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

SAIFE expert workshop on
Content Governance in Times of Crises: Conflicts, COVID, and Climate Change

11 October 2022

CONCEPT NOTE

Spotlight on the impact of artificial intelligence on freedom of expression (SAIFE)

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 pluralism but, ultimately, for international peace and security. Content governance determines the availability of information, the accessibility of public interest content, and the flow of information, including across borders. As online platforms deploy artificial intelligence (AI) and automation to support the prioritization and dissemination of content as well as to filter and take down illegal, harmful, or otherwise unwanted content, AI-led processes provide the basis for how society interacts with information online today.

Putting a spotlight on the impact of AI on freedom of expression (SAIFE), the OSCE Representative on Freedom of the Media published the SAIFE Policy Manual in January 2022, the culmination of two years of research and several workshops with over 120 experts from diverse backgrounds. The Policy Manual provides human rights-centric recommendations for states on how to safeguard freedom of expression and media pluralism in the context of automated content governance of online platforms.

Content governance in crises or crises of content governance

While human rights-centric content governance is key at all times, it becomes particularly relevant in times of crises, where the right of society to know and to be informed becomes ever more essential. The aim of this expert workshop is to understand the specific challenges of content governance in crisis situations and to contextualize existing recommendations.

Online platforms play a growing role in crises, be they health emergencies, conflicts, or natural disasters. The expert workshop will explore whether and how platforms' services and business practices prioritizing user engagement and ad revenues over accuracy, diversity and public interest can contribute to polarizing public discourse and to increasing societal tensions during emerging crises. And whether and how they can contribute to inciting violence and suppression during crisis, and to widening divisions hampering post-crises reconciliation.

Plenty of examples illustrate how platforms and the rules and tools of content governance can be weaponized by powerful actors in a bid to conduct information (and disinformation) operations, shrink civic space, propagate harmful speech, and target dissenting voices. Common practices of digital authoritarianism include disinformation and smear campaigns, surveillance and internet shutdowns. State requests for removal of content can lead to censorship, including of news content, and careless content moderation ('overblocking'), including through reacting to notifications by actors involved in coordinated inauthentic behavior, can result in account suspensions of those subjected to targeted harassment or silencing attempts. Automated content governance can be exploited to target individuals, groups and communities that are already disadvantaged, marginalized or otherwise oppressed in society, as well as to silence prominent dissenting or opposing voices to restrict public's access to information. In parallel, targeted hatred and violent content on online platforms have been instrumentalized to further agendas of systemic discrimination and the persecution of minorities and dissenting voices. Moreover, platform's ubiquitous data collection and analyses pose risks of surveillance that can become particularly dangerous in crisis contexts.

The workshop will explore online platforms' role in providing spaces for public debate particularly in contexts of censorship, suppression, media capture and state propaganda. Platforms may provide the last remaining access to independent news and supply essential information. They can also play a vital role in organizing civic movements, fact-checking and bringing together communities to overcome tension.

In the context of conflicts, content governance policies often struggle to deplatform war propaganda without blocking content about the conflict *per se* which contributes to authoritarian attempts to limit access to accurate information. The workshop will look into automated content governance and whether they can undermine the public's access to information, or amplify global networks of disinformation. The workshop will also discuss automated content governance's contribution to the removing of evidence of human rights violations that is essential for accountability and reconciliation, as well as of reporting on war crimes and counter speech. Overall, the workshop aims to assess what standards and practices would apply or could be replicated to protect and enable the right to freedom of expression, and promote the role of independent news media in times of crises.

The expert workshop will explore whether rising tensions, turmoil and crisis situations require context-specific content governance, and whether contemporary policies sufficiently extend to conflict sensitivities. Typically, resources spent and policy focus depend on the market size and economic or reputation considerations rather than on comprehensive impact and risk assessments. Consequently, online platforms may fall short of providing contextualization, ensuring local language-knowledge and allocating sufficient resources to various regions across the world – with a disproportionate impact on disadvantaged groups and dangerous potential for detrimental consequences during crises. Whistleblower revelations have highlighted how content governance priorities and effectiveness differs between regions, and how little some platforms have invested in understanding political nuances of local contexts and stopping the spread of harmful and illegal content in specific crisis-struck contexts.

