Content Standards Forum
March 26, 2019

TOPICS: White Nationalism & Separatism, Tier 3 Exclusion
<table>
<thead>
<tr>
<th></th>
<th>Agenda</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RECOMMENDATION: WHITE NATIONALISM AND WHITE SEPARATISM</td>
</tr>
<tr>
<td>2</td>
<td>RECOMMENDATION: TIER 3 EXCLUSION</td>
</tr>
</tbody>
</table>
Our discussion with 20 experts across the globe and our own research shows that white nationalism & white separatism is tied to organized hate groups.
This slide illustrates the way that we currently handle white nationalism and white separatism on our services.

Under our current hate speech policy, we allow content that promotes nationalism and separatism. That said, under our dangerous orgs policy, we consider promotion of an ethno-state as a signal in determining whether a person or group amounts to a hate org. The signal alone will not lead to designation, but it’s one factor in our analysis.

We also remove symbols and slogans that are connected to designated hate entities.
This is a sample of the signals we consider when determining whether a person, group, band, or institution amounts to a hate entity.

- Level one signals are the strongest, and level three signals are the least strong. In determining whether someone amounts to a hate entity, we require a combination of signals. We don’t share the specific requirements externally because we don’t want people to be able to game the system or subvert our policies.

As part of this policy development process, we reviewed hate figures and organizations – as defined by our Dangerous Individuals & Organizations policy – which revealed the overlap between white nationalism and white separatism and white supremacy and organized hate groups.
Option 1: Status Quo

Allow praise, support, and representation of white nationalism and white separatism.

**PROS:**
- Allows for the argument that white nationalism – like any form of nationalism – could be about something other than hate.
- Matches with our hate speech policy of allowing calls for a space limited to one PC only.
- Bright line rule that can be enforced consistently.

**CONS:**
- Allows what most experts think is organized hate.
Option 2: (Recommendation for discussion) Ban Explicit Promotion of White Nationalism & White Separatism

Remove explicit praise, support, and representation of “white nationalism” and “white separatism” under the Dangerous Organizations and Individuals policy.
Example: “I am a proud white nationalist.”

**PROS:**
- Accounts for expert view that WSN is organized hate.
- Will not affect other types of nationalism/separatism that aren’t inherently hateful.
- Easier to implement and enforce globally than Option 3.

**CONS:**
- Disagreement on whether WSN is inherently hateful.
- Does not capture implicit expressions of white nationalism and separatism.
Both examples on the left are similar: they are explicitly promoting or representing white nationalism – they would come down. But, if the group featured in the lower left image were titled, “I love America” though, it wouldn’t be implicated by this policy option.

The content on the right may be deeply offensive to some, but because neither of those two examples explicitly promote white nationalism or separatism, they wouldn’t go against Option 2.
Option 3: Ban Promotion of White Nationalism & White Separatism Where There is Criticism of a PC

Remove content that praises, supports, or represents “white nationalism” and “white separatism” (WNS) if the content (a) explicitly promotes WNS; and (b) criticizes a group/individual based on a protected characteristic.

Example: “Asians are taking over all of our jobs, the only way to fix this is through white separatism...”

**PROS:**
- Targets WNS content where hateful intent is clear.
- Removes content that is most often associated with designated hate entities.
- Scope of policy is more precise.

**CONS:**
- Testing shows that this would have minimal impact – WNS posts often don’t include this context. Especially noticeable in regions that have known WNS groups.
- Does not address implicit expressions of these ideas.
- Disagreement on whether WNS are inherently hateful.

This option is more targeted than Option 2 in that it requires promotion of white nationalism and white separatism be accompanied by criticism of a group or individual based on a protected characteristic. Under this policy proposal, criticism would be defined as negative sentiment (we would, of course, have to be more specific if we were to write operational guidelines on the basis of this policy).

As noted among the cons, this policy option would have a very limited impact because it doesn’t reflect the way people actually talk about these concepts.
The example of the left would violate under this policy proposal because two things are present: 1) explicit representation of white nationalism, and 2) criticism of a group of people who share a protected characteristic (“colored people and women”).

