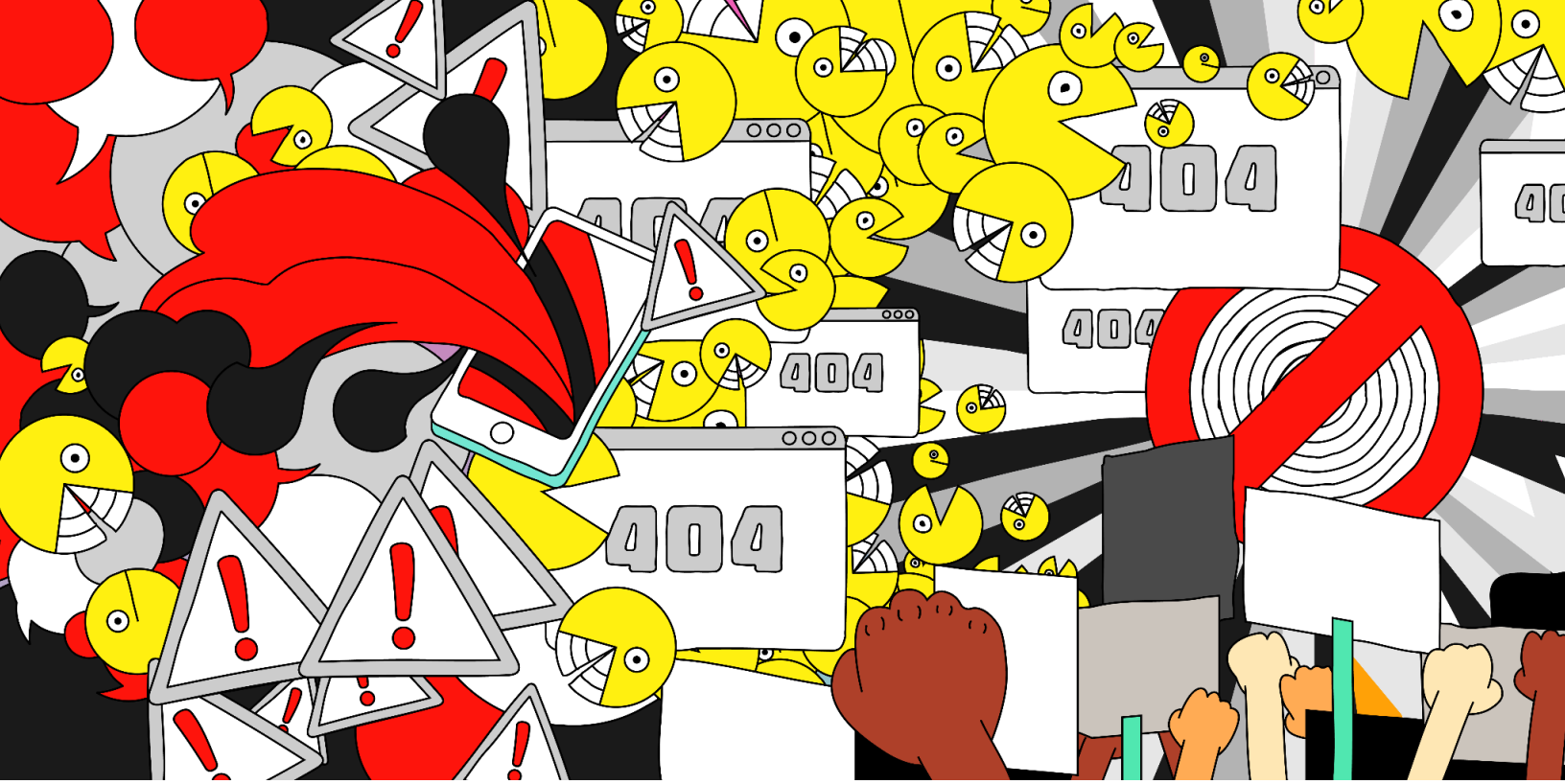

Declaration of principles for content and platform governance in times of crisis

For over a decade, large online platforms have played a significant role during armed conflicts and other crises. People use social media to report and document human rights abuses and war crimes, access information, mobilize for national and global action, and crowdsource humanitarian assistance and relief. State and non-state actors use these same platforms to spread disinformation and hate speech, incite violence, and attack or surveil activists, journalists, and dissidents.

