POLICY FORUM

facebook

March 9, 2021

Recommendation:
Attacks on Concepts v. People

Organic Content Policy

WARNING
OFFENSIVE CONTENT

Attacks on Concepts v. People

Overview

Issue Statement

We are exploring options to remove hateful speech against institutions, ideas, and general concepts to enhance safety in certain environments. While this could help address situations where there is a real threat of offline violence, it could censor expression, such as a person sharing their personal views about a religion they chose to leave.

Source

- External feedback suggesting that attacks on concepts results in harm, intimidation, and exclusion of people
- Inconsistent enforcement on concepts and PCs due to language nuance
- Content escalations on conversion therapy content, caricatures of Prophet Muhammad, LGBTQ-free zones etc.

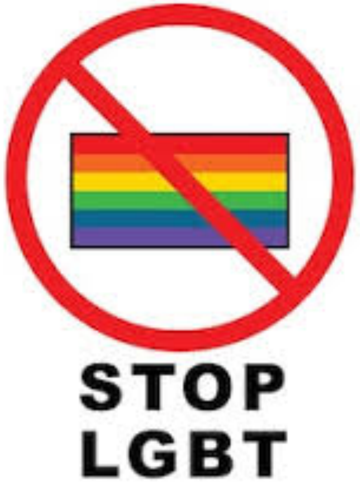
Policy Development

- 6 Working Groups
- 115 External Engagements

Status Quo Enforcement

Example (Allow)

✓



Attack on LGBT Flag

Example (Allow)

✓

Burning of the American flag

Attacks on Concepts v. People

Research and External Engagement

Research

Key Points:

- Other social media companies largely **do not prohibit** hate speech attacks with conceptual targets.
- Globally, attitudes about issues like insulting depictions of Muhammed and national flag burning **vary substantially** across countries.
- Research suggests that certain concepts can **more plausibly represent or signify** "people"; for others, the expansiveness of the concept would suggest things other than people, including political speech.

External Engagement

Key Points:

- Stakeholders agree that **sexual orientation and religion** are key concepts for this policy discussion -- but are **divided** when it comes to removing attacks on them.
- Some stakeholders express support for removing **Tier 1 attacks** on concepts, though this approach proves difficult to scope and other stakeholders push back
- Stakeholders broadly support a policy to remove attacks on concepts where there's a **serious risk of real world harm.**
- To identify risks of real world harm, experts in dangerous speech and atrocity prevention recommend that FB develop a set of **"community trigger points"** based on context, though there is no formula.

Option Analysis

Option	**Rationale** what is the main reason for supporting this option?	**WG XFN Feedback** What feedback has the XFN provided on the impact of this option on their stakeholders?	**Major Concern** What is the major concern or risk related to this option?
Option 1 Status Quo - Allow all attacks on Concepts	Ensures most room for voice and critical expression of the concepts. Easiest to operationalize.	Criticism from LGBTQ+ stakeholders on treating speech about sexual orientation and gender identity as concepts	Allows speech that may intimidate or exclude people associated with the concepts
Option 2 [Rec] Context-specific policy framework	Takes local context into account and assesses the risk of harm associated with the speech	Context is key in assessing the risks of harm and a globally consistent line is difficult to draw	Inherently inequitable in application as it would be enforced only on escalation-basis, and does not preempt harm or risks
Option 3 Remove Tier 1 attacks against all concepts	Removes attacks that are most likely to lead to real world harm at scale	May lead to over-enforcement as even severe attacks on concepts do not have the same impact as on people	Does not take local context into account and treats all concepts the same
Option 4 Carve-out to remove all attacks against specific concepts - sexual orientation, gender identity, and religion	Addresses attacks on most frequently attacked concepts, which have a heightened risk of harm	Responds to external criticism, which is mainly focused on the concepts identified	Applies disproportionately to specific PC groups, harder to operationalize.

Attacks on Concepts v. People

Content-specific Policy Framework (Option 2)

SIGNALS = A range of signs to determine whether there is a threat of harm in the content.

