Briefing to Canada's Standing Senate Committee on Human Rights: Meta's approach to Hate Speech and Islamophobia

Name of the organization submitting the brief: Meta Platforms, Inc.

Name of the Senate committee to whom the brief is being submitted: The Standing Senate Committee on Human Rights

Summary of the main points

Introduction

Meta's platforms, including Facebook and Instagram, are places for people to share content with their friends, families, communities, or customers. Every day, billions of pieces of content are shared on our platforms and many people have a positive experience with the people and communities they engage with. When someone signs up for an account with our platforms, they agree to Facebook's Community Standards and Instagram's Community Guidelines, which are rules for what is and what is not allowed to be posted on our platforms. Our Community Standards consist of over 20 policies including, notably, policies against Hate Speech, Bullying & Harassment, Violence & Incitement, and Dangerous Organizations & Individuals.

However, despite our best efforts to prevent it, we know that there are people who abuse our platforms. We strive continuously to improve on our industry-leading record of identification and removal of this content as soon as it comes to our attention.

We take very seriously our role in keeping abuse off our platforms. Our Community Standards have evolved over time to reflect changes in form and substance of posted material, and we are constantly iterating on where to draw the line in terms of permissible content. Our Community Standards apply to everyone, all around the world, and to all types of content.

The goal of our Community Standards is to create a place for expression based on our core values: voice, safety, privacy, dignity, and authenticity.

- **Voice**: A commitment to expression is paramount, but we recognize the internet creates new and increased opportunities for abuse. For these reasons, when we limit expression (voice) we do it in service of one or more of the following values.
- **Privacy**: We are committed to protecting privacy and personal information. Privacy gives people the freedom to be themselves, and to choose how and when to share on Facebook and to connect more easily.
- **Safety**: Our content policies focus on safety and mitigating harm. We are committed to making Facebook a safe place; expression that threatens people has the potential to intimidate, exclude, or silence others and isn't allowed on Facebook.
- **Authenticity**: We want to make sure the content people are seeing on Facebook is authentic. We believe that authenticity creates a better environment for sharing, and that's why we don't want people using Facebook to misrepresent who they are or what they're doing.
- **Dignity**: We believe that all people are equal in dignity and rights. We expect that people will respect the dignity of others and not harass or degrade them.

We recognize that building a community and bringing the world closer together depends on people's ability to share diverse views, experiences, ideas and information. These standards are based on feedback from people and the advice of experts in fields such as technology, public safety and human rights. We believe that everyone's voice is valued, so we take care to create standards that include different views and beliefs, especially from those from marginalized communities.

Our policies cover a broad range of content that can be found on our platforms. Below we will explore our approach to human rights, content policies and engagement with Muslim communities across the globe in more detail.

Meta's Human Rights Framework

We implement our commitment to human rights using approaches set out in the United Nations Guiding Principles on Business and Human Rights (UNGPs). This framework includes (1) applying human rights policies; (2) conducting human rights due diligence and disclosure; (3) providing access to remedy; (4) maintaining oversight, governance, and accountability; and (5) protecting human rights defenders.

Meta's content policies take into account the goal of conforming with the highest global standards. The policies have evolved such that we ensure they now overtly align with IHRL standards. For example, our hate speech policies seek to implement ICCPR Art. 20, which reads: "Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law".

Meta strives to have the most comprehensive set of protections against hate speech and discrimination in the technology industry. Our iterative approach to content policy development also allows swift and explicit policy changes to adapt to evolving concerns

relating to hate speech. For instance, we've recently updated policy areas including outing risk and bullying and harassment policies with insights provided by the Law Commission of Ontario, among others.

Meta's Human Rights Team, which is part of our Trust & Safety organization, supports leadership in making decisions that respect human rights, wherever they arise in the company. Since 2019, the Human Rights Team has sought to "show not tell" human rights actions – ranging from offering product advice, to conducting extensive content policy due diligence, to developing the company's approach to global regions in crisis.

