March 2025

GUIDE

FOR RISK MANAGEMENT IN THE CONTEXT OF EMERGENCIES, ARMED CONFLICTS AND CRISES

Based on contextual analysis of the situation in Ukraine

unesco

From the People of Japan

IMS

INTERNEWS UKRAINE

Executive summary

1. **Adaptation of community standards and moderation principles**
   Social platforms are encouraged to consider or further develop a flexible application of their global policies and standards, as well as moderation practices, in regions experiencing emergencies, crises, and armed conflict, guided by international humanitarian law, human rights standards, and local expertise. This means considering unified approaches for clusters of countries with common risks, particularly related to the right to self-defence and self-determination. Such flexibility should be based on engaging relevant local stakeholders and monitoring events in conflict regions, like Ukraine, allowing platforms to update algorithm settings promptly and avoid unjustified blocking or reduction in the visibility of content.

2. **Creation of regional crisis teams**
   Platforms are encouraged to consider or integrate into crisis protocols the formation of specialised teams for rapid response to contexts affected by crises, emergencies, and armed conflict. These teams, working in collaboration with local organisations, can ensure quick decision-making and adaptation of policies and community standards to the context. Ukrainian partners have demonstrated readiness to actively participate in combating disinformation and hate speech.

3. **Regulation of the use of commercial tools**
   Platforms are encouraged to increase transparency in relation to the use of advertising tools to prevent manipulation and abuse. Increasing transparency regarding ad funding, analysing its content and sources, and thereby reducing the risks of disinformation spreading will contribute to strengthening trust in social media platforms.

4. **Support for and strengthening of fact-checking efforts**
   Platforms are encouraged to strengthen their fact-checking efforts and collaborations. Independent research[1] and experts[2] have concluded that fact-checking matters in countering the spread of disinformation. It would be beneficial not only to maintain existing fact-checking mechanisms but also to actively develop and adapt them to local needs, ensuring cooperation with local fact-checkers and experts to account for regional specifics.

[1] van Erkel, P. F. A. et al., (2024).
[2] European Commission, (2018).

Guide for Risk Management in the context of emergencies, armed conflicts and crises

Based on contextual analysis of the situation in Ukraine

The intention of the Guide for Risk Management is to provide guidance and mitigation recommendations for social media platform companies (hereafter 'companies'). The guide offers a context-specific framework enabling companies to safeguard human rights in the Ukrainian information ecosystem, particularly freedom of speech and access to information, whilst placing a specific emphasis on groups such as women, minorities and marginalised groups that are commonly targeted. This is interlinked with the second scope of the guide, which is to inspire action beyond Ukraine by outlining how the framework can be adapted to other countries affected by crisis, emergency, and armed conflict.

This guide is a living document which will be subject to change according to the contextual developments within and outside Ukraine.

Introduction

Safe access to reliable information can be the difference between life and death in crisis, emergency and armed conflict. Just like crisis protocols for the physical world, protocols for the digital sphere can save lives. Social media platforms have become crucial digital infrastructures for accessing and sharing information and play a key role in promoting reliable information and combating disinformation and hate speech in conflicts like the full-scale invasion of Ukraine.

For Ukraine, social media platforms are critically important amid the full-scale invasion; 84% of Ukrainians use social media as their primary news source and 42% view it as their only channel for information.[3]

Since 2014, and especially since 2021, Ukraine has been subject to hybrid warfare where information warfare tactics are intertwined with kinetic warfare, driven by the Russian Federation on an unprecedented scale. However, Ukraine is far from a standalone case when it comes to hybrid warfare. On the contrary, information warfare tactics and foreign interference are becoming a global issue.

The full-scale war has seriously increased the targeting and vulnerability of women, minorities and marginalised groups and led to the emergence of new vulnerable groups. Women, minorities and

[3] Schafer, Bret et al., 2022.

marginalised groups have faced a heightened risk of violence, displacement, and economic hardship.

Emergency, crisis and armed conflict as well as disinformation and hate speech disproportionately affect women, minorities and marginalised groups. Representatives in the public eye like politicians, journalists, servicewomen and activists are at high risk of becoming targets of online attacks. Sensitising efforts and initiatives to the needs and rights of women, minorities and marginalised groups not only helps identify risks but also enables the development of effective mitigation measures, contributing to the protection of human rights and democratic values.

