July 29, 2022

Office of the High Commissioner for Human Rights
Palais Wilson
52 rue des Pâquis
CH-1201 Geneva
Switzerland

To whom it may concern,

Please find enclosed Meta's response to the UN Special Rapporteur on freedom of expression's call for submissions for her upcoming report on disinformation in times of conflict. We remain at your disposal if you have any questions and look forward to engaging with you and your office on this important issue.

Yours faithfully,

Gabrielle Guillemin

Meta

Gabrielle Guillemin
Human Rights Policy Manager
Facebook UK Ltd
10 Brock Street,
London, NW1 3FG

Meta submission to the UN Special Rapporteur on freedom of expression on challenges and approaches to misinformation, disinformation and propaganda in times of conflict.

Introduction

Meta welcomes the opportunity to contribute to the UN Special Rapporteur on freedom of expression's call for inputs to her report on addressing misinformation, disinformation and propaganda in times of conflict. Her report could not be more timely. Over the last few years, Meta has developed a number of approaches to deal with the challenges raised by disinformation and misinformation, including in countries where conflict takes place. At the same time, the war in Ukraine has thrown into sharp relief the need to address propaganda for war or the meaning of the prohibition on incitement to violence when lives are being destroyed by enemy forces. We hope that sharing what we've learnt in our submission will be useful to the UN Special Rapporteur's work.

The UN Special Rapporteur is well aware that freedom of expression is one of the essential tenets of democratic societies. In times of war, the free flow of reliable information could not be more vital as all warring parties seek to control the narrative and gain the support of public opinion to justify their actions. Meta is deeply committed to our responsibility to respect human rights as set out in our Corporate Human Rights Policy. Meta's family of apps and services (Facebook, Instagram, Messenger, WhatsApp) give people a voice and help build community: these rights are core to our mission. We look forward to engaging with the UN Special Rapporteur on this critical issue and on how best to address these challenges.

In this submission, Meta sets out:

1. Key challenges raised by disinformation/misinformation/propaganda in times of conflict;
2. Our approach to combating misinformation;
3. Our approach to at risk countries;
4. Case study: our recent actions in the context of the Russian invasion of Ukraine.

1. Key Challenges Raised by Disinformation, Misinformation, and Propaganda in Times of Conflict

Disinformation and misinformation present significant conceptual and definitional challenges. We have previously outlined our understanding of those terms in our response to the UN Special Rapporteur's call for contributions to an earlier report where she examined the threats posed by disinformation to human rights and democratic institutions (A/HRC/47/25). In summary, we have adopted the following definitions:

- Misinformation: refers to misleading content, including provably false information and manipulated videos. It is often shared without an intent to mislead.
- Disinformation: refers to provably false information shared with an intent to deceive.
- Influence operation: coordinated effort to manipulate or corrupt public debate for a strategic goal.

We do not have a separate definition for "propaganda." This concept is not well- or consistently defined across diverse countries' domestic legislation or under international law. This is partly because "propaganda", on its own, says nothing about the substance of the message. If we were to ban "propaganda" under Facebook's (FB) Community Standards or Instagram (IG)'s Community Guidelines,[1] this would invite subjective interpretations and could lead to the removal of legitimate political content. The lack of a working operational definition of "propaganda for war" also means that the concept of disinformation is more relevant, where the focus is more specifically on actors and behaviors.

Our Community Standards remain applicable in times of conflict. Relevant policies include: Misinformation, Inauthentic Behavior, Hate Speech, Violence and Incitement, Violent and Graphic Content, Bullying and Harassment, Coordinating Harm and Promoting Crime. If militarized social movements or hate groups take part in the conflict, we rely on our Dangerous Individuals and Organizations Policy. We also have policies in place to deal with the second order effects of conflict on our platforms, including Human Exploitation, Sexual Solicitation, and Child Sexual Exploitation, Abuse and Nudity.