On the right, however, the examples featured would stay up, in large part because the second of the two criteria isn’t met.
White Nationalism & White Separatism
External Input on “White Nationalism” & “White Separatism”

• Almost every external expert agreed that “White Nationalism” & “White Separatism” are synonymous with organized hate.

• All experts agreed that the term “white nationalism” currently takes on negative connotations.

• All experts agreed that the negative connotations connected to both terms are hateful and may lead to offline harm.
Calls for Exclusion

External Outreach

We spoke to 25 experts globally, including academics, civil society & legal experts.
White Nationalism & White Separatism
Snapshot of External Outreach

**Option 1: Status Quo**
- Civil Society (US)
- Civil Society (US)
- Civil Society (US)
- Civil Society (EMEA)

**Option 2: Remove Explicit White Nationalism/ Separatism**
- Civil Society (US)
- US Legal Consultant
- Civil Society (US)
- Civil Society (US)
- Academic (EMEA)
- Civil Society (EMEA)
- Civil Society (EMEA)
- Academic (EMEA)

**Option 3: Remove Explicit White Nationalism/ Separatism + Criticism of PC’s**
- Academic (US)
- Legal Expert (US)
- Academic (EMEA)
**Discussion**

**Question:** Did experts weigh in on whether white nationalism and separatism is different than black nationalism and separatism?

**Answer:** Experts noted that historical context is very important. Black nationalism and separatism is often the product of groups/people, who have dealt with a history of oppression, trying to support one another. We don’t see the same thing for white nationalism and separatism.

**Additional follow-up:** In thinking through this recommendation – there are actually three strands of research we identified internally that track with external feedback:
1) The framing of white nationalism is grounded in separatist conceptions which are themselves premised on the notion of removal of individuals who are non-conforming (i.e. non-white individuals). There’s a lot of scholarship on this and most
recently, JM Berger has a book on extremism that covers some of this in theory and associated history explicitly. [SOURCE: Perlman 2017, Hawley 2017, Daniels 2018, Berger 2019]

2) Separate from the history though, the behavior of white nationalist groups is actually very similar to dangerous orgs – rather than political movements. This line of research finds group ideology and behavior consistent with the belief in hostile action against an out-group - more like something between terrorist groups and gangs; rather than movements for political rights. [SOURCE: Reid & Valasik 2018a; 2018b; Simi 2006; Reitman 2018]

3) These groups are really distinct from other types of separatists movements and so it’s worth looking at them separately. Some of the broader research covers exclusion and separatism in the context of deprivation (i.e. preventing social/economic/political rights) vs using it for support. Natalie will cover this work more later. [SOURCE: Kaber, 2000; Pearce, 2001; Leader 2009; Chakraborti, 2015]

Comment: I see APAC was consulted on this policy change, but I don’t see anyone from our ANZ teams on the list. In the wake of the horrible tragedy in New Zealand, I would suggest we reach out to regional public policy to get their input

Response: Yes we can follow up. As I mentioned, our external engagement focused on the parts of the world where white nationalism and white separatism has historically been more prevalent, but you’re right that we should talk to people in ANZ as we move forward with next steps.

Question: Right now, white nationalism isn’t something we account for in our designation signals. Is this something we’re considering, particularly after New Zealand?

Answer: Yes, and I think this is a good segue way into our next steps slide because we talk about the work we want to continue to do to build upon today’s policy recommendation. We’ll be
undertaking three specific next steps here:

1) To the question just asked, we’ll be assessing how the new policy will affect our hate designation analysis. We already account for the call to create an ethno-state as one of our signals in designating hate figures and organizations, but should we also be accounting for white nationalism in some way.

2) We’ll also be working to understand the hate slogans and symbols affiliated with nationalism and separatism so we can further strengthen enforcement of the policy.

3) And we’ll be identifying other hateful ideology that are associated with designated hate entities to see if they should also fall within our dangerous orgs policy.

**Question:** Does this policy extend beyond white nationalism and separatism?

**Answer:** This policy is specific to white nationalism and separatism. As part of the third of our “next steps, we’ll be looking at other hateful ideologies to see if they also qualify for treatment under our dangerous orgs policy.

**Question:** What steps have you taken to remediate against white nationalism when they use different terms?

**Answer:** As part of next steps, we will be looking at updating slogans and symbols, which may reveal use other words.

**Comment:** Earlier you said, the term “white nationalism” has evolved. I think the term has always meant the same thing, but our understanding of it has evolved.

