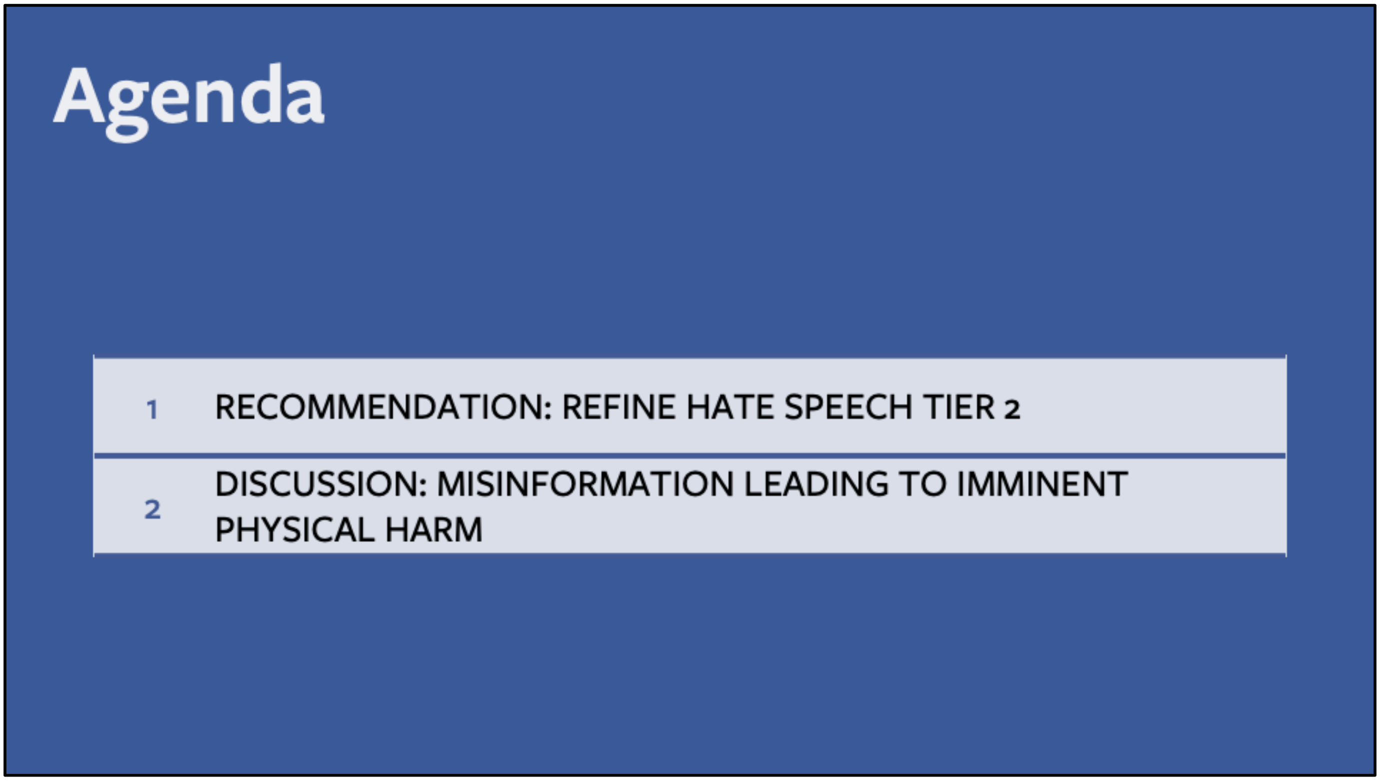
Product Policy Forum

April 23, 2019

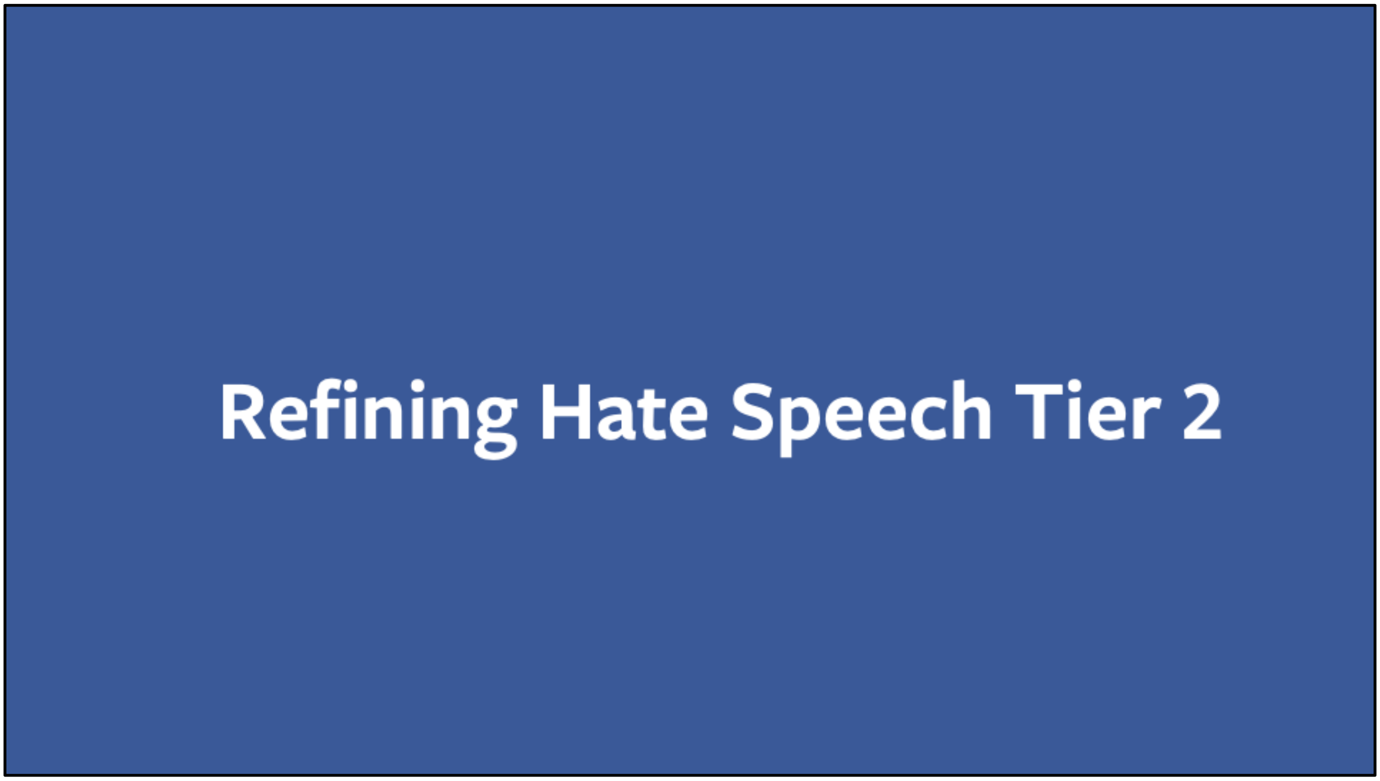
TOPICS: *Refine Tier 2 Hate Speech, Discussion on Harmful Misinformation*



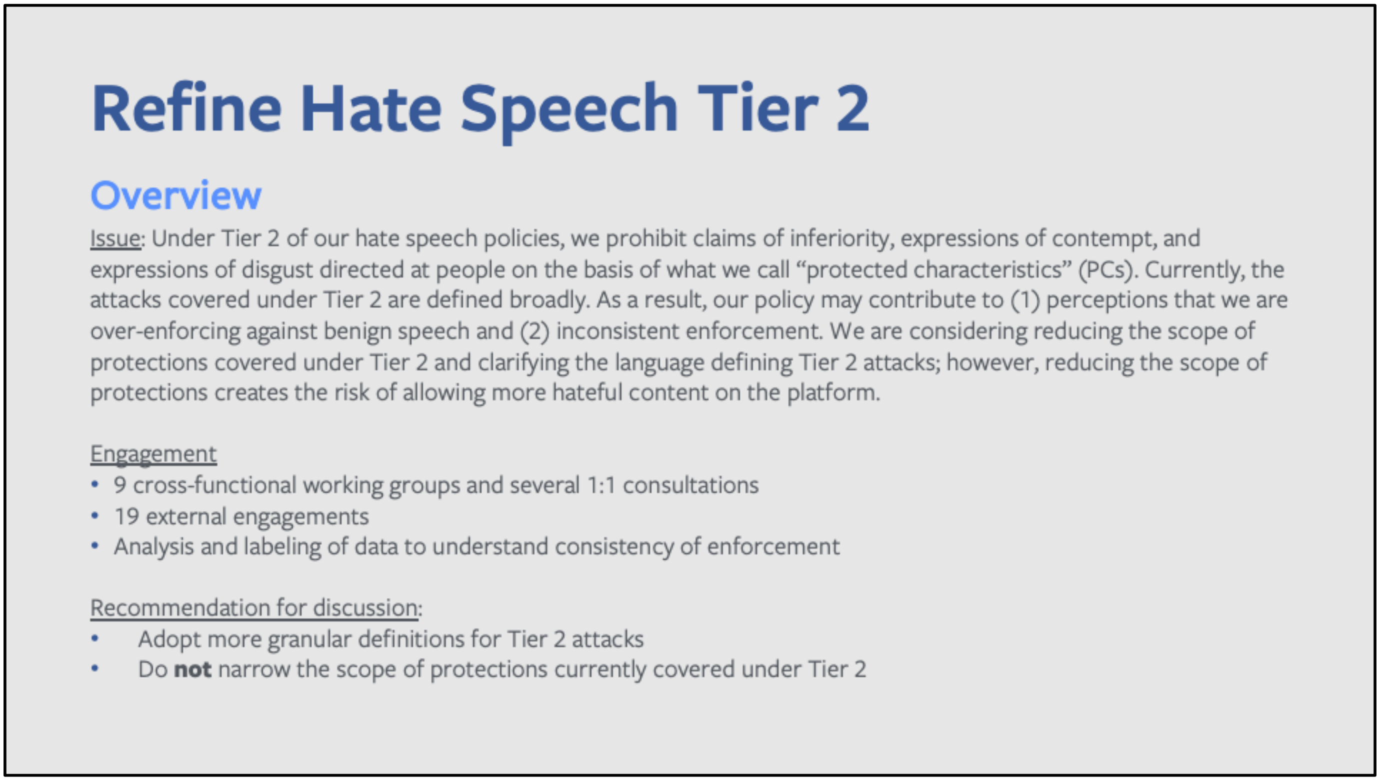
Agenda

1 RECOMMENDATION: REFINE HATE SPEECH TIER 2

2 DISCUSSION: MISINFORMATION LEADING TO IMMINENT PHYSICAL HARM



Refining Hate Speech Tier 2



Refine Hate Speech Tier 2

Overview

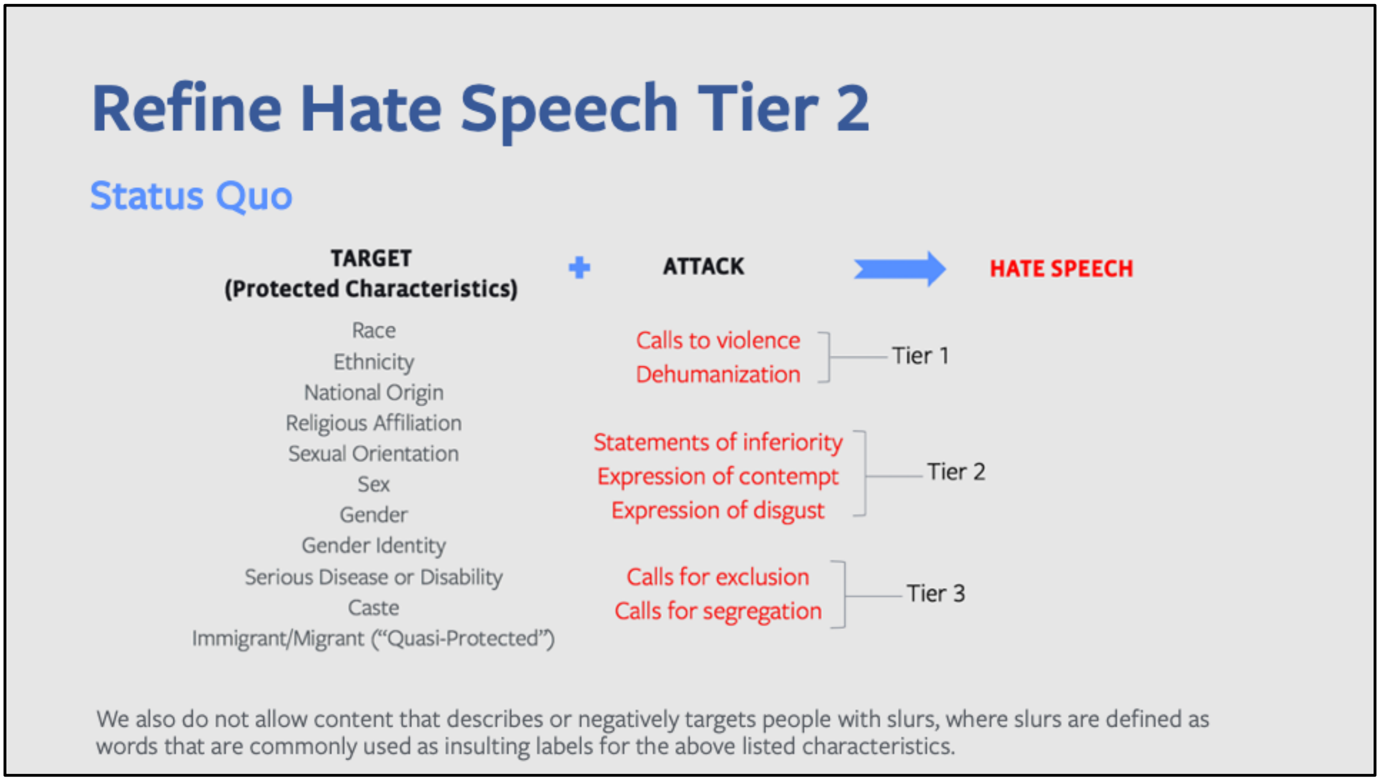
Issue: Under Tier 2 of our hate speech policies, we prohibit claims of inferiority, expressions of contempt, and expressions of disgust directed at people on the basis of what we call “protected characteristics” (PCs). Currently, the attacks covered under Tier 2 are defined broadly. As a result, our policy may contribute to (1) perceptions that we are over-enforcing against benign speech and (2) inconsistent enforcement. We are considering reducing the scope of protections covered under Tier 2 and clarifying the language defining Tier 2 attacks; however, reducing the scope of protections creates the risk of allowing more hateful content on the platform.