Meta's Development of Community Standards

We currently have 2.88 billion daily active users on Meta's family of applications, and billions of pieces of content uploaded daily. We receive millions of reports every week flagging content under our policies. Of our 2.88 billion daily users, nearly 90% are outside of the US & Canada; our policy development and enforcement have a similar global focus.

At Facebook and Instagram's scale, content review requires the partnership of multiple teams and subject matter experts across policy, product, and operations. All of these teams work in concert together. For example, if content reviewers do not have the right tools, they cannot serve the community effectively. If the right global policies are not in place, content reviewers will not be able to enforce on our policies effectively.

- **The Content Policy Team** develops and refines our Community Standards. They work with internal teams and outside experts, academics, NGOs and policymakers to get feedback on our Community Standards and make improvements. Given the global focus of our policies, the Content Policy Team sits in 11 offices around the world. Having a diverse team situated across the world helps us to be responsive to our users and enhances our language, cultural, and regional expertise to better understand the situation on the ground.
- **The Community Integrity Team** builds products to help keep people safe and secure. They do this in two ways: (1) building the user experiences that enable people to report harmful content, send feedback, and appeal decisions, as well as tools that better protect users from sensitive information or allow them to block, hide, or unfollow accounts; and (2) developing artificial intelligence algorithms that surface potentially prohibited content and either take it down or escalate it for review. This team includes engineering, data science, product, program management, and data engineering.
- **The Global Operations Team** enforces our Community Standards through human review. We work with companies who are experts in this type of staffing and in locations all over the world to do the job of supporting our community.

Gathering input from our external stakeholders is an important part of how we develop Facebook's Community Standards. Engagement makes our policies stronger and more nuanced and inclusive. It brings our stakeholders more fully into the policy development process, introduces us to new perspectives, allows us to share our thinking on policy options, and roots our policies in sources of knowledge and experience that go beyond Meta.

In order to do this work meaningfully and effectively, the Stakeholder Engagement Team, which sits within the wider Content Policy team, engages regularly with external stakeholders in an effort to ensure that our content policy development process is informed by the views of outside experts and people who use our platform. The Stakeholder Engagement Team is organized into subject matter experts who are directly tasked with developing and iterating on the Community Standards. These experts include lawyers, human and civil rights experts, political scientists and tech professionals. In addition to being responsible for the policy development process, the Stakeholder Engagement team also builds relationships with the broadest possible spectrum of NGOs, academics and other thought leaders, and civil society organizations around the world, guided by our core principles of inclusivity, expertise, and transparency.

Our policy development process involves different teams across the organization and across all regions. The need for a policy re-evaluation may be surfaced to us through our reviewers, users, policy teams, research teams, or other stakeholders. We then initiate a process of better understanding the issue through gathering data and content examples, and conducting research and engagement with a wide variety of internal and external stakeholders. The Content Policy team, in collaboration with other teams, develops options for improved policies, analyzes them, and aligns on a path forward.

When a new policy is adopted, the launch process involves training our content reviewers and updating tools and materials. After a policy has launched, we monitor how it works in practice. This process can take between 2-4 weeks to 6 months, or even a year, depending on the complexity of the policy. This multi-step effort allows us to account for a range of perspectives and opinions across the globe, and ultimately to develop stronger policies. When our policies are written or updated, we share those updates on our Transparency Center website.

Hate Speech Policy

Our Hate Speech policy is part of our Community Standards. We understand that hate, whether expressed online or in-person, may have significant negative effects. The robust policies and practices we have in place to address hate on the platform is something we closely monitor and work with stakeholders on. We do not allow hate speech on Facebook because, although we want to provide people with a voice, hate speech creates an environment of intimidation and exclusion and in some cases may promote violence.

We define hate speech as attacks on an individual or group based on one or more protected characteristics which include race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability.

Hate speech can include different types of attacks. We define ‘attack’ as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, or calls for exclusion or segregation. It includes any violent speech or support for death or harm and statements of inferiority, contempt or disgust. It also includes speech that is dehumanizing, like describing a group of people with protected characteristics in negative ways. Another kind of hate speech is exclusionary speech - for instance, saying people in a protected group should be kicked out of the country or that they shouldn’t be allowed to hold office.

