Platform and Content Governance in Times of Crisis

Author: Julia Haas
Matthias Kettemann

osce
the representative on freedom of the media
THERE IS NO SECURITY WITHOUT MEDIA FREEDOM

This publication is part of the project “Spotlight on Artificial Intelligence and Freedom of Expression” (#SAIFE).

The report consolidates the views and recommendations expressed by the experts and conclusions of several roundtable discussions. They do not necessarily represent the official position of the OSCE and/or its participating States.

6a Wallnerstrasse
1010 Vienna, Austria
Phone +43-1-514-36-68-00
e-mail: pm-fom@osce.org
https://www.osce.org/fom/ai-free-speech

ISBN: 978-92–9234-741-3

Spotlight on artificial intelligence and freedom of expression

The way content is governed by dominant online platforms that have become **gatekeepers** to information is not only relevant for the **realization of freedom of expression** and **media freedom** but, ultimately, for **international peace** and **security**. Content governance determines the availability of information, the accessibility of public interest content, and the administration of information, including across borders. As online platforms deploy artificial intelligence (AI) to support the curation and dissemination of content as well as to filter and take down unwanted content, AI-led processes provide the basis for how society interacts with information online today. The emergence of technologies like generative AI raises additional questions concerning content production and the broader effects on the way information is distributed and perceived.

Putting a **spotlight on AI and freedom of expression** (SAIFE), the OSCE Representative on Freedom of the Media (RFoM) published the SAIFE Policy Manual in 2022, the culmination of two years of research and workshops with over 120 experts from diverse backgrounds including media, technology, and security. It provides **human rights-centric recommendations** to states on safeguarding free speech and media pluralism in the context of automated content governance on online platforms.

Content governance in crises

While **human rights-centric content governance** is key at all times, its significance becomes particularly evident in **times of crisis**. In view of contextualizing the SAIFE recommendations to the specific challenges arising in crisis situations, the OSCE RFoM initiated several expert workshops and roundtable discussions in cooperation with leading civil society organizations including Access Now, Digital Security Lab Ukraine, and the European Center for Not-for-Profit Law, as well as international partners such as the World Health Organization and the International Science Council, and independent researchers and academia from across the OSCE region. The main workshop was held in autumn 2022 with Professor Matthias Kettemann as rapporteur, the annex enlists all roundtable discussions and key contributors.

Based on these expert discussions, this report provides **five key considerations** for state policy and regulatory frameworks towards **healthy information spaces in times of crisis**, building on the core principles of robust transparency, human rights due diligence, and accountability mechanisms. Given legitimate concerns over information manipulation and interference, propaganda (for war) and access to reliable information as well as a related risk of hasty, unbalanced state measures in emergency situations, it is essential that any interference with freedom of expression and media freedom must be lawful, legitimate, necessary and proportionate, and that any introduced derogation from state human rights obligations is limited to the extent strictly required by the exigencies of the situation. With **human rights-based governance** being essential at all times, this report aims to highlight additional considerations relevant to crisis contexts by accentuating **crisis-specific components** to complement the core principles. Thereby, the report intends to provide guidance on how human rights-centric content governance can contribute to safeguarding, and even promoting, free speech and media pluralism in crisis situations and throughout the conflict cycle, which, in turn, contributes to sustainable democracy, comprehensive security, and lasting peace.

Crises of content governance?

Already in 2011, the **UN Guiding Principles on Business and Human Rights** (UNGPs) recognized that the scope of corporate responsibility is broader in conflict contexts and that businesses should respect the standards of international humanitarian and human rights law (Principle 12). The Guiding Principles note that during conflict, business practices may increase the risk of becoming complicit in gross human rights abuses committed by others, which necessitates "enhanced" due diligence (Principle 23). Yet, when recent crises emerged, online platforms did generally **not** have comprehensive frameworks in place or **crisis protocols** with clear definitions, checklists and measures to take. Policy changes have remained fragmented and reactionary, invoking questions of the need for more human rights-friendly and systemic responses. Given the increasingly central **role** they play **during crises,** there is a growing call particularly on **dominant online platforms** with considerable market power to adopt crisis-sensitive policies and practices.

