Product Policy Forum

May 21, 2019

TOPICS: *Hate Speech Behavior vs. Generalizations , Marijuana and Cannabis Products, Manipulated Media*

Agenda

1 RECOMMENDATION: HATE SPEECH BEHAVIOR VS. GENERALIZATIONS

2 RECOMMENDATION: MARIJUANA AND CANNABIS PRODUCTS

3 HEADS UP: MANIPULATED MEDIA

Hate Speech
Behavior vs. Generalizations

Behavior vs. Generalizations

Overview

Issue: We remove attacks, including generalizations, against people based on protected or quasi-protected characteristics (PC or QPC), but at scale, we've been inconsistent in the way we enforce on statements about the behavior of people. We want to remove speech that ascribes violent or injurious behaviors (murder, rape, etc.) to people who share protected characteristics; however, we do not want to over-enforce on speech that may have public interest, such as discussions of crimes and atrocities or debates around immigration policy.

Engagement:

- 8 working groups
- Collected and labeled examples from 8 countries
- 21 external engagements

Recommendation:

- Continue to remove generalized statements against PCs and QPCs, as well as certain behavior statements related to Tier 1 attacks.
- Pursue policy development on hateful stereotypes as an immediate follow-up.

Behavior vs. Generalizations

Status Quo

Tier 1

Dehumanizing speech or imagery, such as **reference or comparison** to:

- Insects (e.g., cockroaches)
- Animals (e.g., pig)
- Filth, bacteria, disease, or feces
- Sub-humanity (e.g., savages, primitives)
- Sexual predators
- Violent Criminals
- Other criminals

Tier 2

Statements of inferiority (a **statement or term or image implying a person's or a group's physical, mental, or moral deficiency**):

- Physical (e.g., deformed, undeveloped)
- Mental (e.g., retarded, stupid)
- Moral (e.g., greedy, slutty)

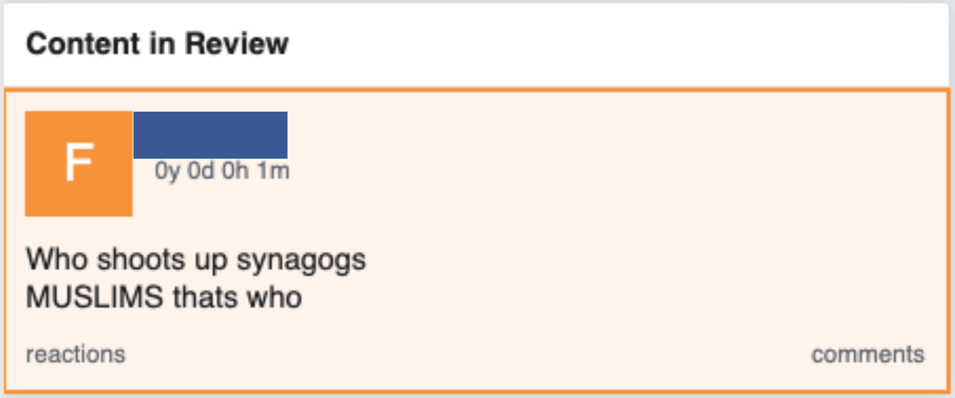
Behavior vs. Generalizations

Status Quo

REMOVE	GRAY AREA	ALLOW
Men are rapists.	Men are raping people all over America.	This man raped me.
Migrants are terrorists.	Migrants are raping, killing, stealing and terrorizing.	An Afghani asylum applicant stabbed a pregnant Polish woman last Friday.
Woman are liars.	Women lie.	The women who accused R Kelly are lying.

Behavior vs. Generalizations

Examples (Currently Allowed)

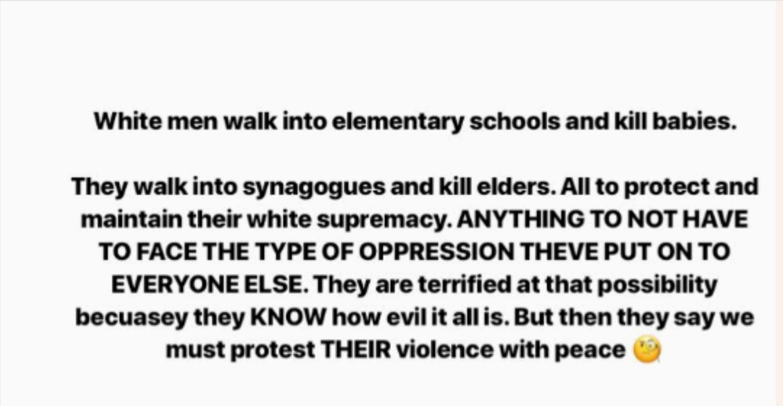


Content in Review

F [redacted]