We also do not allow content that describes or negatively targets people with slurs - words that are inherently offensive and commonly used as insulting labels for some of these protected characteristics.

We separate attacks into three tiers of severity. Using a tiered approach, we're able to provide protections based upon the type of attack. For instance, we forbid attacks against refugees, migrants, immigrants and asylum seekers, though we do allow commentary and criticism of immigration policies.

Bullying and Harassment Policy

Bullying and harassment happen in many places and come in many different forms, from making threats and releasing personally identifiable information to sending threatening messages and making unwanted malicious contact. Under our Bullying and Harassment Policy, we prohibit this kind of behavior because it prevents people from feeling safe and respected.

Under this policy, we distinguish between public figures and private individuals because we want to allow discussion, which can include critical commentary of people who are featured in the news or who have a large public audience. For public figures, we remove direct attacks as well as certain attacks where the public figure is directly tagged in the post or comment. For private individuals, we also remove content that's meant to degrade or shame, including, for example, claims about someone's personal sexual activity. We recognise that bullying and harassment can have a significant impact on minors, which is why our policies provide heightened protection for users between the ages of 13, which is the minimum age for our products, and 18.

Context and intent matter, and we allow people to post and share if it is clear that something was shared in order to condemn or draw attention to bullying and harassment. In certain instances, we require self-reporting because it helps us understand that the person targeted feels bullied or harassed. We prohibit content that targets anyone maliciously by repeatedly contacting someone in a manner that is

unwanted, sexually harassing, directed at a large number of individuals with no prior solicitation, or attacking someone through derogatory terms related to sexual activity.

In addition to reporting such behavior and content, we encourage people to use the tools available on Facebook to help protect against it. We also have a Bullying Prevention Hub, which is a resource for teenagers, parents and educators seeking support for issues related to bullying and other conflicts. It offers step-by-step guidance, including information on how to start important conversations about bullying.

Violence and Incitement Policy

Our Violence and Incitement Policy aims to prevent potential offline harm that may be related to content on Facebook, and we remove language that incites or facilitates serious violence. We remove content, disable accounts and work with law enforcement when we believe there is a genuine risk of physical harm or direct threats to public safety. Under this policy, we remove threats that could lead to death or serious injury such as statements of intent to commit violence or statements advocating violence or calls for violence, including content where no target is specified but a symbol represents the target.

We also consider language and context in order to distinguish casual or non-serious statements from content that constitutes a credible threat to public or personal safety. In determining whether a threat is credible, we may also consider additional information such as a person's public visibility and the risks to their physical safety.

Dangerous Organizations and Individuals Policy

We prohibit entities, including organizations or individuals, that proclaim a violent mission or are engaged in violence from having a presence on our platforms. This is set out in our Dangerous Organizations & Individuals policy.

A key part of our strategy is identifying dangerous individuals and groups, banning and removing them from our platforms, and removing praise, support, or representation of these groups on our platforms. For example, a hate group like the KKK is a designated entity. We would also remove content that proclaims praise or support for a designated individual.

We have a robust process to assess cases for possible designation. This process is based on two pillars: one is structured review, which means that we ensure that all relevant teams have an opportunity to provide input and feedback at defined phases of the review process. The second pillar is strong and reliable information; we build cases based on strong and corroborated information that can be sourced and cited.

In 2019, Meta designated multiple well-known Canadian organizations and figures as organized hate per our Dangerous Individuals and Organizations policy, including Faith Goldy, Kevin Goudreau, Canadian Nationalist Front, Aryan Strikeforce, Wolves of Odin, and Soldiers of Odin (Canadian Infidels), banning them from having any further

presence on our services and removing affiliate representation for these entities, including linked pages and groups. Proud Boys was designated in October 2020.

