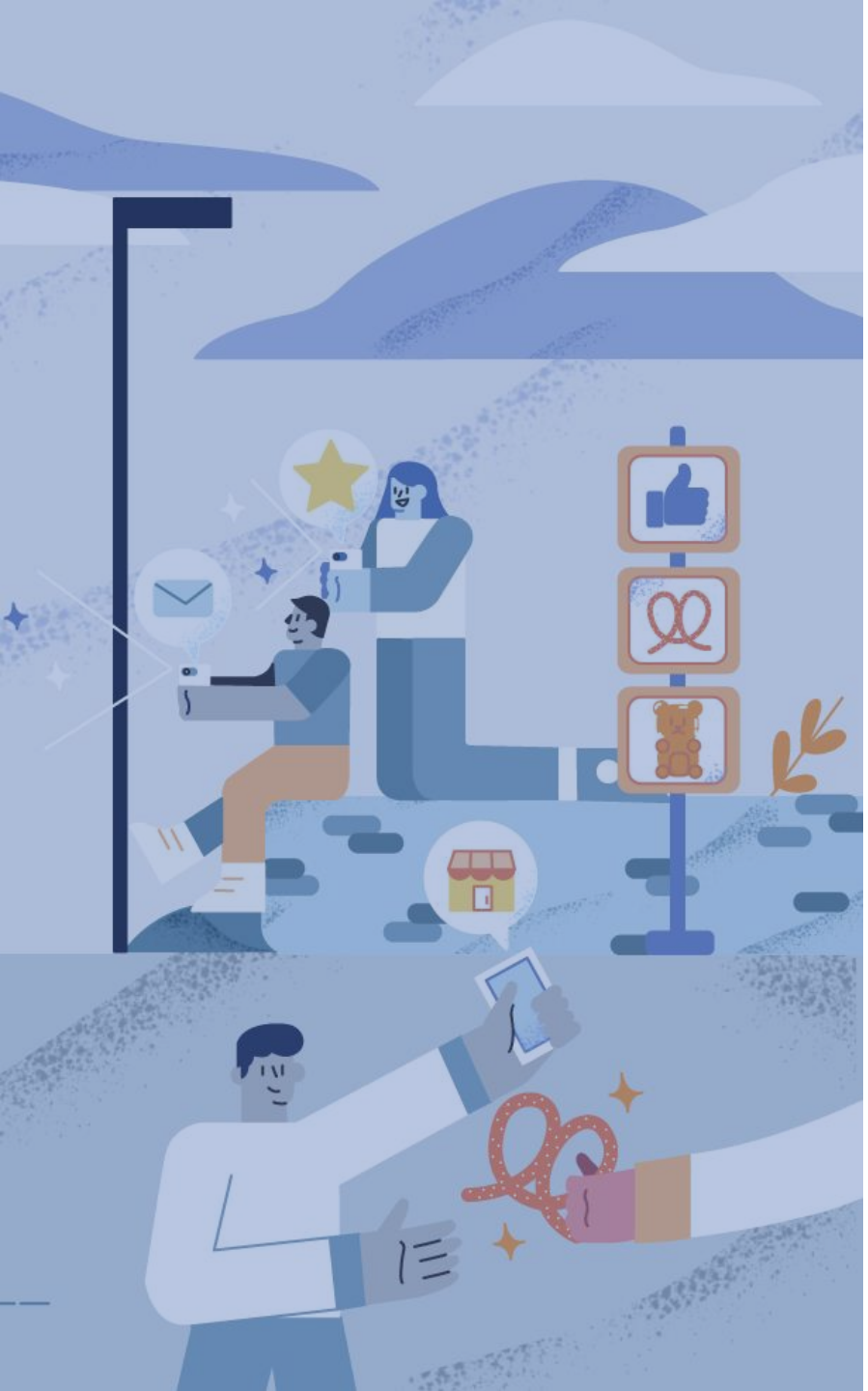
PRODUCT POLICY FORUM

facebook

August 11, 2020



- RECOMMENDATION: Harmful Stereotypes

Recommendation: Harmful Stereotypes

Organic Content Policy

Issue

Our hate speech policy aims to protect people from explicit attacks based on race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease.