In crisis situations, online platforms often provide **crucial spaces for public debate**. They can be critical in ensuring access to trustworthy, up-to-date and potentially lifesaving information. By facilitating information flows, they can **mitigate crisis** impacts. During crises such as natural disasters, health emergencies or conflict, promoting authoritative sources increases exposure to critical information, be it on protective measures, equipment, and facilities, or on treatment, humanitarian aid, or areas at risk. In contexts of censorship, suppression, media capture or state propaganda, platforms can enable access to independent news, secure communication and fact-checked information. Platforms also often play a vital role in **organizing** civic movements, bringing together polarized communities or drawing attention to emerging tensions. Recent examples in the context of pandemic responses, elections, evacuations, and emergency assistance illustrated platforms' potential to promote and proactively disseminate **information of public interest**. Next to prioritizing authoritative messages vetted by independent experts in some of these instances, various platforms have also strengthened their efforts in combatting the spread of falsities and deception. At the same time, however, recent crises such as the global pandemic, conflicts, and accelerated climate crisis have exposed risks of hasty and reckless content moderation as well as a growing societal susceptibility to disinformation and conspiracy theories, which is frequently seen as intersecting with challenges in the digital realm such as exposure to independent quality media and public interest information on online platforms. The **information disorder** emerging in the last years and increased mistrust in news reporting is regularly linked to the attention-driven business practices of dominant online platforms. Of late, their content governance has been referred to as being in crises itself.

Business practices prioritizing user engagement over human rights considerations risk to intensify societal tensions in emerging crises and the spread of incitement and prejudice during crises as well as to jeopardize post-conflict reconciliation. The facilitated spread of **mis- and disinformation** can fuel radicalization and violent extremism. In the context of **conflicts**, platforms often struggle to de-platform propaganda for war without blocking or shadow-banning content by certain communities or about the conflict per se. This can, in turn, contribute to authoritarian attempts to limit access to accurate information or create information voids that can be filled with malicious information and state propaganda. In general, AI's acceleration, amplification and scalability capabilities can carry particularly dire implications during crises. Challenges also arise in the context of **evidences of human rights violations**. While they may fall within the category of impermissible content, platforms are often the only easily accessible option to find and store proof essential for achieving accountability in the future.

Over the past years, online **platforms** have faced criticism for **falling short** of adequately providing contextualization, ensuring local language competences, and allocating sufficient resources to regions around the world – with reportedly **detrimental consequences in crisis situations**. In certain contexts, platforms have implemented **ad hoc carve-outs** to their content governance policies to address crisis-specific challenges, while in others, they have not. Typically, the allocation of resources and policy

attention is influenced by **market size, economic and reputation considerations**, and prioritized regional focus, rather than comprehensive impact and risk assessments. Whistleblower revelations exposed regional disparities in policy priorities and effectiveness, and how little some platforms have invested in addressing certain situations and their potential contribution to crisis dynamics.

Beyond a narrow understanding of what constitutes a crisis, platforms and their automated content governance systems have regularly been **weaponized** by authoritarian actors in a bid to shrink civic space, to propagate harmful and/or illegal speech, or to target independent, dissenting and scrutinizing voices. Common practices of digital authoritarianism include online violence, disinformation and smear campaigns, surveillance, and internet shutdowns. In this context, platform's ubiquitous data collection and analyses can be **exploited** for **surveillance**, and algorithmic logics can be **instrumentalized** to target individuals and groups that are already disadvantaged, marginalized or oppressed in society.

As **violations of human rights online** can lead to an escalation of suffering and exacerbation of systemic injustice, any spike in tensions or turmoil necessitates particularly **sound and context-specific governance of content**. While concerns over disinformation and access to reliable information intensify during crises, so do questions about balancing these concerns with safeguarding the right to freedom of expression and media freedom, including in the digital realm.

The role of the state

Platforms' substantial role in shaping information spaces prompts **state action to uphold and safeguard** freedom of expression and media freedom online. This general responsibility is framed, inter alia, by **Article 19** of the International Covenant on Civil and Political Rights (ICCPR). Regulation of information flows has to be clearly and precisely established by law, pursue a legitimate aim, be proportionate and necessary. Otherwise, regulation may lead to illegitimate restrictions, **adversely impacting** internet freedom, independent media, and public debate more broadly. Striking a balance considering various concerns and rights is challenging at all times, and becomes even more complex in intricate contexts of crisis. In times of public emergency which threatens the life of a nation, a state can – under an additional specific legal regime – exceptionally and temporarily derogate from their human rights obligations.

