
facebook

Content Standards Forum

January 29, 2019

TOPICS: *Human Trafficking, Gender and Hate Speech*

Agenda

1. RECOMMENDATION: HUMAN TRAFFICKING
2. RECOMMENDATION: GENDER AND HATE SPEECH

RECOMMENDATION:
Human Trafficking

Human Trafficking

Overview

Issue: Our Community Standards prohibit human trafficking but do not account for the full scope of prohibited human trafficking-related behavior. This means there's some harmful activity (e.g. illegal adoptions, domestic servitude) that we haven't captured in our policies and aren't enforcing on. A more explicit policy with corresponding operational guidelines will capture a broader ranger of harmful activities; however, too much nuance in the policy could create confusion and result in unintended enforcement.

Summary to date:

- Held 4 internal cross-functional working groups
- Consulted with 11 external stakeholders

Recommendation:

Create a new standalone Human Exploitation Policy and establish clear operational guidelines that cover the wider spectrum of harmful behaviors associated with human trafficking and human smuggling.

Human Trafficking

Status Quo

Dangerous Organizations and Individuals Policy

- Human trafficking groups and their leaders may not have a presence on the platform
- We remove content that praises supports and represents these groups and individuals

Coordinating Harm Policy

- We remove statements of intent, calls to action, or content coordinating human trafficking

Our definition of "human trafficking"

Human trafficking includes the recruiting, transporting, or harbouring of people by means of threat, coercion, or fraud for the purpose of exploitation. That exploitation can come in many different forms, including sexual exploitation, slavery, servitude, or the removal of organs.

Our definition of "human smuggling"

Human Smuggling is the procurement or facilitation of illegal entry into a state across international borders by a person that is neither a citizen nor a resident of that state for the financial or material gain of others.

In addition to the fact that our policies don't currently account for the full spectrum of trafficking-related behaviors, human trafficking is included in multiple places in our Community Standards, which may contribute to enforcement challenges. There's also room for more granularity in our definitions of human smuggling and trafficking.

Human Trafficking

Recommendation: A new dedicated policy section that covers all forms of human exploitation.

Pros

- Prevents potential offline harm.
- Enables robust enforcement against more forms of human exploitation.
- Grouping policy in this way will make it easier to understand our policy.
- Shows improved understanding of human trafficking and the forms it takes on the platform.
- Opportunity to increase awareness and reporting of human exploitation.

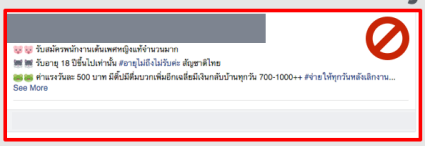
Cons

- A more expansive and detailed definition may be more challenging for reviewers to retain and may contribute to operational challenges.

Human Trafficking

Examples

Recruitment for the sex industry



Child/organ selling

Baby selling

Domestic servitude

Forced marriage

All of these examples violate our current policy against human trafficking, but may cause confusion among content reviewers because our definitions of human trafficking and smuggling don't explicitly reference things like domestic servitude, the sale of children and organs or forced marriage.

Human Trafficking

Option 1 – No changes to policy language but expand on operational guidelines

Pros

- Expansion of operational guidelines would enable reviewers to more accurately assess potentially harmful content.

Cons

- Updating our operational guidelines, without simultaneous update of our Community Standards creates a disconnect and undermines efforts at transparency.
- A more expansive and detailed definition may be more challenging for reviewers to retain and may contribute to operational challenges.
- If we don't expand on our definitions of human trafficking and smuggling (and only focus on operational guidelines), we may end up enforcing on content inconsistently.

As part of our internal and external working group process, we evaluated several policy options. Under Option 1, we would maintain the status quo policy, but provide content reviewers with a more comprehensive set of operational guidelines to help them identify additional forms of human exploitation. We decided against this option because it would create a disconnect between the Community Standards and operational guidelines, thereby undermining our efforts at transparency with updated Community Standards.

Human Trafficking

Option 2 – Expand existing policy language to include all forms (but not stages) of human exploitation under Coordinating Harm

Pros

- Shows improved understanding of human trafficking and the forms it takes on the platform.
- Improves enforcement because of parity between policy and operational guidelines.