0y 0d 0h 1m

Who shoots up synagogs
MUSLIMS thats who

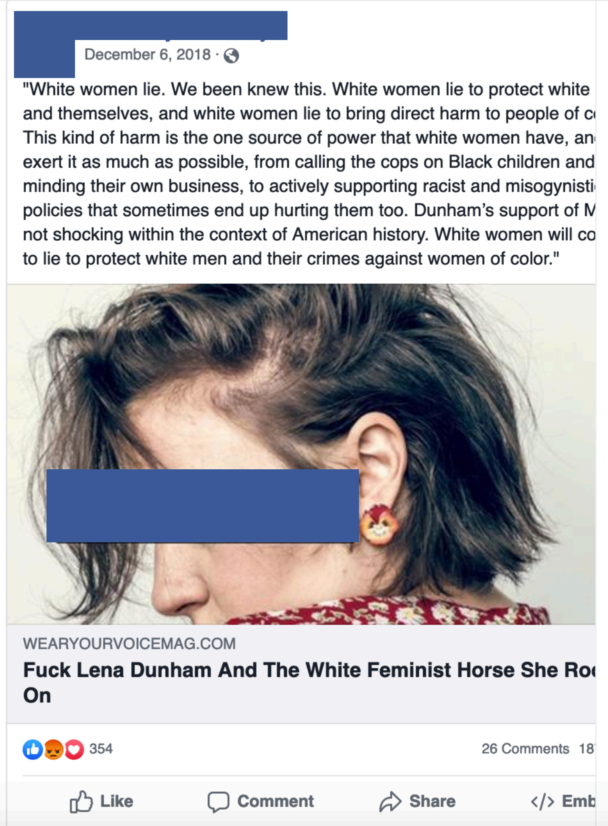


White men walk into elementary schools and kill babies.

They walk into synagogues and kill elders. All to protect and maintain their white supremacy. ANYTHING TO NOT HAVE TO FACE THE TYPE OF OPPRESSION THEVE PUT ON TO EVERYONE ELSE. They are terrified at that possibility becuasey they KNOW how evil it all is. But then they say we must protest THEIR violence with peace 🧐

Behavior vs. Generalizations

Examples (Currently Allowed)



[redacted]

December 6, 2018 ·

"White women lie. We been knew this. White women lie to protect white
and themselves, and white women lie to bring direct harm to people of c
This kind of harm is the one source of power that white women have, an
exert it as much as possible, from calling the cops on Black children and
minding their own business, to actively supporting racist and misogynisti
policies that sometimes end up hurting them too. Dunham's support of M
not shocking within the context of American history. White women will co
to lie to protect white men and their crimes against women of color."

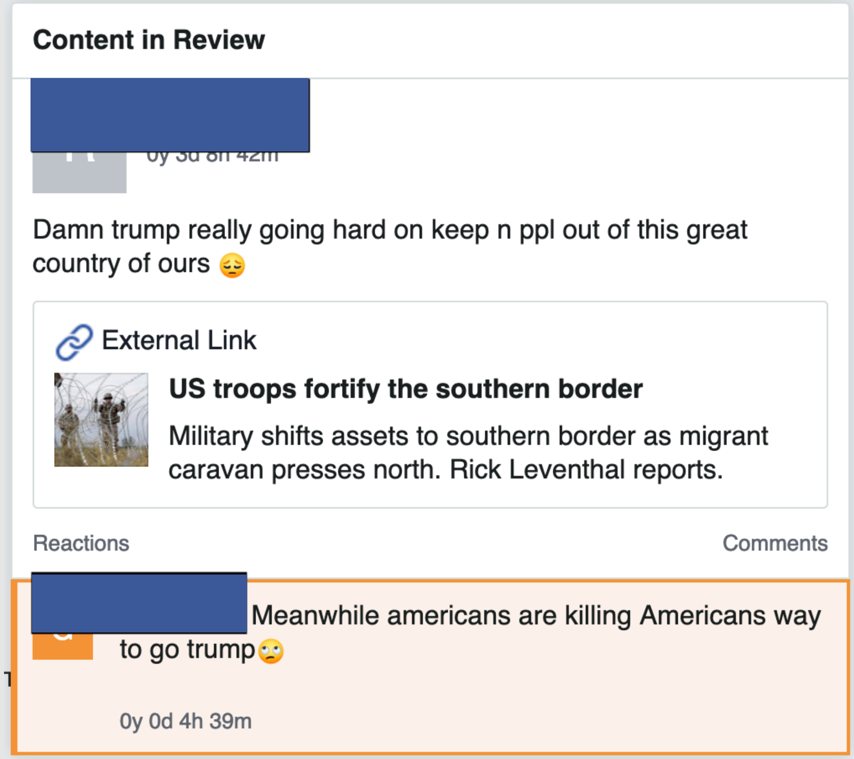
[redacted]

WEARYOURVOICEMAG.COM

**Fuck Lena Dunham And The White Feminist Horse She Ro
On**

354

26 Comments 18



Content in Review

[redacted]

Damn trump really going hard on keep n ppl out of this great country of ours 😔

External Link

US troops fortify the southern border

Military shifts assets to southern border as migrant caravan presses north. Rick Leventhal reports.