Generally, platform regulation should focus on processes rather than content, i.e., on reach, not speech. International human rights standards do, however, require certain content-related interferences: **Article 20** of the ICCPR mandates states to prohibit propaganda for war and incitement to discrimination, hostility, and violence. Similar provisions exist regarding racial and gender-based discrimination and incitement to genocide and other international crimes, which are equally relevant in crisis situations. In these cases, **regulatory frameworks** are necessary as part of states' positive human rights obligations.

To this day, most states do not have **specific procedures or policies** in place regarding platform governance during crises. Yet, **preparing** for crises built on an inclusive public debate can help to identify balanced responses. At the same time, states should be wary about **excessive regulation** or preserving an everlasting sense of emergency. Perpetually adopting a crisis mode can prove counterproductive, especially in the context of the widespread **securitization of human rights**. A clear regulatory and policy framework, however, can contribute to more effective, consistent and human rights-friendly measures while preventing a disproportionate restriction of information.

There are **proliferating regulatory initiatives** aimed at ensuring more anticipatory and sustainable responses to challenges of content governance in times of crisis. The draft **UNESCO guidelines** on regulating digital platforms, for example, stipulate that risk assessment and mitigation policies should be in place for emergencies, crises, and conflict, as well as other significant changes in the operating environment. EU Member States, moreover, will soon have to fully apply the **Digital Services Act**, which entitles the European Commission to demand very large online platforms to take **specific measures** during crises, such as modifying AI tools to display authoritative information more prominently, and to initiate **voluntary crisis protocols** that define (1) parameters determining extraordinary circumstances; (2) the role of participants and measures to be put in place; (3) a procedure to determine when the protocol is to be activated; (4) a process to define the period of measures; (5) safeguards to address negative human rights effects; and (6) public reporting.

Human rights-based approach

Identifying a **human rights-based** approach to platform and content governance during crises necessitates additional considerations complementing the generally required safeguards as identified in the SAIFE Policy Manual. These supplementary components depend on the various **stages** and **types** of crises. **Before** a crisis unfolds, different measures may be appropriate and needed than **during crisis management**, or in **post-crisis** settings. While **crises differ** depending on context and nature of the threat and emergency, recognizing similarities – and differences – can be relevant for identifying proportionate responses. Determining **patterns** can provide useful lessons and guidance for the prevention and handling of future crises. In this regard, as vetted in the various expert forums, it may be valuable to categorize crises as **short-**, **medium-** or **long-term**. Short-term crises include, for example, terrorist attacks. While a medium-term crisis could be an ongoing pandemic or natural disaster, long-term crises include climate change and protracted conflicts. Ongoing acts of war or armed conflict can be of medium- or long-term nature depending on the context.

A sustainable and effective international crisis response requires **clear terminology** and **criteria** for determining what circumstances constitute a crisis, when crisis protocols should be activated, which specific measures they should entail at which stage, and for how long. Generalizable factors can serve as a foundation for crisis-specific regulation and protocols, which should further provide for appropriate contextualization and localization. Once a crisis unfolds, the adequate contextual grounding requires close cooperation with local **civil society** and experts. A certain level of coordination with states is also crucial, depending on their role and whether a crisis is linked to state repression or to external factors.

The stage and type of the crisis also impacts the level of appropriate use of AI for content governance. **Short-term crises** call for prompt actions and may justify the use of AI to a greater extent, for example to instantaneously remove the livestreaming of violent acts or to provide information on rapid relief. While AI can reduce the visibility of harmful content, its shortcomings can equally result in an over-blocking of legitimate content and thus reduce exposure to public interest information.

Human rights due diligence is essential in order to assess the human rights impact and risks of platform services, content governance and related policies during crises, and to timely understand where tensions may arise and require an adaption of policies. While due diligence is a general requirement for human rights-based governance, considering **conflict sensitivities** in a regular and transparent manner provides for faster and more flexible responses to emerging risks, in particular if built on **multi-stakeholder engagement** and local expertise. When crises **protract**, robust due diligence to continuously assess and adapt policies becomes more important, relying less on algorithmic solutions alone. Risks stemming from under- and over-enforcement, surveillance-based advertising, market concentrations, and the broader internet infrastructure also need to be considered.