**Response:** That’s a great point. You’re right that the meaning of the term hasn’t evolved, but people’s use of it has – and people on the platform (and in offline contexts) are steering away from the term “white supremacy” and isn’t speaking of white nationalism. This is
where the evolution has been.

**Comment:** Some organizations do see this as progress, but they won’t be completely satisfied. There is still a difference between option 2 and what they want to see on the platform.

**Response:** We recognize that, and we’re going to continue looking at the edges of our policy to see where we can make changes.

**Additional follow-up:** Also worth noting that we aren’t just tackling this with policy. When people search for terms associated with white nationalism/separatism/supremacy, they will be redirected to a counter speech organization.

**Comment:** In general, it sounds like people are in agreement on the policy recommendation. Any concerns? Any questions? If not, let’s consider this policy passed. Please feel free to reach out to the content policy subject matter experts as they take on next steps.
Refine: Calls for Exclusion
Calls for Exclusion

Overview

**Issue:** We've historically not allowed people to exclude based on a protected characteristic ("no Jews allowed") because of the often-hateful context, but we've allowed limited-inclusion groups ("Religious study group – Muslims only") because of the many positive examples we see. The line is not perfect, and we'd rather focus our policy on the intent of the exclusion or limited inclusion; however, a more nuanced policy could be hard to enforce or lead to claims of bias.

**Working Groups**
- 6 XFN working groups
- 27 external engagements
- Enforcement consistency data collection

**Recommendation**
- Continue to allow limited inclusion but not where there is negative context.
- Create categories to clarify the limits of exclusion for more consistent enforcement.
- Expand the carve-out that allows gender exclusion in the context of health or support groups.
Calls for Exclusion
Status Quo – Hate Speech Tier 3

• We do not allow calls to exclude or segregate a person or group of people based on their protected characteristics.
• We have, however, created a carve-out that allows for gender-based exclusion in health and positive support groups.
• We allow criticism of immigration policies and arguments for restricting those policies.
• We allow statements of limited inclusion based on protected characteristics.

<table>
<thead>
<tr>
<th>Leave Up</th>
<th>Remove</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Zulus only in this group</td>
<td>• No Zulus allowed in this group.</td>
</tr>
<tr>
<td>• Candid talk about breastfeeding – Women only</td>
<td>• No women allowed</td>
</tr>
<tr>
<td>• Candid talk about breastfeeding – No men please!</td>
<td>• Muslims should not be allowed to vote.</td>
</tr>
<tr>
<td>• Migrant refugees should not be allowed in Germany.</td>
<td></td>
</tr>
</tbody>
</table>
Calls for Exclusion
Gender carve-out in health & positive support groups

• Allow gender-based exclusion only in groups discussing health or positive support related issues:

  • **Health** is defined as:
    - Serious diseases or disabilities issues (e.g. cancer, HIV)
    - Birthing/Postpartum issues “Lesbian Doula Group. No trans allowed”
    - Gender confirmation issues

  • **Positive Support** is defined as:
    - Groups discussing parenting issues, for example “Mothers’ Group. No Men allowed”
    - Groups discussing addiction recovery issues, for example “Mens’ Alcoholics Anonymous. No Women”
    - Groups discussing issues in the workplace, for example: “Trans Managers Empowerment Group. No Men Allowed”
These are all examples that we don’t allow under our status quo policies:

- Upper left: in response to an article about transgender individuals in the military, a commenter writes that no queers should be in the military, which violates as a call for exclusion.
- Upper right: a post calling for the deportation of all white people from South Africa.
- Lower right: a post calling for the exclusion of Muslims from public office.
Calls for Exclusion
Product Policy Research Findings

- Social Science research-based definitions of exclusion define it in terms of economic, social, and political deprivations intended to negatively impact intrinsic rights, security, and/or well-being.

- However, external research finds that some private and exclusionary groups are beneficial:
  - Specific empirical examples in the context of female collective action and safety as well as LGBT+ identity and support
  - Exclusion definitions should consider the distinction between exclusion to address marginalization as distinct from exclusion to aggravate or encourage marginalization.