Cons

- Limited scope of policy could leave up content leading to real world harm.
- A more expansive and detailed definition may be more challenging for reviewers to retain and contribute to operational challenges.

We also considered the option of expanding our policy language to include all forms and stages of human exploitation, but to leave human trafficking and smuggling under the Coordinating Harm section of our Community Standards. We decided against this option because it doesn't convey our understanding of the complexity and nuance of human trafficking behaviour. Having a distinct policy in place will be easier for people to understand and easier for reviewers to make sense of as they make decisions about content.

Human Trafficking

External Outreach

- All external experts were in favor of creating a new section within the Community Standards to cover human exploitation. They indicated that it would bring clarity to a complex problem and highlight the diverse and nuanced nature of trafficking.

- Many experts believe that we have a responsibility to educate the public about the nature of human trafficking, and think that a standalone policy would help us achieve that goal. It may also result in increased reporting of exploitative content.

- Experts also pointed to the varying definitions of human trafficking in different countries around the world, and felt like Facebook should strengthen its own definition and promulgate its own narrative in this space.

Human Trafficking

Timeline/Next Steps

- Draft new policy language
- Work with Community Operations on updated operational guidelines

Discussion

Any additional questions from this group?

Question: How exactly will we lay out the policy language in the Community Standards. Many reporters and other people we speak to think that things like virtual kidnapping should fall within the ambit of human exploitation.

Answer: Something like virtual kidnapping would fall under our policies against fraud. We can work to be more explicit across our policies so people understand what's covered. To your point though, if we're adding an entirely new section to the Community Standards, it's something that we should communicate about beyond just the addition of policy language to the site.

RECOMMENDATION:
Hate Speech and Gender

Hate Speech and Gender

Overview

Issue: Under our hate speech policy, we remove attacks targeting people on the basis of protected characteristics, including gender. But some have suggested certain gender-based hate speech is less intense than other hate speech. Changing our policy to reduce protections might expand users' ability to discuss sensitive issues, including gender, sexuality, and sexual assault; however, it will likely lead to more misogynistic content. Alternatively, a distinction between sexes in our policy or its enforcement may raise questions of fairness and whether our policies should account for other social power dynamics, including those that exist among races, ethnicities, or nationalities.

Summary to Date:

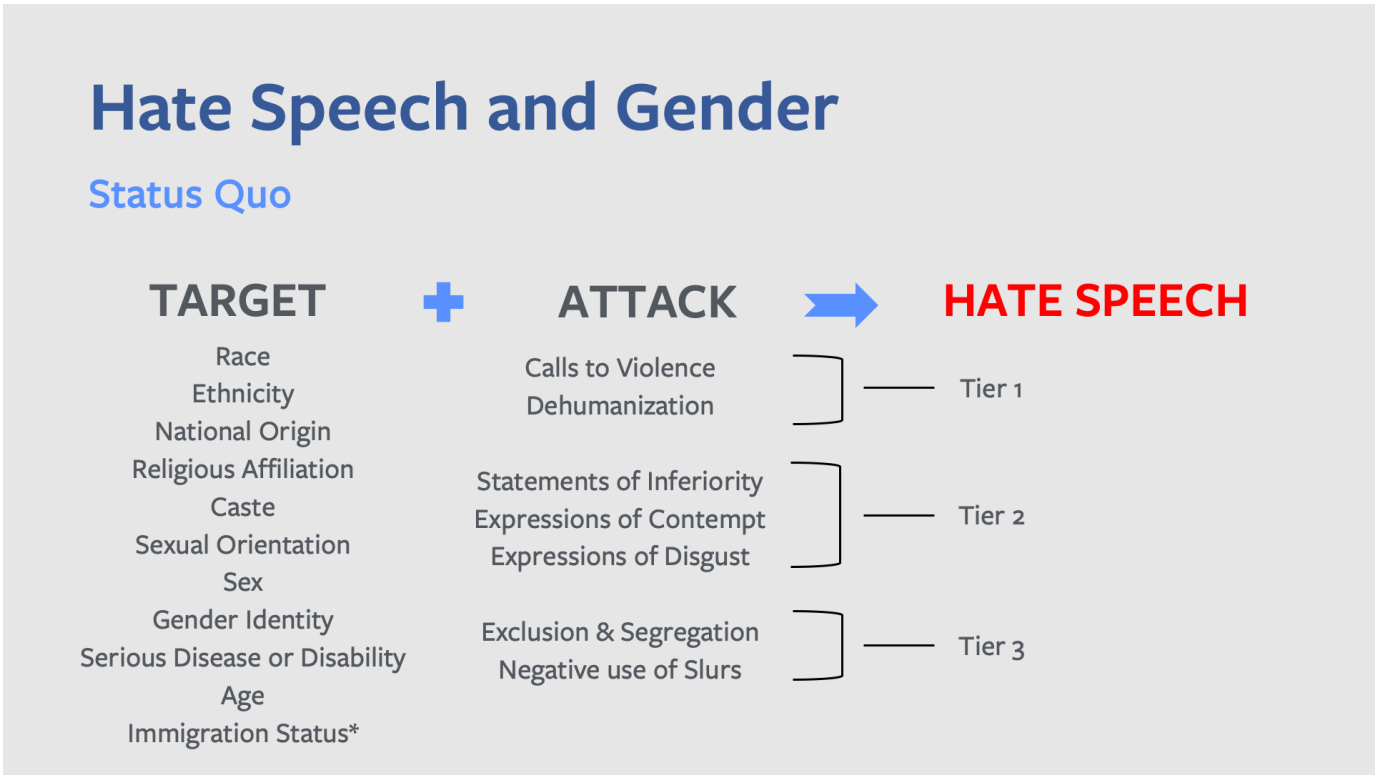
- Consulted with 24 external stakeholders.
- Convened 4 working group meetings.

