Policy Forum

Meta

February 28, 2023

Policy Recommendation: Hate Entity Designation Signals

Organic Content Policy

Issue Statement

Recognizing that our current hate entity designation policy does not always match our practices, we have drafted a new approach that we think better aligns with our practices and underlying values. However, we recognize that any changes to our policies and the increased transparency into our process might result in an incorrect perception that we are taking a lighter approach to hate entities, so we want to test the issue.

Status Quo Examples

Proud Boys (Designated Hate Org.)

The Proud Boys is a designated Hate Organization as it meets the threshold for designation because its: 1) Founder, Gavin McInnes, is a designated Hate Figure; 2) leaders and prominent members have made public statements using hate speech; and 3) the organization and leaders have engaged in praise or support of a designated Hate Entity.

Myanmar National Organization (Designated Hate Org.)

The Myanmar National Organization is a designated Hate Organization as it meets the threshold for designation because it: 1) Fundraised for a designated Hate Entity; 2) engaged in PSR of other designated entities; 3) organized an event with other designated entities; and 4) repeatedly used hate speech.

Policy Development Process

48

Stakeholder engagements in 25 countries

14

Regional and functional working groups

12

Countries in original survey research
+
Comprehensive literature review

17

Entities tested against proposed signals

Hate Entity Designation Signals

External Engagement

Key Points:

- Stakeholders agree on expanding the signals to encompass more coded hate speech, narratives and behaviors related to hateful disinformation, conspiracy theories, and doxxing based on protected characteristics.

- Stakeholders agree that the definition of hate entities should reflect the growing trend towards “post-organization” to be able to capture less structured, ideologically incoherent, or/and leaderless hate entities.

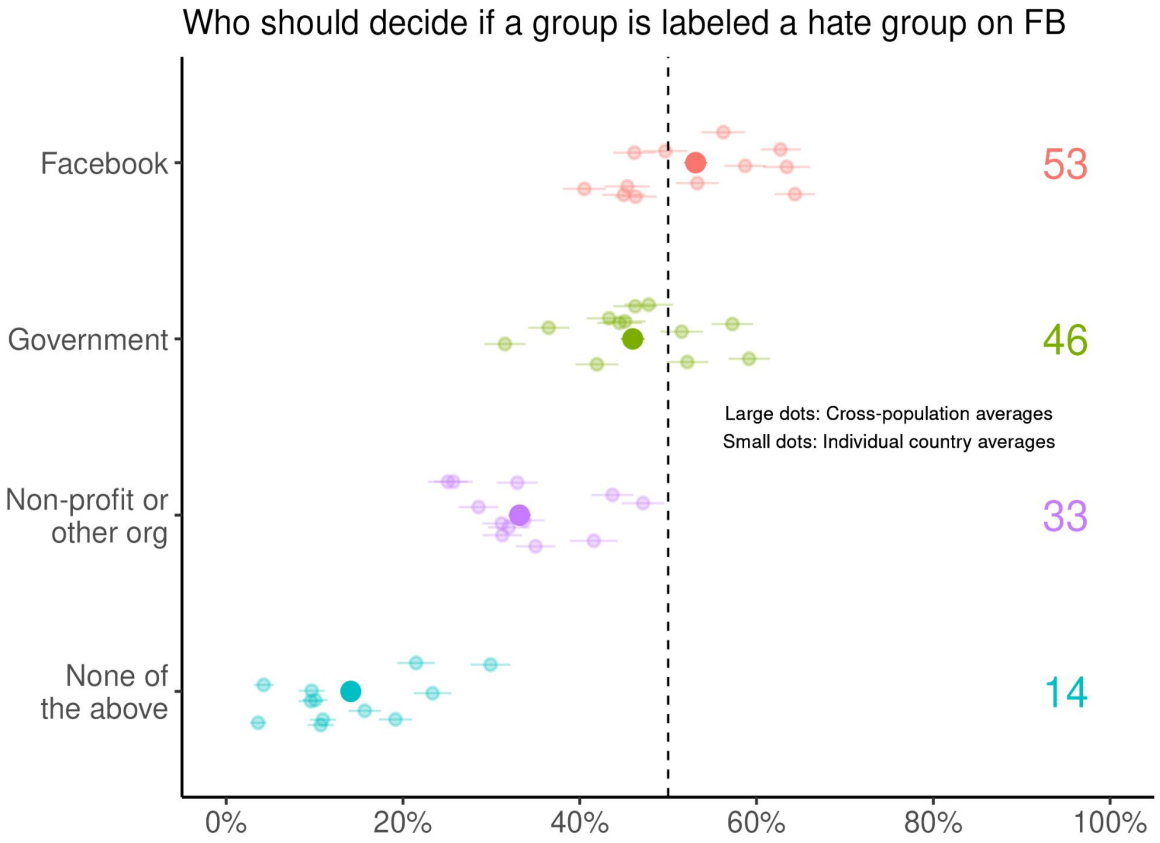
- Stakeholders are divided on the extent of importance regarding the concepts of recency and frequency. However, they agree that recency should be longer than 6 months.

Hate Entity Designation Signals

Research

Key Points:

- A 12-country survey fielded for this development finds:
  - **Most people choose Facebook**—ahead of the government or an NGO—**when asked which organization(s) should get to decide if a group is designated as a hate group on Facebook**.
  - People did not perceive a hateful post from 3 years ago differently from a more recent post.
  - A majority of people in the 12 countries surveyed said that a Facebook post representing or praising a known hate-entity counts as spreading hatred.



Note: Respondents could choose more than one organization.

N=21,600. YouGov adult samples from SE, UK, BR, DE, TR, FR, US, NG, MX, IN, ID, JP. Estimates weighted to represent the online populations of each country and cross-countries population. Question wording: Which organization(s) do you think should get to decide if a group is labeled as a hate group on Facebook? Select all that apply. Facebook, The government, A non-profit organization, other, none of the above.

Recommendation

Proposed Hate Entity Definition

A Hate Entity is an organization or individual that spreads and encourages hate against others based on their protected characteristics. The entity's activities are characterized by at least some of the following behaviors:

- Violence, threatening rhetoric, dangerous forms of harassment targeting people based on their protected characteristics;
- Repeated use of hate speech;
- Representation of Hate Ideologies or other designated Hate Entities; and/or
- Glorification or substantive support of other designated Hate Entities or hate ideologies.

Prerequisite: Influence and Spread of Hate

Proposed Behavior Categories

- **Violence and Harassment**
- **Hate Speech**
- **Representation**
- **Support and Glorification**

Prerequisite: Influence and Spread of Hate

The entity's influence and spread of hate is a prerequisite to designation that will be discussed in all evidentiaries.

Requirements:

Entity is an Organization

OR

Individual is a Public Figure

OR

Individual has Large Influence in Local Community

Other Considerations for Prioritization

- **On-Platform Influence**
- **Recency of Harmful Behavior**

Impact of Recommendation

Transparency and Explicability

- Better aligns the public-facing **Hate Entity definition** with the **behavioral signals** we use to designate those entities.
- **Relies on existing internal policy definitions** within signals.

Focus on Violence and Harm

- Emphasizes the strength of **violence or threats of violence** as a key signal, and expands the **range of behaviors** we consider for designation.
- Expands signals to **better capture implicit hate speech and behaviors**, such as coded hate speech proved by the whole of the evidentiary.
- **Avoids designation by association** by requiring some degree of hateful conduct by the entity under consideration.

Efficiency and Clarity

- **Simplifies** the complex status quo policy to make designations not only more efficient but also to more clearly show to internal decisionmakers how an entity's conduct falls within identified harms.


Meta