We believe that our policies provide an appropriate framework to address the vast majority of issues that arise on our platforms. We constantly keep our policies under

[1] We will refer to them collectively as the "Community Standards" throughout the rest of this document.

review and adapt them as appropriate. Nonetheless, as a social media company committed to respecting human rights, we face a number of challenges:

Legal considerations

- **Internet shutdowns or other severe restrictions on access to our platforms:** In times of conflict, governments may seek to shutdown the Internet entirely. They may also throttle access to our platforms when they consider that we have failed to comply with their demands, which may be inconsistent with international human rights standards. These measures severely restrict the free flow of reliable information and freedom of expression. The lack of access to our platforms also has practical implications for the ability of our trusted partners and third-party fact-checkers to help us identify content that may be in violation of our policies as detailed further below. We track and disclose Internet disruptions (intentional restrictions on connectivity that limit people's ability to access the internet or specific websites and apps) on our Transparency Center: https://transparency.fb.com/data/internet-disruptions/.
- **Government takedown requests:** Some governments may call for us to remove content in circumstances where the legal basis for those actions or their compatibility with international human rights norms is unclear but public pressure to act is overwhelming. We track content restrictions based on local law on our Transparency Center: https://transparency.fb.com/data/content-restrictions/.
- **Preservation of evidence of war crimes:** We have policies and procedures in place to respond to preservation requests from governmental agencies. We are further exploring steps to preserve data that may be relevant for potential use in future international accountability proceedings.

Policy challenges

- **How the enforcement of our Community Standards should be adapted in times of conflict:** Our Community Standards are global and we take into account local context when we enforce them. We do so based on feedback from Trusted Partners and taking into account international human rights law, including the Rabat Principles. In practice, it is extremely challenging to determine the likelihood of content leading to offline harm under our Violence and Incitement or Misinformation policies in the ever evolving complexity of conflict settings. Further guidance is needed on the interplay between international human rights law and international humanitarian law in the age of social media at different stages of conflict situations. For example, should social media platforms allow for calls of violence against active combatants that would be legitimate under international humanitarian law during an armed conflict? What about calls for violence against armed occupying forces?

- **Lack of clear international standards on propaganda for war:** We believe that most content that may be understood as "propaganda" is likely to be covered under our existing Misinformation and Coordinated Inauthentic Behavior policies. That said, we would welcome guiding principles on propaganda for war that could help guide our response when being asked by governments to remove content, e.g. symbols that are used in support of one side of a conflict or how to address calls for the initiation of armed hostilities that, if carried out, would violate the fundamental principles of the UN Charter, including Article 2 (4) of the UN Charter;
- **State-sponsored disinformation:** State actors may themselves be behind disinformation campaigns and promote their own false narratives on social media, for example denying atrocity crimes despite overwhelming evidence to the contrary. While our Coordinated Inauthentic Behavior policies may help us address some of these issues, we would welcome additional guidance on applicable international standards in this area;
- **Prisoners Of War:** Video content showing prisoners of war (POWs) involves a balancing exercise between the norms of both international human rights and international humanitarian law. The rules of international humanitarian law may weigh in favour of removing videos of prisoners of war (POW) to protect them from public curiosity but there may also be a public interest in this kind of content, including by showing that a POW is alive. Any additional guidance on how to balance these interests would be welcome.

Practical constraints

- **Constrained information landscape:** In situations of violent conflict, determining what information is true or false becomes more challenging for journalists, third-party fact-checkers and local communities alike. Access to primary sources may become difficult or impossible, and security risks to operating in conflict-affected environments also increase. These constraints are even more acute when the Internet is shut down in conflict areas.
- **Polarisation:** Where society becomes polarised, the number of sources of information regarded by all elements of society as being credible may become more limited. This in turn makes providing access to reliable information more challenging.
- **Potential contribution to harm:** The causal relationship between misinformation and violence is not well understood. However, it appears that the risk of some false claims contributing to violence likely increases at times of conflict.

