assembly, Countering Disinformation for the Promotion and Protection of Human Rights and Fundamental Freedoms, A/RES/76/227 (2021) https://undocs.org/A/RES/76/227

19 World Health Organization. *WHO Pandemic Agreement* (Resolution WHA78.1), adopted 20 May 2025, 78th World Health Assembly. The agreement, along with the resolution, sets out the final text agreed by member states and establishes the process for negotiating and adopting the Annex on Pathogen Access and Benefit Sharing. https://apps.who.int/gb/ebwha/pdf_files/WHA78/A78_R1-en.pdf

20 The 2022 Council of Delegates Resolution 12: Safeguarding Humanitarian Data (CD/22/R12) acknowledged that cyber operations, data breaches and disinformation pose serious risks to the trust and functioning of impartial humanitarian organizations. The Movement's Appeal for Respect for Neutral and Impartial Humanitarian Action highlighted how disinformation and misinformation jeopardize the safety of humanitarian workers, influence public perceptions and hinder humanitarian response. The Council of Delegates is one of the Statutory Meetings of the International Red Cross and Red Crescent Movement.

21 The spread of misinformation and disinformation is addressed in the Council of Delegates resolution 'Call for respect and support for principled humanitarian action' (CD/24/R5). The impact of disinformation on humanitarian organizations is included in the ICT resolution of the International Conference. International Conference of the Red Cross and Red Crescent, Resolution 2: 'Protecting civilians and other protected persons and objects against the potential human cost of ICT activities during armed conflict' (34IC/24/R2), adopted 2024. https://rcrcconference.org/app/uploads/2024/11/34IC_R2-ICT-EN.pdf

22 Protocol I additional to the Geneva Conventions, Article 51(2), (1977) and Article 13(2) of Protocol II (1977).

23 Rodenhäuser, T. and D'Cunha, S. Foghorns of war: IHL and information operations during armed conflict. ICRC Law and Policy Blog. (2023) https://blogs.icrc.org/law-and-policy/2023/10/12/foghorns-of-war-ihl-and-information-operations-during-armed-conflict

24 See the IFRC Disaster Law website for more information: https://disasterlaw.ifrc.org/why-disaster-law

25 See Richter, S. 'True or false? Here are five ways to manage false information about risk.' UN Office for Disaster Risk Reduction. Feature. 16 December 2022. www.undrr.org/news/true-or-false-here-are-five-ways-manage-false-information-about-risk.

26 See European Commission. *Information manipulation and misinformation: A threat for EU civil protection and humanitarian aid.* https://civil-protection-humanitarian-aid.ec.europa.eu/resources-campaigns/information-manipulation-and-misinformation. See also PreventionWeb. *Common Challenge: Misinformation and disinformation on disaster risk communication – Practical tips.* www.preventionweb.net/hubs/disaster-risk-communication-hub/do/misinformation-disinformation

27 Although the IFRC's Disaster Risk Governance Guidelines (https://disasterlaw.ifrc.org/DRMguidelines) do not address harmful information in DRM directly, they offer useful guidance on relevant topics such as enhancing disaster risk knowledge and education, protecting the rule of law in disasters, and enhancing coordination across all elements of the DRM continuum.

28 Singer, PW. and Brooking, ET. *LikeWar: The Weaponization of Social Media.* (2018) pp.3, 16, 19, 62

29 UN General Assembly Resolution A/78/L.49: Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development, adopted 21 March 2024.

30 Ibid, section g.

31 UN. *Information Integrity on Digital Platforms (Our Common Agenda Policy Brief 8)*, Executive Office of the Secretary-General. (2023) www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-information-integrity-en.pdf, www.un.org/en/information-integrity

32 UN. *Governing AI for Humanity: Final Report.* (2024) p.29 www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf The UN High Level Advisory Body on Artificial Intelligence (HLAB-AI) established by the UN Secretary General was composed of 32 experts and aims to align AI development with human rights and the SDGs. High-Level Advisory Body on Artificial Intelligence. UN Office of the Secretary-General's Envoy on Technology. www.un.org/digital-emerging-technologies/ai-advisory-body

33 Ibid. p.8. The report shows seven prominent non-UN AI initiatives. Seven countries are parties to all the sampled AI governance efforts, whereas 118 countries are parties to none (primarily in the global South). The sample comprised OECD AI Principles (2019), G20 AI Principles (2019), Council of Europe AI Convention drafting group (2022–2024), GPAI Ministerial Declaration (2022), G7 Ministers' Statement (2023), Bletchley Declaration (2023) and Seoul Ministerial Declaration (2024). Canada, France, Germany, Italy, Japan, UK and USA are parties to all sampled initiatives/instruments.