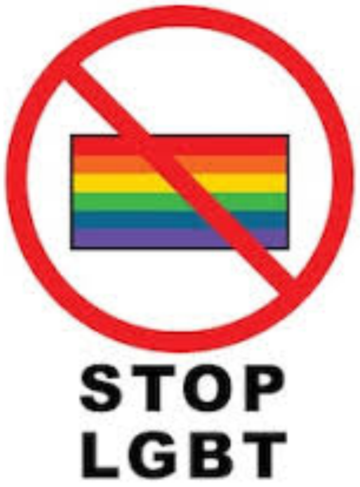
1. Does the content pose a risk of inciting imminent offline violence, intimidation, or discrimination against the PC group associated with the concept?
2. Is there a period of heightened tension, such as an election, ongoing conflict, ongoing protests, etc. ?
3. Is there is a recent history of violence or discrimination against the target PC group associated with the concept in country/region the speech is originating from (or being widely shared in)
4. Are there documented past instances of similar speech linked to offline violence, intimidation, or exclusion against the PC group associated with the concept
5. Does the speaker occupy a position of formal or informal power or authority (i.e., are they able to order or inspire action against others)?
6. Does the speaker have a large following, reach, or platform? Example- Is the speech public or in a group of over XX members?
7. Does the speaker have a history of violations of our community standards on Hate Speech, Violence & Incitement, or Dangerous Individuals or Organizations?

WARNING
OFFENSIVE CONTENT

Attacks on Concepts v. People

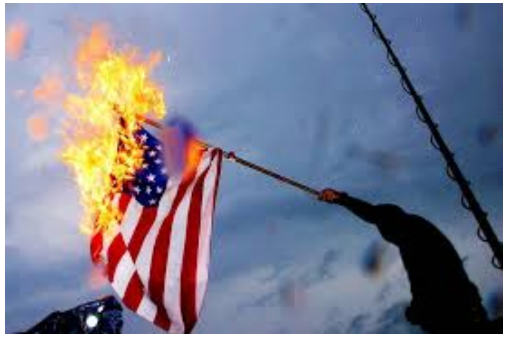
Examples

Example



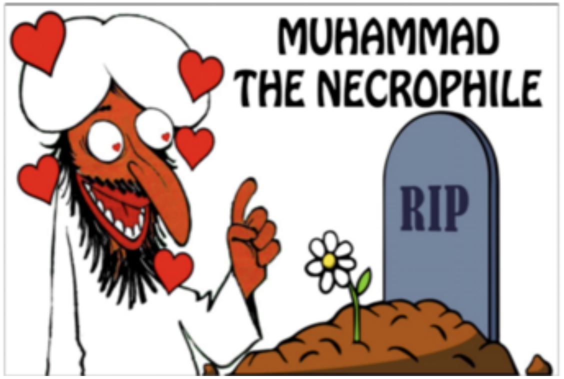
Option 1: Status Quo	✓
Option 2: Context-specific policy	*
Option 3: Tiered approach	✓
Option 4: Carve-out for specific concepts	✗

Example



Option 1: Status Quo	✓
Option 2: Context-specific policy	*
Option 3: Tiered approach	✗
Option 4: Carve-out for specific concepts	✓

Example



Option 1: Status Quo	✓
Option 2: Context-specific policy	*
Option 3: Tiered approach	✗
Option 4: Carve-out for specific concepts	✗