The Ukraine Risk Management Guide is the result of a multi-stakeholder collaboration between a cross-sectoral national expert working group (civil society organisations, government bodies, regulatory agencies, academia and media); individual local, regional, and international experts; and companies.

The risks identified in the guide are based on experiences from Ukraine from February 2022 to January 2025 but particularly focus on the time between February and December 2022. War fluctuates and changes – sometimes by day and by hour – hence mitigation techniques ought to always be flexible and adaptable, requiring ongoing assessments of the context. Guiding risks and recommendations for mitigations building on best practice can function as vital tools particularly during times of escalation and heightened armed conflict, allowing stakeholders to act swiftly. This guide is inspired by and follows the principles outlined in "Guidelines for the Governance of Digital Platforms"[4] by UNESCO and the "Declaration of Principles for Content and Platform Governance in Times of Crisis"[5] by Access Now, which emphasises the importance of flexibility, proportionality, and consideration of local context in content moderation and information dissemination.

Given the hybrid nature of the conflict, it is important to consider the high risk of external informational influences across the region and beyond that may intensify during periods of ceasefire and post-conflict times. Therefore, a tailored approach is necessary not only for the countries directly involved in the conflict but also for potentially-threatened countries in the region and globally.

In the context of crisis, emergency and armed conflict, downscaling or abandoning fact-checking efforts and company collaboration with independent fact-checking organisations can significantly increase the risk of disinformation and hate speech, especially in crisis regions. Experiences from Ukraine have shown that systematic collaborations between companies and trusted partners have been important to protect the Ukrainian population from disinformation and hate speech and to support access to information. In fragile contexts, any discontinuation or weakening of such efforts would be cause for serious concern.

[4] UNESCO, 2023.
[5] Access Now, 2022.

Risk matrix based on experiences from the armed conflict in Ukraine

The risk matrix is shaped through the lens of the full-scale war against Ukraine, which defines the specificity of risk assessment. It is based on documented cases of information attacks, algorithmic biases, content manipulation, and the use of social platforms as instruments of hybrid warfare.

During different phases of crisis, approaches to risk prioritisation may differ. The risks identified in the matrix are based on experiences from Ukraine from February 2022 to January 2025 but particularly focus on the time between February and December 2022.

Ten key risks were identified by the local expert working group. Each risk is classified as either medium or high as low-level risks were not included in the table.

Key risk	Impact	Priority
1. Blocking and/or reducing the reach of war-related content that does not violate community standards.	High	High
2. Removal of content that documents war crimes.	High	Medium
3. The use of bots and fake accounts to spread disinformation and hate speech.	High	High
4. Presence of false, misleading, and malicious content (including AI generated).	High	High
5. Ineffective tools for searching reliable war-related information on platforms.	High	High
6. Moderation policies and practices lacking consideration of contextual linguistic, social, political, historical, and cultural understanding, including of gender, minorities and marginalised groups.	Medium	High
7. Users are recommended harmful content, including disinformation and hate speech.	High	High
8. Lack of ability to reach users in temporarily occupied territories of Ukraine.	High	High
9. Abuse of commercial tools for political and military purposes that violate community standards.	High	Medium
10. Company policies, standards and practices are not adapted to a crisis context.	Medium	Medium

Ten key recommendations based on experiences from the armed conflict in Ukraine

These 10 recommendations have been identified by local experts of actors operating in the Ukrainian context before and during the armed conflict.

1. **Support the sharing of war-related content in the public interest that follows community standards**
   Focus could be strengthened to ensure that war-related content that adheres to community standards is not removed or limited in reach.

2. **Preserve content documenting war crimes**
   Efforts could be enhanced and further communicated to protect and preserve content that documents war crimes to support accountability.

3. **Combat disinformation from bots and fake accounts**
   Efforts could be strengthened to prevent bots and fake accounts from spreading disinformation and hate speech.

4. **Remove false or harmful content**
   Focus could be furthered on removing misleading or harmful content, including AI-generated material and content targeting vulnerable groups like women, minorities, and marginalised communities.