Under the UN Guiding Principles on Business and Human Rights ("UN Guiding Principles"), social media companies have a responsibility to respect human rights, to prevent risks stemming from their systems, and to remedy abuses wherever they operate. These Principles specifically highlight the heightened risk of gross human rights abuses in conflict affected areas, and call on companies to not only respect international human rights treaties, but to also "respect the standards of international humanitarian law" when operating in such situations.

To mitigate the risk of contributing to gross human rights abuses or being complicit in exacerbating existing tensions or conflicts, companies must take extra care and conduct heightened human rights due diligence, as well as treating the risk as a legal compliance issue "arising from extraterritorial civil claims, and from the incorporation of the provisions of the Rome Statute of the International Criminal Court in jurisdictions that provide for corporate criminal responsibility."

However, social media companies have responded inadequately and inconsistently in situations of armed conflict, fragile governance, and crises – such as those in Ethiopia, Syria, Israel/Palestine, and Myanmar, among others. Companies have often failed to respect human rights or to mitigate their activities' adverse human rights impacts, and have been slow or ineffective in removing or restricting hate speech, disinformation, and incitement to violence in real time. Rather, their responses (or lack thereof) have disproportionately impacted marginalized communities and historically oppressed groups, and have facilitated serious human rights abuses.

Civil society organizations have long documented platforms' unequal, nontransparent, and inconsistent approach to platform and content governance, and have repeatedly called on social media companies to properly invest in moderation and curation in chronically deprioritized non-English speaking countries beyond the U.S. and Western Europe. The discrepancy in companies' responses was further highlighted during the recent Russian invasion of Ukraine, which showed how swiftly platforms can roll out measures when they have the interest or will to do so.

The following **Declaration of principles for content and platform governance in times of crisis** ("Declaration") was born in response to the ad hoc and inconsistent approach to handling crises. To define the Declaration's scope, we referred to the E.U.'s definition of conflict affected and high risk areas: "areas in a state of armed conflict, fragile post-conflict areas, as well as areas witnessing weak or non-existing governance and security, such as failed states, and widespread and systematic violations of international law, including human rights abuses."

Building on the Santa Clara Principles on Transparency and Accountability in Content Moderation and the continuous efforts of civil society organizations and content moderation experts, **this Declaration seeks to advance consistent and rights-respecting principles for platforms to adhere to in times of crises.**

Principles and protocols for engagement *before* a crisis

1. **Conduct human rights due diligence (HRDD) to address the lifecycle of crises, situations of conflict, and human vulnerabilities:**

- **Conduct regular ex ante human rights impact assessments (HRIAs), as outlined in the UN Guiding Principles for Business and Human Rights, and take all necessary steps to address and mitigate any identified, adverse human rights impacts.** Social media companies should identify any actual or potential adverse impact on human rights in conflict affected and high risk areas, or at significant moments, such as the lead up to, during, and immediately after elections. The assessment should, at a minimum, specifically consider:

    a) Foreseeable and negative impact on the enjoyment of human rights as defined in international human rights law and treaties;
    b) Foreseeable and negative impact on civic discourse and electoral processes;
    c) Foreseeable and negative impact on in-country employees and third party contractors, including content reviewers.

In situations of armed conflict, the impact assessment should also carefully examine whether any platform's operations, product design, policies, or services contribute to gross human rights violations, exacerbate conflict, lead to evidence of human rights breaches being destroyed, or fail to respect other international humanitarian law standards.