Recommendation for consideration :

- Maintain status quo and continue to remove all hate speech attacks targeting people on the basis of sex/gender.

At the crux of this specific policy proposal is the question of whether gender should be treated differently than other protected characteristics. Our own research reveals that "boys are gross" is perceived as less severe than "the trans community is gross," but both statements currently violate our standards as attacks on the basis of gender. Removal of charged speech that targets gender (e.g., "men are scum") has also led to accusations that Facebook doesn't account for social dynamics.

I also want to draw attention to the fact that we're simultaneously in the middle of a few other hate speech-related working groups - namely, a proposal to re-evaluate what is and isn't included under Tier 2 attacks, gender-based exclusion in the context of groups, and cursing when it's directed at gender.



Under our hate speech policies, immigration status is considered a quasi-protected characteristic, which means Tier 1 attacks directed at someone on the basis of immigration status would be removed. We do not, however, take down calls to limit immigration.

Hate Speech and Gender

Recommendation: Remove all tiers of attack targeting people on the basis of sex/gender (Status Quo)

Pros

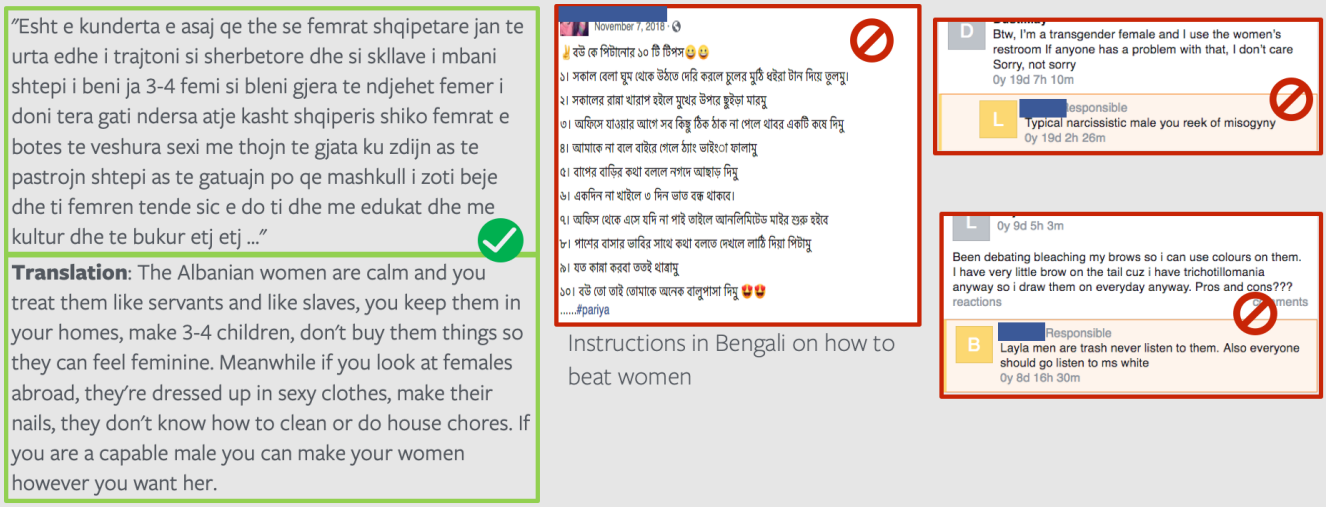
- Protects the most people from attacks.
- Accounts for data that shows there is an equal volume of attacks directed at men as women
- Treats gender the same as other protected characteristics.
- We can continue to assess newsworthiness and provide necessary guidance to our content review teams.