Example

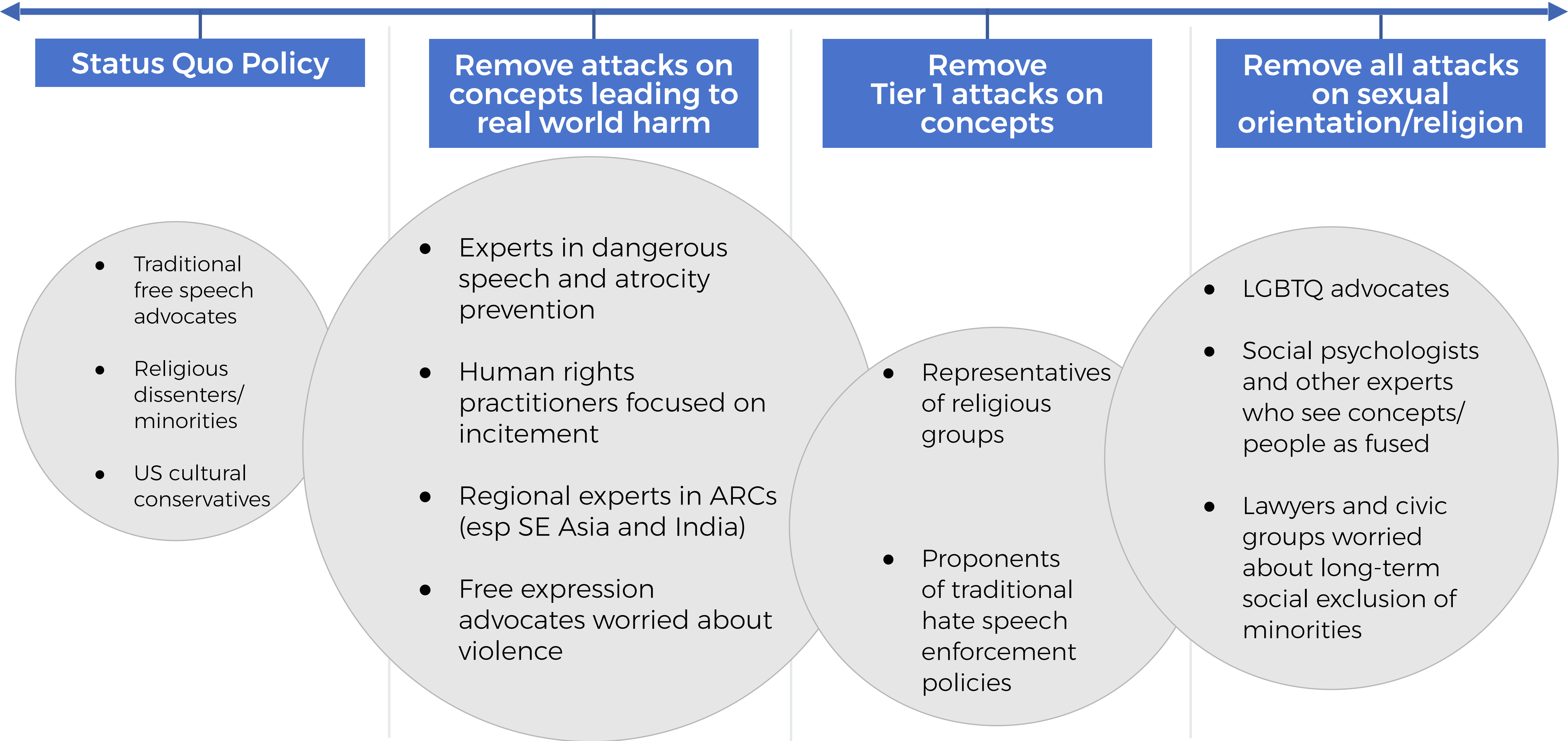
“We are not step-children in this country; White arrogance has no place in our democracy”

Option 1: Status Quo	✓
Option 2: Context-specific policy	*
Option 3: Tiered approach	✓
Option 4: Carve-out for specific concepts	✓

***Allow or Remove based on policy framework assessment under local context**

Attacks on Concepts v. People

Stakeholder Engagement Overview



Status Quo Policy

- Traditional free speech advocates
- Religious dissenters/ minorities
- US cultural conservatives

Remove attacks on concepts leading to real world harm

- Experts in dangerous speech and atrocity prevention
- Human rights practitioners focused on incitement
- Regional experts in ARCs (esp SE Asia and India)
- Free expression advocates worried about violence

Remove Tier 1 attacks on concepts

- Representatives of religious groups
- Proponents of traditional hate speech enforcement policies

Remove all attacks on sexual orientation/religion

- LGBTQ advocates
- Social psychologists and other experts who see concepts/ people as fused
- Lawyers and civic groups worried about long-term social exclusion of minorities