Engagement

- 9 cross-functional working groups and several 1:1 consultations
- 19 external engagements
- Analysis and labeling of data to understand consistency of enforcement

Recommendation for discussion:

- Adopt more granular definitions for Tier 2 attacks
- Do **not** narrow the scope of protections currently covered under Tier 2



Refine Hate Speech Tier 2

Status Quo

TARGET (Protected Characteristics) + **ATTACK** → **HATE SPEECH**

Race
Ethnicity
National Origin
Religious Affiliation
Sexual Orientation
Sex
Gender
Gender Identity
Serious Disease or Disability
Caste
Immigrant/Migrant ("Quasi-Protected")

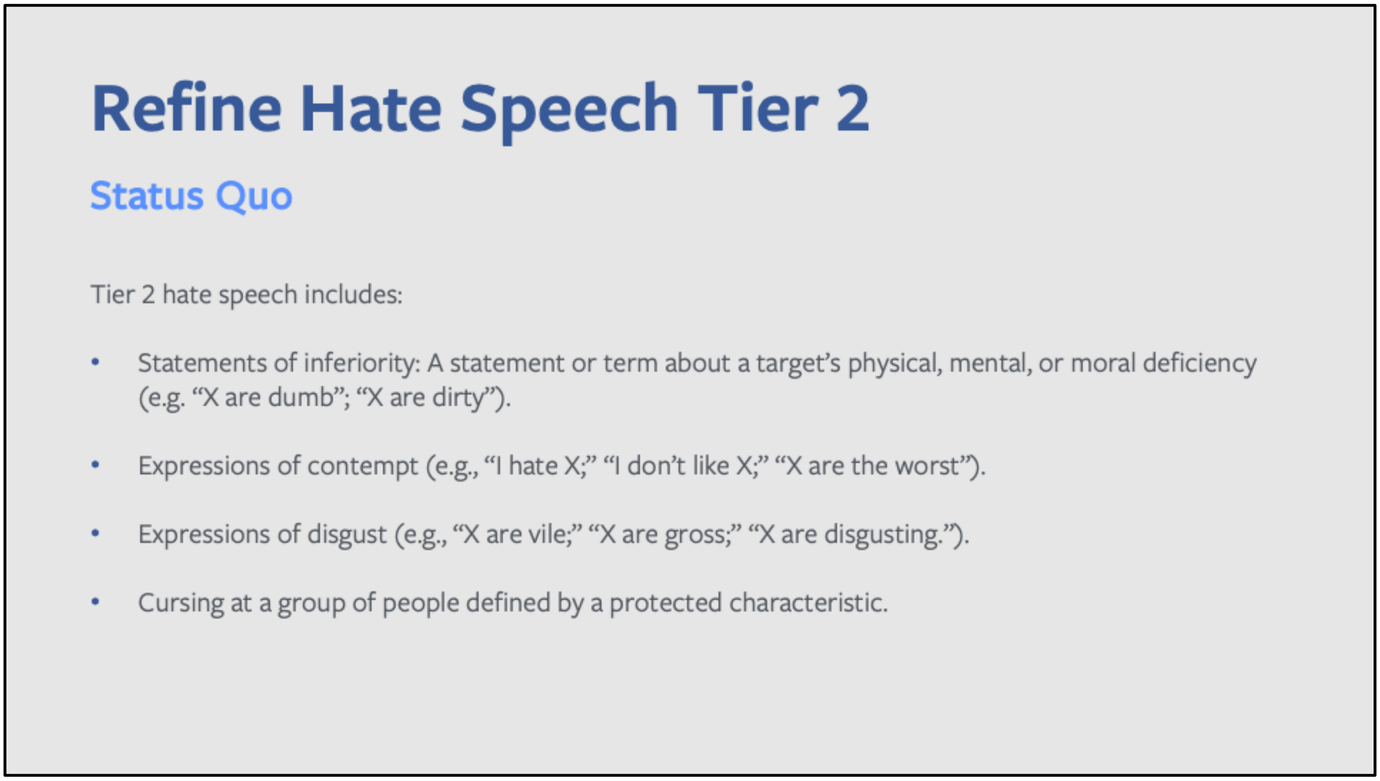
Calls to violence
Dehumanization
Tier 1

Statements of inferiority
Expression of contempt
Expression of disgust
Tier 2

Calls for exclusion
Calls for segregation
Tier 3

We also do not allow content that describes or negatively targets people with slurs, where slurs are defined as words that are commonly used as insulting labels for the above listed characteristics.

Right now, the technology we use to proactively detect hate speech focuses specifically on violent and dehumanizing speech. By adopting more granular definitions for the types of attacks included under Tier 2, we hope to be able to expand use of our technology to a broader range of hate speech. It's important to note the technology merely works to proactively detect. Thereafter, potentially violating content is sent to our content review teams so that people with the appropriate language skills and subject matter expertise can help determine whether something does, in fact, violate our hate speech policies.

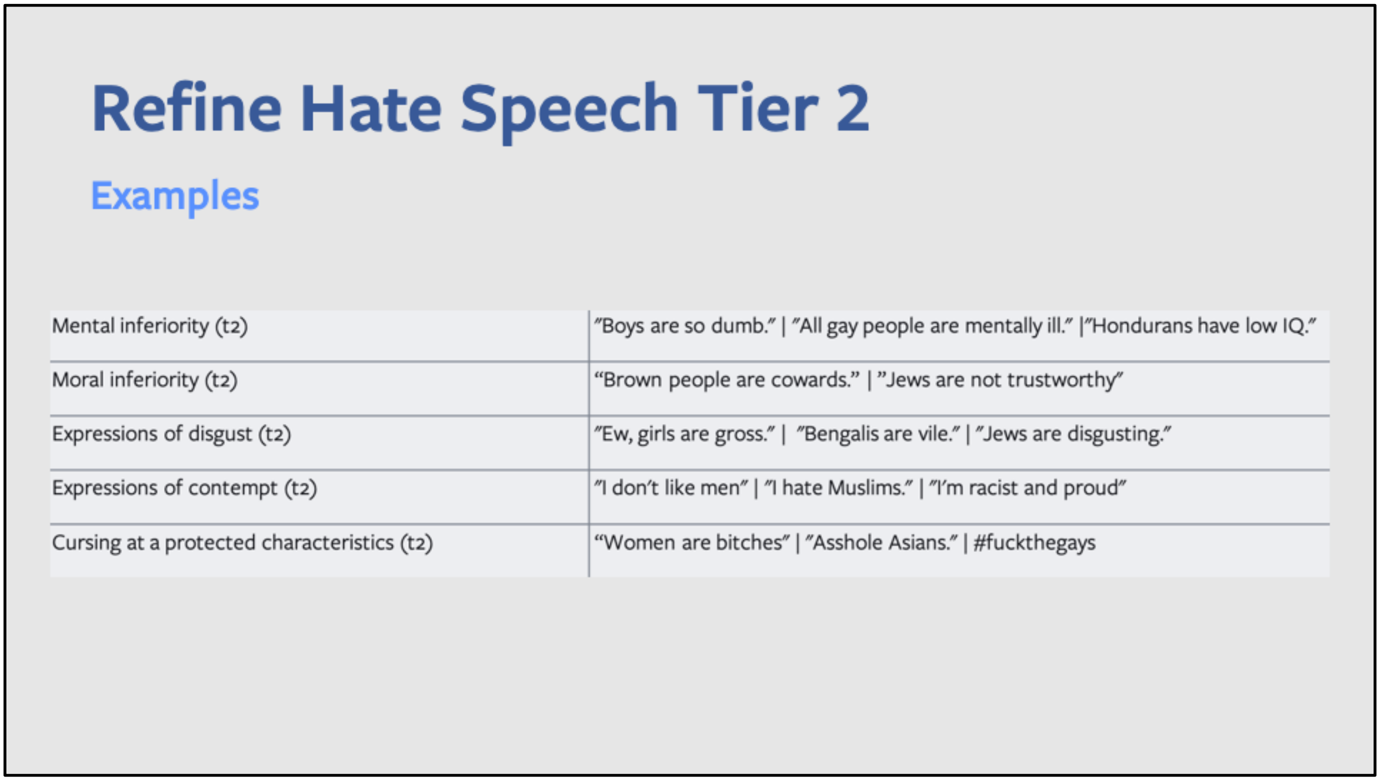


Refine Hate Speech Tier 2

Status Quo

Tier 2 hate speech includes:

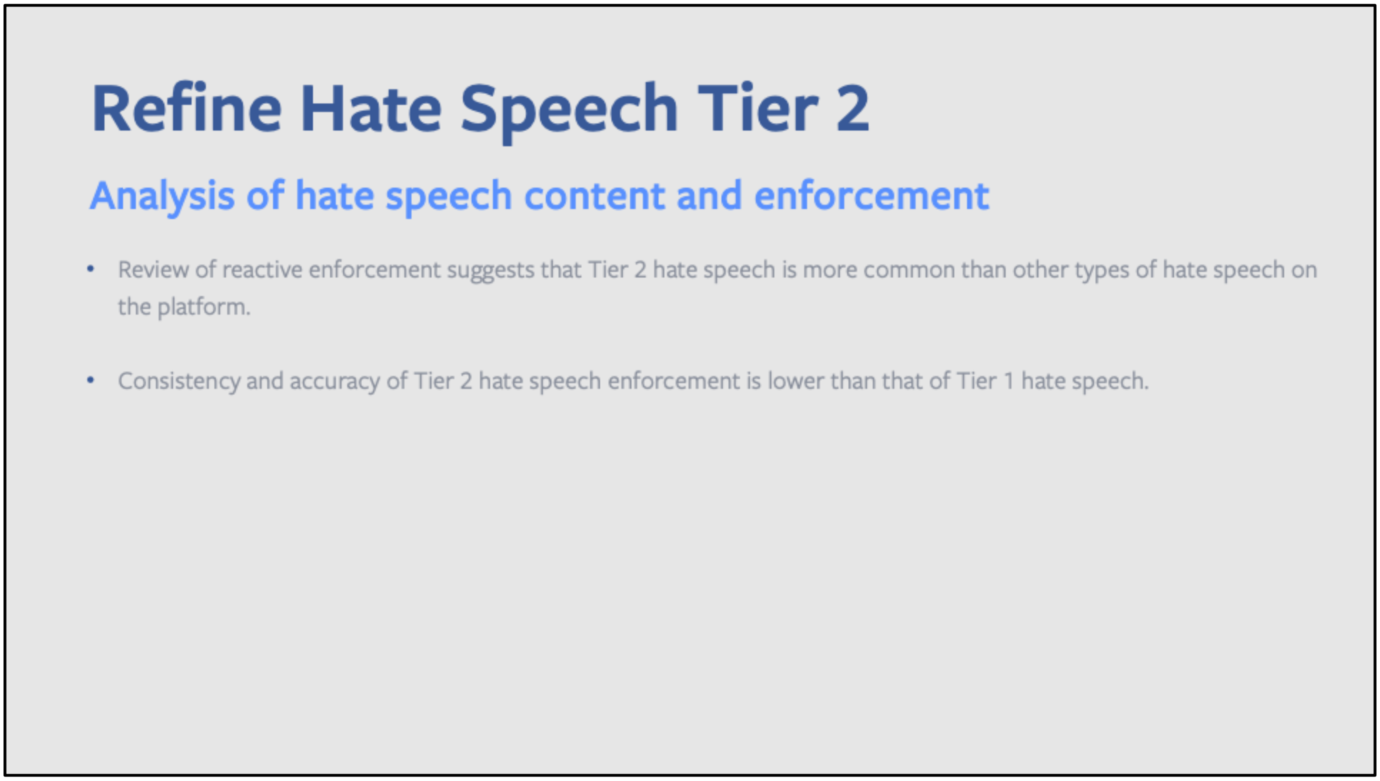
- Statements of inferiority: A statement or term about a target’s physical, mental, or moral deficiency (e.g. “X are dumb”; “X are dirty”).
- Expressions of contempt (e.g., “I hate X;” “I don’t like X;” “X are the worst”).
- Expressions of disgust (e.g., “X are vile;” “X are gross;” “X are disgusting.”).
- Cursing at a group of people defined by a protected characteristic.