Cons

- Overly restrictive. False positives/edge cases suggest that Facebook doesn't account for social dynamics (e.g., "men are trash" and other #metoo takedowns).

Hate Speech and Gender

Examples



On the left: This example doesn't include a specific attack, as defined under our hate speech policies, and therefore wouldn't violate.

In the middle: Instructions in Bengali on how to beat your wife. This would violate as a Tier 1 attack (violent speech) directed at women.

On the top right: Here, "narcissistic male" would violate as a Tier 2 attack (statement of inferiority) directed at men.

On the bottom right: "Men are trash" is a Tier 1 attack (dehumanizing speech).

Hate Speech and Gender

Severity Across Tiers

	Type	Examples
1	Violent speech (t1)	"Women should die." | "Let's beat up some migrants." | "Kill the Gays."
2	Comparison to filth/disease (t1)	"Men are trash." | "Trannies are a disease." | "Koreans are scum of the earth."
3	Comparison to animals (t1)	"Men are pigs" | "Hutus are cockroaches." | "Iraqi women are cows."
4	Designated dehumanizing comparisons (t1)	"Muslims have sex with goats" | "Jews are vermin." | "Black people are monkeys"
5	References to subhumanity (t1)	"Men are the less evolved sex." | "Intersex are freaks" | "White devils." | "Native savages"
6	Mocking hate crimes (t1)	"Nothing to see here, just the shoes of people waiting to vote at Dachau death camp."
7	Physical inferiority (t2)	"Girls are ugly." | "Trans women are hideous." | "Jews are deformed."
8	Mental inferiority (t2)	"Boys are so dumb." | "All gay people are mentally ill." |"Hondurans have low IQ."
9	Moral inferiority (t2)	"Dating black women will ruin your life." | "Brown people are thieves." | "Never trust a Jew"
10	Expressions of disgust (t2)	"Ew, girls are gross." | "Bengalis are vile." | "Jews are disgusting."
11	Expressions of contempt (t2)	"I don't like men" | "I hate Muslims." | "I'm racist and proud"
12	Cursing at a protected class (t2)	"Women are bitches" | "Asshole Asians." | #fuckthegays
13	Calls for exclusion or segregation (t3)	"Women drivers should stay the fuck off the road" | "I don't want Ethiopians in our pool."
14	Slurs	"You towelhead." | "Look at the tranny." | "Faggots can't force me to do anything."

All of this content would be removed, but we've highlighted the gender-based attacks so you can see the type of content we considered as part of this policy proposal.

From the beginning, we wanted to be very careful not to allow more attacks on trans individuals, so we initially tried to limit the scope here to speech targeting "men" or "women." That said, the Community Operations team who identified and labeled content did find some attacks on transgender people. The examples highlighted here are very clear attacks on specific targets, but the Community Operation's team work indicated that it can be difficult to distinguish some attacks on trans or nonbinary individuals from other attacks on gender. This vulnerability was among the reasons that the working group ended up recommending that we stick with the status quo policy.

Hate Speech and Gender

Research Findings

Past Research

- People think hate speech against women is worse than men.
- Hate speech directed at men is only slightly less upsetting.
- Hate speech in general elicits an upsetting reaction from people.