5. **Ensure access to reliable public information during conflict**
   Efforts could be increased to make sure users can easily access vital public information during times of conflict.

6. **Consider local context in moderation decisions**
   Moderation policies could better take into account contextual linguistic, social, political, historical, and cultural specifics, including of gender, minorities and marginalised groups.

7. **Reduce the spread of harmful content through recommendations**
   Mechanisms could be strengthened to prevent harmful content such as disinformation and hate speech from being recommended to users.

8. **Enhance collaboration with local representatives in vulnerable regions**
   Collaborations could be furthered with local representatives to ensure that reliable information is accessible in vulnerable regions and temporarily occupied territories.

9. **Prevent abuse of commercial tools for harmful purposes**
   Efforts could be increased to ensure that commercial tools are not used for political or military purposes that violate community standards, mislead or are harmful in other ways.

10. **Adapt company policies to crisis situations**
    Policies and practices could be further adjusted to respond effectively during crises, considering the unique challenges they bring.

Crisis protocol

This crisis protocol offers recommendations for mitigating disinformation and hate speech by companies based on experiences from the armed conflict in Ukraine. The efficiency and impact of these measures should be regularly evaluated and reviewed, as needs and urgency can fluctuate during different phases of conflict. For instance, the size of moderation teams or the vulnerability of certain groups may vary.

It is crucial to understand that steps recommended for the initial phases are not confined to those phases alone but should be continuously reviewed and adapted as the crisis evolves.

Before emergency, crisis and armed conflict: preparation and prevention

1. **Develop and regularly review crisis protocols**
   - Companies should develop context-specific protocols that can be activated in the case of crisis and regularly assess the protocol in collaboration with local experts to ensure that the protocol evolves in line with emerging threats and is adapted to the local context, including being sensitive to risks and harms related to gender, minorities and marginalised groups. Protocols could include a risk matrix as exemplified above.
   - Companies should ensure that resources are set aside to ensure that internal staff have the contextual understanding and capacities to appropriately address risks and harms of disinformation and hate speech in relation to a given crisis. This includes all levels from moderators to policy-level decision-makers.

2. **Establish a crisis management team**
   - Companies should create a dedicated crisis management team consisting of internal staff and local stakeholders with expertise in crisis response and deep knowledge of local contexts, including linguistic, social, political, historical, legal, and cultural aspects of the region.
   - Companies should ensure units and mechanisms that specifically monitor, analyse and implement initiatives to counter gender-based disinformation and hate speech and protect women, minorities, marginalised and vulnerable groups, including efforts to protect representatives in the public eye.

3. **Continuously review policies, algorithms and moderation processes**
    - Companies should involve relevant experts, local stakeholders and representatives from affected communities and vulnerable groups in regular reviews to address the everchanging nature of emergency, crisis and armed conflict, including in the build-up, duration and aftermath of these situations.

4. **Collaborate with local experts and stakeholders**
    - Companies should initiate and strengthen collaboration with local organisations, for example, trusted partners, fact-checkers, local independent media organisations, civil society groups, and organisations specialising in emergency, crisis and armed conflict. Companies should do so in order to contextualise efforts to counter disinformation and hate speech, for example, through informing moderation policies, reviewing crisis protocols, monitoring and evaluation efforts, and contributing to developing effective responses to disinformation and hate speech.

5. **Develop verified user lists**
    - Companies should compile lists of verified local users, including fact-checkers and independent media and other stakeholders who have expertise in combating disinformation and hate speech and providing reliable information. They should ensure that these experts have fast-track appeal processes. These lists should be produced in collaboration with local stakeholders.

6. **Collaborate with fact-checkers and researchers**
    - Companies should partner and/or strengthen partnerships with both local, regional and global fact-checkers and researchers to create effective systems and channels for information-sharing, knowledge-sharing and rapid exchange of data regarding developments within disinformation and hate speech.

7. **Strengthen transparency and accountability efforts**
    - Companies should ensure transparency about how and why specific content is removed or flagged. They should provide regular reports showing the volume of removed posts and the rationale behind those actions.
    - Companies should increase meaningful transparency regarding advertising tools and ad funding to ensure that commercial tools do not become a loophole for disinformation and hate speech.
    - Companies should disclose how they are using AI tools to monitor content and the limitations of such tools, particularly in emotionally charged, complex situations like armed conflict.
    - Companies should keep users informed about the evolving content moderation strategies and updates tailored to crisis situations.