Refine Hate Speech Tier 2

Examples

<table>
<tr><td>Mental inferiority (t2)</td><td>”Boys are so dumb.” | ”All gay people are mentally ill.” |”Hondurans have low IQ.”</td></tr>
<tr><td>Moral inferiority (t2)</td><td>“Brown people are cowards.” | ”Jews are not trustworthy”</td></tr>
<tr><td>Expressions of disgust (t2)</td><td>”Ew, girls are gross.” | ”Bengalis are vile.” | ”Jews are disgusting.”</td></tr>
<tr><td>Expressions of contempt (t2)</td><td>”I don’t like men” | ”I hate Muslims.” | ”I’m racist and proud”</td></tr>
<tr><td>Cursing at a protected characteristics (t2)</td><td>“Women are bitches” | ”Asshole Asians.” | #fuckthegays</td></tr>
</table>

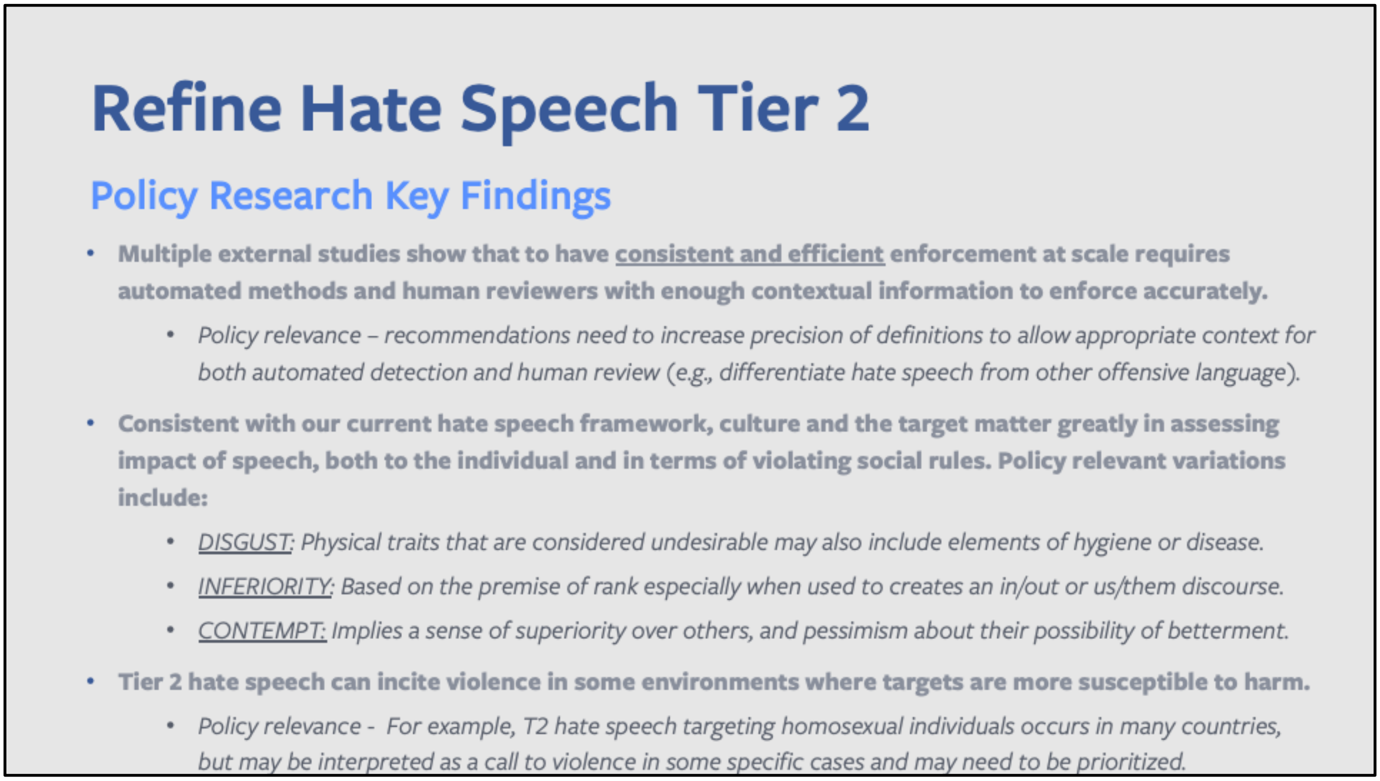


Refine Hate Speech Tier 2

Analysis of hate speech content and enforcement

- Review of reactive enforcement suggests that Tier 2 hate speech is more common than other types of hate speech on the platform.

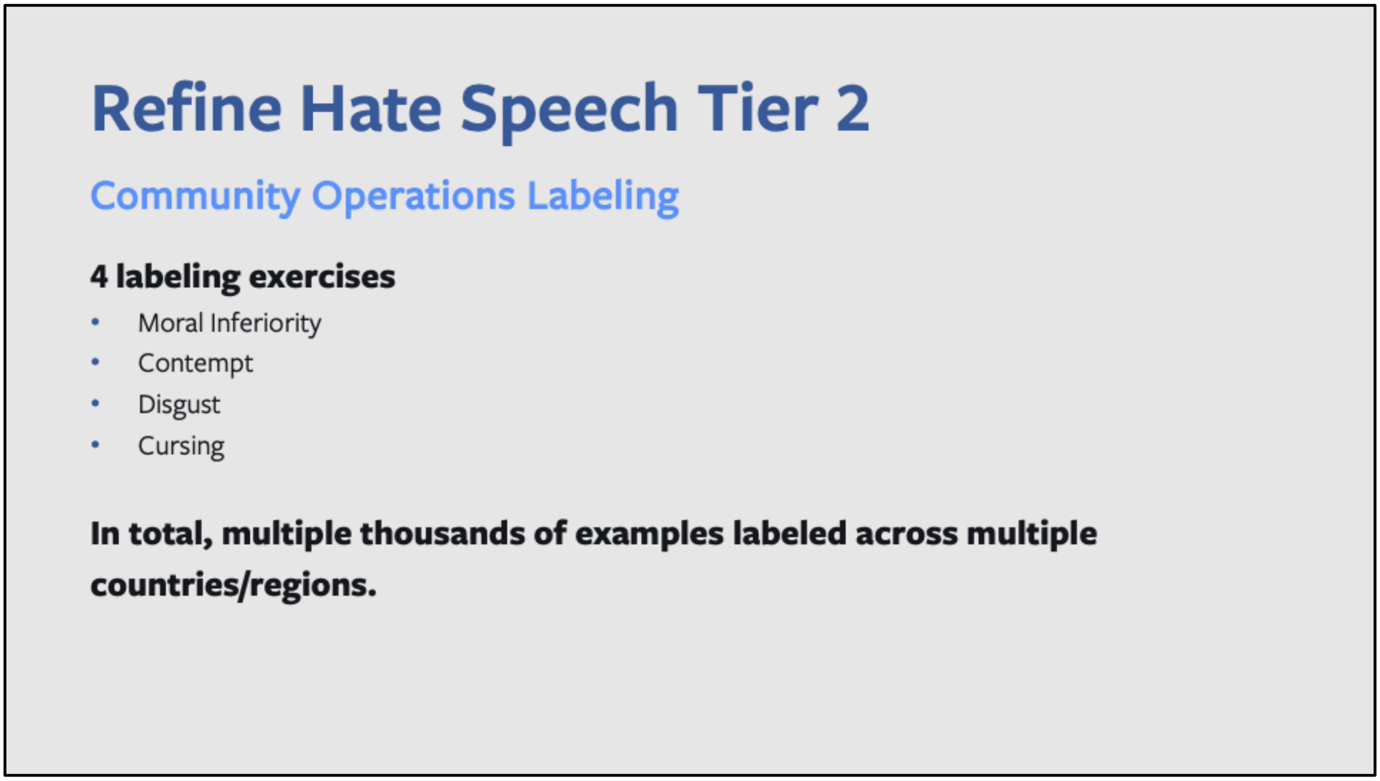
- Consistency and accuracy of Tier 2 hate speech enforcement is lower than that of Tier 1 hate speech.



Refine Hate Speech Tier 2

Policy Research Key Findings

- **Multiple external studies show that to have <u>consistent and efficient</u> enforcement at scale requires automated methods and human reviewers with enough contextual information to enforce accurately.**
  - *Policy relevance – recommendations need to increase precision of definitions to allow appropriate context for both automated detection and human review (e.g., differentiate hate speech from other offensive language).*
- **Consistent with our current hate speech framework, culture and the target matter greatly in assessing impact of speech, both to the individual and in terms of violating social rules. Policy relevant variations include:**
  - *<u>DISGUST:</u> Physical traits that are considered undesirable may also include elements of hygiene or disease.*
  - *<u>INFERIORITY:</u> Based on the premise of rank especially when used to creates an in/out or us/them discourse.*
  - *<u>CONTEMPT:</u> Implies a sense of superiority over others, and pessimism about their possibility of betterment.*
- **Tier 2 hate speech can incite violence in some environments where targets are more susceptible to harm.**
  - *Policy relevance - For example, T2 hate speech targeting homosexual individuals occurs in many countries, but may be interpreted as a call to violence in some specific cases and may need to be prioritized.*



Refine Hate Speech Tier 2

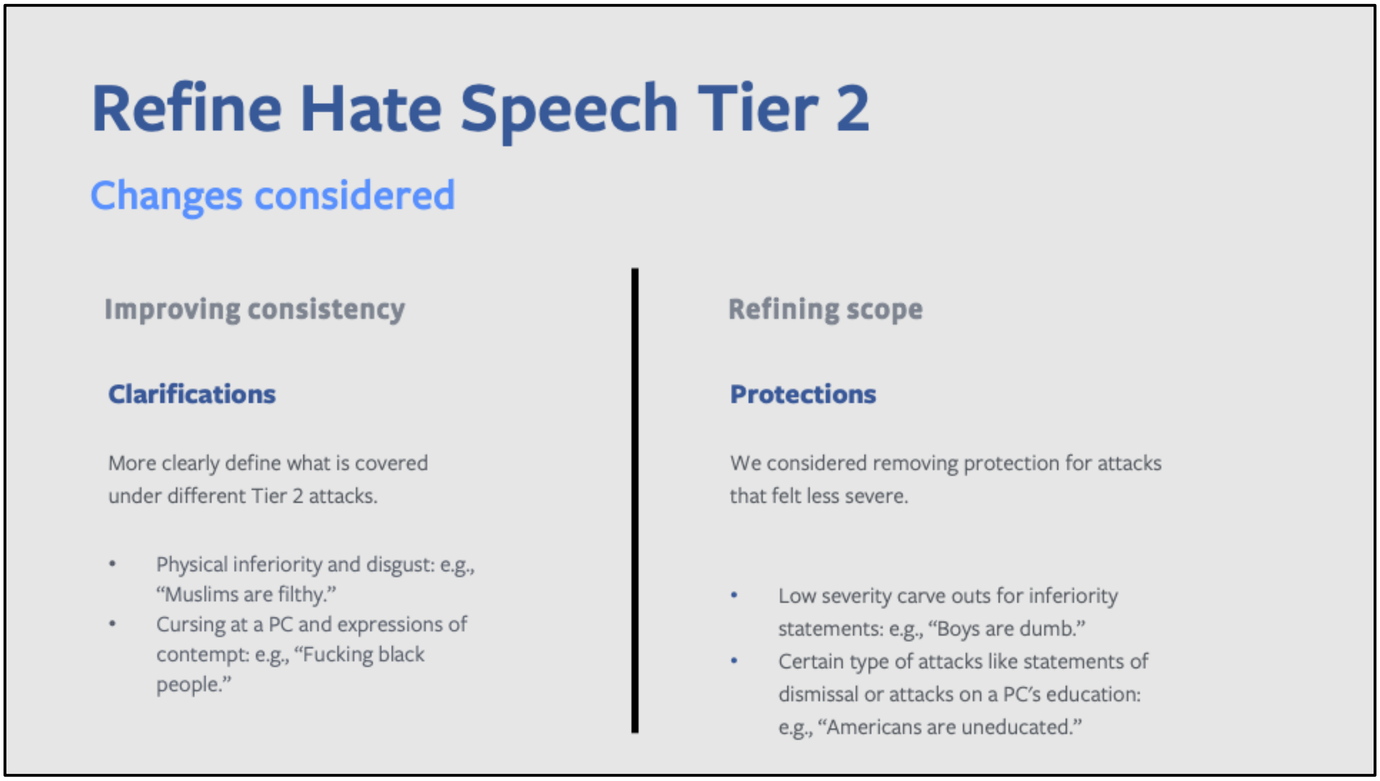
Community Operations Labeling

4 labeling exercises

- Moral Inferiority
- Contempt
- Disgust
- Cursing

In total, multiple thousands of examples labeled across multiple countries/regions.

One of the things we do when we’re considering changes to our policies is label real examples of content on the platform to understand how the content would be affected if we were to adjust the policy line. This helps us see – in very real terms (e.g. what kind of content will remain up and what will come down as a result of a change to policy).



Refine Hate Speech Tier 2

Changes considered

Improving consistency

Clarifications

More clearly define what is covered under different Tier 2 attacks.