While our hate speech policies apply equally to all people, we recognize that statements about specific groups of people may pose unique harm because of the way they've historically been used to attack, intimidate or exclude. That's why we prohibit oft-used hateful comparisons whether by text or image (e.g. comparing Black people to farm equipment). We want to consider the possibility of extending our hate speech policy to account for stereotypes that are similarly used in a harmful way; however, we're cognizant of the fact that it may be difficult to write and enforce on that speech without capturing speech that is not harmful.

Harmful Stereotypes

Overview

Recommendation: Expand Designated Dehumanizing Comparisons (DDCs) list to add stereotypes based on evidence that they are linked to harm or are likely to incite harm

External Outreach: 60 External Engagements

Working Groups: 9 Working Groups

Harmful Stereotypes

Status Quo: Designated Dehumanizing Comparisons

Remove

Designated dehumanizing comparisons, generalizations, or behavioral statements (in written or visual form) that include:

- Black people and apes or ape-like creatures
- Black people and farm equipment
- Jewish people and rats
- Muslim people and pigs
- Muslim person and sexual relations with goats or pigs
- Mexican people and worm like creatures
- Women as household objects or referring to women as property or "objects"
- Transgender or non-binary people referred to as "it"

Harmful Stereotypes
Status Quo - Examples

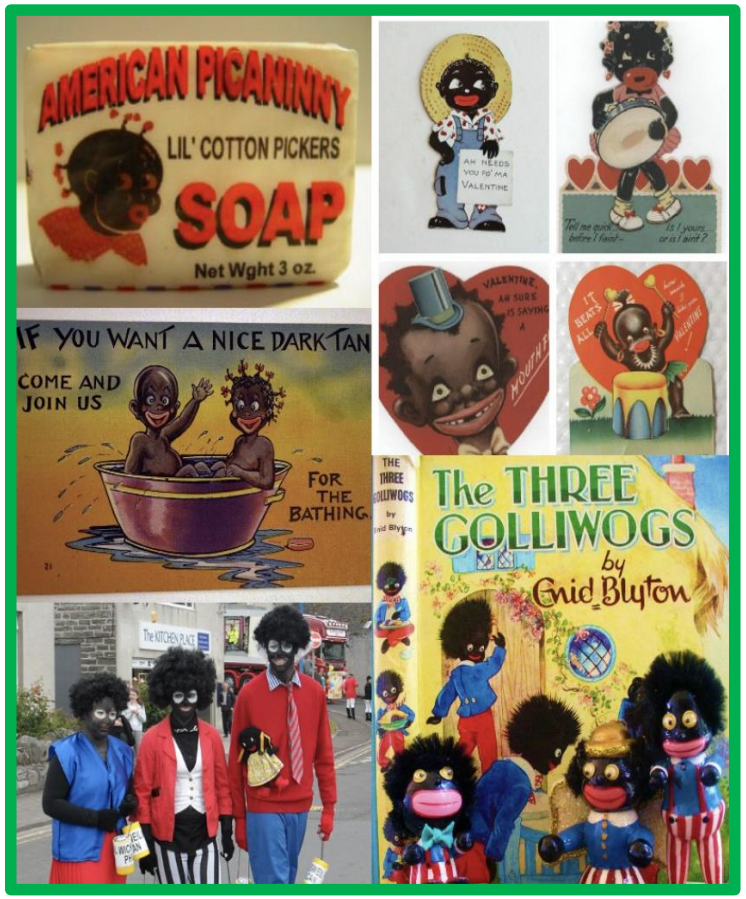
Dalits as cleaners

Allow ✓

"When your maid is on leave and you do not feel like cleaning up, invite your Dalit friend to your house"