8. **Initiate platform-to-platform cooperation**
    - Companies should, in collaboration with local stakeholders and independent researchers, develop efficient communication lines with other companies to prevent disinformation and hate speech from spreading from one social media platform to another.

9. **Initiate efforts to preserve content documenting war crimes**
    - Companies should identify and develop crisis protocols with relevant archival stakeholders to prepare for and initiate efforts to ensure retention of content documenting war crimes.

10. **Adapt private policies and data handling practices**
    - Companies should address vulnerabilities that could expose users to surveillance, targeting, or manipulation, while providing users with tools to protect their data and privacy. Special attention should be given to prisoners of war and their networks.

During emergency, crisis and armed conflict: real-time response and protection

1. **Real-time monitoring and threat detection**
    - Companies should monitor real-time developments and trends on their platforms related to conflict-related disinformation and hate speech, including gendered disinformation and surges in inauthentic activities. These efforts should involve both automated systems and human moderators and be informed by fact-checkers, researchers and local stakeholders.

2. **Collaborate with local experts for real-time updates**
    - Companies should engage local experts, media, civil society groups, and stakeholders particularly in active conflict zones and temporarily occupied territories to ensure that content moderation decisions are safe, contextually appropriate, and protect groups vulnerable to and/or are likely targets of disinformation and hate speech campaigns.

3. **Collaborate with trusted partners for rapid fact-checking**
    - Companies should strengthen collaboration with trusted fact-checking organisations, local media, and other information integrity stakeholders to ensure accurate and swift verification of information related to the crisis.

4. **Strengthen transparency and accountability**
    - Companies should publicly report on their actions, including their moderation decisions and policy changes, to local stakeholders and platform users to gain trust and support them in understanding how to use the platforms.

5. **Promote accurate and safe information locally related to crisis, emergency and armed conflict**
    - Companies should promote humanitarian and security-related information that is essential to the safety and well-being of the local population.

6. **Provide support for independent fact-checkers**
    - Companies should continue providing technical and financial support for independent fact-checkers to ensure their impartiality and objectivity during the conflict. Open methodologies and transparent reporting should be emphasised to build user trust.

After emergency, crisis and armed conflict: recovery, evaluation, and long-term monitoring

1. **Evaluate and review crisis response**
   - After the crisis, companies should conduct post-crisis evaluations to assess the effectiveness of crisis response, including how well disinformation and hate speech were handled. This review should involve local, regional and global experts; media; fact-checkers; and other relevant stakeholders to gather learnings for future crises.

2. **Reinforce collaboration for long-term stability**
   - Companies should continue collaborate with local stakeholders, civil society groups, fact-checkers and independent researchers to monitor the ongoing recovery process and prevent the resurgence of disinformation and hate speech.

3. **Monitor post-conflict disinformation and hate speech**
   - Companies should continue monitoring and responding to the spread of disinformation and hate speech for no less than two years after the immediate crisis has ended to promote stability and support peace processes.

4. **Ensure transparency and accountability**
   - Companies should publicly report on their actions, including their content moderation decisions, disinformation removals, and how local expertise informed those decisions. This transparency helps rebuild trust and shows accountability for the platform's role during the crisis.

5. **Continuous improvement of crisis protocols**
   - Based on the evaluation of the conflict and its aftermath, companies should refine their crisis management protocols. Regular involvement of local experts and stakeholders is crucial to improving policies and practices for future crisis scenarios.

Recommended Ukrainian organisations who could be approached by companies for further collaboration:

This list is not extensive but could be a helpful starting point for further collaboration and to establish relationships with relevant local stakeholders.

CEDEM (Centre for Democracy and Rule of Law)
Areas of cooperation: Analytics and expertise on freedom of speech issues, media legislation reform, advocacy campaigns for transparent regulation of online space.

Institute of Mass Information (IMI)
Areas of cooperation: Monitoring and analysis of violations of journalists' rights, fact-checking disinformation, training for media representatives on security issues and professional standards.