- In the context of setting hate speech rules, we believe that providing individuals who violate our rules with greater transparency and information on our rules can improve outcomes: it can lead to an increase in future rule compliance (i.e.: a reduction in future violations) and reduce negative experiences on the platform.
Three Part Approach

1. Part One – Limited Inclusion
   
   **Recommendation: (Option 3) Allow limited inclusion except in negative context**

2. Part Two – Refined definition of exclusion
   
   – *Divide exclusion into 4 categories for consistency*

3. Part Three – Gender carve-out
   
   – *Continue to allow the Carve-out for Gender in health/support*
Part 1 – Limited Inclusion
OPTION 1: Ban Limited Inclusion Unless Meets Specific Carve-Out

Treat limited inclusion the same as exclusion and ban both except in the following allowable contexts:

- **Health support** (e.g., “Breast cancer support. Women only. No men allowed”)
- **Places of worship and religious communities** (e.g., “Prayer Requests - Christians Only”)
- **Dating** (e.g., Christian Mingle, JDate, black singles group)
- **Support groups** (e.g., “Ladies who Launch. Women's only tech support group”)
- **Parenting groups/conversations** (e.g., “Mommies' Group - women only”)

**Pros:**
- Treats PCs the same.
- Expansion of current carve-out policy that allows gender-based exclusion in the health and positive support context.

**Cons:**
- Allowable carveouts may seem arbitrary or not relevant to certain speech (E.g., Israelis who want to talk about Palestinian conflict / Indians who want to talk about Kashmir conflict).
- Carve-outs allow for bias.
- We won't capture all acceptable use cases globally.
OPTION 2: Ban Limited-Inclusion Groups and Pages Where There is Negative Context or No Context

- **No context**: “Straights only” = Remove
- **Negative* context**: “Straights only, because we know we are better” = Remove
- **Positive context**: “Straights only, to discuss infertility with your partner” = Allow

*“Negative” context means it falls within the spirit of our hate speech policy, such as claims of PC superiority

**Pros:**
- Focus would be on context

**Cons:**
- Will be hard to define “negative context” to cover all types of hateful content globally
- We’ll have to define what we mean by “context” and what context can be taken into account (other content on Group/Page/account? Market knowledge? etc.)
- Default to removal without context may remove non-hate content
- Operationally challenging
OPTION 3: Ban Limited-Inclusion Groups and Pages Where There is Negative Context [RECOMMENDATION]

- **No context**: “Straights only” = **Allow**
- **Negative context**: “Straights only, because we know we are better” = **Remove**
- **Positive context**: “Straights only, to discuss infertility with your partner” = **Allow**

**“Negative” context means it falls within the spirit of our hate speech policy, such as claims of PC superiority**

- On Escalation: Remove Inclusion Groups and Pages when the implied intent is PC superiority

**Pros:**
- Focus would be on negative context rather than the PC so more objective.
- Limited to Groups and Pages would ensure benign content like “Girls night out! We are the best” is not removed.
- Default to allow without context reduces erroneous non-hate removals.

**Cons:**
- Will be hard to define “negative context” to cover all types of problematic content globally.
- We’ll have to define what we mean by “context” and what context can be taken into account (other content on Group/Page/account? Market knowledge? etc.).
- Operationally challenging.
This gets to the core of content that many of the internal/external stakeholders we engaged with were concerned about

- Groups/Pages that appeared to be promoting/advocating white supremacy or WNS (implicitly if not explicitly
- One message we heard repeatedly in our external engagements was that exclusion/limited inclusion just feels different depending on the groups/power dynamics (e.g. there are reasons why a “black only parenting group” might be benign/necessary, but ”whites-only parenting group” didn’t feel the same

Ultimately, we will be handling “whites only” content under our Dangerous Orgs policy as indicative of a hateful ideology, so we decided not to move forward with this option
In our working groups, we also discussed the need to be more precise when it comes to impermissible exclusion, which is what part 2 covers.
Calls for Exclusion
Inconsistent Enforcement

- Worked our operations team to understand consistency and accuracy in enforcement across all types of hate speech.

- For tier 3 exclusion, consistency for is lower than it is for other types of hate speech, and the rate of false positives is higher as compared to other types of hate speech.