Blackface

Allow ✓



Jews run the world

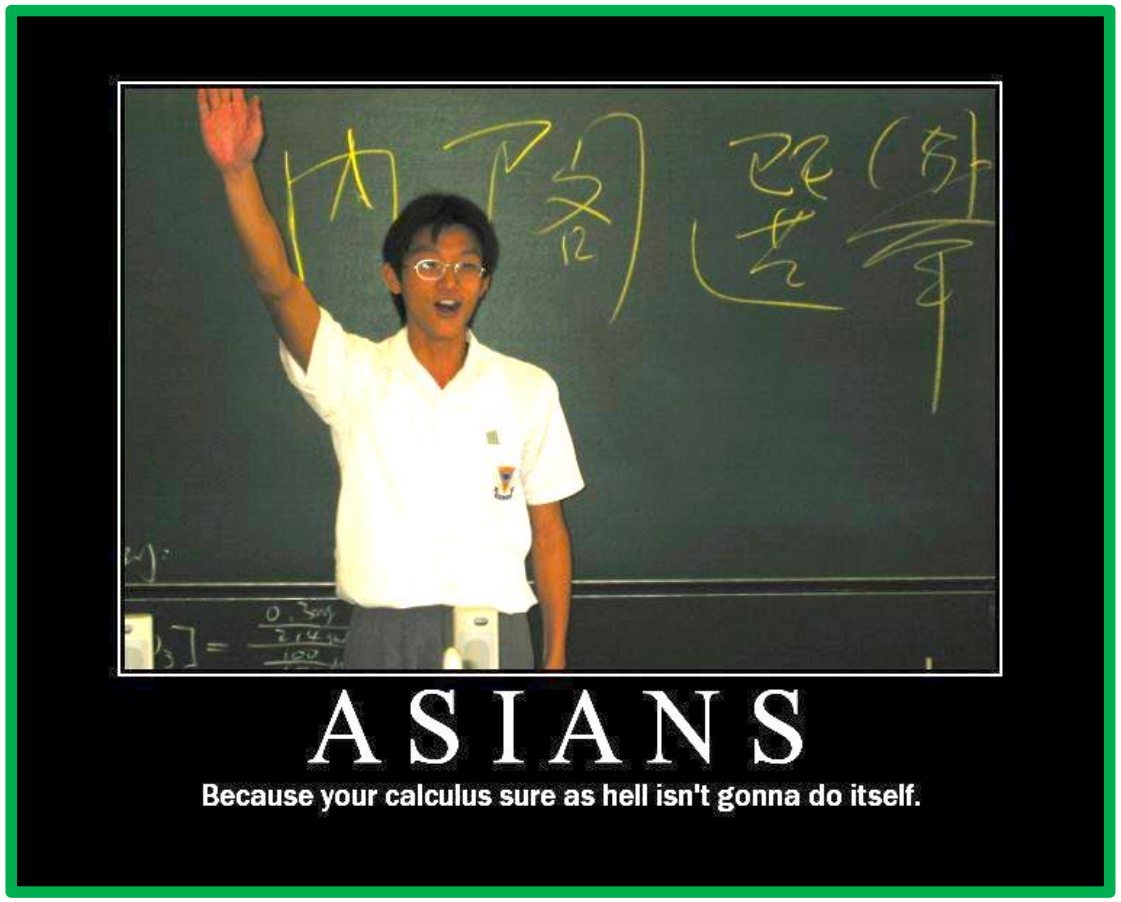
Allow ✓

Harmful Stereotypes

Status Quo - Examples

Allow

Allow



Harmful Stereotypes
Research Findings

- While there is no widely agreed upon definition of harmful stereotypes, research does suggest that they:
  - Are "cognitive schemas used by social perceivers to process information about others"
  - Often generate expectations around an individual's behavior based on observable or perceived qualities (such as race or religion) which can generate harm

- Harmful Stereotypes can contribute to othering, exacerbate bias and at worst, stoke violence against an out-group
  - In a political context, it has been associated with offline harm though direct causality is uncertain
  - Due to cultural variation, identifying global stereotypes should consider vulnerability both along PC (which ethnic group is vulnerable?) and location (in which locations are risks higher?)