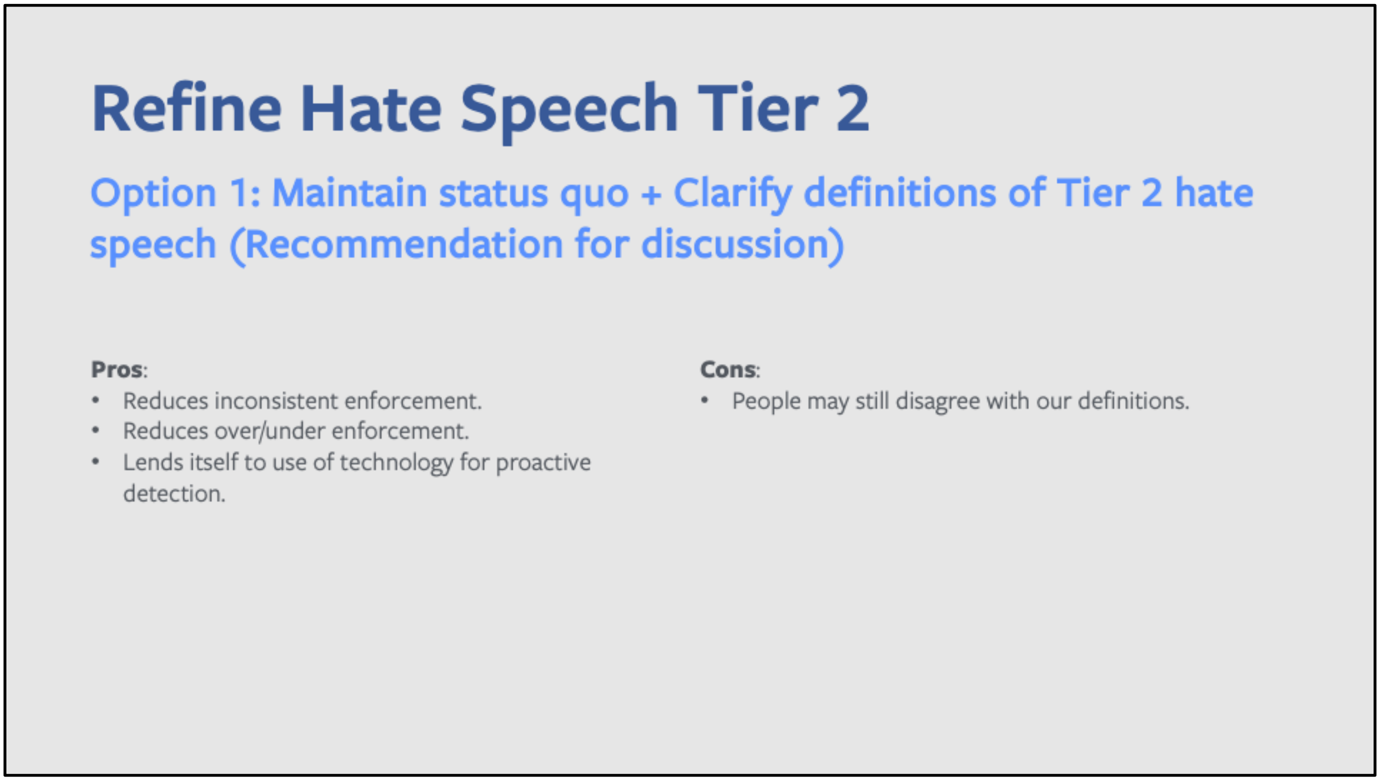
- Physical inferiority and disgust: e.g., “Muslims are filthy.”
- Cursing at a PC and expressions of contempt: e.g., “Fucking black people.”

Refining scope

Protections

We considered removing protection for attacks that felt less severe.

- Low severity carve outs for inferiority statements: e.g., “Boys are dumb.”
- Certain type of attacks like statements of dismissal or attacks on a PC’s education: e.g., “Americans are uneducated.”

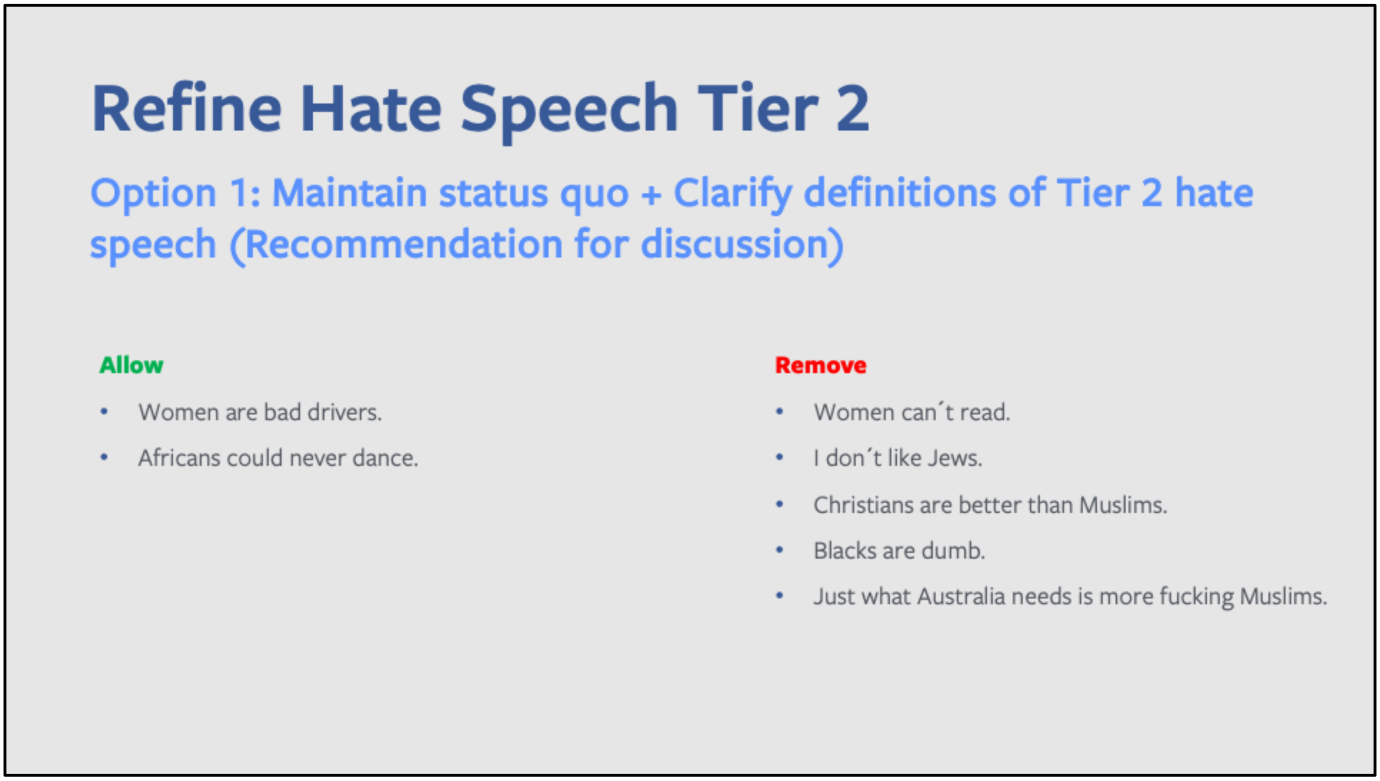


Refine Hate Speech Tier 2

Option 1: Maintain status quo + Clarify definitions of Tier 2 hate speech (Recommendation for discussion)

Pros:
- Reduces inconsistent enforcement.
- Reduces over/under enforcement.
- Lends itself to use of technology for proactive detection.

Cons:
- People may still disagree with our definitions.



Refine Hate Speech Tier 2

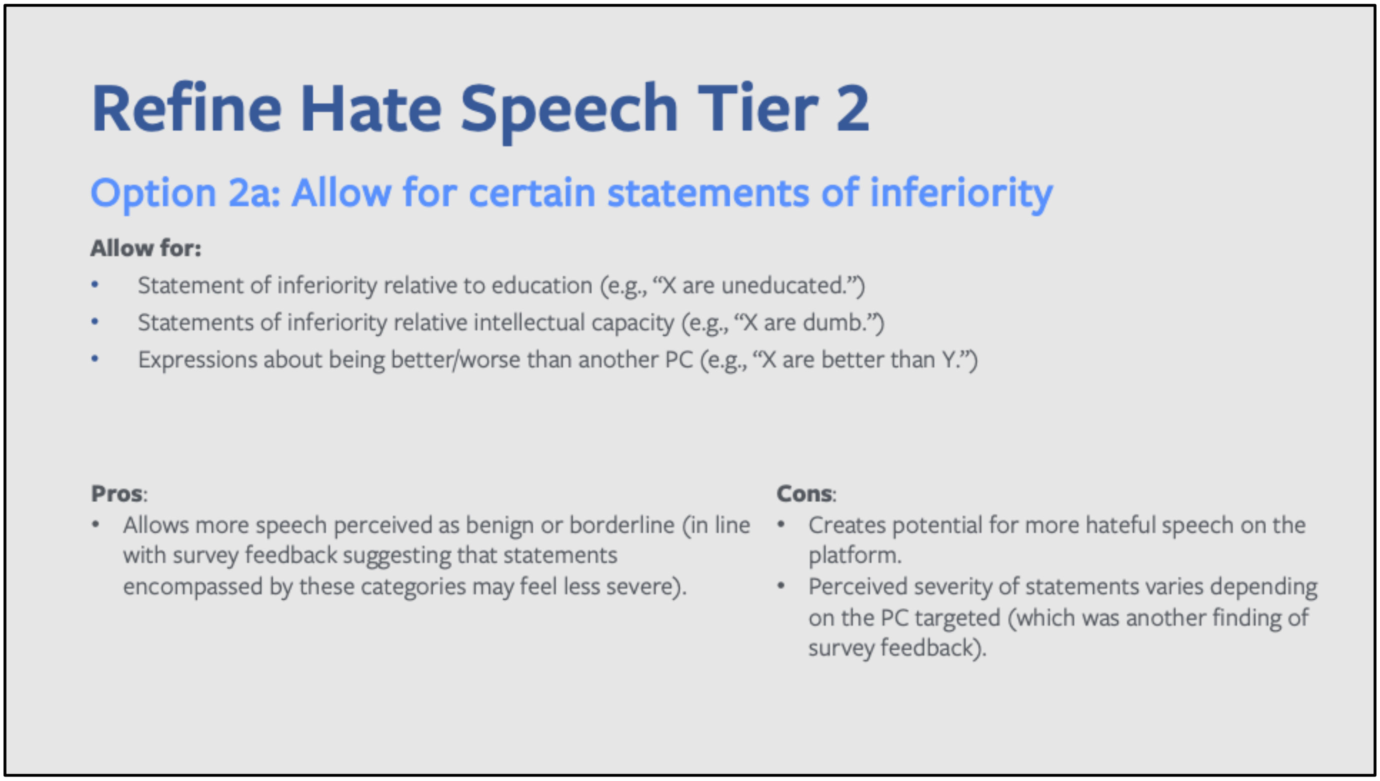
Option 1: Maintain status quo + Clarify definitions of Tier 2 hate speech (Recommendation for discussion)

Allow

- Women are bad drivers.
- Africans could never dance.

Remove

- Women can´t read.
- I don´t like Jews.
- Christians are better than Muslims.
- Blacks are dumb.
- Just what Australia needs is more fucking Muslims.



Refine Hate Speech Tier 2

Option 2a: Allow for certain statements of inferiority

Allow for:

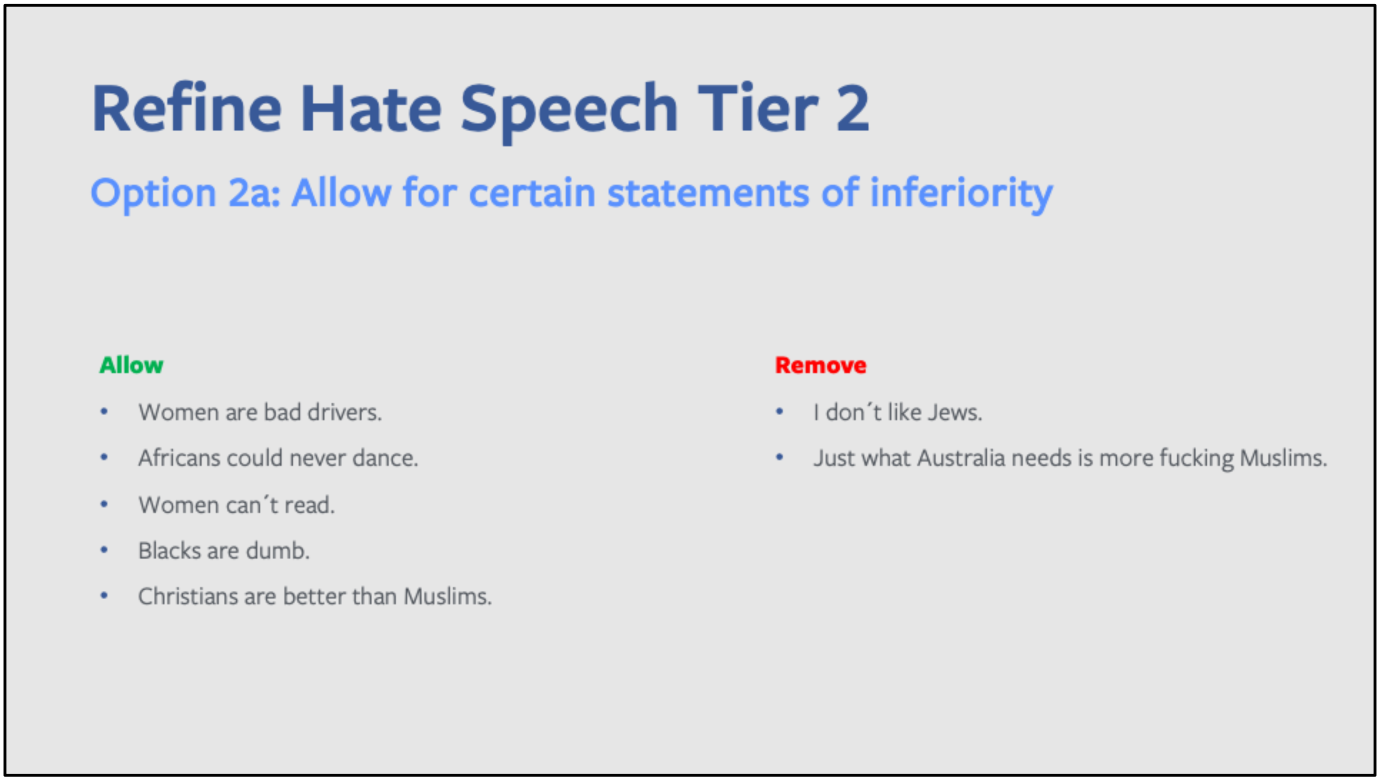
- Statement of inferiority relative to education (e.g., “X are uneducated.”)
- Statements of inferiority relative intellectual capacity (e.g., “X are dumb.”)
- Expressions about being better/worse than another PC (e.g., “X are better than Y.”)