A democratic **public debate** on content governance is crucial, in particular in extended crises, including to ensure **predictability** of applicable rules and overall consistency of their application. As always, applicable changes and notice-and-review mechanisms need to be clearly communicated to users who should be provided with effective remedies. The longer a crisis lasts, the clearer and more transparent exceptional rules and measures need to be, the more inclusive decision-making processes should become, and the more resources should be invested in ensuring content governance is aligned with human rights. As adequate transparency and accountability mechanisms still lack for content governance generally, it becomes even more vital to ensure **oversight** of crisis-related policies, protocols and carve-outs.

Key considerations

The OSCE RFoM would like to highlight **five key considerations** for states in the context of platform governance in times of crises to safeguard freedom of expression and media freedom. Based on several expert forums, these five elements refine the **SAIFE Policy Manual** guidance on a human rights-based use of AI for content governance by accentuating crisis-specific components.

1. States should prepare for crisis, and mandate online platforms to have certain crises-specific measures in place. This should include developing comprehensive human rights-based crises protocols that ensure responses are not ad hoc but coherent, while ensuring contextualization, and clear timeframes.

If states issue human rights **derogations** during crises, they must be limited to the extent strictly required by the exigencies of the situation. **States of emergency** or martial law should be declared only in case of genuinely exceptional circumstances, be limited in time, and be introduced in line with the rule of law.

With exception to what is required by international law, states should not criminalize crisis-related information unless legality, legitimacy, proportionality and necessity criteria are met, but to the contrary **promote access to crisis-linked public interest information**. Crisis-specific content regulation must be carefully designed to avoid media capture and undue restrictions of the free flow of information.

States should establish **clear** and comprehensive rules and procedures to **determine** what constitutes a crisis, and ensure that state regulation as well as platform policies governing information are formulated in an easily accessible way (including in local languages).

States should ensure that their and platforms' criteria for crisis responses consider the unique context of each setting and locality. Crisis protocols should be flexible to recognize the **diverse nature** of crisis contexts, while ensuring **consistency** within an overall framework of human rights due diligence, transparency, accountability and safeguards against misuse.

States should ensure **online connectivity** and open **communication channels** during crises, including by refraining from imposing **internet shutdowns**, throttling or fragmentation, and strengthening efforts to bridge digital divides that may enlarge during crisis situations.

Generally, states should strengthen democratic resilience and enable an **independent**, **pluralistic media landscape**. By providing accurate and timely information, giving voice to the marginalized, and offering a platform for dialogue, the media can play an essential role in mitigating risks and impacts of crisis situations.

2. States should mandate platforms to undertake crisis-sensitive human rights due diligence. Each crisis requires a different customized approach depending on its type (short-, medium-, long-term) and the phase it is in (before, ongoing, after).

States should mandate online platforms to conduct **crisis-sensitive human rights risks and impact assessments**. Due diligence should consider the unique cultural, linguistic and political dimension of each crisis as well as the prevention, mitigation, and management of risk factors, and aspects of **marginalization** and power imbalances. It should be embedded in an overarching human rights by design approach.

Crisis-sensitive due diligence should be conducted ex ante and **regularly** throughout the lifecycle of a crisis. Emergency measures should be tested before being deployed. Due diligence should be conducted in an inclusive and transparent manner, while the level of transparency can vary according to the different stage of the crisis (the longer a crisis persists, the greater the need for transparency).

Due diligence should consider proportionality and reliability on **AI tools** and automated measures.

States should mandate platforms to take all necessary steps to **address and mitigate** the identified adverse human rights impacts and to allocate sufficient resources to crisis-sensitive content governance (grounded in linguistic and context-specific competencies and capacities), **remedies** and independent crisis-response **audits**. States should incentivize platforms to activate crisis-specific safety features such as encryption, cybersecurity measures or account locking, as well as to increase user agency and empowerment.

States should encourage research into how **AI tools** can be utilized to ensure a human rights-based approach to managing and overcoming crises, and how the use of AI, for example through **value-oriented content recommender systems**, can promote access to independent quality media and public interest content in times of crisis.

3. States should ensure meaningful multi-stakeholder engagement in decision-making processes on crisis protocols and deployed tools. In order to ensure legitimacy and that content governance is rooted in human rights and local context, inclusive coordination is needed.