In some contexts, platforms have provided for ad hoc carve-outs and changes in content governance policies to address crisis-specific challenges. The expert workshop will explore such policy adjustments, as well as the process thereto. To date, there are no comprehensive frameworks or crisis protocols in place, with clear definitions, checklists or measures to undertake once an emergency situation is unfolding. Crises-induced content governance changes have remained fragmented and reactionary, prompting the workshop's question of the need for human rights-friendly and systemic responses.

The expert workshop will also identify to what extent crises differ depending on the context and nature of the emergency, and whether specific patterns and generalizable factors are useful for crisis protocols and deployable tools that could be contextualized and localized when needed, and how communication with platform can be improved.

Principle 12 of the UN Guiding Principles on Business and Human Rights states that the scope of corporate responsibility to respect human rights is broader in conflict contexts, given that the likelihood and severity of harm is greater and the threshold for hatred, incitements to violence and other conflict drivers comparatively low. Moreover, Principle 23 recognizes that business practices may risk complicity in human rights abuses, which necessitates extra care and due diligence.

Regular, transparent and comprehensive human rights impact and risks assessments could contribute to understanding where tensions or crises may rise, and allow for the adaption of policies.

The workshop will explore whether acknowledging conflict sensitivities, based on multi-stakeholder engagement and close coordination with civil society, could provide for fast and flexible responses to emerging risks and crises.

The expert workshop particularly aims to explore the role of states based on their positive human rights obligations. States can set regulatory frameworks for human rights-centric content governance, built on robust transparency, oversight and accountability mechanisms. Given the increased risk of hasty, unbalanced state measures in times of crises, coupled with legitimate concerns over disinformation and (war) propaganda, the imperative to protect freedom of expression and media freedom highlights that any interference must be lawful, legitimate, necessary, proportionate and time-bound.

This workshop aims to explore how human rights-centric content governance can contribute to safeguarding freedom of expression and media freedom throughout the conflict cycle, and, ultimately, contribute to peacebuilding efforts, alleviating conflict drivers, and thus to comprehensive security.

SAIFE expert workshop on
Content Governance in Times of Crises: Conflicts, COVID, and Climate Change

11 October 2022
10am-2.30pm CEST

AGENDA

10.00-10.10 CEST

Welcome

- Welcome remarks by **Teresa Ribeiro**, OSCE Representative on Freedom of the Media and SAIFE team
- Brief introduction of the SAIFE Policy Manual
- Introducing the agenda and objectives of the workshop
- Housekeeping rules

10.10-11.00

Tour de table

- Name and affiliation (and favorite vacation spot)
- *What 3 key words come to mind when you think of the role that content governance plays in crisis situations (in the run-up to crisis, during crisis and post-crisis settings)?*

11.00-11.30

Introductory presentations

- **Tetiana Avdieieva** (CEDEM Ukraine) and
  Maksym Dvorovyi (Digital Security Lab Ukraine)
  Content governance in the context of the war against Ukraine
- **Arzu Geybullayeva** (journalist)
  Content governance in the context of Armenia-Azerbaijan/Nagorno-Karabakh
- **Brian Yau** (WHO)
  Infodemics and content governance during health emergencies
- **Alison Meston** (International Science Council)
  Content governance in the context of climate change
- **Marwa Fatafta** (Access Now)
  Content governance in crises

11.30-11.45

Coffee break

11.45-12.30

Session I
Role of rules and practices of content governance by online platforms in times of crisis in enabling and violating the right to freedom of expression

- Introduction by rapporteur Prof. **Matthias Kettemann** (University of Innsbruck)

- Discussion among experts
  - *What purpose should content governance serve in times of crisis? How can online harms be reduced, and public interest information be promoted online?*
  - *Do crises necessitate new/different rules for automated content governance? And if yes, what, and who is to set them?*
  - *What are potential positive contributions of automated decision-making systems in protecting human rights, and in particular freedom of expression and the free flow of information in times of crisis?*