Ongoing Research

- What type of hate speech is the most intense or severe?
- How do speaker versus audience dynamics change the perception of speech?
- What can we do to better capture the most hateful content on our services?

Discussion

It's important to note that the findings here focused on men and women, but gender-based protections under our hate speech policy also protect non-binary and trans people.

Hate Speech and Gender

Community Operations Conclusions

Evaluated ~6,800 pieces of content across 17 countries and found:

- 6 - 8 % of hateful attacks in the sample target gender, including trans and non-binary people.
- Of these:
  - 53% are dehumanizing speech
  - 18% are statements of inferiority
  - 15% are violent or supporting death/disease/harm.
- Men are more often targeted with dehumanizing speech.
- Women are more often targeted with statements of inferiority and exclusion.
- Sex is often combined with another protected characteristic or quasi-protected characteristic (e.g., Chinese men or young women).

Hate Speech and Gender

Option 1: Provide different protections to men and women

Under this proposal, we would:

- Remove all tiers of attack targeting women.
- Remove only Tier 1a (violence) attacks targeting men.

Pros

- Treating men and women differently recognizes historic imbalance and power dynamics between the sexes.
- Allows more speech on sensitive issues.

Cons

- Doesn't protect men against the hate speech that is most commonly directed at them (dehumanizing speech).
- Doesn't take into account non-binary gender attacks
- Enforcement may be perceived as inconsistent/unfair.
- May raise questions about why we treat attacks against men and women differently but don't introduce same nuance to attacks against other protected characteristics (e.g., white vs. black).

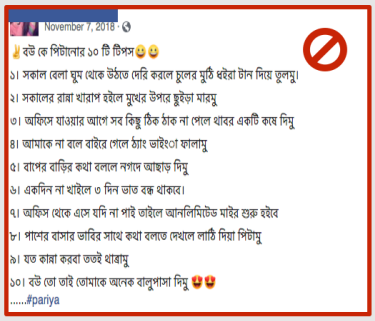
These are the options we looked at when doing external outreach and internal working groups.

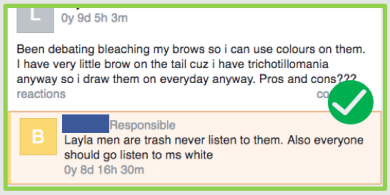
Hate Speech and Gender

Examples – Option 1

"Esht e kunderta e asaj qe the se femrat shqipetare jan te urta edhe i trajtoni si sherbetore dhe si skllave i mbani shtepi i beni ja 3-4 femi si bleni gjera te ndjehet femer i doni tera gati ndersa atje kasht shqiperis shiko femrat e botes te veshura sexi me thojn te gjata ku zdijn as te pastrojn shtepi as te gatuajn po qe mashkull i zoti beje dhe ti femren tende sic e do ti dhe me edukat dhe me kultur dhe te bukur etj etj ..."

Translation: The Albanian women are calm and you treat them like servants and like slaves, you keep them in your homes, make 3-4 children, don't buy them things so they can feel feminine. Meanwhile if you look at females abroad , they're dressed up in sexy clothes, make their nails, they don't know how to clean or do house chores. If you are a capable male you can make your women however you want her.





Here, you can see how Option 1 would apply to the pieces of content we previously evaluated under our status quo policy. As you can see, we would be more permissive under this option, leaving up the comment that attacks someone as a "narcissistic male" and the comment that refers to men as trash."

Hate Speech and Gender

Option 2: Treat sex/gender as a quasi-protected characteristic

- Remove Tier 1 attacks targeting people on the basis of sex/gender but do not remove Tier 2 or Tier 3.