Social media companies should publicly release complete findings of human rights impact assessments containing all relevant metrics listed above, as well as the steps taken in response. If they contract external third parties to conduct independent reviews, HRIAs, or other forms of due diligence, they should disclose the terms and conditions of their involvement and cooperation.

- **Identify and consistently monitor conflict affected and high risk areas.** Social media companies should develop a crisis matrix, updated on at least a quarterly basis, to flag areas for heightened due diligence to be conducted in partnership with local or expert teams and relevant independent stakeholders, including civil society organizations. They should rely on proven and existing classification of fragile and conflict affected situations, such as the OECD States of Fragility or the World Bank Classification of Fragile and Conflict-Affected Situations, and refer to ongoing statements, resolutions, and/or reporting by UN bodies and the International Committee of the Red Cross. When identifying these areas, companies should consider not only the intensity of hostilities or conflict on the ground, but also internal statistics related to content moderation in these areas, such as the number of content removal requests on conflict-related issues, which might indicate rising tensions or an escalation of violence. The crisis matrix should be publicly available on a dedicated and centralized interface, and should map existing integrity and safety measures at different levels of each crisis or conflict.

- **Build teams with strong local and regional expertise and language skills.** Social media companies should take a careful and conflict sensitive approach to the hiring and appointment of team members working or engaging in conflict affected and high risk areas (in-country or overseas), especially where current or former affiliation with political parties, government or military entities, may influence company decisions and actions. Teams should be ethically and fairly contracted across all markets where the companies' platforms are used, ideally in proportion to the number of end users and each region's specific conflict and human rights risks. They should be adequately staffed, to ensure local and regional teams can respond effectively to emerging conflicts or risks, even where such risks were not previously identified.

- **Conduct human rights and conflict sensitive risk assessments specifically tailored to potential crises' national specificities or the impacted area's context.** Social media platforms should identify, analyze, and assess all risks stemming from their products, services, operations, and design, including algorithmic content moderation and content curation tools, which they deploy in conflict affected and high risk areas, and that may contribute to an escalating crisis. Such risk assessments should be conducted periodically before any crisis emerges, to assess the company's integrity, as well as their products, systems, and functions' readiness and effectiveness. At a minimum, the risk assessment should consider any foreseeable and negative consequences of the following:

    a) Design of the platform's recommender systems and ad delivery systems;
    b) Geographical application of restrictions, including regional and local limitations on dangerous content;
    c) Content moderation systems, both human and automated;
    d) Application and enforcement of terms and conditions in conflict affected and high risk areas;
    e) Privacy and data protection policies and enforcement;
    f) Existing company resources, including staffing capacity;

g) Resources, strategies, and actions to mitigate and manage identified risks, and to remedy actual harm.

- **Subject their crisis response mechanisms to yearly independent audits**. Through annual independent audits, social media platforms should regularly assess their crisis response mechanisms' effectiveness, identify gaps in policy response, enforcement, and resources, and ensure the proper implementation of lessons learned or recommendations. Social media companies should make audit reports publicly available, only removing information whose disclosure may negatively impact stakeholders or personnel, in line with UNGP 21(c). The release and communication of such reports should also be done in a conflict sensitive manner.

2. **Create channels for meaningful and direct engagement with relevant independent stakeholders, including civil society organizations operating in conflict affected and high risk areas.** Social media companies should initiate, create, and maintain direct and meaningful engagement with trusted partners at grassroots and regional levels. The cooperation should include a communication platform or forum that enables the regular exchange of information. Notably, external experts, relevant stakeholders, and representatives from affected communities should be able to inform, shape, and review platforms' new and existing content moderation and content curation policies, such as:

- **Developing strong and continuous cooperation with trusted partners, independent media organizations, individuals, and flaggers,** especially if activities are likely to escalate violence and exacerbate tensions. Cooperation and communication with trusted partners should be ongoing and not limited to moments of escalating tension or crisis. Companies must have readily available resources to mobilize effectively, in response to guidance provided by civil society organizations and other independent stakeholders.