- **Transparency and engagement:** At the onset of conflict, decisions may need to be made urgently without time for external stakeholder engagement. Generally, we have a robust and consistent process of evaluating and updating our Community Standards, with proper disclosure on updates to our policies in our Transparency Center. Yet, in times of crisis, we may need to update policies without providing robust user disclosure, due to concerns about alerting bad actors to changes in policy that they could abuse in real-time.

2. Our Approach to Combating Misinformation

Misinformation is different from other types of speech addressed in our Community Standards because there is no way to articulate a comprehensive list of what is prohibited. With graphic violence or hate speech, for instance, our policies specify the speech that we prohibit, and even persons who disagree with those policies can follow them.

With misinformation, however, we cannot provide such a line. The world is changing constantly, and what is true one minute may not be true the next minute. People also have different levels of information about the world around them, and may believe something is true when it is not. A policy that simply prohibits "misinformation" would not provide useful notice to the people who use our services and would be unenforceable, as we don't have perfect access to information.

Instead, our policies articulate different categories of misinformation and try to provide clear guidance about how we treat that speech when we see it. For each category, our approach reflects our attempt to balance our values of expression, safety, dignity, authenticity and privacy.

Misinformation & Harm

We remove misinformation or unverifiable rumours that expert partners have determined are likely to directly contribute to a risk of imminent violence or physical harm to people. We define misinformation as content with a claim that is determined to be false by an authoritative third party. We define an unverifiable rumour as a claim whose source expert partners confirm is extremely hard or impossible to trace, for which authoritative sources are absent, where there is not enough specificity for the claim to be debunked, or where the claim is too incredulous or too irrational to be believed.

We know that sometimes misinformation that might appear benign could, in a specific context, contribute to a risk of offline harm, including threats of violence that could contribute to a heightened risk of death, serious injury or other physical harm. We work with a global network of non-governmental organisations (NGOs), not-for-profit organisations, humanitarian organisations and international organisations that have expertise in these local dynamics.

In the countries where the risk of misinformation contributing to imminent violence or physical harm appears to be the greatest, we have also proactively consulted expert local partners to determine if there are specific categories of false claims which have an ongoing elevated risk of contributing to imminent harm. On the basis of their recommendations, we have designated ‘persistently harmful claims’ in countries including Bangladesh, Burkina Faso, the Democratic Republic of Congo, Ethiopia, Iraq, Kenya, Libya, Mali, Myanmar, Niger, and Sri Lanka - removing, for example, out-of-context media purporting to depict acts of violence or weapons, and false claims about intentional damage to or destruction of religious iconography, places or worship, or other holy buildings or structures.

In addition to misinformation that could contribute to the risk of imminent violence or physical harm, we also remove harmful health misinformation, voter or census interference, and manipulated media. You can read more about these policies in the relevant section of our Community Standards, here.

Fact-Checking

For all other misinformation, we focus on reducing its prevalence or creating an environment that fosters a productive dialogue. We know that people often use misinformation in harmless ways, such as to exaggerate a point ("This team has the worst record in the history of the sport!") or in humour or satire ("My husband just won Husband of the Year.") They also may share their experience through stories that contain inaccuracies. In some cases, people share deeply held personal opinions that others consider false or share information that they believe to be true but others consider incomplete or misleading.

Recognising how common such speech is, we focus on slowing the spread of hoaxes and viral misinformation, and directing users to authoritative information. As part of that effort, we partner with third-party fact-checking organisations to review and rate the accuracy of the most viral content on our platforms (see here to learn more about how our fact-checking programme works). We also provide resources to increase media and digital literacy so people can decide what to read, trust and share themselves.

Through our fact-checking programme, we partner with more than 80 independent third-party fact-checkers around the world covering over 60 languages. This includes many countries affected by violent conflict, with the full global coverage of our fact-checking partners detailed here. Facebook’s fact-checking partners all go through a rigorous certification process with the non-partisan International Fact-Checking Network (IFCN), ensuring their independence and transparency.