Pros

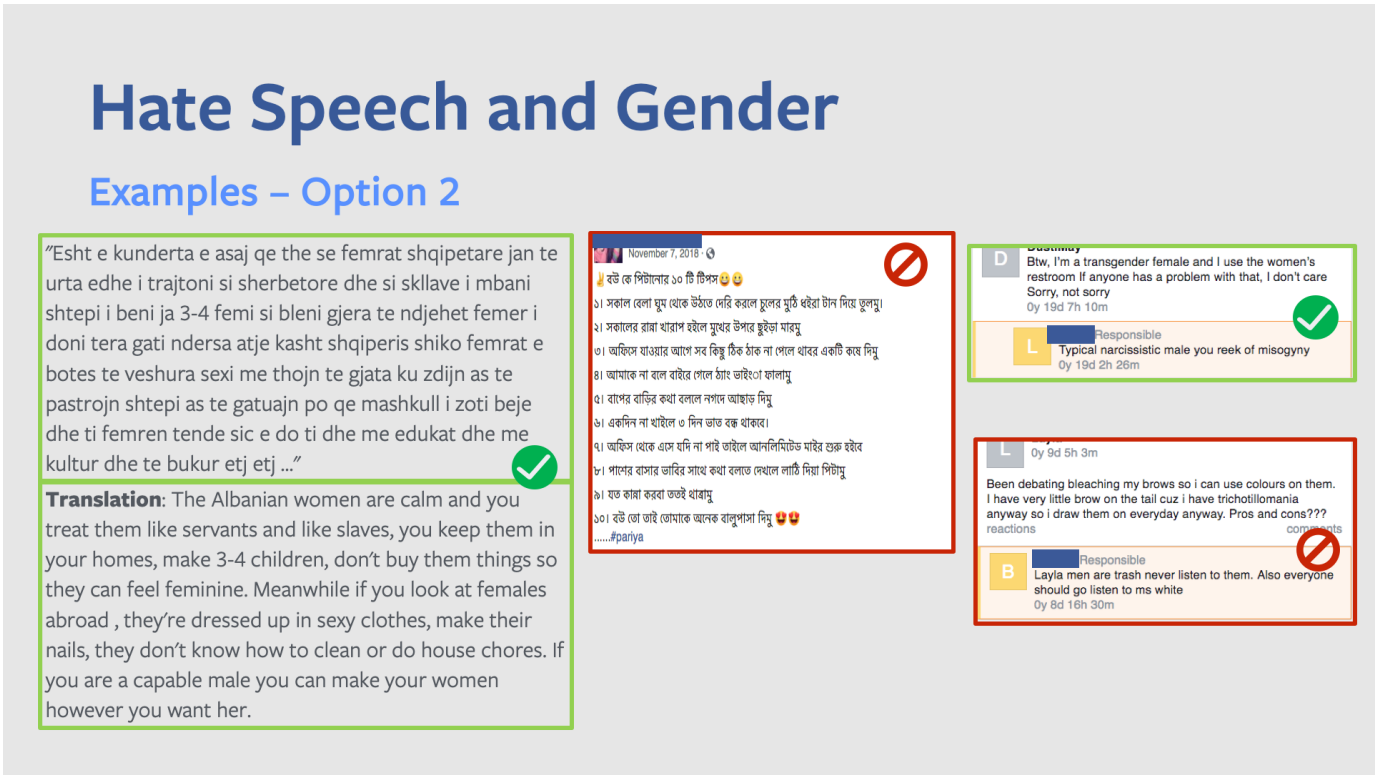
- This option mitigates some of the risk of removing edge cases (e.g., "contempt," "moral inferiority").
- Aligns with different protections (e.g., "intermediate scrutiny") that gender receive in some legal systems.

Cons

- Allows more hate speech against women (Tier 2 & 3), thus protecting women less.
- In some cultures, Tier 2 or 3 attacks are worse than Tier 1.
- Enforcement could appear overbroad.
- Could increase risk of harmful speech targeting vulnerable groups, including trans and intersex individuals.

As previously noted, quasi-protected characteristics are protected from Tier 1 attacks, but not Tier 2 or 3 attacks.

Based on the data we evaluated, women are more often the victims of Tier 2 and 3 attacks, while men are subject to a higher volume of Tier 1 attacks. As such, under this option, we would end up taking down more attacks directed at men than we would women. We would also end up leaving up Tier 2 and 3 attacks against non-binary and trans individuals, both of whom are routinely subject to these kinds of attacks (i.e. statements of physical, mental, and moral inferiority, expressions of contempt, expressions of disgust, and calls for exclusion).