Reactions

Comments

[redacted] Meanwhile americans are killing Americans way to go trump🙄

0y 0d 4h 39m

Behavior vs. Generalizations

On-Platform Research

Exercise:

- Subject matter experts labeled hundreds of pieces of content across multiple countries.
- Labeling included reference to attack type (Tier 1, 2 or 3) and protected characteristic targeted.

From this, we were trying to understand:

- How prevalent are behavioral attacks?
- Who is most frequently targeted with behavioral attacks?
- Do behavioral attacks tend to fall into Tier 1, Tier 2 or Tier 3?
- And what is the geographic spread of these statements?

Behavior vs. Generalizations

Policy Research Key Findings

Relevant Internal and External Research:

- Generalizations can be linguistically distinguished by quantity words (i.e. millions), religious group names (i.e. Jews), & lethal words (i.e. kill)
- Behavior specific language, or more direct speech, is informal, angrier, and often attacks a target using words to hinder their action
- Generalized attacks can be associated with a fixed mindset whereas statements about behavior may be associated with a growth mindset because behaviors are changeable; for this reason, psychologists warn that generalizations can be more harmful
    - Fixed mindset: believes that human attributes are fixed traits
    - Growth mindset: believes that all people, no matter who they are, can take steps to develop over time

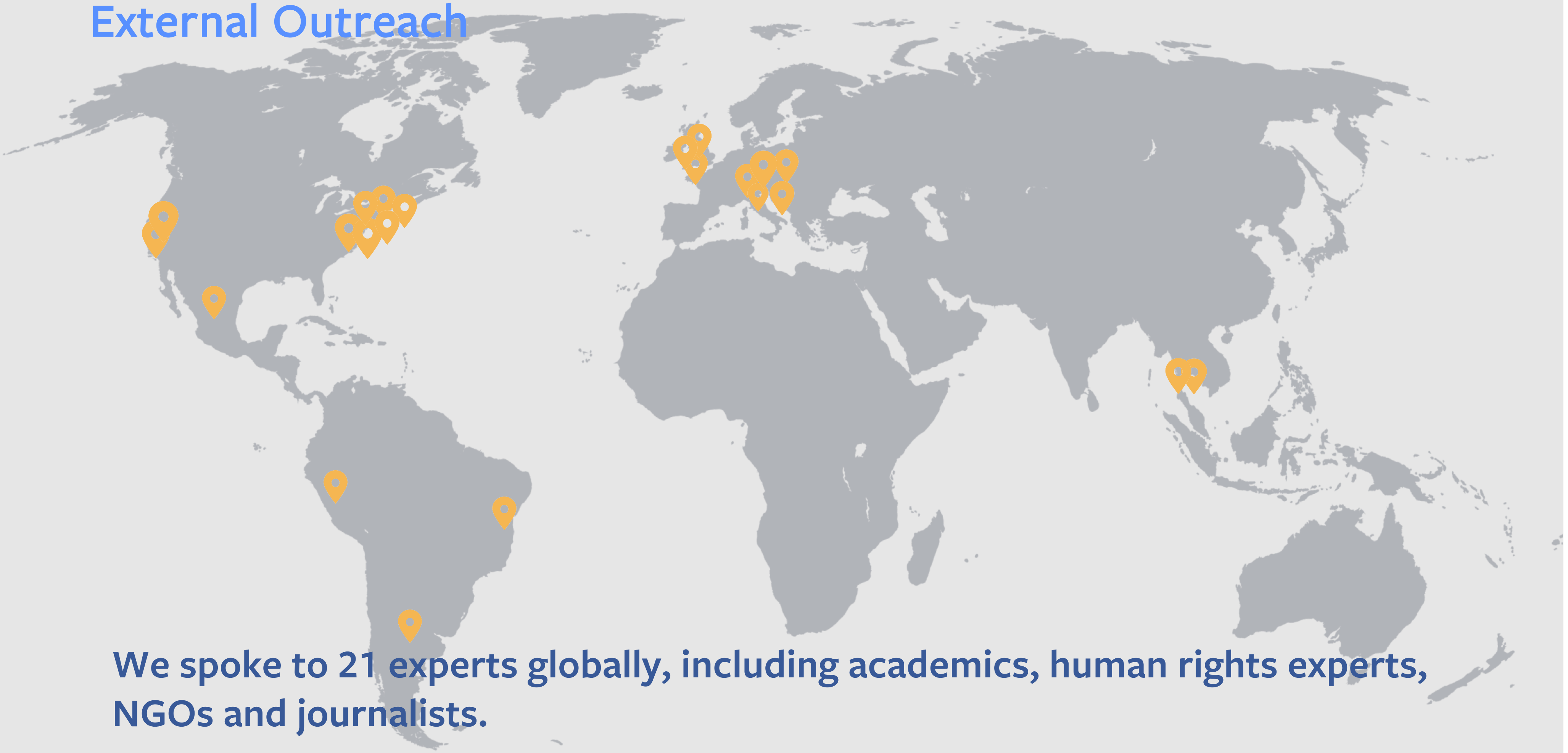
Policy Relevance: Generalizations and certain stereotypes can have negative and sustained consequences and may be linguistically distinguished from behavioral statements