Pros:

- Allows more speech perceived as benign or borderline (in line with survey feedback suggesting that statements encompassed by these categories may feel less severe).

Cons:

- Creates potential for more hateful speech on the platform.
- Perceived severity of statements varies depending on the PC targeted (which was another finding of survey feedback).



Refine Hate Speech Tier 2

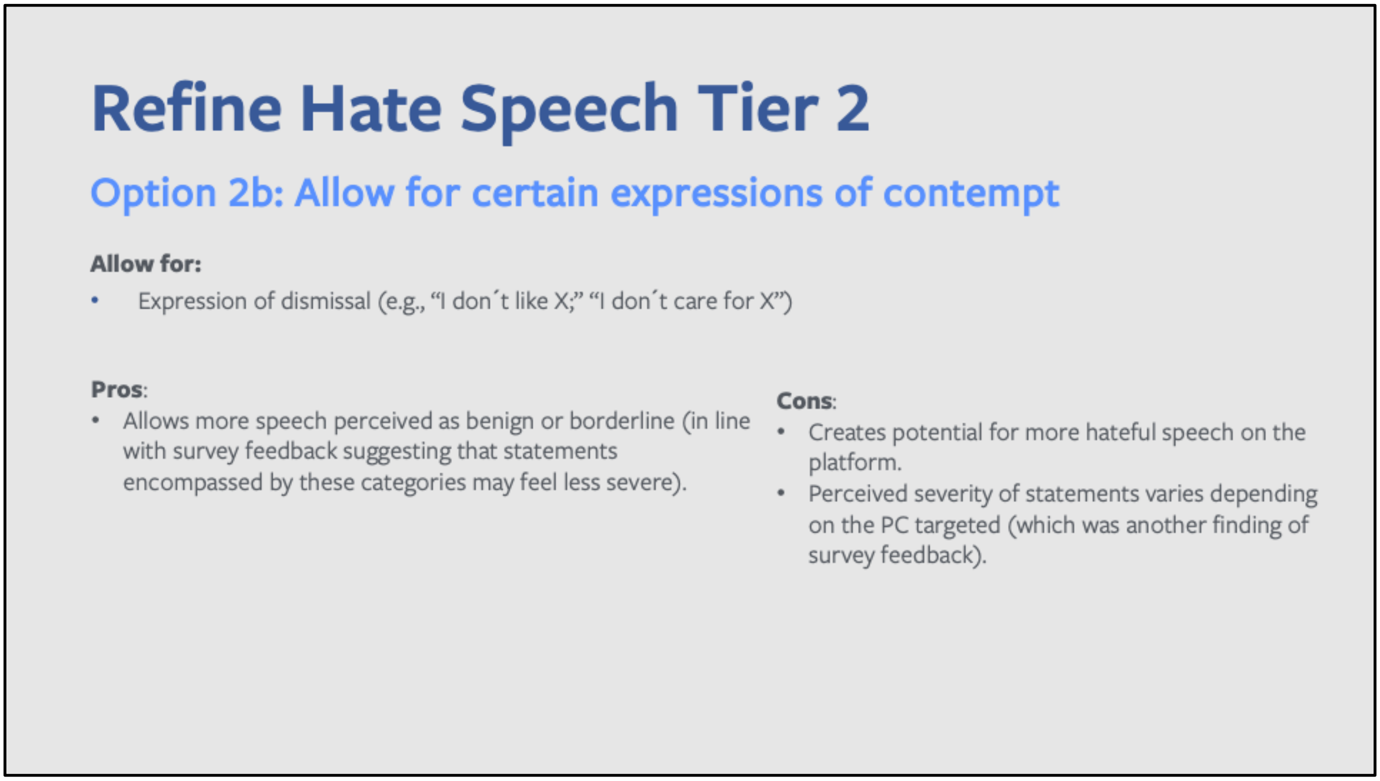
Option 2a: Allow for certain statements of inferiority

Allow

- Women are bad drivers.
- Africans could never dance.
- Women can´t read.
- Blacks are dumb.
- Christians are better than Muslims.

Remove

- I don´t like Jews.
- Just what Australia needs is more fucking Muslims.



Refine Hate Speech Tier 2

Option 2b: Allow for certain expressions of contempt

Allow for:

- Expression of dismissal (e.g., “I don´t like X;” “I don´t care for X”)

Pros:

- Allows more speech perceived as benign or borderline (in line with survey feedback suggesting that statements encompassed by these categories may feel less severe).

Cons:

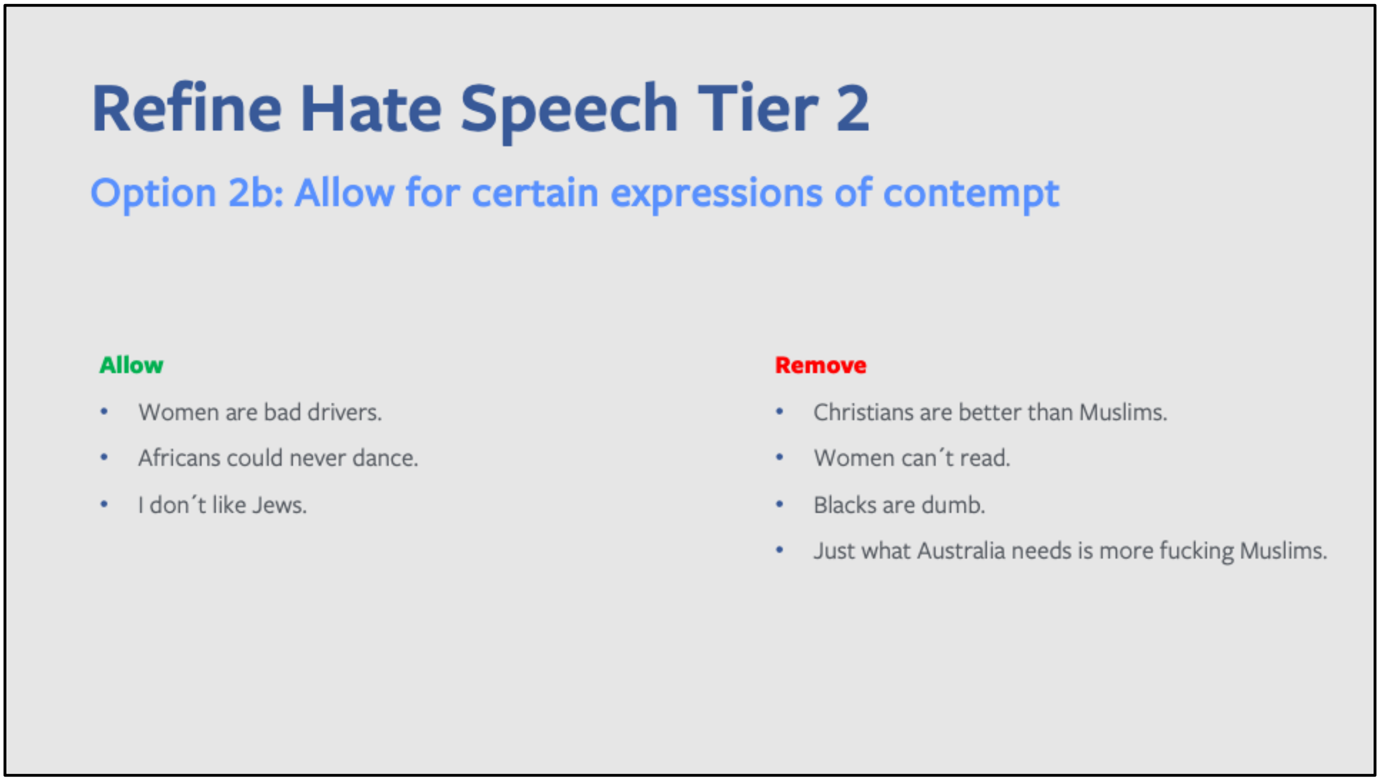
- Creates potential for more hateful speech on the platform.
- Perceived severity of statements varies depending on the PC targeted (which was another finding of survey feedback).

Discussion

Question: Do we ever see speech that speaks to sexual preference, such as “My parents can’t get it through their heads. I don’t like men; I like women.”

Answer: Yes, we do, and this is potentially a place where we’ll be over-enforcing. The challenge here is that scaling back protection would apply to all protected characteristics, and we know that the same statement applied to a different protected characteristic suddenly feels more problematic. We do, however, have a narrow policy carve-out for gendered-statements of contempt made in the context of a romantic break up.

Follow up: Also worth noting that when we analyzed data in the context of the narrow policy carve-out just mentioned, we found that expressions of contempt in the context of a break-up aren’t very common.



Refine Hate Speech Tier 2

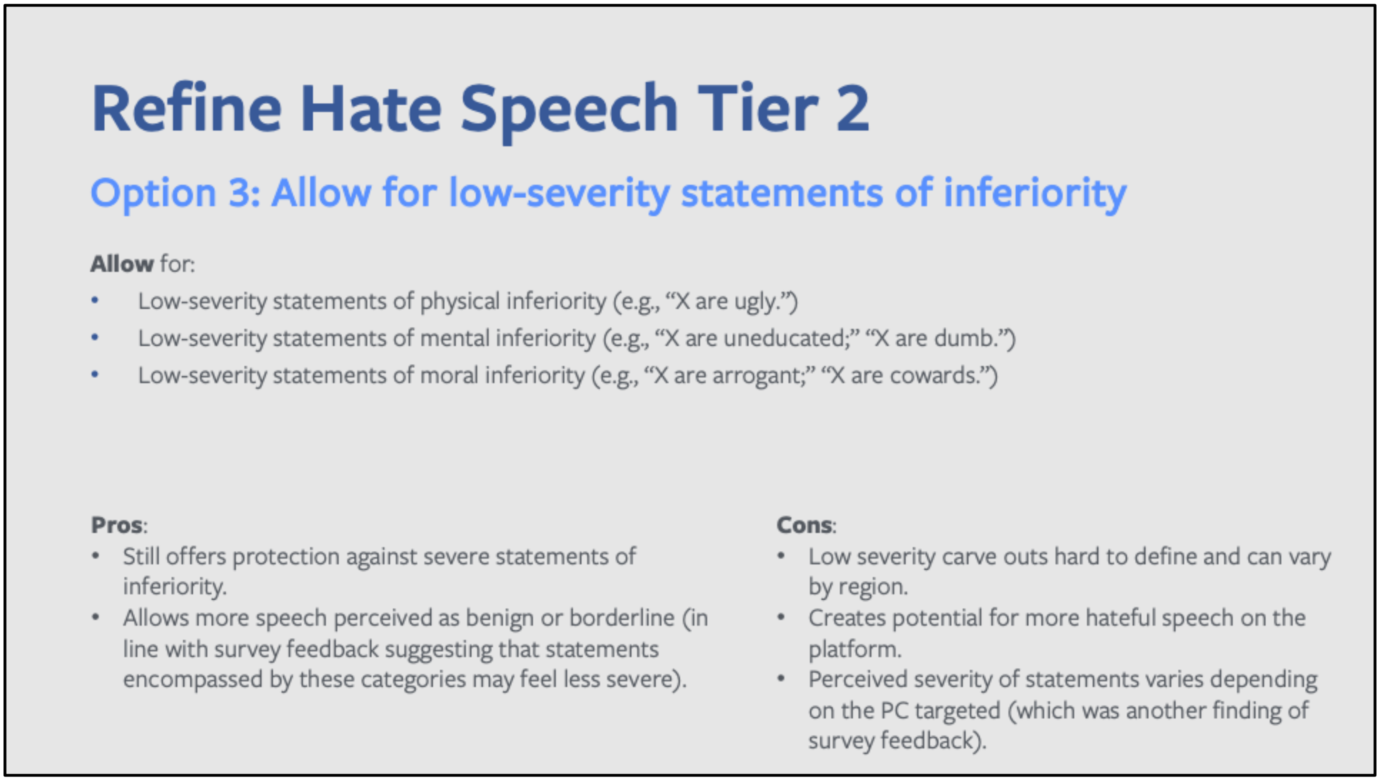
Option 2b: Allow for certain expressions of contempt

Allow

- Women are bad drivers.
- Africans could never dance.
- I don´t like Jews.