3. Our Approach to Countries At Risk

Since 2018, we've had dedicated teams spanning product, engineering, policy, research and operations to better understand and address the way social media is used in countries experiencing conflict. Many of these individuals have experience working on conflict, human rights and humanitarian issues, as well as addressing areas like misinformation, hate speech and polarization. Many have lived or worked in the countries we've identified as highest risk and speak relevant languages. They are part of the over 40,000 people we have working on safety and security, including global content review teams in over 20 sites around the world reviewing content in over 70 languages.

In the last couple of years, we've hired more people with language, country and topic expertise. For example, we've increased the number of team members with work experience in Myanmar and Ethiopia to include former humanitarian aid workers, crisis responders and policy specialists. And we've hired more people who can review content in Amharic, Oromo, Tigrinya, Somali and Burmese. Adding more language expertise has been a key focus area for us. In 2021, we hired content moderators in many new languages, including Haitian Creole, Kirundi, Tswana and Kinyarwanda.

Evaluating Harm in Countries at Risk

Our teams have developed an industry-leading process for reviewing and prioritizing which countries have the highest risk of offline harm and violence every six months. We make these determinations in line with the UN Guiding Principles on Business and Human Rights and following a review of these factors:

- **Long-term conditions and historical context**: We rely on regional experts, platform data and data from more than 60 sources like Varieties of Democracy (V-Dem), Uppsala Conflict Data Program, the United States Holocaust Memorial Museum's Early Warning Project, the Armed Conflict Location & Event Data Project, and the World Bank to assess the long-term conditions on the ground. These can include civic participation and human rights, societal tensions and violence, and the quality of relevant information ecosystems.
- **How much the use of our products could potentially impact a country**: We prioritize countries based on a number of factors, including: where our apps

have become most central to society, such as in countries where a larger share of people use our products; where there's been an increase in offline harms; and where social media adoption has grown.

- **Current events on the ground**: We also give special consideration to discrete events that might magnify current societal problems, such as local risk or occurrence of atrocity crimes, elections, episodes of violence and COVID-19 vaccination and transmission rates. For example, we recently conducted rapid human rights due diligence in relation to Bosnia in light of the country's recent history of war and atrocity crimes, increased tensions in the country and upcoming elections in October 2022.

Strategies for Helping to Keep People Safe in Countries At Risk

Using this prioritization process, we develop longer-term strategies to prepare for, respond to and mitigate the impacts of harmful offline events in the countries we deem most at risk. This allows us to act quickly to remove content that violates our policies and take other protective measures while still protecting freedom of expression and other human rights principles. Recent examples include our preparations for elections in Myanmar, Ethiopia, India, Philippines and Mexico.

- **Understanding and engaging with local contexts and communities:** We know that working with the people and organizations on the ground with firsthand information and expertise is essential. Over the past few years, we've expanded these relationships with local civil society organizations to support country-specific education programs and product solutions, and to ensure our enforcement accounts for local context. We've also expanded our global network of third-party fact-checkers. Additionally, we have invested significant resources in more than 30 countries with active conflict or societal unrest. Together with UN partners and dozens of local and global NGOs, we have developed programming, including through global digital literacy initiatives such as We Think Digital or programs to make online engagement safer, such as Search for Common Ground's program in central Africa.
- **Developing and evaluating policies to prohibit harmful content and behavior**: We are constantly evaluating and refining our policies to address evolving nuances of hate speech, identify groups at heightened risk of violence or perpetrators of atrocities and human rights abusers, or the potential for rumors and misinformation to contribute to offline physical harm, particularly in countries where ethnic and religious tensions are present.

- **Improving our technology and enforcement to help keep our community safe**: During moments when the risk of harm is greater, we may take more aggressive action. For example, ahead of elections and during periods of heightened unrest in India, Myanmar and Ethiopia, we significantly reduced the distribution of content that likely violated our policies on hate speech or incitement of violence while our teams investigated it. Once we confirmed that the content violates these policies, we removed it. We also significantly reduced the distribution of content posted from accounts that have repeatedly posted violating content — in addition to our standard practice of removing accounts that frequently violate our Community Standards. To protect people in Afghanistan following the Taliban takeover, we launched a feature that allows them to lock their profile to provide an extra layer of privacy, security and protection for their information.