34 UNESCO. *Guidelines for the Governance of Digital Platforms: Safeguarding Freedom of Expression and Access to Information*. (2023) www.unesco.org/en/internet-trust/guidelines

35 Gleicher, N. 'Removing Coordinated Inauthentic Behavior.' Meta 8 July 2020. https://about.fb.com/news/2020/07/removing-political-coordinated-inauthentic-behavior; Clegg, N. 'What We Saw on Our Platforms During 2024's Global Elections.' Meta. 3 December 2024. https://about.fb.com/news/2024/12/2024-global-elections-meta-platforms

36 Microsoft. *Microsoft Digital Defense Report 2024.* Section: AI Threats and Ecosystem-Level Risks. (2024) p.88 www.microsoft.com/en-us/security/security-insider/intelligence-reports/microsoft-digital-defense-report-2024

37 Ibid. Section: Recommendations on Limiting Foreign Influence Operations. p. 93 www.microsoft.com/en-us/security/security-insider/intelligence-reports/microsoft-digital-defense-report-2024

38 Singer, PW. and Brooking, ET. *LikeWar: The Weaponization of Social Media* (2018) pp.221–233

39 US Communications Decency Act of 1996. 47 U.S.C. § 230. Protection for private blocking and screening of offensive material www.law.cornell.edu/uscode/text/47/230

40 US. Digital Millennium Copyright Act of 1998, Pub. L. No. 105–304, 112 Stat. 2860 (1998), codified in Title 17 of the US Code. The DMCA implements international copyright treaties, criminalises circumvention of digital rights management technologies, and establishes a 'safe harbour' shielding online service providers from liability for infringing content uploaded by users. www.copyright.gov/legislation/dmca.pdf

41 European Union. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act). (2022) p. 1–102. http://data.europa.eu/eli/reg/2022/2065/oj

42 Ibid. See in particular articles 33–44 on systemic risk management for 'very large online platforms' and 'very large online search engines', articles 14–23 on transparency and accountability, article 36a on the integration of the Code of Practice on Disinformation (February 2025), article 16 on user reporting and redress mechanisms, and article 36 on the crisis response mechanism. https://eur-lex.europa.eu/eli/reg/2022/2065/oj

43 European Commission. Code of Practice on Disinformation (2018) and Strengthened Code of Practice on Disinformation (2022) https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

44 Schiffrin, A. 'How Europe fights fake news.' *Columbia Journalist Review*. 26 October 2017. www.cjr.org/watchdog/europe-fights-fake-news-facebook-twitter-google.php

45 International Panel on the Information Environment. *Trends in the Global Information Environment: 2023 Expert Survey Results* (2023) www.ipie.info/research/sr2023-3

46 Schiffrin, A. 'How Europe fights fake news.' *Columbia Journalist Review*. 26 October 2017. www.cjr.org/watchdog/europe-fights-fake-news-facebook-twitter-google.php; Schiffrin, A. *The Pursuit of Truth: Fixes for the Spread of Online Mis/Disinformation* (2023) https://igp.sipa.columbia.edu/sites/igp/files/2023-12/IGP_Anya_Schiffrin_The_Pursuit_of_Truth-Fixes_for_the_Spread_of_Online_Mis_Disinformation.pdf

47 UNESCO, *Journalism,"Fake News" & Disinformation: Handbook for Journalism Education and Training*, (2021); OECD. *Facts Not Fakes: Tackling Disinformation, Strengthening Information Integrity*. (2022) https://doi.org/10.1787/5b7e3c1c-en. European Commission, *Report on the Implementation of the 2022 Strengthened Code of Practice on Disinformation* (2023).

48 Brezzi, M. Gonzalez, S., Nguyen, D. et al. An updated OECD framework on drivers of public trust in public institutions to meet current and future challenges, OECD Working Papers on Public Governance No. 48 (2021) p.9 www.oecd.org/content/dam/oecd/en/publications/reports/2021/12/an-updated-oecd-framework-on-drivers-of-trust-in-public-institutions-to-meet-current-and-future-challenges_bfa20b1b/b6c5478c-en.pdf

49 As an OECD study on COVID-19 measures found, digital prompts reduced people's intent to share false headlines by 21% compared to a control group, especially among frequent online users. OECD. *Misinformation and Disinformation: An International Effort Using Behavioural Science to Tackle the Spread of Misinformation*. Policy Paper No. 21 (2022) www.oecd.org/en/publications/an-international-effort-using-behavioural-science-to-tackle-the-spread-of-misinformation_b7709d4f-en.html See also MIT Initiative on the Digital Economy. *Reducing Misinformation Sharing with Accuracy Prompts*. Research Brief. (2024), reporting on field experiments showing that content-neutral accuracy prompts reduce the sharing of misinformation. https://ide.mit.edu/wp-content/uploads/2024/04/RB__3-31-24.pdf

50 Ibid (OECD, 2022)