Sources: El Sherief et al., 2018; Dweck, 2012; & Brigham, 1971; Berkman Klein Center, 2017

External Outreach

Behavior vs. Generalizations

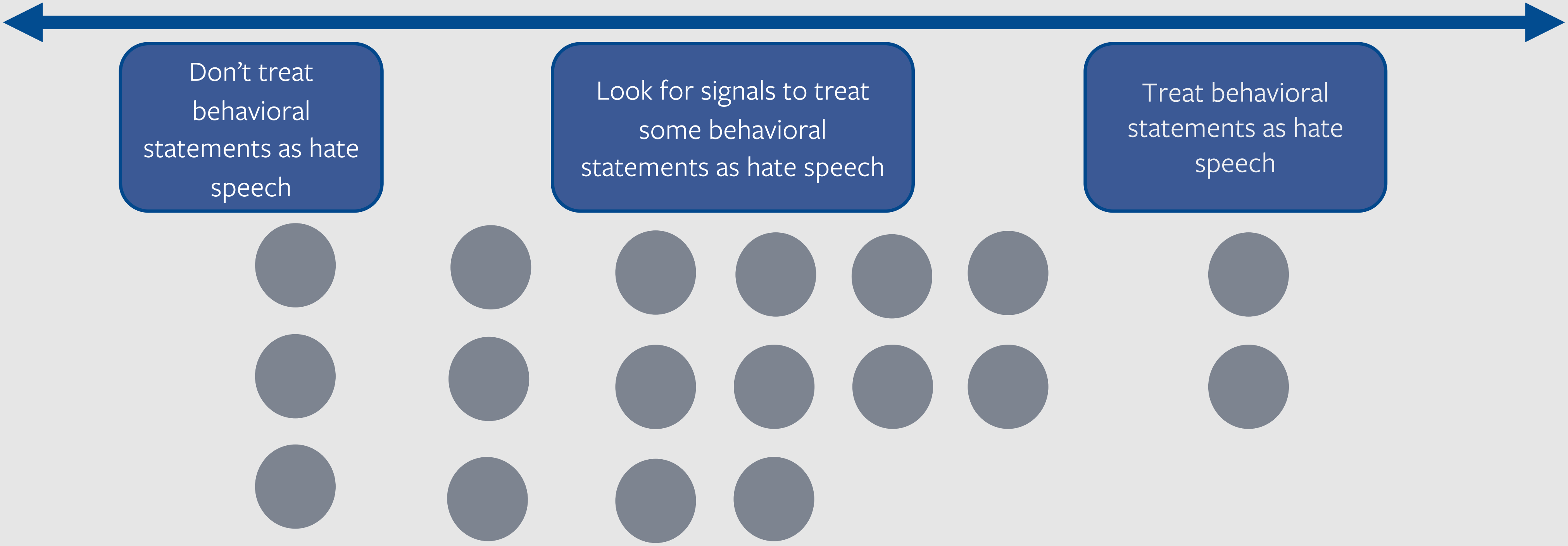
External Outreach



We spoke to 21 experts globally, including academics, human rights experts, NGOs and journalists.

Hate speech behavior vs. generalizations

Mapping of External Outreach



Don't treat behavioral statements as hate speech
Look for signals to treat some behavioral statements as hate speech
Treat behavioral statements as hate speech

Recommendation

Behavior vs. Generalizations

Recommendation

PART ONE – Address behavioral statements

- Option 2 – Remove some negative behavioral statements (Tier 1 attacks only)

PART TWO – Further policy development on hateful stereotypes

- Pursue a separate policy update to address “hateful stereotypes”

Options

Behavior vs. Generalizations

Option 1: Maintain status quo

- Our policy prohibits attacks, including generalizations, made against people on the basis of PCs and QPCs
- However, we don’t clearly distinguish between or define behavioral statements and generalizations.
- On escalation, Content Policy provides guidance for specific cases.

Pros:

- Protects speech that has public interest value (i.e. speech about crimes)

Cons:

- More room for discretion in enforcement creates room for bias
- Inconsistent, and therefore inequitable, enforcement
- High volume of escalations/questions

Behavior vs. Generalizations

Option 1: Maintain status quo

REMOVE	GRAY AREA	ALLOW
• “Muslims are rapists.”	• “White men walk into elementary schools and kill babies. They walk into synagogues and kill elders.” • “White women lie, [They] lie to protect white men and themselves...” • “Who shoots up synagogues? Muslims that’s who.” • “Muslim migrants are raping women in the streets.”	• “This man raped me.”