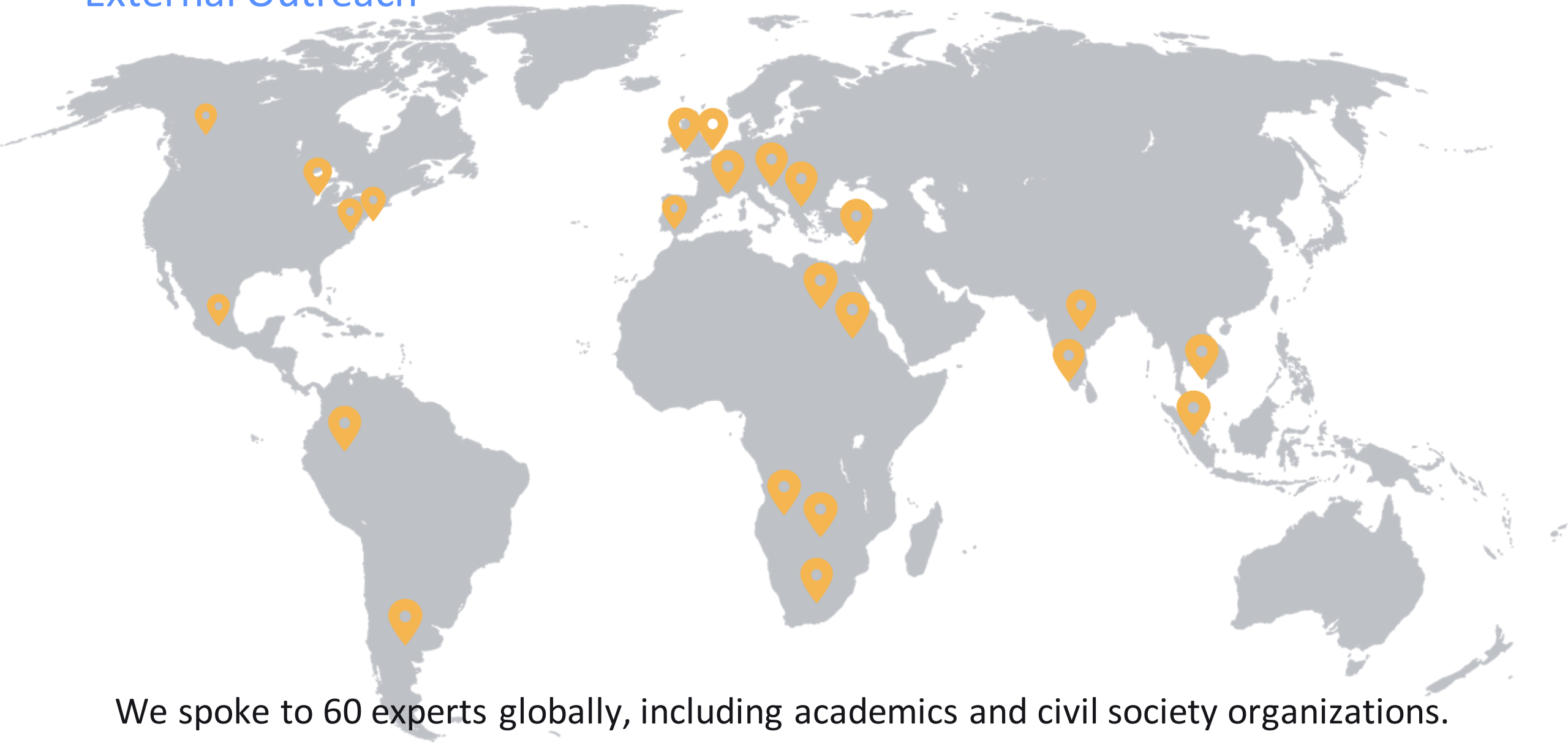
- Harmful stereotypes on the platform manifest in different formats (visual vs. text) in different countries and detection capability varies based on content, format, and language

Policy Relevance: For 'worst of the worst' examples of stereotypes, harm can be significant and safety may be improved by limiting exposure to this content. Policies should consider both dimensions of vulnerability (characteristics, location) and format in which speech occurs (text, images, memes) to support effective implementation

Sources: Dovidio et al (2010); Arendt, F., Marquart, F., & Matthes, J. (2015); Jackson, April 2018; Leavitt, July 2018

Harmful Stereotypes

External Outreach



We spoke to 60 experts globally, including academics and civil society organizations.