- **Allocating sufficient financial and human resources to content moderation efforts.** Social media companies should ensure that they recruit a sufficient number of content reviewers with the language skills relevant to conflict affected and high risk areas. Local language content reviewers should demonstrate sufficient understanding of each area or country's political, social, historical, and cultural context. The ratio of platform moderators must be proportionate to the number of country users and to the human rights risks in a specific region. This context sensitive approach should guide the hiring process as well as accounting for any potential bias or perception of bias.

- **Establishing early warning systems and clear escalation systems for emergency situations to help detect imminent harm to individuals' physical safety.** Social media platforms should develop early warning systems and escalation systems by cooperating with all relevant grassroots stakeholders, including civil society organizations and human rights experts. Platforms should ensure these systems enable trusted national partners to evaluate their performance regularly.

- **Coordinating global, regional, and local offices and staff efforts, to allow timely and coherent decision making, led by human rights officers well-versed in the dynamic context.** There are clear gaps in information sharing and decision making between engineers, human rights teams, and public policy teams who engage with trusted partners and civil society at national and regional levels. The integration of human rights experts across key decision making processes should be a priority, with crossfunctional reach to product, engineering, marketing, emergencies, elections, and other teams as necessary.

3. **Develop crisis protocols across all levels and likelihood of risks, designed to prevent and mitigate potential harms.** Social media companies should work closely with local and regional teams, as well as independent stakeholders, to raise risk awareness and co-create conflict sensitive prevention and response measures.

- **Develop and adopt user centric, conflict sensitive measures that adequately mitigate all identified and foreseeable risks of future or ongoing crises. These should specifically focus on protecting the human rights of groups and individuals.** Social media platforms should design, develop, and deploy measures that can mitigate future and current risks stemming from their operations and systems in high risk areas and conflict affected countries. At minimum, these should include:

    a) Modifications (carve outs) of terms and conditions and their consequent enforcement;
    b) Design features of their service (interface);
    c) Adapting content moderation processes;
    d) Decision making processes and dedicated resources for content moderation; and
    e) Modifying content recommender systems and advertising systems (especially ad delivery systems responsible for how advertisements are presented).

- **Companies should develop and test crisis protocols before a crisis breaks out, focusing specifically on the risks for groups, individuals, and their rights.** These protocols should be informed by previous human rights impact and risk assessments. Crisis protocols and their application should be tested prior to any crisis occurring, in diverse environments, and taking into consideration the impact on end users.

- **Social media platforms should display and regularly update information about the crisis situation** provided by relevant international bodies (see above).

- **Social media platforms should appoint a dedicated crisis management team for each identified conflict affected and high risk area.** This could also be an electronic point of contact. This team should include content governance and human rights experts, and should include individual team members who thoroughly understand the national context and have the necessary language skills.

Core principles applicable to all scenarios, with an emphasis on protocols for ongoing crises

1. **Conduct rapid and conflict sensitive human rights due diligence (HRDD).** In times of crisis, social media platforms must conduct rapid HRDD to identify and mitigate any actual or foreseeable negative impact on human rights. The outbreak of armed conflict triggers the application of international humanitarian law, which introduces additional rules and protections for social media platforms to comply with when developing and deploying their products and services, content moderation and curation systems, designs, and actions. This HRDD should be ongoing and should evolve as the situation does, taking into consideration any invocation of martial and emergency laws by national governments, often used in contexts of crisis and armed conflict to censor opposition or political speech.

2. **Activate meaningful, direct, and concurrent engagement with local and regional civil society organizations and experts.** When a crisis or armed conflict breaks out, social media platforms must immediately engage with local and regional human rights experts, civil society organizations, and other relevant stakeholders to advise on and monitor each platform's response and the impact of crisis measures on affected communities. Platforms must immediately, regularly, and collectively update stakeholders about the ongoing situation and the company's consequent measures and actions. Platforms should take all measures needed to ensure the safety of people they engage with in high risk contexts.