Behavior vs. Generalizations

Option 2: Remove some behavioral statements related to Tier 1 attacks (Recommendation)

Remove behavioral statements related to a Tier 1 attack UNLESS the content includes:

- qualifiers that limit the scope of the PC or QPC targeted, OR
- indicators that the statement references a specific event, pattern of events, or discussion of criminal behavior, OR
- language that describes direct experience or reporting (e.g., "I saw," "I experience," "this happened to me")

Allow behavioral statements related to a Tier 2 attack against members of a PC.

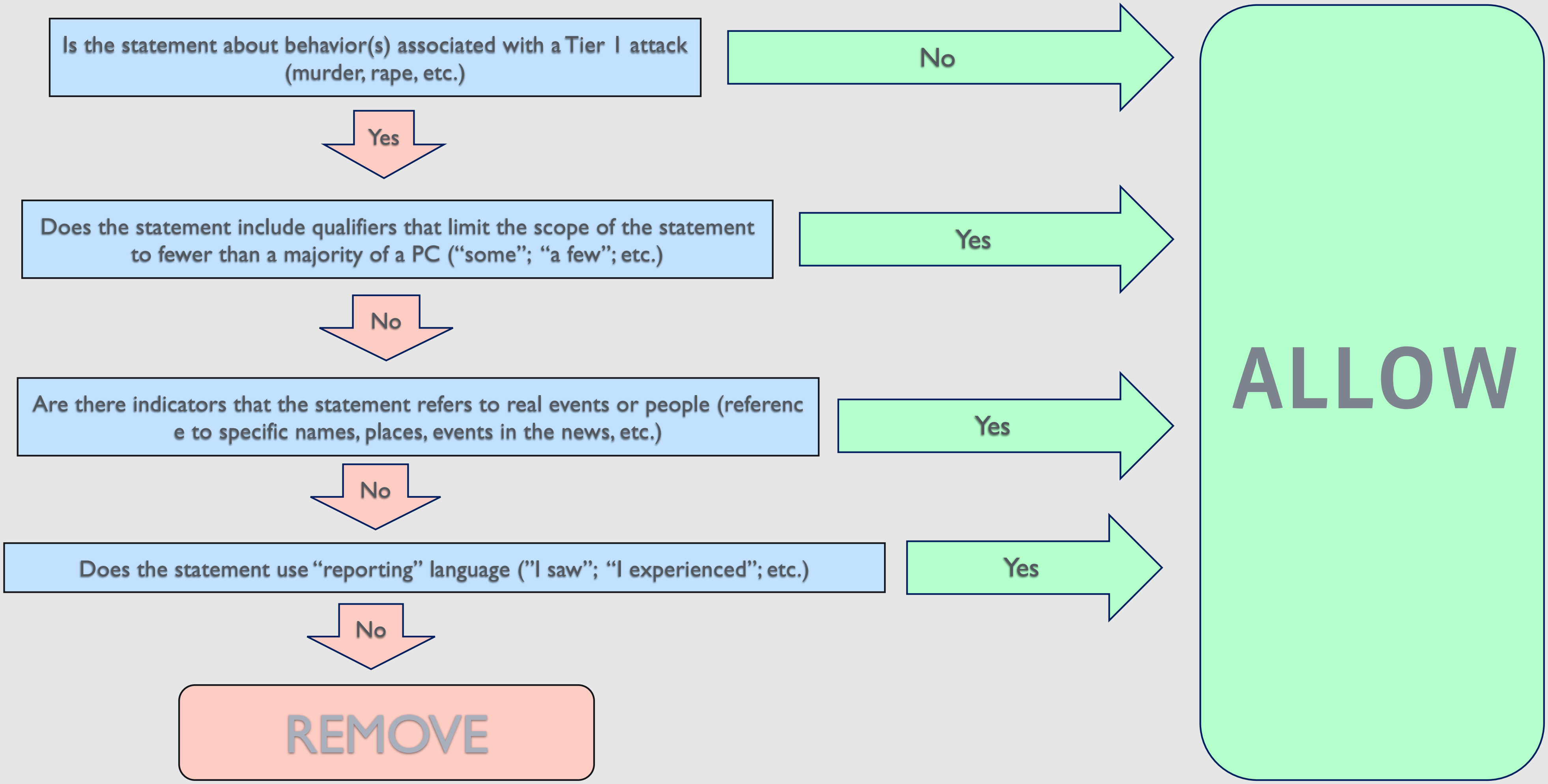
Pros:

- Captures more content widely perceived as hateful

Cons:

- Potentially overbroad
- May be challenging to operationalize consistently

Behavior vs. Generalizations



Is the statement about behavior(s) associated with a Tier 1 attack (murder, rape, etc.)

No → ALLOW

Yes ↓

Does the statement include qualifiers that limit the scope of the statement to fewer than a majority of a PC (“some”; “a few”; etc.)