Harmful Stereotypes
External Outreach

Key themes:

1. Stereotypes that are harmful are the ones that cause people not to feel safe in the public domain, not to fulfill their roles and expectations as citizens, and discourage participation

2. Any assessment of stereotypes quickly confronts the issue of historical discrimination and minority vs. majority treatment; negative stereotypes are much more likely to be associated with out-group or minority members

3. Perceptions and standards change over time and considering the current public discourse and how it relates to the stereotype can help assess if the stereotype has harmful potential.

Harmful Stereotypes
Evaluating Content on our Services

We conducted a labeling exercise that accounts for reported, **non-violating speech referencing a protected characteristic** from five different countries/regions

- **Key Findings:**
  - Top targets for stereotypes **vary by country/region**
  - Most stereotypes are **regional**
  - Stereotypes are **concepts** – visuals/statements that require cultural context to understand

- **Operational Challenges:**
  - **Designation:** A thorough vetting process is required to identify the most toxic stereotypes with minimal bias
  - **Cognitive overload on reviewers:** A long list of stereotypes will be difficult to memorize
  - **Trade-off with accuracy:** The conceptual nature of stereotypes may lead to subjective interpretation, tooling solutions challenging

- **Recommendations:**
  - Designation of a limited list of the most toxic stereotypes
  - Clear guidelines to differentiate between designated dehumanizing comparisons, hateful stereotypes, and dehumanizing characteristics

Harmful Stereotypes
Definitions

1. A stereotype is a depiction of people that serves as a mechanism to articulate perceived threats to a community — arousing emotions and possible action
2. A stereotype is harmful when it:
    1. Targets people based on protected characteristics AND
    2. Encourages or is linked to harms perpetrated against “out-groups” in the form of intimidation, exclusion, or violence

Harmful Stereotypes
Framework

- **Nature of stereotype**:
  - Does the speech serve as a narrative to articulate how a group defined by Protected Characteristic(s) presents a threat/loss of status to another group?
  - Does the speech "lower the cost" of perpetrating intimidation, exclusion, or violence against the group by depicting them as sub-human, inferior, or threatening?
- **Historical context**:
  - Has the speech been linked to episodes of intimidation, exclusion, or violence perpetrated against a group defined by Protected Characteristic(s)?
- **Link to incitement of harm**:
  - Is the speech being used during a period of heightened societal tension such as war, ethnic conflict, economic crisis, or election?
  - Is the speech exceptionally prevalent or has it been affirmed/used by public figures, social influencers, or mass media?

Harmful Stereotypes
Status Quo

Option 1

On escalation, Organic Content Policy can make relevant spirit of the policy exceptions to address content that is leading to harm

Pros

- **Operability**: Easy to operationalize

Cons

- **Safety/Dignity**: Allows many types of harmful content that may/have lead to harm

Harmful Stereotypes
Expand Existing List of DDCs

Option 2 (Rec)

Expand our list of Designated Dehumanizing Comparisons (DDCs) to capture commonly used global stereotypes, beginning with:

1. Blackface
2. Jews run the world or control its major institutions

Pros

- **Safety/Dignity**: Removes more content that leads to harm and incitement to harm
- **Explicability**: Additions to the list are grounded in stakeholder feedback

Cons

- **Voice**: Removes more speech, creates the possibility of increased false positives, and may sweep up cultural images that are significant in different countries

Harmful Stereotypes
Create a Global List

Option 3

- Create a new list of Designated Dehumanizing Comparisons (DDCs) to capture global stereotypes and rename to “Designated Dehumanizing or Inciteful Speech”
- Build out a vetting and designation process for Organic Content Policy to assess items for inclusion on an ongoing basis