3. **Take an equitable, fair, and consistent approach to engaging in situations of armed conflict and crises.** Social media companies must ensure they address the risks and adverse impacts of their platforms' usage during armed conflict and crises in a systematized and equitable manner across the globe and in accordance with international human rights standards. They should prioritize crisis response and allocate resources based on salience, scale, and scope of human rights threats and violations, and not on market value or profit share. They must equitably invest in and prioritize non-English speaking countries and areas, by hiring staff and content reviewers with the cultural and linguistic knowledge to effectively enforce companies' policies in all operating markets, and by creating a standardized crisis response protocol to be enforced as required.

4. **Provide full transparency on content moderation policy design and enforcement, both human and automated:**

    - **Make any content moderation policy carve outs or extraordinary measures public, clear, specific, predictable, and time limited.** Companies' content moderation policies or standards are designed for business-as-usual operating environments, which may make them harmful or counterproductive when enforced in times of crises, war, or armed conflict. Where platforms want to create exceptions to existing standards or apply policy carve outs, these must be proactively and publicly announced, published on platforms' websites in the languages spoken by the affected communities, and with sufficient detail, guidelines, and clear examples for users to understand how they can and should conduct themselves. Any decision making on carve outs should be done in direct consultation with affected communities, local civil society organizations, and human rights experts. Social media companies must also rigorously assess the impact of these extraordinary carve outs and measures, through ex ante and ex post HRDD, to ensure that they do not exacerbate existing tensions, contribute to violence, or violate international human rights or humanitarian law.

    - **Take a context dependent approach to geo-moderation of content in conflict affected and high risk areas.** Social media platforms should only opt to geo-block or otherwise withhold a specific piece of content if, upon assessment of the content's legality, it falls under the threshold determined by Articles 19 and 20 of the ICCPR or if it amounts to a violation of applicable international humanitarian law. Besides the transparency criteria of content moderation policies and tools listed in point five of this Declaration, social media companies should rely on international human rights instruments when conducting any assessment of content's legality, such as the Rabat Plan of Action. Blocking or withholding content within specific countries or areas may not be the most effective approach to content moderation in times or locations of crisis, and should be considered as a last resort measure. Hence, this approach should be tailored not only to the region, but also to the gravity of the content, its source, and its potential impact.

- **Disclose government requests made to social media platforms and their responses, including through voluntary reporting channels, to the extent that it does not expose employees to serious risk of personal harm and to the extent allowed by legal frameworks.** Platforms should be proactively transparent about the number of government requests received, the types of content requested for removal, and the platform's action or response to the requests. Such transparency reports must comprehensively capture the numbers and types of content removal requests received from state and non-state parties to an armed conflict, including those that come via voluntary reporting mechanisms such as internet referral units. When faced with government demands that may lead to a violation of international legal frameworks, social media companies should comply with international humanitarian law and international human rights law to the greatest extent possible.

- **Disclose whether any request issued by public authorities has led to tweaks or changes in automated decision making** developed to moderate or curate content related to the conflict or crisis, including criteria for optimizing content recommender systems, and outline the practical impact of such changes on users.

- **Evaluate the operations of automated decision making systems to mitigate and address the risks and harms of overenforcement (false positives) or underenforcement (false negatives).** Automated content moderation tools are context-blind, imply significant human rights risks, and carry a risk of arbitrary and discriminatory censorship that disproportionately impacts marginalized or historically oppressed communities. Social media platforms must therefore evaluate the automated decision making tools they have deployed, and make necessary updates or impose restrictions to mitigate such risks. Platforms should restrict machine learning and deep learning so that, in times of conflict, such systems cannot update themselves without human review.

- **When deploying automated content moderation and curation tools for non-English languages, ensure that a human always reviews the outputs.** There is severe underinvestment in sourcing sufficient training data sets to build, train, and test classifiers with a high level of accuracy in a real world setting. In conflict scenarios, platforms must prioritize human review of user-generated content.

- **Disclose any implemented automated models designed for blanket content de-amplification, "shadow banning," and deranking content.** Any disclosure should be done in a way that is understandable and accessible for all users.