Yes → ALLOW

No ↓

Are there indicators that the statement refers to real events or people (reference to specific names, places, events in the news, etc.)

Yes → ALLOW

No ↓

Does the statement use “reporting” language (”I saw”; “I experienced”; etc.)

Yes → ALLOW

No ↓

REMOVE

Behavior vs. Generalizations

Option 2: Remove some behavioral statements related to Tier 1 attacks (Recommendation)

REMOVE	ALLOW
• “Muslims are rapists.” • “Muslim migrants are raping women in the streets.”	• “This man raped me.” • “White men walk into elementary schools and kill babies. They walk into synagogues and kill elders.” • “Who shoots up synagogues? Muslims that’s who.” • “White women lie all the time. [They] lie to protect white men and themselves...”

Behavior vs. Generalizations

Option 3: Remove some behavioral statements

Remove behavioral statements UNLESS the content includes:

- qualifiers that limit the scope of the PC targeted, OR
- indicators that the statement references a specific event, pattern of events, or discussion of criminal behavior, OR
- language that describes direct experience or reporting (e.g., “I saw,” “I experience,” “this happened to me”)

Pros:

- Captures more content widely perceived as hateful

Cons:

- May be challenging to operationalize consistently
- May remove counter-speech and/or content widely perceived as benign

Behavior vs. Generalizations

Option 3: Remove some behavioral statements

REMOVE	ALLOW
• “Muslims are rapists.” • “Muslim migrants are raping women in the streets.” • “White women lie all the time. [They] lie to protect white men and themselves…”	• “This man raped me.” • “White men walk into elementary schools and kill babies. They walk into synagogues and kill elders.” • “Who shoots up synagogues? Muslims that’s who.”

Behavior vs. Generalizations

Option 4: Define and allow behavioral statements

- Behavioral statements about people belonging to a PC will not violate our hate speech policies.
- In cases where we might have more context available to us that indicates content serves to generalize the behavior of members of a PC or QPC, the content policy team can make the decision to remove.

Pros:

- Less likelihood of mistakes or removals of content widely perceived as benign

Cons:

- Certain languages do not make this distinction
- Difficult to explain in some cases
  (difference between people who rape and rapists?)
- Allows content that is widely perceived to be highly toxic/hateful

Behavior vs. Generalizations

Option 4: Define and allow behavioral statements

REMOVE	ALLOW
• “Muslims are rapists.”	• “This man raped me.” • “White men walk into elementary schools and kill babies. They walk into synagogues and kill elders.” • “White women lie all the time. [They] lie to protect white men and themselves…””” • “Who shoots up synagogues? Muslims that’s who.” • “Muslim migrants are raping women in the streets.”

Behavior vs. Generalizations

Option 5: Remove behavioral statements linked to designated dehumanizing comparisons

- Remove behavioral statements ONLY when linked to a designated dehumanizing comparison
- Expand the list of designated dehumanizing comparisons to capture some of the more prevalent and toxic types of behavioral statements (e.g., migrants as violent/sexual criminals, Jews controlling the world)

Pros:

- Captures “worst of the worst”
- Narrowly targets problematic content, limiting the likelihood of mistakes
- Easier to operationalize

Cons:

- May not remove content widely perceived as hateful
- Potential overlap with other violation types
- Limits our ability to quickly respond to changing trends in speech

Behavior vs. Generalizations

Option 5: Remove behavioral statements linked to designated dehumanizing comparisons

REMOVE	ALLOW
• “Muslims are rapists.” • “Muslim migrants are raping women in the streets.” • “Who shoots up synagogues? Muslims that’s who.”	• “This man raped me.” • “White men walk into elementary schools and kill babies. They walk into synagogues and kill elders.” • “White women lie all the time. [They] lie to protect white men and themselves…”

Behavior vs. Generalizations

Options matrix

	Option 1	***Option 2 Recommendation***	Option 3	Option 4	Option 5
"Muslims are rapists."	X	X	X	X	X
"Muslim migrants are raping women in the streets."	?	X	✓	X	X
"Who shoots up synagogues? Muslims that's who."	?	✓	✓	✓	X
"White men walk into elementary schools and kill babies. They walk into synagogues and kill elders."	?	✓	✓	✓	✓
"White women lie all the time. [They] lie to protect white men and themselves..."	?	✓	✓	X	✓
"This man raped me."	✓	✓	✓	✓	✓

Behavior vs. Generalizations

Further policy development on hateful stereotypes

- Historical and social context are important to consider for certain attacks against PCs.
- With this in mind, we will undertake a separate policy development process focused on “designated hateful stereotypes”:
  - Hateful stereotypes is broader than the issue we are trying to solve for with this recommendation.
  - Developing a policy to address hateful stereotypes globally will require a thorough process to ensure the policy is principled, operable and explicable.