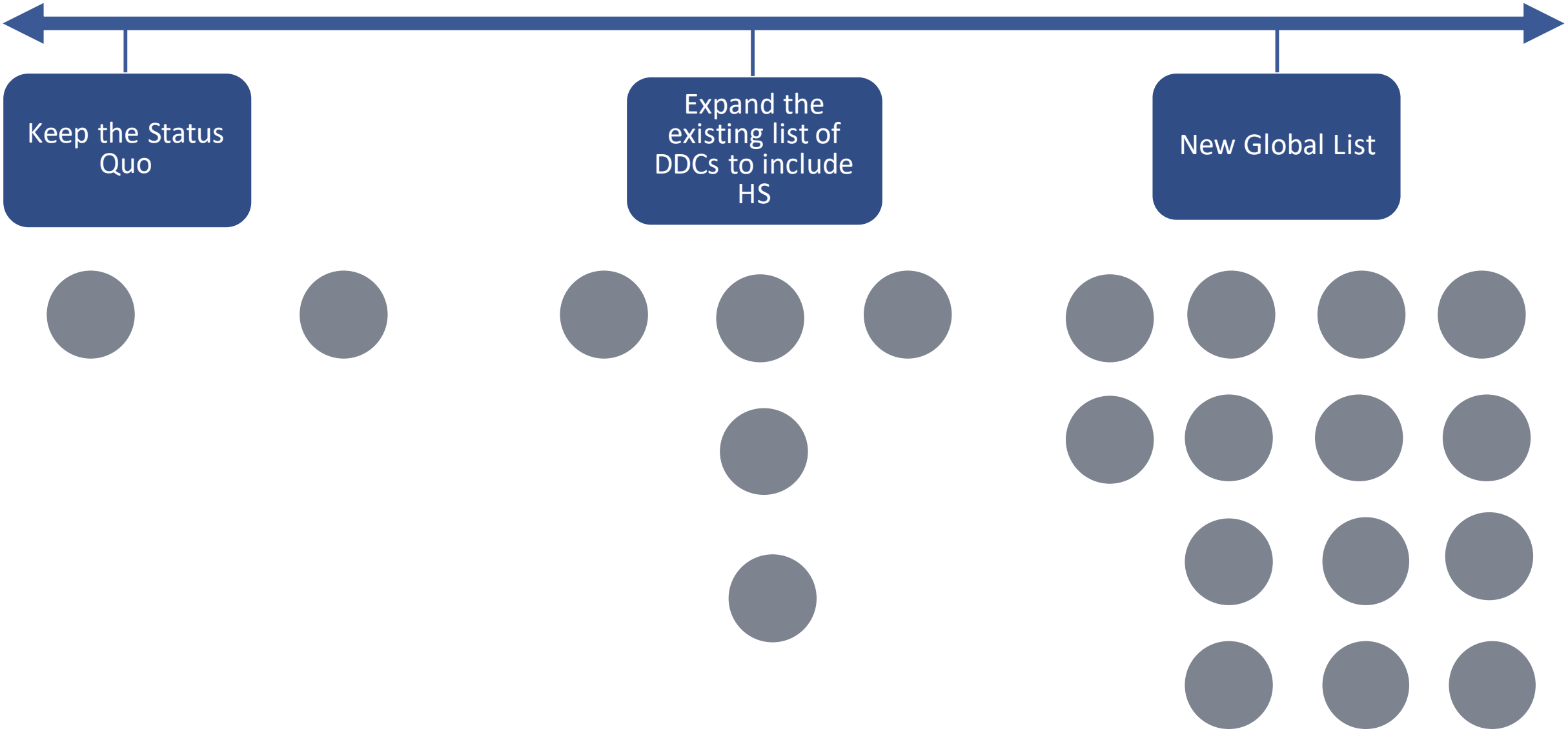
Pros

- **Safety/Dignity**: Will help ensure the list is robust, up to date, and reflective of issues facing our global user base

Cons

- **Voice**: Removes more speech, and creates possibility of increased false positive
- **Operability**: Enforcement complexity may override benefits

Harmful Stereotypes
External Outreach



Next Steps

Developments

Fine-tune operational language associated with each stereotype

Update messaging sent to users who violate policy

Comms

Share policy publicly and update Community Standards

Launch

Proposed launch date of Q2-Q3 2020

facebook