- We think that by more clearly defining exclusion and segregation under tier 3 of our hate speech policies, we can help improve consistency and accuracy.

- To do this, our Community Operations team undertook a labeling exercise to help us categorize the types of exclusion we see on the platform, which are discussed in detail on the next slide.
There are, of course, pros and cons to coming up with a more specific definition of exclusion.

**Pros:**
- Gives clear guidance as to what constitutes exclusion
- Removes the worst types of exclusion
- Allows for many types of socially acceptable exclusion (”no kids allowed

**Cons:**
- Opens the door to many types of exclusionary speech.
• May still remove many types of political speech

We are starting here now, but will continue to talk to external experts and evaluate content on the platform to see if we need to add categories of exclusion to this list.
Part 3 – Gender Carve-Out
Calls for Exclusion
Part 3: Gender Carve-Out in health & positive support groups

- Allow Gender based exclusion only in Groups discussing **health or positive support** related issues:
  - Health is defined as:
    - Serious diseases or disabilities issues (e.g. cancer, HIV)
    - Birthing/Postpartum issues “Lesbian Doula Group. No trans allowed”
    - Gender Confirmation issues
  - Positive Support is defined as:
    - Groups discussing parenting issues, for example “Mothers’ Group. No Men allowed”
    - Groups discussing addiction recovery issues, for example “Mens’ Alcoholics Anonymous. No Women”
    - Groups discussing issues in the workplace, for example:
      - “Trans Managers Empowerment Group. No Men Allowed”
      - **Groups discussing domestic/emotional abuse, for example “Men’s Domestic Abuse Support Group, No Women allowed”**

As we discussed, our status quo policy currently includes a carve-out for gender-based exclusion in health and positive support groups. We have defined health and positive support very clearly, but over the course of our working groups, we identified an additional use case that amounts to “positive support” – groups discussing domestic/emotional abuse - and want to include it in the carve-out.
Calls for Exclusion
External Outreach

We spoke to 27 experts globally, including academics, LGBTQ advocates, digital rights and women’s safety organizations, and Group admins.
Calls for Exclusion
Snapshot of External Outreach

Most permissive:
Allow broad range of limited inclusion

Intermediate:
Allow minorities more room for limited inclusion

Least permissive:
Treat in-group behavior as suspect

Group Admin (US)
Group Admin (APAC)
Academic (US)
Civil Society (LATAM)
Civil Society (EMEA)
Academic (EMEA)

Group Admin (US)
Group Admin (APAC)
Academic (US)
Civil Society (LATAM)
Civil Society (EMEA)
Academic (EMEA)

Group Admin (US)
Civil Society (US)
Group Admins (US)
Academic (US)
Civil Society (LATAM)
Civil Society (EMEA)

Civil Society (US)
Academic (EMEA)

Civil Society (US)
Civil Society (EMEA)

Civil Society (LATAM)
Discussion
Any questions/feedback on this specific policy.

Question: Did we consider using the *bona fide occupational qualification* standard that applies in employment discrimination cases. For example, if X airline posted a flight attendant description that said, “Women only” or “Women only, to ensure passengers have a pleasant experience,” that would satisfy the recommendation because there was no negative context. But legally, it would be insufficient, as passengers can have a pleasant experience with men.

Answer: Determining whether or not there is a "bona fide" reason for limiting a space to people of a particular protected characteristic would require reviewers to make a subjective interpretation about whether or not there is a
"reason" for the exclusion (which opens the door to bias, inconsistency, etc.). But that’s why we have things like carve-outs for situations in which we feel pretty confident that there is a bona fide reason for limiting membership, however, such as Groups/Pages discussing health issues.

**Question:** Was there consensus internally?
**Answer:** Overwhelmingly, yes, particularly when looking at the three parts of the recommendation as a whole.

**Question:** Was the research we did commissioned for this particular recommendation, or is it ongoing? And if the latter, how do we plan to use it.
**Answer:** The research we presented today was analyzed specifically for this policy proposal, but the underlying internal research (on how we define exclusion in context and what that means in different settings) is continuing both in the context of hateful/harmful speech research internally and through supporting external research in this area.

**Comment:** Thank you all for your participation and discussion. Based on general consensus in the room, let’s consider this recommendation passed. Please reach out with any questions.
facebook