Behavior vs. Generalizations

Next steps

- Launch policy update
- Policy development: Hateful Stereotypes
- Partner with Community Operations to further test and analyze examples that fall within ambit of policy to understand whether refinement is necessary in future

Recommendation:
Marijuana and Cannabis Products

Marijuana and Cannabis Products

Issue Statement

Issue: Our policies do not allow for the sale of marijuana in organic content. We want to consider whether we should adapt this policy in light of shifting trends in certain countries. However, since marijuana remains subject to different social, cultural, political, and legal environments around the world, changes to policy must be assessed for operational feasibility and appropriateness globally, including in regions with a more restrictive treatment.

Summary to Date:

- 4 working groups
- 13 external engagements

Recommendation: Maintain status quo; Further define permissible language for marijuana content, and prioritize training to improve accuracy on content related to marijuana and cannabis products.

Marijuana and Cannabis Products

Status Quo

Remove:

Content that depicts the sale or attempt to purchase marijuana:

- Mentions or depicts marijuana or pharmaceutical drugs, and
- Makes an attempt to donate or sell or trade

What is prohibited under our Community Standards?

- Marijuana & marijuana products
- “Spice” or any other synthetic cannabis that is used to achieve a “high”
- Other forms of THC such as “hash oil,” “weed oil,” or “liquid cannabis”

What is not prohibited under our Community Standards?

- Marijuana seeds
- Marijuana paraphernalia (e.g. bongs, pipe)
- CBD or cannabidiol products

Marijuana and Cannabis Products

Research Findings

Internal and External Research:

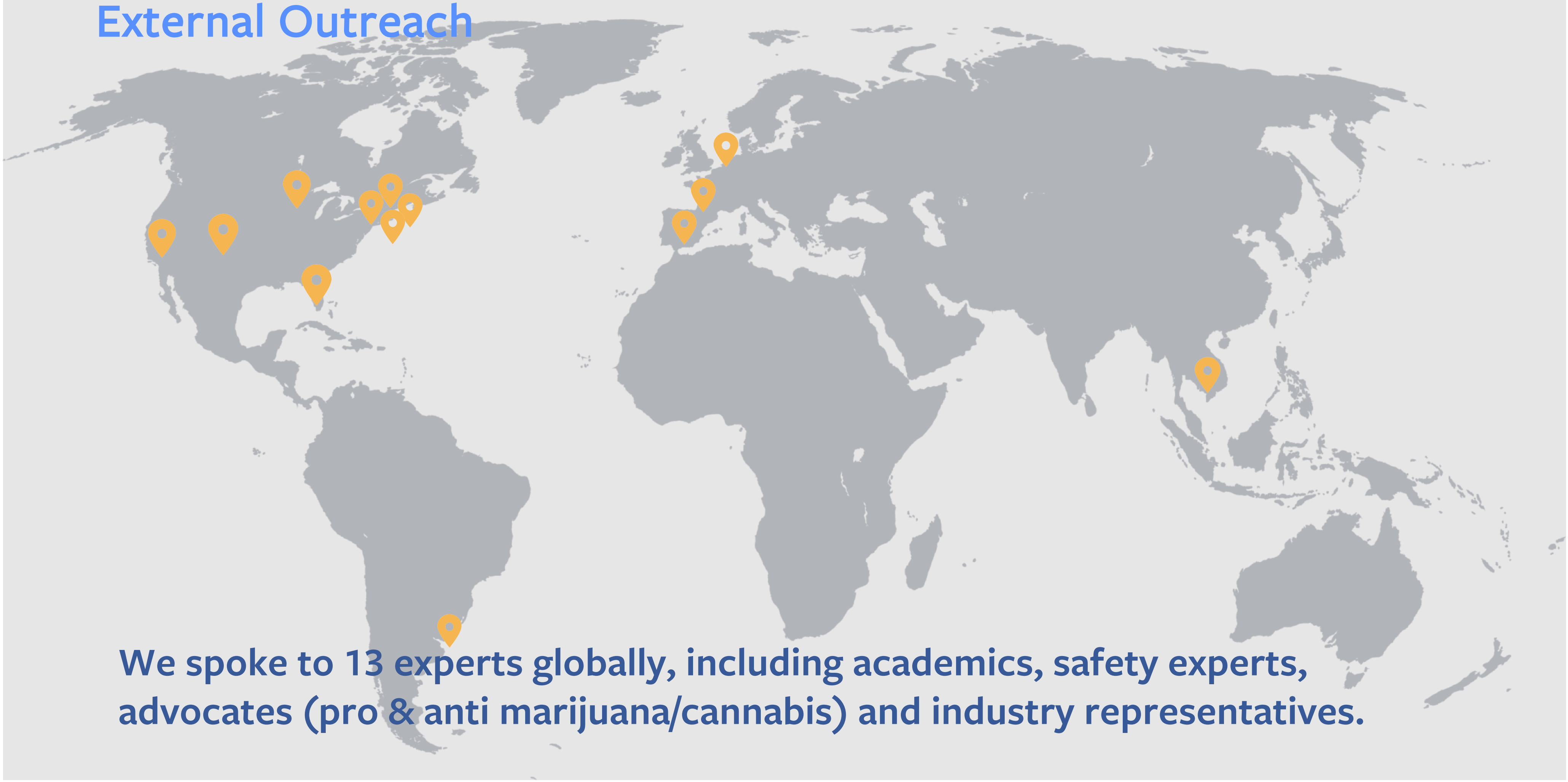
- Marijuana is the most commonly used drug in the world, but global policies vary.
- Facebook is not in a position to verify or vet sellers of marijuana.
- A change in policy will be positive for some people, neutral for others, and negative for still others.