  In some circumstances, we may reduce the distribution of potentially violating content even when our systems predict that a specific post has a very low probability of violating our policies, while our teams investigate it. The extent of that reduction is based on the confidence of our systems' prediction, as well as local conditions. For example, we may take less action to reduce a piece of content in News Feed that is determined to have a 25% chance of violating our policies on hate speech versus a piece of content that has a 50% percent chance of violating.

In a crisis, we will determine what kind of support and teams we need to dedicate to a particular country or language, and for how long we need to keep them in place. This might include deploying our crisis operations center model to monitor and respond to threats in real time. It can also include seeking to ensure our integrity systems and resources are robust and ready where there may be ongoing risk of political unrest, or building temporary product levers ahead of a protest or a culturally sensitive event — all while ensuring that we have teams ready to support unplanned events, such responding to the coup in Myanmar.

We know that we face a number of challenges with this work and it is a complex and often adversarial space — there is no one-size-fits-all solution. Many of these offline issues have existed for decades or longer, and media services have a long history of being abused by those seeking to assert or maintain power or incite violence. But, we know our work to keep our global community safe will never be finished and it requires ongoing vigilance and investments. That's what we've done for many years and we will continue doing it going forward. As we do so, our work will continue to be grounded in the UN

Guiding Principles on Business and Human Rights, including as it relates to conflict situations.

More information about our at risk countries program and related human rights due diligence work, including in the context of the conflicts in Myanmar and Ethiopia, can be found in our Annual Human Rights Report (2022, p. 65 ff).

4. Case Study: Actions Taken By Meta related to the Russian Invasion of Ukraine

Disinformation, misinformation, and war propaganda are major challenges that social media platforms have had to face in the wake of Russia's invasion of Ukraine. We are alert to emerging threats and ways social media platforms may be targeted to sow confusion, post false narratives around the conflict, and attempt to compromise people's accounts. We seek to respond to these challenges in a timely manner.

We are acutely aware of the role that our apps play in this conflict as we witness ordinary Ukrainians and Russians using our platforms to shine a light on what's happening every day. Social media allows people from outside Ukraine to reach out to those inside the country.

With this in mind, we detail the efforts we have undertaken as a company to address the different risks and threats that we've identified during these exceptional times.

Responding to government requests

In February, we received several takedown requests from the Russia's Federal Service for Supervision of Communications, Information Technology and Mass Media (Roskomnadzor) which did not violate Facebook Community Standards or other policies, so we took no action on the reported contents:

- 7 items calling for participation in anti-war protests in the context of the war in Ukraine.
- 1 item condemning the war in the context of the war in Ukraine.
- Request to remove fact checking information labels and demotions applied to content posted on Facebook by 4 Russian media organisations. These posts were about the conflict in Ukraine and had been assessed by independent fact checking organisations to be false.

At the request of the government of Ukraine, we have restricted access to several accounts in Ukraine, including those belonging to some Russian state-controlled media organisations. In response to government sanctions, we've also restricted access to Russia Today (RT) and Sputnik, including their subsidiary Pages and Accounts, across the EU and the UK.

Enforcement of our Community Standards

At the start of the war, we established a special operations center in order to quickly and appropriately react to the ongoing evolving situation. Special operation centers are staffed by experts from across the company who work around the clock to monitor and respond to this rapidly evolving conflict in real-time. This allows us to remove content that violates our Community Standards faster and serves as another line of defence against misinformation. We have teams of native Russian- and Ukrainian- speaking content reviewers to help us review potentially violating content. Besides, many of the violating pieces of content such as graphic content can be identified regardless of the language-expertise. We're also using technology to help us scale the work of our content review teams, so we can take down more violating content before it goes viral.