Removing Violating Content

We take action to address and reduce the prevalence of content that violates our policies such as bullying and harassment, hate speech, and violence and incitement. When we enforce against any piece of content, we take a three-part approach: remove, reduce and inform. This strategy includes actions like removing accounts, groups and events that violate our Community Standards, filtering problematic groups and pages from recommendations, and reducing the distribution of certain content, including applying warning screens. Moreover, any user who encounters content that he or she believes is in violation of our policies has the option to report that content to Meta for review.

We use a combination of human review and AI to prioritize and review content. When a human reviewer assesses content, they will review the content against all of our policies. Once content is reviewed, if it violates our policies, we will remove the content and the user will receive a notification in their inbox that it was removed. Users can also appeal our decisions relating both to content that has been removed and content that has not been removed. Appeals are submitted to our Oversight Board which considers appeal submissions and can issue final written decisions.

Each quarter in our Community Standards Enforcement Reports, we report on prevalence, which is the amount of hate speech people actually see on the platform. This metric is important because it helps us to measure how violating content impacts people. We also report on our record in removing particular forms of problematic content, so that we can track progress over time.

Global engagement with Muslim Communities

As part of our ongoing efforts to ensure that our Community Standards are informed by external experts and the communities we serve, the Content Policy Stakeholder Engagement Team recently developed and implemented a strategy to engage more deeply with Muslim communities across Europe, Sub-Saharan Africa, and and the Asia-Pacific regions, as well as the US and Canada. Two overarching goals that guided our strategy were: (1) building sustainable and long-term relationships with diverse Muslim communities; and, (2) creating a knowledge base regarding issues faced by Muslim communities. In the process, we sought to raise awareness and share learnings and knowledge gained through engagements with Muslim communities with internal stakeholders.

Over the course of 2021 and 2022, we developed a program of engagements with stakeholders from a range of Muslim communities globally, including in Canada, which exceeded 50 groups and individuals. Each engagement was designed to achieve a specific purpose and ultimately support the strategy's goal of meaningfully including more Muslim voices in the content policy development process.

Broadly, we utilized four modes of engagement: roundtables, 1:1 engagements, policy development discussions, and trend monitoring. These different engagement structures allowed us to gather input on specific policies under review or in development, understand broader trends and challenges faced by specific Muslim communities, and learn from academics studying online speech-related trends facing Muslim communities.

With respect to the **roundtables**, we carried out a series of six roundtables (some were country specific and other across regions) in partnership with different internal teams, external partners, academics and NGOs. Through this series of engagements, we made meaningful progress in bringing Muslim voices into internal teams' work, and developed an important foundation for future engagements and discussions with key stakeholders in the regions mentioned above. Across the roundtables, Muslim stakeholders wanted to learn more about Meta's efforts in identifying and moderating anti-Muslim hate speech, and measures taken by Meta other than removal of violating content.

With respect to **1:1 engagements**, we continue carrying out 1:1 engagements with academic experts and NGOs from the Muslim communities with the aim of better understanding current trends and research on key issues. Our 1:1 engagements provide us with more tailored opportunities to discuss specific insights and trends, particularly from organizations or scholars who have carried out extensive research and online monitoring.

With respect to **trend monitoring**, we analyzed existing academic literature, continued to monitor online platforms for anti-Muslim hate speech and discussed current trends with Muslim stakeholders. This helped us revise and improve our engagement with Muslim communities. We also brought academic experts to speak about their research with internal teams, mainly focussed on online anti-Muslim hate speech.

With respect to **policy development discussions**, Stakeholder Engagement recently consulted with over 20 stakeholders from Muslim communities on specific policies as part of our content policy development process. These policy development discussions brought Muslim voices into the development of the policies that might impact them, allowed us to build closer relationships with stakeholders from Muslim communities, provided communities with more transparency on content moderation at Meta, and helped them give us more actionable and informed feedback on our policies.

Conclusion

Our efforts to engage meaningfully with Muslim communities and address Islamophobia and anti-Muslim hate speech are a work in progress, and we have a long road ahead to achieve our goal of building sustainable and long-term relationships with diverse Muslim communities. We will be continuing our engagement with Muslim organizations and individuals to ensure that our policies are informed by those we serve.