Remove

- Christians are better than Muslims.
- Women can´t read.
- Blacks are dumb.
- Just what Australia needs is more fucking Muslims.



Refine Hate Speech Tier 2

Option 3: Allow for low-severity statements of inferiority

Allow for:

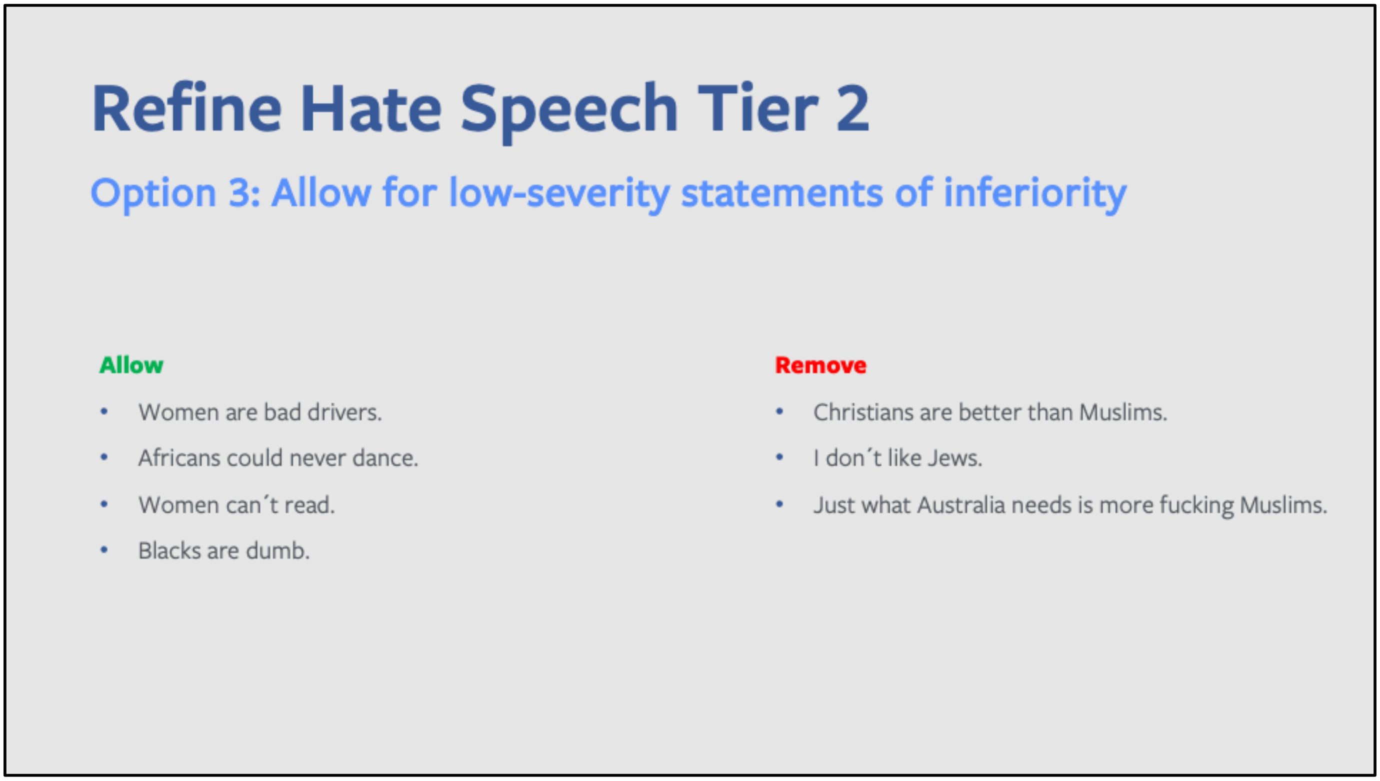
- Low-severity statements of physical inferiority (e.g., “X are ugly.”)
- Low-severity statements of mental inferiority (e.g., “X are uneducated;” “X are dumb.”)
- Low-severity statements of moral inferiority (e.g., “X are arrogant;” “X are cowards.”)

Pros:

- Still offers protection against severe statements of inferiority.
- Allows more speech perceived as benign or borderline (in line with survey feedback suggesting that statements encompassed by these categories may feel less severe).

Cons:

- Low severity carve outs hard to define and can vary by region.
- Creates potential for more hateful speech on the platform.
- Perceived severity of statements varies depending on the PC targeted (which was another finding of survey feedback).



Refine Hate Speech Tier 2

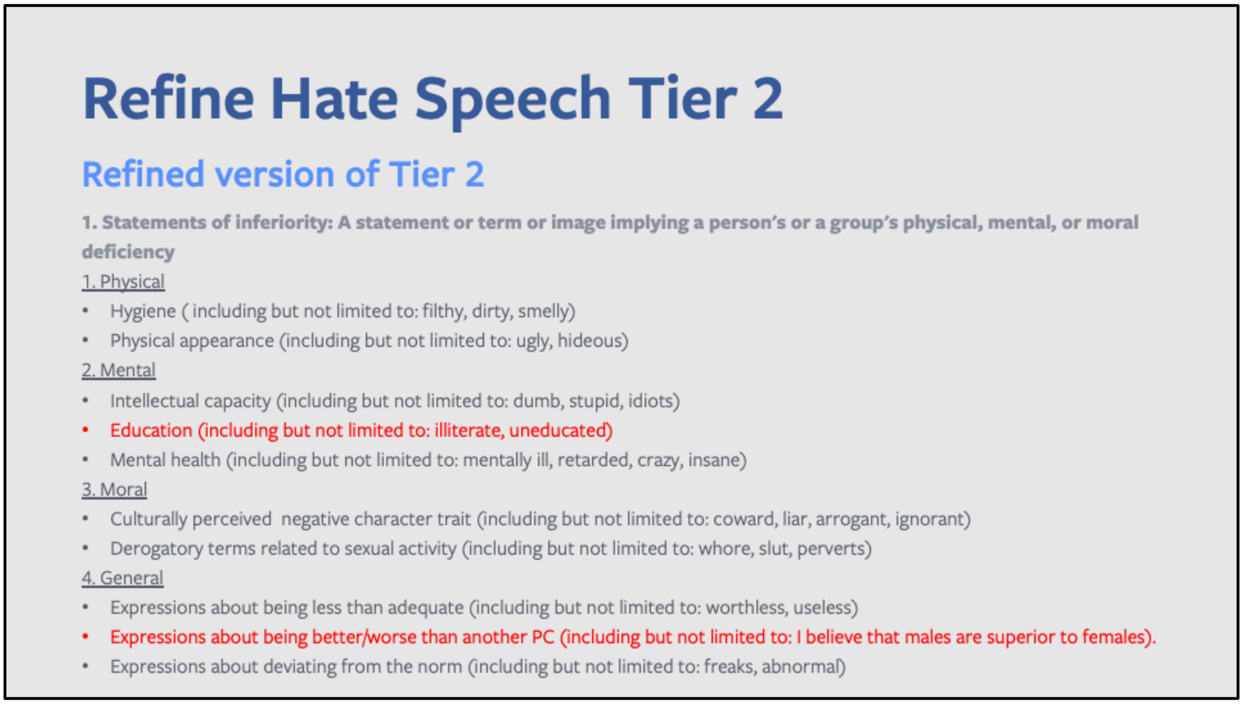
Option 3: Allow for low-severity statements of inferiority

Allow

- Women are bad drivers.
- Africans could never dance.
- Women can´t read.
- Blacks are dumb.

Remove

- Christians are better than Muslims.
- I don´t like Jews.
- Just what Australia needs is more fucking Muslims.



Refine Hate Speech Tier 2

Refined version of Tier 2

1. Statements of inferiority: A statement or term or image implying a person's or a group's physical, mental, or moral deficiency

1. Physical

- Hygiene (including but not limited to: filthy, dirty, smelly)
- Physical appearance (including but not limited to: ugly, hideous)

2. Mental

- Intellectual capacity (including but not limited to: dumb, stupid, idiots)
- Education (including but not limited to: illiterate, uneducated)
- Mental health (including but not limited to: mentally ill, retarded, crazy, insane)

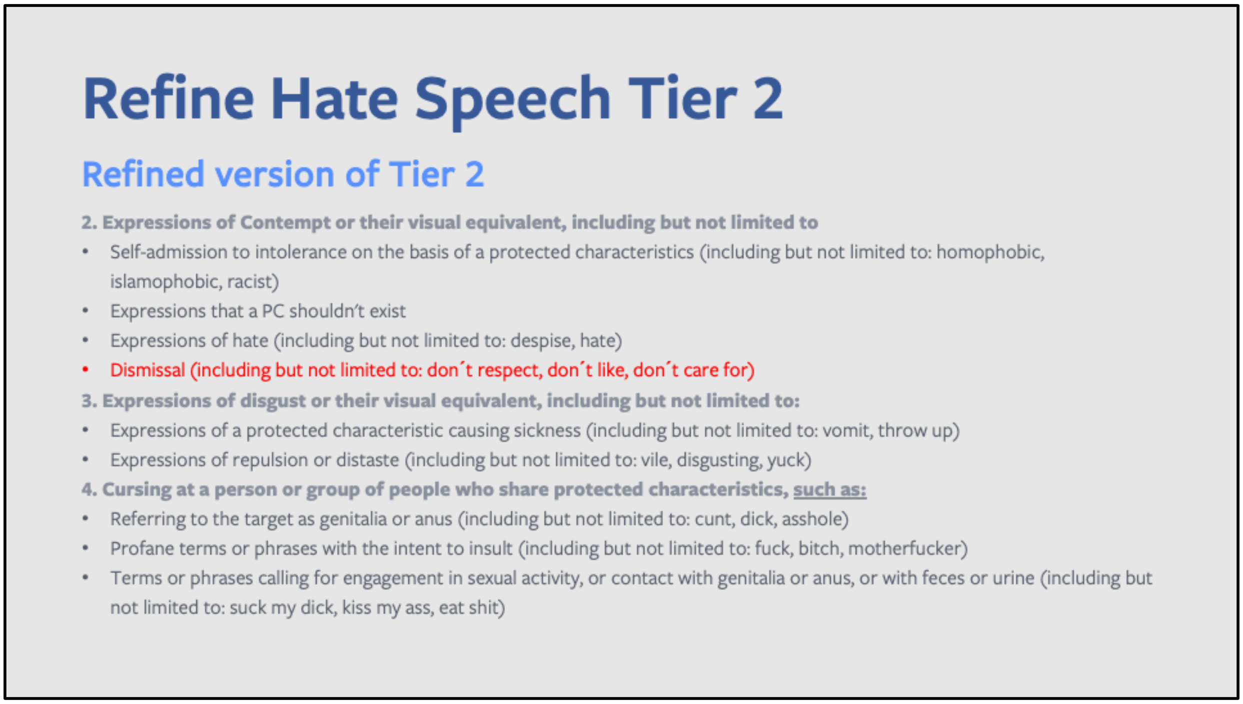
3. Moral

- Culturally perceived negative character trait (including but not limited to: coward, liar, arrogant, ignorant)
- Derogatory terms related to sexual activity (including but not limited to: whore, slut, perverts)

4. General

- Expressions about being less than adequate (including but not limited to: worthless, useless)
- Expressions about being better/worse than another PC (including but not limited to: I believe that males are superior to females).
- Expressions about deviating from the norm (including but not limited to: freaks, abnormal)

Red text = areas we considered scaling back protections



Refine Hate Speech Tier 2

Refined version of Tier 2

2. Expressions of Contempt or their visual equivalent, including but not limited to

- Self-admission to intolerance on the basis of a protected characteristics (including but not limited to: homophobic, islamophobic, racist)
- Expressions that a PC shouldn’t exist
- Expressions of hate (including but not limited to: despise, hate)
- Dismissal (including but not limited to: don´t respect, don´t like, don´t care for)