Internews-Ukraine
Areas of cooperation: Training specialists in media literacy and digital security, conducting research on the information space, promoting quality journalism standards and combating disinformation.

Digital Security Lab
Areas of cooperation: Consultations on data protection and privacy, responding to cyber threats and coordinating efforts in crisis situations.

Ministry of Digital Transformation of Ukraine
Areas of cooperation: Initiating regulatory and legislative changes in the field of digital technologies, providing consultations on state priorities in the field of information security, collaborating with large technology companies to form strategic approaches to data protection and countering disinformation.

Mnemonic/Ukrainian Archive
Areas of cooperation: Professional collection, archiving, and cataloguing of data related to war crimes and human rights violations; consultations on long-term storage, processing, and verification of digital materials' authenticity; support for international investigations by providing access to digitised evidence.

StopFake
Areas of cooperation: Fact-checking and debunking false information about events in Ukraine, analysing Kremlin propaganda, and conducting educational activities to enhance media literacy and critical thinking in society.

VoxUkraine

Areas of cooperation: Fact-checking statements made by politicians, businesspeople, bloggers, and other public figures; analysing and debunking disinformation in public discourse; providing analytical materials on strategies to combat misinformation.

Women in Media

Areas of cooperation: Analysis of gender balance in the field of media, collaboration with women in journalism and media management, expertise in gender-sensitive policies in the media industry, research into gender stereotypes in news coverage.

Literature list

Access Now. (2022, November 29). *Declaration of principles for content and platform governance in times of crisis.* https://www.accessnow.org/wp-content/uploads/2022/11/Declaration-of-principles-for-content-and-platform-governance-in-times-of-crisis.pdf

Alaphilippe, A., Machado, G., Miguel, R., & Poldi, F. (2022, September 27). *Doppelganger: Media clones serving Russian propaganda.* EU Disinfo Lab. https://www.disinfo.eu/wp-content/uploads/2022/09/Doppelganger-1.pdf

Aleksejeva N., Osadchuk, R., Gelava, S., Le Roux, J., Caniglia, M., Suárez Pérez, D., & Kann, A. (2022, September 27). *Russia-based Facebook operation targeted Europe with anti-Ukraine messaging.* Digital Forensic Research Lab (DFR Lab). https://medium.com/dfrlab/russia-based-facebook-operation-targeted-europe-with-anti-ukraine-messaging-389e32324d4b

Centre for Democracy and Rule of Law (CEDEM). (2024, January 25) *Recommendations for improving the moderation of Ukrainian content about Russia's armed aggression on Meta platforms.* https://cedem.org.ua/library/rekomendatsiyi-sotsmerezhi/

Centre for Strategic Communications and Information Security, & Centre for Democracy and Rule of Law (CEDEM). (2024, April 24). *Informational attacks in social networks: Research on Russian disinformation influence through advertising on Facebook.* https://spravdi.gov.ua/en/yak-rosiya-atakuye-ukrayinu-dezinformacziyeyu-cherez-reklamu-v-facebook-doslidzhennya-czentru-strategichnyh-komunikaczij/

Counter Disinformation Network. (2024, September 3). *Fool Me Once: Russian influence Operation Doppelganger continues on X and Facebook.* https://alliance4europe.eu/wp-content/uploads/2024/09/CDN-Report-%E2%80%93-Fool-Me-Once_-Russian-Influence-Operation-Doppelganger-Continues-on-X-and-Facebook-%E2%80%93-September-2024.pdf

CyberPeace Institute. (2024, October). *Report of Second Expert Meeting on Harms Methodology.* https://cyberpeaceinstitute.org/wp-content/uploads/2024/10/Second-Expert-Meeting-Harms-Methodology-2024.docx.pdf

CyberPeace Institute. (2023). *Cyber dimensions of the armed conflict in Ukraine - Quarterly analysis report Q3 July to September 2023.* https://cyberpeaceinstitute.org/wp-content/uploads/2023/12/Cyber-Dimensions_Ukraine-Q3-2023.pdf

European Commission. (2018). *A multi-dimensional approach to disinformation: Report of the independent High level Group on fake news and online disinformation.* https://digital-strategy.ec.europa.eu/en/library/final-report-high-level-expert-group-fake-news-and-online-disinformation