States should actively engage and collaborate with **international partners** and local civil society to identify, implement and promote international standards for human rights-centric content governance in times of crisis, fostering a coordinated and cohesive response.

States should adopt an **evidence-based**, **interdisciplinary** and **inclusive** approach to regulating AI-based content governance, considering its role in the context of emerging and recurring crisis. They should provide for meaningful engagement in their **decision-making processes** on crisis regulation.

States should ensure online platforms create **communication** and engagement **channels** with relevant stakeholders for crisis contexts.

In particular, states should mandate platforms to create **crisis protocols** based on multi-stakeholder engagement. These should provide for close coordination with local experts and civil society throughout the crisis cycle and on all aspects, including risk assessments, policy adjustment processes, the determination of the ending of crises-specific measures as well as deployed AI tools. Protocols should incorporate mechanisms for monitoring, adjustments, and accountability for crisis-specific policies.

States should urge platforms to provide **access** to **data** on crisis-specific policies and practices to support independent research, documentation, and audits, and fund independent crisis-specific research.

States should encourage multi-stakeholder initiatives that explore how crisis-specific considerations and lessons learnt can inform **overall platform and content governance** and enable healthier and more vibrant information spaces in the digital realm.

4. Human rights-centric responses to the weaponization of information become ever more important in crisis contexts. States should exercise their obligation to respect, protect and fulfil human rights in a transparent manner, based on the rule of law.

- States should refrain from **spreading or sponsoring disinformation** or deploying malign information operations at all times. States should not use information as a weapon to manipulate, deceive, or sow confusion within their own territory or across borders.
- While states should **not** generally **prohibit information based on accuracy**, they should **prohibit and combat propaganda for war**, incitement to genocide or other international crimes, as well as **advocacy of hatred** that constitutes incitement to discrimination, hostility or violence.
- Particularly in times of crisis, states should apply maximum transparency of public administration, **proactively disclose** information of public interest, encourage fact-checking, and support information and digital literacy.
- States should request online platforms to **allocate adequate resources** to combat the weaponization of information and coordinated inauthentic behaviour in particular in crisis contexts, based on human rights due diligence and identified risks.
- States should encourage platforms to **preserve** content removed during crises to support **accountability mechanisms** with evidence of human rights violations and international crimes.
- Particularly in times of crisis, states should limit platforms' ability to prioritize profit maximization at the expense of human rights or public interest.
- States should explore **incentives** beyond legislative approaches to promote a healthier online information ecosystem.
- States should recognize, protect and promote the role of **independent quality media** in combating disinformation and the weaponization of information during crises.

5. Any crisis protocol and preventive or responsive measure should include intersectional considerations with a strong gender perspective. Specific attention should be given to the risks and consequences of structural discrimination, historical marginalization, global power imbalances and vulnerability factors.

ANNEX
Expert Roundtables and Discussions Platform and Content Governance in Times of Crisis

Expert Roundtables and Discussions

- SAIFE expert workshop on "Content Governance in Times of Crises: Conflicts, COVID, and Climate Change", October 2022
- Session on "Content capture and content control: Spotlight on artificial intelligence and freedom of expression in times of crisis" at the Internet Governance Forum, November 2022
- Panel discussion on "Content governance in times of crisis" at the SAIFE Expedition, December 2023
- Session on "Freedom of Opinion in Times of Crises: the impact of AI-based content governance" at the Mozilla Festival, March 2023
- Expert discussions at the Forum on "Artificial Intelligence 2.0: Regulation and Work during the War" by the Center for Democracy and the Rule of Law (CEDEM) and Digital Security Lab Ukraine, April 2023
- Workshop on "Content governance in crises or crises of content governance? The role of the state in ensuring access to information" at RightsCon, June 2023
- Session on "Ukraine, information interventions and content governance in times of crisis" at the Oxford Media Policy Institute, August 2023

Key contributors (alphabetical order)

Tetiana Avdieieva, Marwa Azelmat, Laura Becana Ball, Maksym Dvorovyi, Marwa Fatafta, Arzu Geybullayeva, Julia Haas, Matthias Kettemann, Elisa Lindinger, Alison Meston, Eliska Pirkova, Courtney Radsch, Zach Rosson, Cristian Vaccari, Marlena Wisniak, Brian Yau


osce
the representative
on freedom
of the media
THERE IS NO SECURITY WITHOUT MEDIA FREEDOM