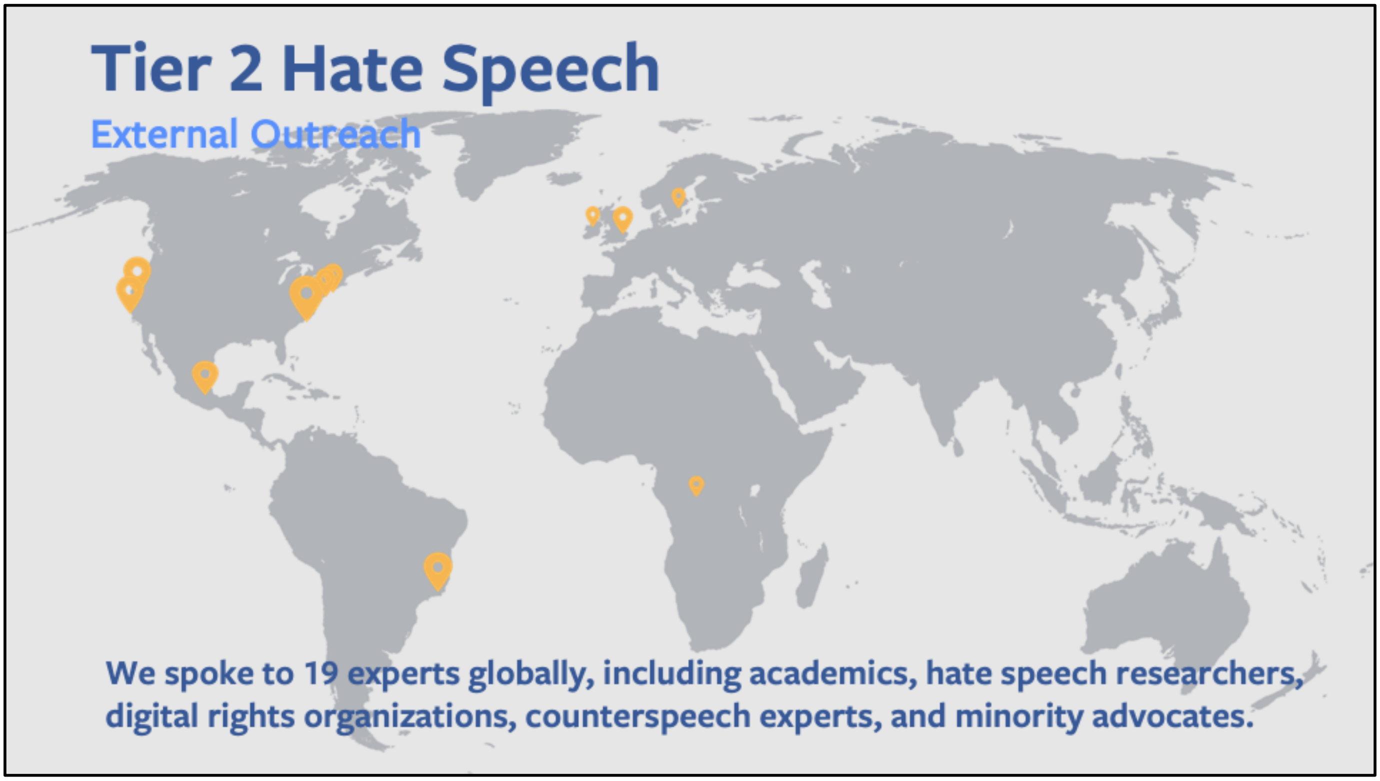
3. Expressions of disgust or their visual equivalent, including but not limited to:

- Expressions of a protected characteristic causing sickness (including but not limited to: vomit, throw up)
- Expressions of repulsion or distaste (including but not limited to: vile, disgusting, yuck)

4. Cursing at a person or group of people who share protected characteristics, such as:

- Referring to the target as genitalia or anus (including but not limited to: cunt, dick, asshole)
- Profane terms or phrases with the intent to insult (including but not limited to: fuck, bitch, motherfucker)
- Terms or phrases calling for engagement in sexual activity, or contact with genitalia or anus, or with feces or urine (including but not limited to: suck my dick, kiss my ass, eat shit)

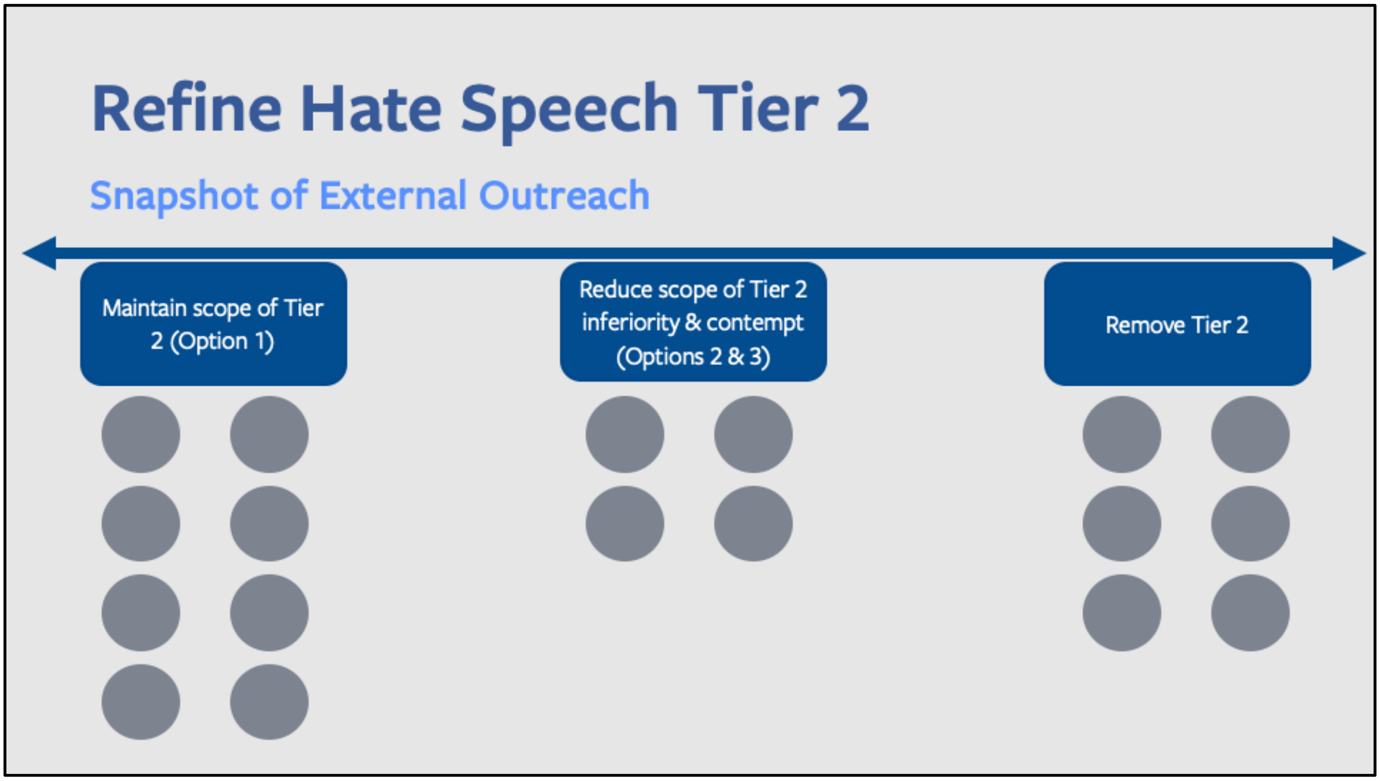
Red text = areas we considered scaling back protections



Tier 2 Hate Speech

External Outreach

We spoke to 19 experts globally, including academics, hate speech researchers, digital rights organizations, counterspeech experts, and minority advocates.



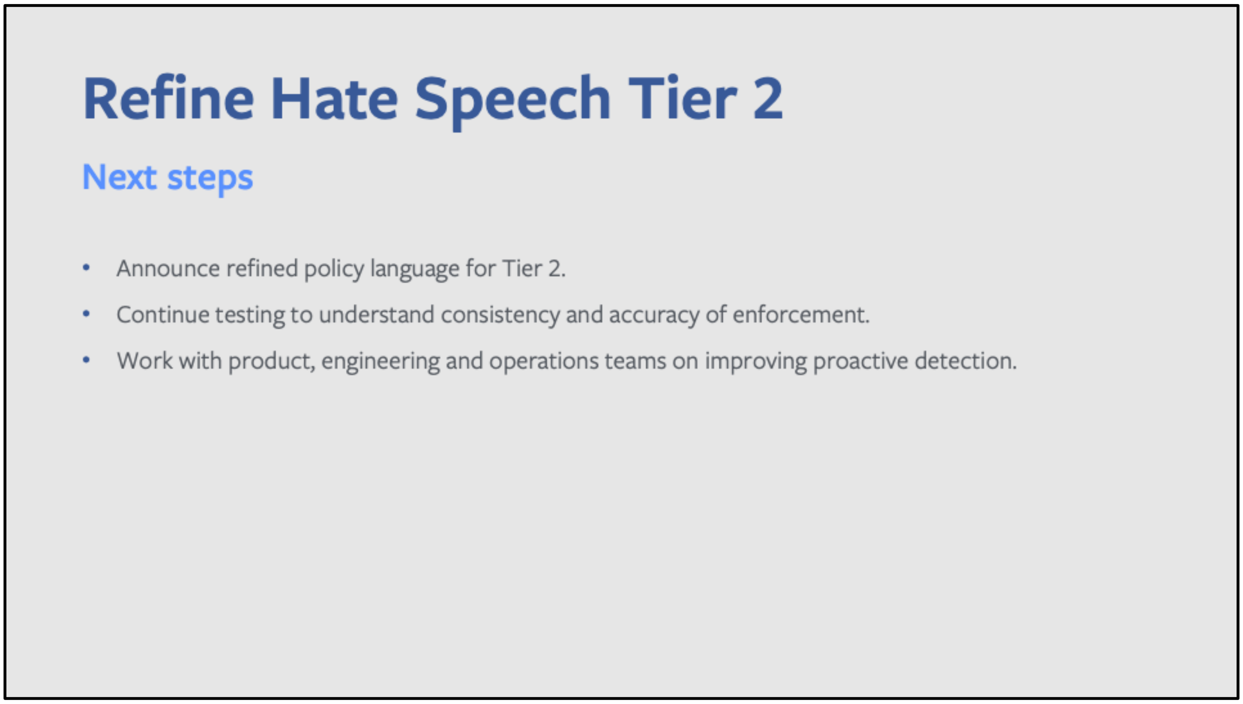
Refine Hate Speech Tier 2

Snapshot of External Outreach

Maintain scope of Tier 2 (Option 1)

Reduce scope of Tier 2 inferiority & contempt (Options 2 & 3)