12.30-12.45 Coffee break

12.45-13.30 Session II
Positive human rights obligations of states in respect to content governance, freedom of expression and media freedom in times of crisis

- Introduction by rapporteur **Matthias Kettemann**
- Discussion among experts
  - *What are the international standards, and practices, that would apply (or could be replicated) to protect and enable the right to freedom of expression, and promote the role of independent news media in times of crisis?*
  - *What obligations should be established by regulatory frameworks in order to enable human rights-centric content governance in times of crises? How can meaningful transparency and accountability look like? How can states ensure an inclusive and participatory approach?*
  - *What developments are currently on the horizon?*
  - *What role does the SAIFE Policy Manual play?*
  - *Do new rules need to be developed?*

13.30-13.45 Coffee break

13.45-14.15 **Closing discussion**
- Areas not covered by this workshop which would need additional attention in the context of crises
- Summarizing the takeaways and seeking to identify operational recommendations

14.15-14.30 **Concluding remarks and way forward**

SAIFE expert workshop on Content Governance in Times of Crisis: Conflicts, COVID, and Climate Change

CONCEPT NOTE

SAIFE expert workshop on Content Governance in Times of Crisis: Conflicts, COVID, and Climate Change

AGENDA

SAIFE expert workshop on Content Governance in Times of Crisis: Conflicts, COVID, and Climate Change

BIOS OF PARTICIPANTS

You can find the concept note, agenda, and bios here

Resources

- SAIFE Policy Manual: https://www.osce.org/files/f/documents/8/f/510332_0.pdf
- Asaf Lubin's book: https://ccdcoe.org/uploads/2022/06/The-Rights-to-Privacy-and-Data-Protection-in-Armed-Conflict.pdf
- Civil society feedback created based on a leaked version of the call, because it was not voluntarily provided to civil society at all. civil society engagement has improved in leaps and bounds since then: https://www.eff.org/files/2019/05/16/community_input_on_christchurch_call.pdf
- https://globalfreedomofexpression.columbia.edu/publications/the-digital-berlin-wall-how-germany-accidentally-created-a-prototype-for-global-online-censorship/
- https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-CRP-MapGap-1.1.pdf

Tour de table

What 3 key words come to mind when you think of the role that content governance plays in crisis situations (in the run-up to crisis, during crisis and post-crisis settings)?

accountability, transparency, trust

Security, Peace and **Accountability**

oversight and remedies

Context & Consistency

Trust & **Transparency**

Vacation spot

close to ancient trees

remote island

mountains

nature

ocean

Unpredictable, Superficial

Policies of Care & Resistance

context

Inclusiveness, Democracy & Public Sphere

important unpredictable superficial

scale, speed, explainability

People, Power, Political Economy

Considerable fundamental rights (i.e. beyond freedom of expression) impact

consistency

about processes not content - centering human rights essential

Scale, Speed, Explainability

Speaking Truth to Power

responsibility, due giligence

Reinforcement of Power Structures

Immediacy & Efficiency

Inform, Disinform, Uniform

Context, Language, Timeliness

Prediction, Hash Matching, Policy Shocks

Checkability and **Governance**

Documentation vs Censorship

Peace, Love, Connectivity vs. Ignorance, Colonialism, Indifference

Propaganda Prevention
Epistemology
Culture
Linguistics
Viewpoint & Ideology
Markets
Infrastructure
Norms
Evidence
Resources

SESSION 1

Role of rules and practices of content governance by online platforms in times of crisis in enabling and violating the right to freedom of expression

- What purpose should content governance serve in times of crisis? How can online harms be reduced, and public interest information be promoted online?
- Do crises necessitate new/different rules for automated content governance? And if yes, what, and who is to set them?
- What are potential positive contributions of automated decision-making systems in protecting human rights, and in particular freedom of expression and the free flow of information in times of crisis?