EEAS. (2024, January). *2nd EEAS Report on Foreign Information Manipulation and Interference Threats: A Framework for Networked defence.* https://www.eeas.europa.eu/eeas/2nd-eeas-report-foreign-information-manipulation-and-interference-threats_en

EEAS. (2023, February). *1st EEAS Report on Foreign Information Manipulation and Interference Threats: Towards a Framework of Networked defence.* https://www.eeas.europa.eu/eeas/1st-eeas-report-foreign-information-manipulation-and-interference-threats_en

Global Network Initiative. (2017, May). *GNI Principles on Freedom of Expression and Privacy.* https://globalnetworkinitiative.org/wp-content/uploads/2018/04/GNI-Principles-on-Freedom-of-Expression-and-Privacy.pdf

Grippo, V. (2024, December 4). *Regulating content moderation on social media to safeguard freedom of expression.* Committee on Culture, Science, Education and Media, Council of Europe. https://rm.coe.int/as-cult-regulating-content-moderation-on-social-media-to-safeguard-fre/1680b2b162

Hearing Before the United States House of Representatives Committee on Energy and Commerce Subcommittees on Consumer Protection & Commerce and Communications & Technology (2021, March 21). *Testimony of Mark Zuckerberg, Facebook, Inc.* https://docs.house.gov/meetings/IF/IF16/20210325/111407/HHRG-117-IF16-Wstate-ZuckerbergM-20210325-U1.pdf

Kyrychenko, Y., Brik, T., van der Linden, S., & Roozenbeek, J. Social identity correlates of social media engagement before and after the 2022 Russian invasion of Ukraine. *Nature Communications 15, 8*127 (2024). https://doi.org/10.1038/s41467-024-52179-8

Mantas, H. (2021, May 13). *Sen. Mark Warner says he is embarrassed by congressional inaction on tech regulation.* Poynter.org. https://www.poynter.org/fact-checking/2021/sen-mark-warner-embarrassed-by-congressional-inaction-on-tech-regulation/

Porter, E., & Wood, T. J. (2021, September 10). *The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom.* PNAS. https://doi.org/10.1073/pnas.2104235118

Schafer, B., Benzoni, P., Koronska, K., Rogers, R., & Reyes, K. (2024, May 30). *Russian propaganda as a nesting doll: How RT layered the digital information environment.* German Marshall Fund. https://www.gmfus.org/sites/default/files/2024-05/Laundromat%20Paper.pdf

Semenyuta, I. (2023, June 26). *Moderation during the war: why social networks delete posts of Ukrainians.* Detector Media. https://ms.detector.media/sotsmerezhi/post/32269/2023-06-26-moderatsiya-pid-chas-viyny-za-shcho-sotsmerezhi-vydalyayut-dopysy-ukraintsiv/

Snopok, O. (2022, November 23). *"Potentially unacceptable." How the Russian war in Ukraine affects content moderation on social networks.* Detector Media. https://ms.detector.media/it-kompanii/post/30718/2022-11-23-potentsiyno-nepryynyatnyy-yak-rosiyska-viyna-v-ukraini-vplyvaie-na-moderatsiyu-kontentu-v-sotsmerezhakh/

Ukrainian Media and Communication Institute. (2023). *Media literacy for senior people (60+).* https://www.jta.com.ua/wp-content/uploads/2023/11/UMCI_MediaLiteracy_60_UA.pdf

UNESCO. (2023). *Guidelines for the governance of digital platforms.* https://unesdoc.unesco.org/ark:/48223/pf0000387339

van Erkel, P. F. A., van Aelst, P., de Vreese, C. H., Hopmann, D. N., Matthes, J., Stanyer, J., & Corbu, N. (2024). When are fact-checks effective? An experimental study on the inclusion of the misinformation source and the source of fact-checks in 16 European countries. *Mass Communication and Society, 27(5), 851–876.* https://doi.org/10.1080/15205436.2024.2321542

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2019). Fact-Checking: A Meta-Analysis of What Works and for Whom. *Political Communication, 37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

The project is implemented by IMS (International Media Support) and NGO Internews Ukraine in partnership with UNESCO and with support from Japan. The project builds on UNESCO's Guidelines for the Governance of Digital Platforms from 2023.