- **Provide transparency on the criteria that social media platforms use to define, detect, review, and remove so-called terrorist and violent extremist content (TVEC),** including content added to the hash-sharing database supported by the Global Internet Forum to Counter Terrorism (GIFTC). In this vein, social media companies should be transparent about:

    a) The criteria used to determine which organizations and groups are considered as "terrorist";
    b) How they define "terrorist" or "extremist" content;
    c) How content added to the database is vetted and verified;
    d) Which extremist or terrorist groups the database includes;
    e) How effective the database is;
    f) How much content or how many accounts have been erroneously removed as a result.

5. **Preserve content removed by the platform for three years and create a secure mechanism granting international accountability mechanisms other than national law enforcement access to this archived material,** including the International Criminal Court (ICC), the International Court of Justice (ICJ), and UN-mandated investigative bodies and commissions:

- To ensure accountability and allow the judiciary sufficient time to review preserved and archived content, ensure it is located and stored outside of high risk countries and conflict affected areas, in accordance with international standards on privacy and data protection.

- Balance protecting individuals using social media services from exposure to graphic or violent content while providing opportunities for eyewitnesses to document human rights violations and atrocities.

6. **Create transparent, clear, rights-respecting, and accessible notice and review mechanisms, and provide access to effective remedy:**

- **Notify users when a moderation decision is made about their content or speech.** At a minimum, the notification should contain adequate information about what sparked the decision, the specific rule that was broken, how content moderation guidelines were interpreted, the action that will be taken, and clear instructions for submitting an appeal.

- **Provide a clear, transparent, predictable, and accessible appeal mechanism for users to request a review of any content moderation decision.** Review or appeal mechanisms must be directly and easily accessible, and be addressed within a reasonable time frame. They may be provided by the company or through recourse to an external entity, such as an oversight board.

- **Notify users when they are subjected to automated processes**. Social media platforms should inform users when automated systems are used to moderate their content, explain how such mechanisms operate, and provide a clear and accessible appeal mechanism for users to request a human review.

- **Provide effective remediation to users affected by a platform's policies, products, or practices.** This includes content moderation decisions, particularly in cases that harm users, such as erroneous or excessive application of the rules. Social media companies should ensure transparent and easy access to remedy mechanisms, and provide sufficient information on timelines for response, as well as tiers of users prioritized.

7. **Address human rights risks related to each platform's business model:**

- **Social media companies should ensure that surveillance based advertisement, i.e. digital advertising targeted at individual segments, usually through tracking and profiling based on personal data, does not contribute to ongoing or future human rights violations.** The use and abuse of individuals' personal data for targeting and monetizing content online leads to manipulation of absolute freedom of thought, discrimination, disinformation campaigns, and security risks. In times of crisis, social media companies should explore alternative forms of digital advertising, such as contextual advertising, which make the supply chain more transparent.

- **Social media companies should ensure that their monetization programs do not channel incomes to actors associated with sanctioned entities, or to foreign and local actors systematically producing and/or distributing disinformation content.** In times of crisis, social media companies should apply heightened due diligence about who they onboard onto their programs and conduct additional reviews of existing monetization partner lists. Social media companies should refrain from applying bulk restrictions and be mindful to not unduly restrict revenue generating opportunities for legitimate content creators and independent media.

8. **Enable account-level safety features for high risk users.** Social media platforms should immediately implement safety measures for users in high risk areas and conflict affected countries, including but not limited to: enabling locking of private profiles for external actors; enabling end-to-end encryption in chat and messaging functions; enabling disappearing messages; rolling out notifications for messages in encrypted chats that are screenshotted; limiting the search function for follower lists; and providing digital safety and security tips in local languages. Platforms should also establish adequate mechanisms, protocols, and partnerships to support the prompt deactivation of accounts belonging to individuals facing arrest, in order to protect these individuals and their networks.

