30 January, 2025

Monika Bickert, Head of Global Policy Management
Kevin Martin, Head of Global Public Policy
Meta Platforms, Inc.
1 Hacker Way
Menlo Park, CA 94025
USA

Dear Monika and Kevin,

*Re: Concerns Regarding Recent Policy Announcements*

The Meta Safety Advisory Council, which, as you know, was established in 2009 to provide expertise and insights to guide Meta's approach to safety, is composed of online safety organisations and experts in 8 countries. As an independent advisory body, we value this opportunity to contribute critical perspectives, particularly when the company's decisions raise concerns about online safety.

We write regarding your recent announcements introducing significant changes to Meta's content moderation policies, including altering the hateful content policy, shifting moderation strategies and ending the fact-checking program. While we acknowledge the complexities of evolving public expectations and the dynamic policy landscape, these shifts carry profound implications that warrant careful scrutiny.

Our concerns include:

**1) Changes to Enforcement and Moderation**

- We commend Meta's ongoing efforts to address the most egregious and illegal harms on its platforms, and we remain committed to supporting this vital work. However, we emphasise the importance of Meta *not* de-prioritising investment in addressing 'borderline' harm – harm that may not meet the threshold of illegality but nonetheless affects significant numbers of young people, women, and other groups representing a substantial portion of Meta's user base. These ubiquitous legal harms can have equally profound impacts, particularly when they involve persistent aggression or exclusion based on protected characteristics.
- We acknowledge the challenge of moderating highly nuanced violative behavior, but removing proactive detection methods for such harm shifts an unreasonable burden onto users, who are now tasked with navigating higher thresholds for content to be deemed in violation of hateful conduct policies. This change disproportionately impacts those most vulnerable to long-term, cumulative harm.
- Many of the most devastating impacts observed on your platforms stem from ongoing hate against individuals or communities. Addressing this harm should remain a top priority for Meta, as its ripple effects extend far beyond your platforms and services.

**2) Changes to Hateful Conduct Policy**

- Groups facing marginalisation offline – including women, LGBTQIA+ communities, immigrants, and others – are disproportionately targeted online. Meta's rollback of protections risk eroding hard-won safeguards that ensure users feel safe and included in online social environments.
- Harassment and cyberbullying are the forms of hateful conduct most commonly experienced by minors, especially those in the groups just mentioned, and we urge you to continue implementing safety policy corresponding to the unique needs of users under 18. De-prioritizing existing safeguards will only embolden harmful behaviours, with repercussions both online and offline. This marks a concerning departure from Meta's history of leadership and innovation in proactive harm prevention.

**3) Ending the Fact-Checking Program**

- Crowd-sourced fact-checking tools like Community Notes can be useful in addressing misinformation. However, independent research raises significant concerns about their effectiveness. Without proper consultation or transparency around their implementation, it is unclear how Meta has weighed these challenges against the potential benefits. For instance, studies of similar initiatives, such as X's program, show that polarising issues often fail to reach consensus, leaving harmful misinformation unchecked.
- Fact-checking serves as a vital safeguard – particularly in regions of the world where misinformation fuels offline harm and as adoption of AI grows worldwide. Meta must ensure that new approaches mitigate risks globally.

**4) Setting a Damaging Precedent**

- This policy shift risks prioritising political ideologies over global safety imperatives. As one of the world's most influential companies, Meta's policies set a powerful signal – not just for online behaviour, but also for societal norms.
- By dialling back protections for protected communities, Meta risks normalising harmful behaviours and undermining years of social progress. These changes send a green light to hate and discrimination both online and offline. Now more than ever, Meta must demonstrate that its commitment to safety transcends politics and reflects its responsibility as a global leader.

We understand that evolving policies are part of Meta's approach. However, the perception of these changes as finalised and ideologically driven has caused concern worldwide. This moment presents an opportunity for Meta to affirm its commitment to safety through decisive action.

We urge Meta to consider the following recommendations:

1. **Prioritise mental health support for young people and marginalised groups.**
   Meta has the opportunity to play a leadership role in fostering cross-industry collaboration to support independent third-party networks and organizations. Strengthening partnerships with youth-serving organisations, mental health services, helplines, "trusted flaggers," and other responders to online harm can provide critical independent infrastructure for safeguarding vulnerable users and addressing harm effectively. Given the significant risks posed by scaling back content moderation, Meta should ensure that trusted safety partners (including those serving on Meta safety advisories) have clear and effective channels to escalate urgent safety concerns, such as threats of violence or doxxing.

2. **Account for the global impact of policy decisions.**
   We call upon Meta to double down on considering the global impacts of U.S.-based policy decisions, recognising diverse markets, languages, and cultural contexts. To ensure equitable protections, we recommend extending the Community Guidelines Enforcement Transparency Report to include moderation data by market and language. This transparency would help Meta address regional disparities and create a more inclusive, responsive approach for its global user base.

3. **Commit to a 'Safety by Design' approach.**
   Safety considerations must be embedded in all future decisions, especially those with clear implications for user safety on a global scale. Adopting a 'Safety by Design' approach will ensure that Meta continues to lead in creating inclusive, secure online spaces that prioritise the wellbeing of all users, particularly the most vulnerable.

4. **Champion media literacy education globally.**
   Beyond digital literacy, media literacy education is essential for equipping AI users with the tools to critically evaluate content and navigate online spaces safely. Meta should lead efforts to promote media literacy on a global scale, empowering users to mitigate risks associated with misinformation, bias, and manipulative content.

We strongly encourage Meta to reconsider the broader impact of these policy changes, prioritising protections for all communities and upholding its role as a leader in creating safer digital spaces.

Sincerely,

**Lucy Thomas OAM,** CEO, PROJECT ROCKIT
**Anne Collier,** Executive Director, The Net Safety Collaborative
**Will Gardner OBE**, CEO, Childnet International
**Dr Ranjana Kumari,** Director, Centre for Social Research
**Stephanie Love-Patterson**, President & CEO, NNEDV
**Janice Richardson**, Director, Insight SA
**Larry Magid**, CEO, ConnectSafely
**Sean Lyons**, Chief Online Safety Officer, Netsafe
**Thiago Tavares,** President, SaferNet Brazil
**June Liu**, Executive Secretary, Institute of Watch Internet Network (iWIN)