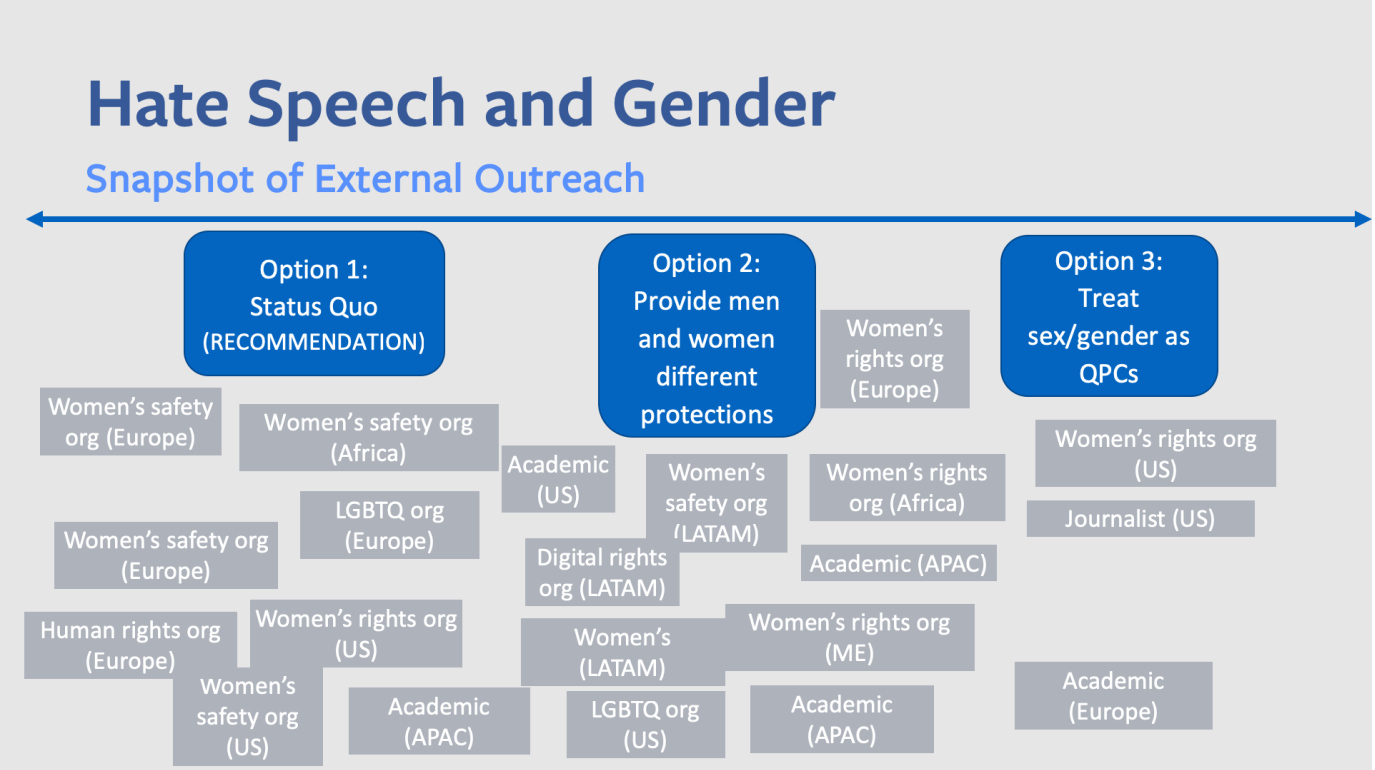


Here, you can see how Option 2 would apply to the pieces of content we previously evaluated under our status quo policy. As was the case with option 1, we would end up being more permissive.