Our investigative teams actively monitor for coordinated inauthentic behaviour in the Ukrainian context from known threat networks as well as proactively identify new operations and threat actors. We saw some evidence of compromised accounts targeting Ukraine during the invasion. We took down a network run by people in Ukraine and Russia targeting Ukraine for violating our policy against coordinated inauthentic behaviour. They ran websites posing as independent news entities and created fake personas across many social media platforms including Facebook, Instagram, Twitter, YouTube, Telegram, Odnoklassniki and VK. Further details of our work to detect and disrupt threats and platform abuse in the cybersecurity space are detailed in our Adversarial Threat Report (April 2022, p. 9 ff)

In addition, we regularly review and adapt our guidance to our content moderators to take into account any relevant changes to the context in which we operate. We also recently updated our Coordinated Harm policy so that we now remove content that exposes the identity or location of a prisoner of war when we become aware of it. We remove this content because it can put these people at risk of harm.

Our Community Standards on inauthentic behavior have also been updated so that we no longer allow governments that have instituted sustained blocks of social media to use their official departments, agencies, and embassies to deny the use of force or violent events in the context of an attack against the territorial integrity of another state in violation of Article 2(4) of the UN charter.

Measures to Tackle Misinformation in Russian and Ukrainian

In response to the invasion of Ukraine, we have expanded our third-party fact-checking program in Russian and Ukrainian languages across the region and are working to provide additional financial support to our Ukrainian fact-checking partners.

We have also taken additional measures to make it easier for all fact-checkers covering the conflict globally to find and rate content related to the war because we recognize that speed is especially important during breaking news events. We use keyword detection to group related content in one place in the fact checking tool used by our partners, making it easy for fact-checkers to find and prioritise content related to the invasion. When content has been rated by one of our third-party fact checking partners, it becomes ineligible for monetization.

Messenger, Instagram and WhatsApp have limits on message forwarding and label messages that haven't originated with the sender so people are aware that something is information from a third party.

We are also supporting reliable access to trusted information through our apps. We've updated Community Help as a central resource on Facebook where Ukrainians and others in the region can find reliable information from local UN agencies and Red Cross societies. This includes where to seek medical help, how to stay safe and how to get assistance — both in Ukraine and once they have crossed into neighboring countries. It also features access to Ukraine's State Emergency Services information helpline on WhatsApp, which connects people with critical information including emergency response procedures. We show a link to Community Help at the top of people's Facebook and Instagram feeds for people in Ukraine or for those who have left recently for neighboring countries, letting them know that this resource is available. It is available globally at facebook.com/community_help_ukraine, as well as at the top of results for relevant searches on Facebook.

Transparency around State-Controlled Media

We label state-controlled publishers because they combine the influence of a media organization with the strategic backing of a state. Our approach therefore seeks to be informational.

Meta has taken unprecedented actions specifically related to Russian state-controlled media entities (‘SCMEs’). In addition to labeling Russian state-controlled media pages and Instagram accounts, in multiple languages, we take enforcement actions globally on all of them in a multitude of ways:

- We have blocked ads from Russian state-controlled media outlets globally;
- We have demonetized Russian SCME Pages and accounts;
- We are demoting content from Russian SCMEs in feed and making it harder to find on Facebook and Instagram;
- We have removed Russian SCMEs from recommendation surfaces;
- We have added SCME labels on posts that contain links to Russian SCME’s off-platform presence;
- We’ve added interstitials that prompt users to confirm their actions when they click on links to Russian SCME content and when they reshare content from these entities.

More information about Meta’s actions with regards to Russia’s invasion of Ukraine can be found in this blog post .

Human Rights Due Diligence in the context of the war in Ukraine

As part of good human rights due diligence practice, Meta has been regularly engaging with both Ukrainian civil society and Russian independent media and civil society. We are also engaging with international institutions such as the Office of the Special Advisor for the Prevention of Genocide, the Council of Europe and other UN institutions. We regularly assess our human rights response to the conflict in light of best practice in this area and the UN Guiding Principles on Business and Human Rights. Our Human Rights team also weighs in on individual content decisions and we have explored measures for evidence preservation work.