Experts' recommendations & main observations and key messages

differentiate between crisis and between pre-durante-post crisis action

not suggesting solutions that only works in democracies

dictotor-proving

Problematic state requests - Internet Referral Units located in law enforcement (not judicial) apparatus often have a privileged and opaque relationship regarding the regulation of speech on popular platforms - very dangerous regarding censorship

context dependent moderation is key

differenteiating between mis- and disinfromation crucial

misinformation vs. disinformaiton - importance of getting the terminology right as these two would call for different level of harshness of content moderation actions

The difference between mis and dis information is intent to harm. Not all opinions are created equal.

Any actions that impact civil liberties and break laws offline should be equally applied online, to all platforms regardless of emergency state.

AI should be seen as a tool not a solution

great point to make the distinction between AI and automation

digital colonialism

Necessary to learn from lessons in other contexts

replicability, clarity, explicability needed

Platforms only respond when there is a reputation crisis and the cost for them not to take an action is high

Access to Information Readiness Participation Scrutiny

set of rules to the platform may be a part of the States' positive obligations

consistent and stable principles for consistent application by the platforms

look not only at created content but creation chain

need to get governance right before discussing AI

too many systemic failures without consequences for platforms

transparency is essential

Huge disparity in platforms' reactions in crises (esp. regarding swiftness, consistency etc.)

Meaningful participatory processes are essential

not call for specific carve outs but for actions to be done transparently, in consultation with communities on the ground and civil society, and ground in human rights

SESSION 2

Positive human rights obligations of states in respect to content governance, freedom of expression and media freedom in times of crisis

- *What are the international standards, and practices, that would apply (or could be replicated) to protect and enable the right to freedom of expression, and promote the role of independent news media in times of crisis?*
- *What obligations should be established by regulatory frameworks in order to enable human rights-centric content governance in times of crises? How can meaningful transparency and accountability look like? How can states ensure an inclusive and participatory approach?*
- *What developments are currently on the horizon?*
- *What role does the SAIFE Policy Manual play?*
- *Do new rules need to be developed?*

Experts' recommendation & main observations and key messages

FoE does not exist in vacuum, but in line with other human rights

States can set up communication infrastructure and impact information

Need to consider digital authoritarianism

Existing rules e.g. Geneva Conventions need to be applied in context of armed conflict also regarding CoGo

Maybe new state obligations under international humanitarian law needed

Moderation should not serve states' political interests (e.g. not steer hate)

two sides: regular due process regulation improving transparency and accountability and in parallel crises

Important to look at who is moderating - humans maybe without context knowledge and support or unrepresentative data training AI tools

transparency

System of checks & balances and rule of law important for proper framework

Media literacy is essential

careful about potentially labelling speech or its regulation as a 'use of force'

Focusing on platforms might not fit during crises - states have an inherent interest to step in

States are ultimately responsible for providing effective and accessible remedies to individuals

Important to consider copycat legislation in authoritarian context, but it cannot prevent proper rule-of-law regulation

Policy shocks & policy responses - multi-stakeholder processes essential

How far you can leave the automated content moderation if you can't keep up with the human review

Multi-stakeholder & multi-disciplinary processes essential, but considering who is at the table (costs etc.)

access to data and independent research important - maybe different at times of crises

States shouldn't hand over responsibilities to private actors

Consistent review of policies important, in participatory processes

Jurisdiction

Conclusion / Key Messages / Take-aways

qualified human review with linguistic, local context and geopolitical nuances understanding

Crises specificaties vs similaries

Crises might shift the main actor of content moderation to states

Relevant to consider authocracies - but democratic governments need to set the agenda / rules of the game

crises literacy next to media literacy

proportionality

Context of mis-/disinformation and evidence of human rights violations provide useful lens

consider positive potential of AI

participatory processes throughout are essential

Due prominence of public interest content

transparency throughout!

difficult relation between platforms and states - even more relevant in state of emergencies

Difference in short-, medium- and long-term crises, impact & efforts --> necessary framework

multidisciplinary whole of society approach

AI adds level of complexity because of the ecosystem and power of a few dominant platforms and their commercial focus / ad prioritization - but AI may offer some beneficial potential re speed and efficiency (e.g. Christchurch)

consider niche platforms and smaller tech

Fair rules and fair implementation of them

crises protocols