Engagement principles and protocols *after* a crisis

1. **Implement a transition phase before winding down companies' operations and notify users of any change in platform functionalities, based on continuous assessment of the conflict's intensity and life cycle.** Social media platforms should phase out their crisis measures gradually, on the basis of HRDD findings, and in correlation with hostilities' intensity levels, as classified in their crisis matrix ranking. Platforms must publicly notify their users of any change or cessation in their content moderation carve outs, policy exemptions, or user safety functions.

2. **Continue to conduct HRDD to identify, mitigate, and address negative human rights impacts throughout the lifecycle of conflicts and crises.** In situations of armed conflict or crises, grievances, violence, and tensions often escalate and decrease in a cyclical manner. Platforms must therefore continue monitoring conflict affected and high risk areas, and update their crisis matrix accordingly. Furthermore, platforms should continue to assess the impact of their products, services, operations, and design on human rights in those areas, even after escalations or violence have decreased or ceased. Platforms should also:

    - **Conduct an audit to review whether their crisis protocols and procedures were adequately followed and implemented.** Social media platforms should conduct a comprehensive human rights audit to fully assess the effectiveness of their crisis protocols measures and their impact on user safety and human rights, based on the initial rapid HRIA or HRDD processes conducted prior to or during the conflict or crisis. As part of this review process, platforms should solicit and collect feedback from local stakeholders, including civil society groups, and human rights groups.

    - **Conduct a public, full, and independent human rights impact assessment.** Social media platforms should commission an independent and conflict sensitive HRIA, particularly when their content moderation actions and crisis measures have severely impacted the human rights of

individuals and communities, exacerbated tensions and conflicts, resulted in or contributed to loss of life and physical harms, or raised a collective grievance among affected individuals and communities. In doing so, platforms should consult with relevant stakeholders including local civil society groups, human rights experts, and activists.

- **Ensure that findings of these audits and assessments result in clearly defined, transparent, measurable, time bound, and public commitments to policy or product change and adjustments.** Social media platforms should publicly commit to ensuring that their content moderation systems and product designs do not adversely affect user safety or human rights in the future. Such learnings can be cross-regional and can feed into the lifecycle of platforms' crisis response toolkits and protocols.

3. **Cooperate with national and international judicial and accountability mechanisms and allow access to preserved and archived evidence.** Social media companies should prioritize processing requests from national and internationally mandated investigative bodies, the ICC, or the ICJ, granting them access to preserved documentation of human rights violations and serious international crimes.

4. **Grant API and data set access to vetted civil society organizations, journalists, and academic researchers.** Social media platforms should allow independent stakeholders conducting research in the public interest to access all information needed to find and archive human rights documentation, and to audit and assess the adequacy and effectiveness of platforms' responses and their impact on conflict or crises.

5. **Conduct quarterly briefings with local and global civil society organizations.** At a minimum, social media platforms should, on a quarterly basis, release public updates and brief relevant stakeholders, including civil society organizations, on the implementation of crisis measures, their effectiveness, and their impact (or lack thereof). This should also be an opportunity to discuss lessons learned and to outline recommendations for the future.

This Declaration has been drafted by Access Now with the contribution and endorsement of our partners:

ARTICLE 19
Center for Democracy and Technology
Centre for Democracy and Rule of Law (CEDEM)
Digital Security Lab Ukraine
JustPeace Labs
Mnemonic
Myanmar Internet Project

Access Now (accessnow.org) **defends and extends the digital rights of people and communities at risk**. As a grassroots-to-global organization, we partner with local actors to bring a human rights agenda to the use, development, and governance of digital technologies, and to intervene where technologies adversely impact our human rights. By combining direct technical support, strategic advocacy, grassroots grantmaking, and convenings such as RightsCon, we fight for human rights in the digital age.

For more information, please contact:
Marwa Fatafta | MENA Policy Manager | marwa@accessnow.org
Eliška Pírková | Europe Policy Analyst and Global Freedom of Expression Lead | eliska@accessnow.org