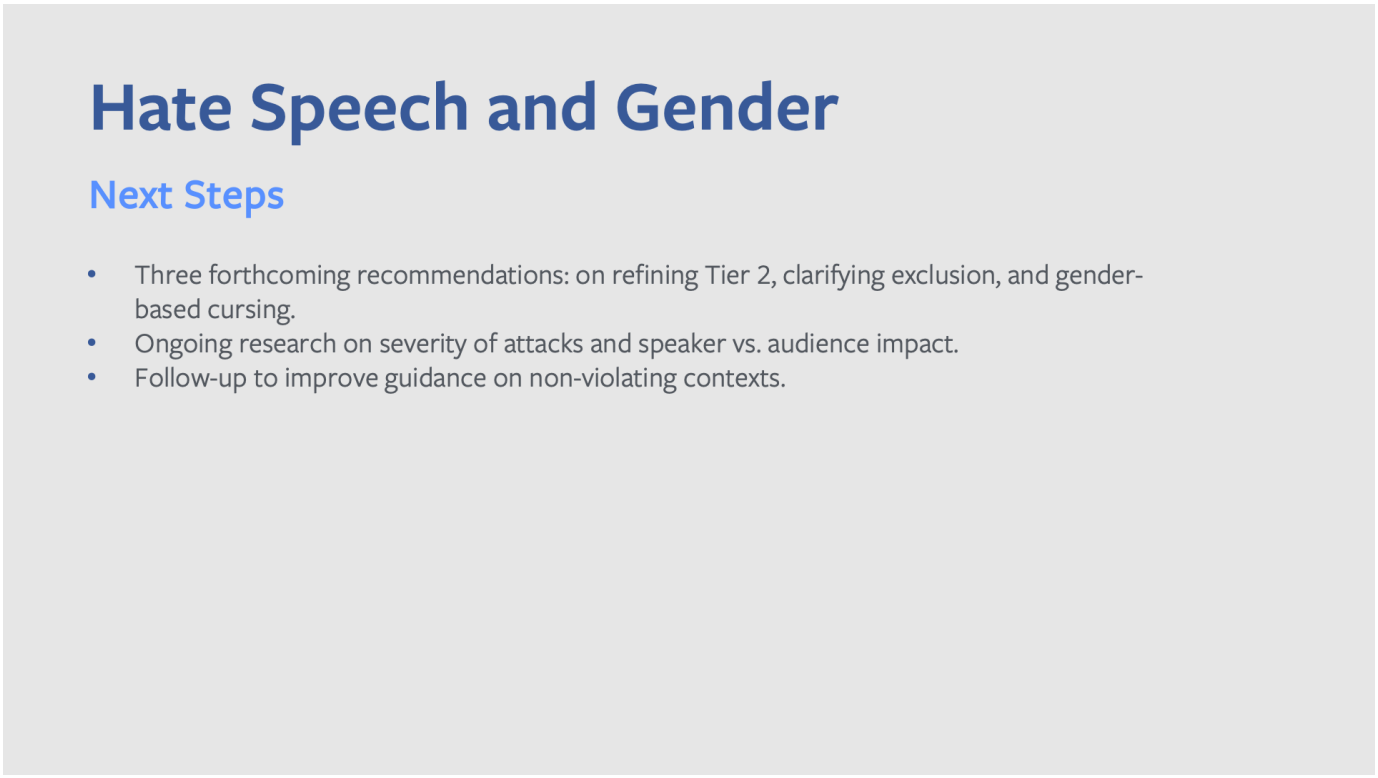
Hate Speech and Gender

External Outreach

We spoke to 24 experts globally, including academics, free expression advocates, LGBTQ advocacy organizations and women's safety organizations.



It was Interesting to see how external expert opinion was clustered. Some people supported our status quo policy, while others think that women should be afforded more protection than men. And there's a group of people who believe that we should be providing less protection altogether. Interestingly, many women's groups supported the status quo policy because they argued that distinct protections for women would prompt an undue backlash.



Discussion

Question: As you build out signals to look for when iterating on our hate speech policies [with current and future working groups], will there be a focus on transgender individuals and attacks directed at people within the trans community?

Answer: Yes. That is a great callout, especially for the policy working groups on gendered cursing.

Comment: Something we heard from external experts was that we need to look more into socially acceptable uses of "hate speech" that may reflect how people are talking and/or account for modern day social movements.

Question: Did we reach out to people who have heavily criticized us for taking down "men are trash" and other

similar content?

Answer: Yes, these groups are the ones that advocated for decoupling speech against men and women, but we also found that many of these people don't necessarily take issue with where we draw the line; they're more upset by inconsistencies in enforcement. As we improve our proactive detection capabilities, we hope to improve enforcement so that we are better addressing these concerns. In fact, our own research shows the importance of proactive detection beyond just Tier 1 hate speech (because of the imbalances between hate speech targeted at men vs. women).


facebook