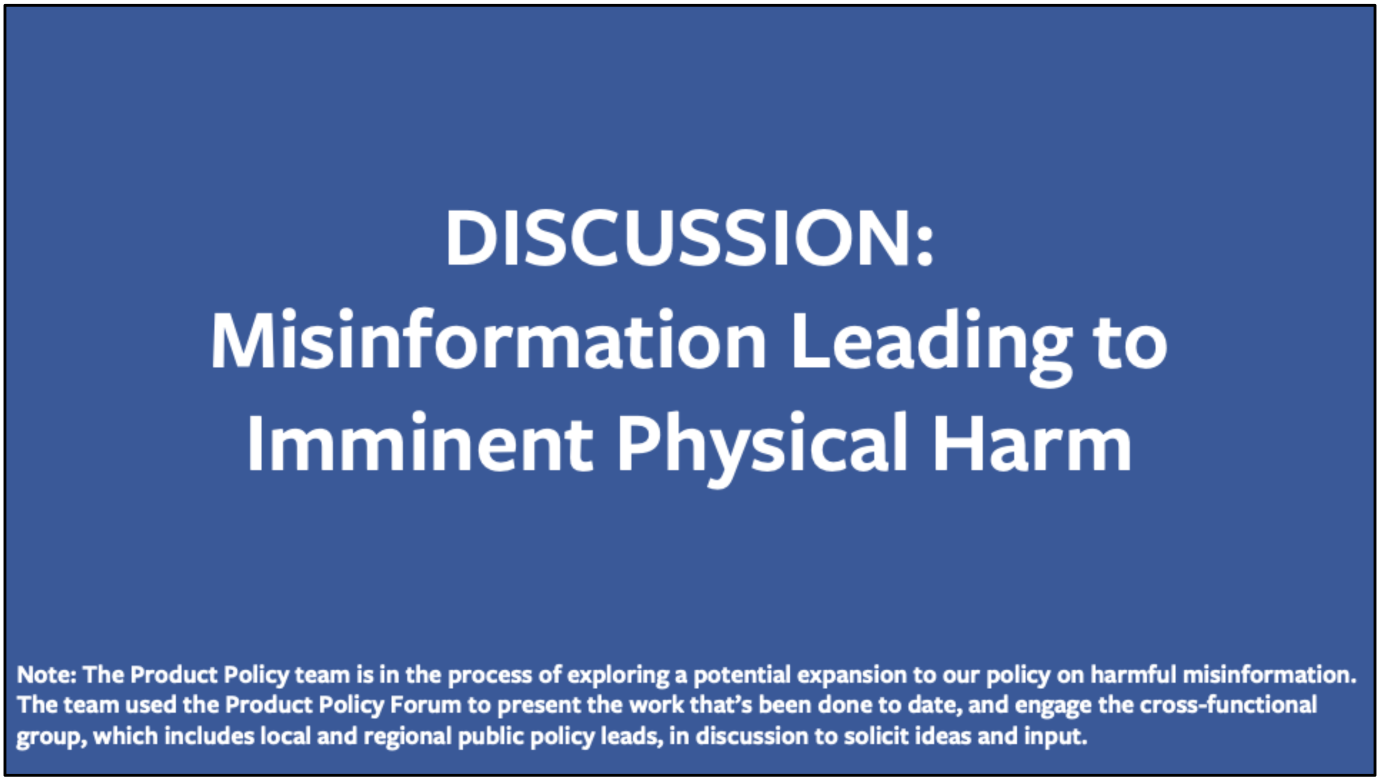
Remove Tier 2



Refine Hate Speech Tier 2

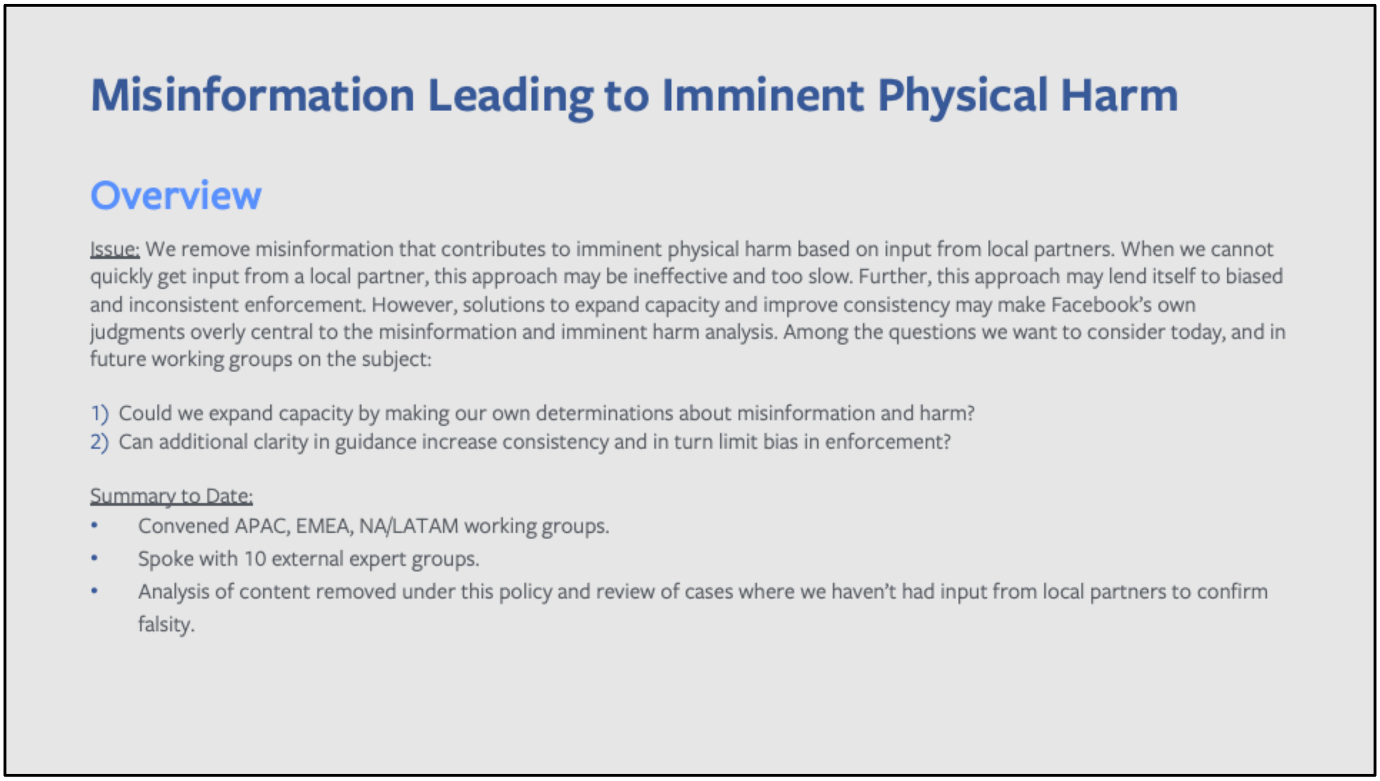
Next steps

- Announce refined policy language for Tier 2.
- Continue testing to understand consistency and accuracy of enforcement.
- Work with product, engineering and operations teams on improving proactive detection.



DISCUSSION: Misinformation Leading to Imminent Physical Harm

Note: The Product Policy team is in the process of exploring a potential expansion to our policy on harmful misinformation. The team used the Product Policy Forum to present the work that's been done to date, and engage the cross-functional group, which includes local and regional public policy leads, in discussion to solicit ideas and input.



Misinformation Leading to Imminent Physical Harm

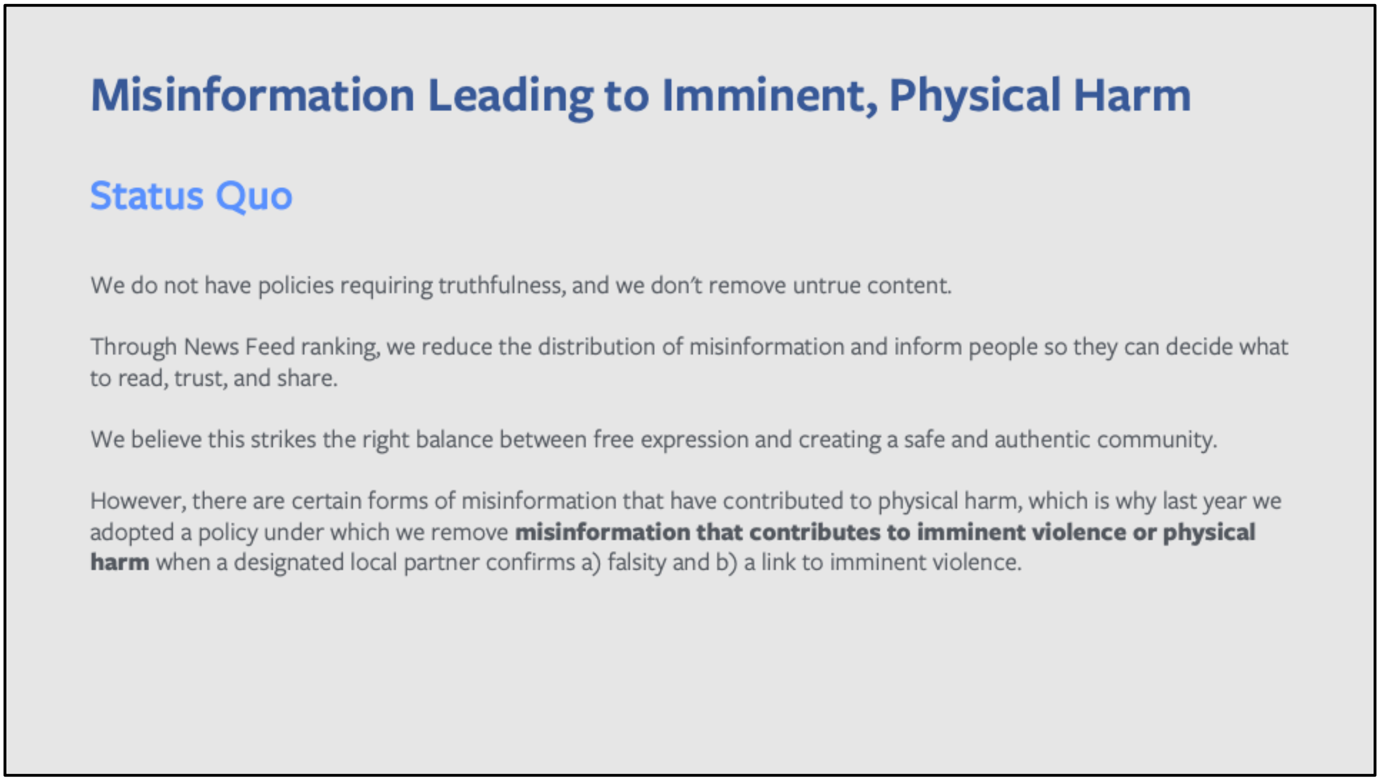
Overview

Issue: We remove misinformation that contributes to imminent physical harm based on input from local partners. When we cannot quickly get input from a local partner, this approach may be ineffective and too slow. Further, this approach may lend itself to biased and inconsistent enforcement. However, solutions to expand capacity and improve consistency may make Facebook's own judgments overly central to the misinformation and imminent harm analysis. Among the questions we want to consider today, and in future working groups on the subject:

1) Could we expand capacity by making our own determinations about misinformation and harm?
2) Can additional clarity in guidance increase consistency and in turn limit bias in enforcement?

Summary to Date:

- Convened APAC, EMEA, NA/LATAM working groups.
- Spoke with 10 external expert groups.
- Analysis of content removed under this policy and review of cases where we haven't had input from local partners to confirm falsity.



Misinformation Leading to Imminent, Physical Harm

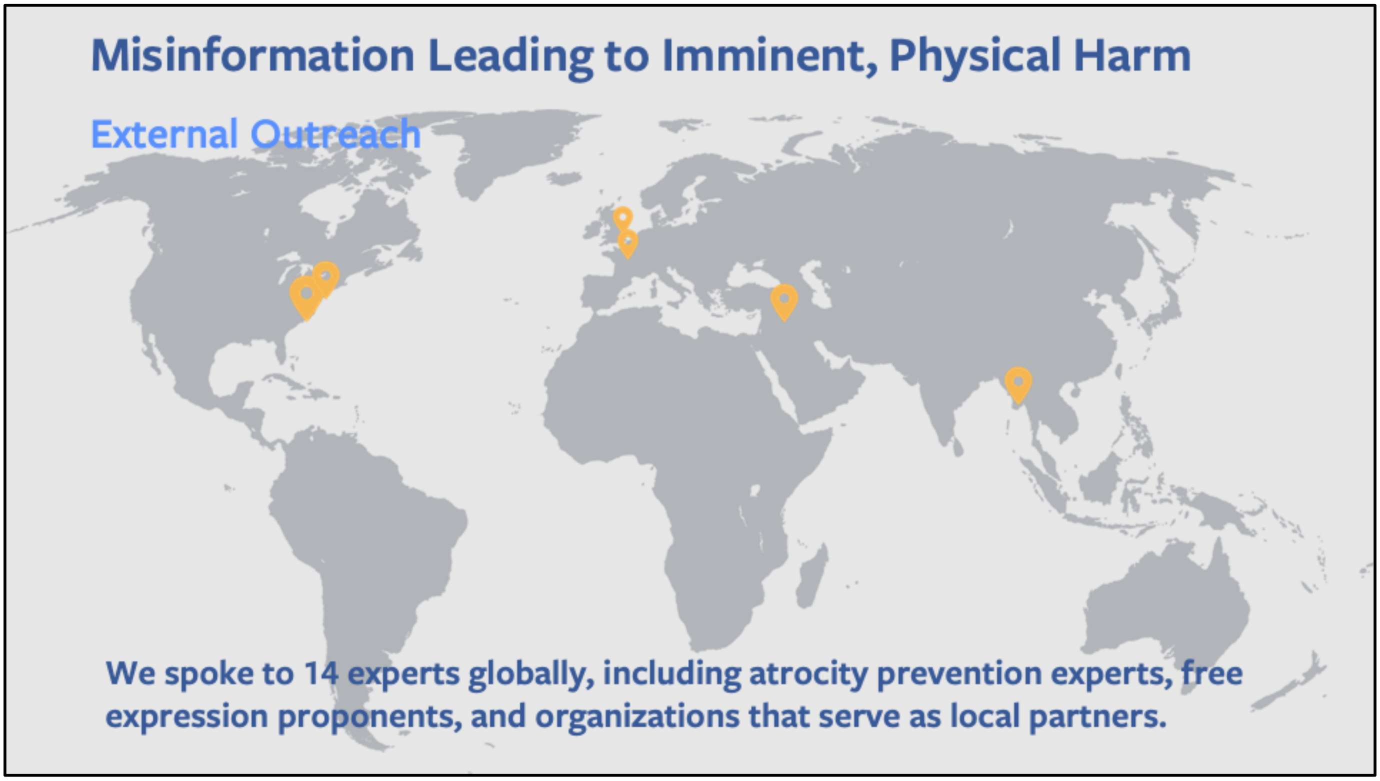
Status Quo

We do not have policies requiring truthfulness, and we don't remove untrue content.

Through News Feed ranking, we reduce the distribution of misinformation and inform people so they can decide what to read, trust, and share.

We believe this strikes the right balance between free expression and creating a safe and authentic community.

However, there are certain forms of misinformation that have contributed to physical harm, which is why last year we adopted a policy under which we remove **misinformation that contributes to imminent violence or physical harm** when a designated local partner confirms a) falsity and b) a link to imminent violence.



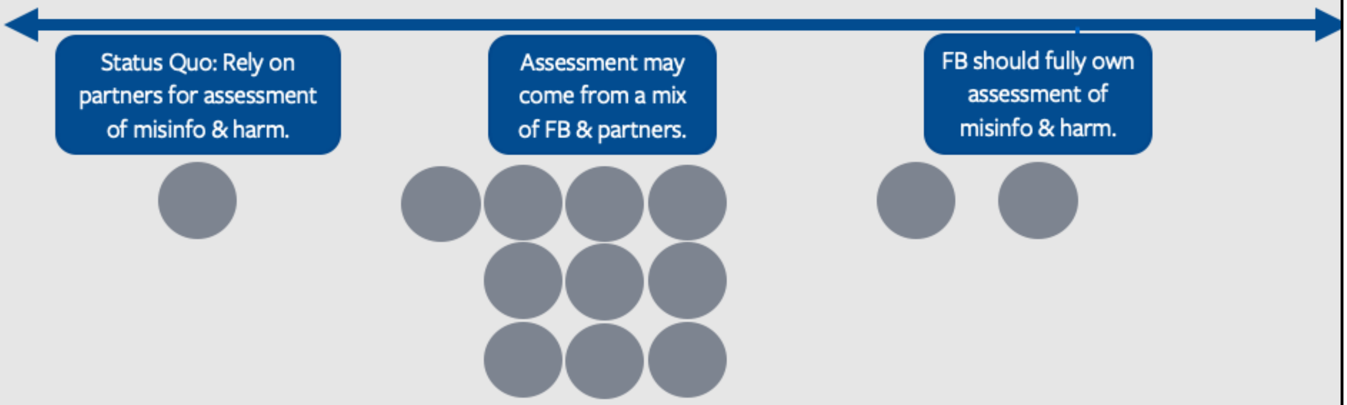
Misinformation Leading to Imminent, Physical Harm

External Outreach

We spoke to 14 experts globally, including atrocity prevention experts, free expression proponents, and organizations that serve as local partners.

Misinformation Leading to Imminent, Physical Harm

Snapshot of External Outreach



Status Quo: Rely on partners for assessment of misinfo & harm.

Assessment may come from a mix of FB & partners.

FB should fully own assessment of misinfo & harm.

facebook