Policy Relevance: any policy change should take into account and adequately address geographical variations and user experiences.

Sources: Internal Research; UN World Drug Report, 2018; Cannabis Business Times, 2017; Business Insider, Jan 2019

Marijuana and Cannabis Products

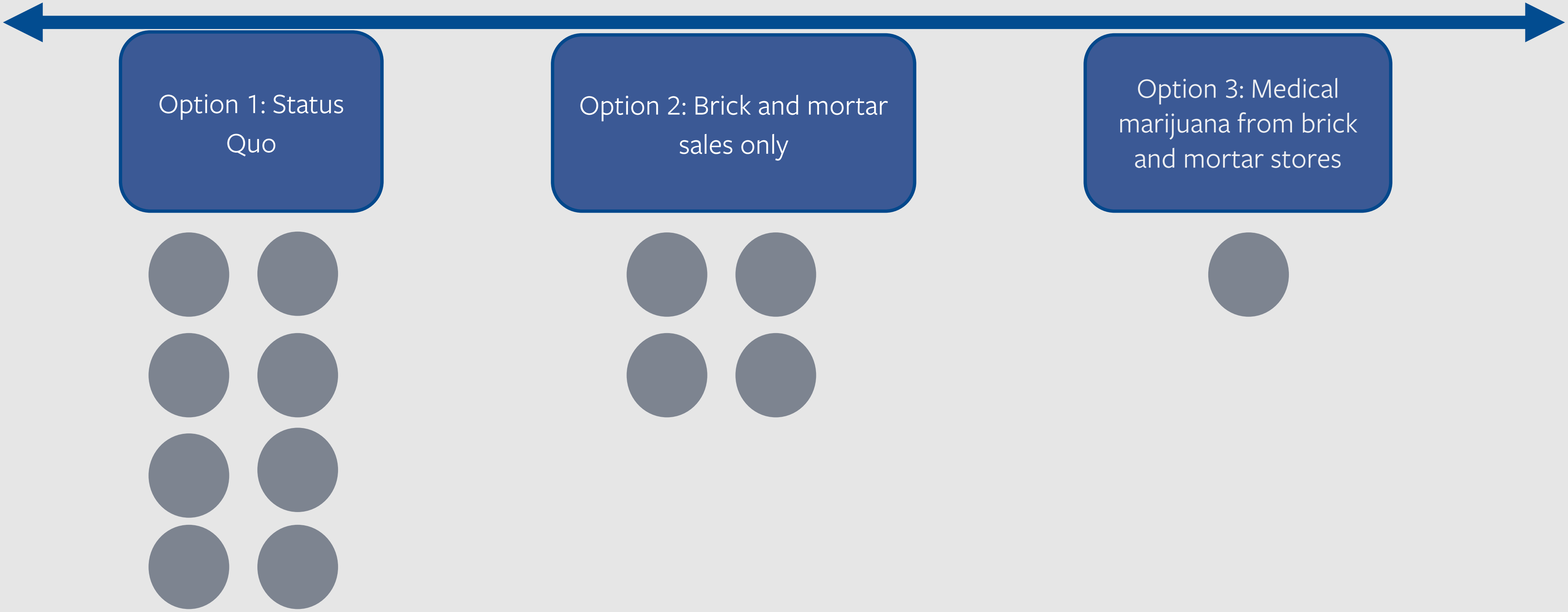
External Outreach



We spoke to 13 experts globally, including academics, safety experts, advocates (pro & anti marijuana/cannabis) and industry representatives.

Marijuana and Cannabis products

Mapping of External Outreach



Option 1: Status Quo

Option 2: Brick and mortar sales only

Option 3: Medical marijuana from brick and mortar stores

Marijuana and Cannabis Products

Option 1: Status Quo Policy + Enforcement Updates (Recommendation)

- Prohibit content that attempts to sell, raffle, gift, transfer, trade or purchase marijuana, regardless of its legal status in the specific jurisdiction. This prohibition also applies commercially to brick and mortar stores.

Pros:

- Aligns with current socio-political and legal environments in many countries.
- Meets operational need for global policy.

Cons:

- As social, cultural, political, and legal environments evolve on a country-by-country basis, Facebook policy may feel restrictive in some places and too permissive in others.

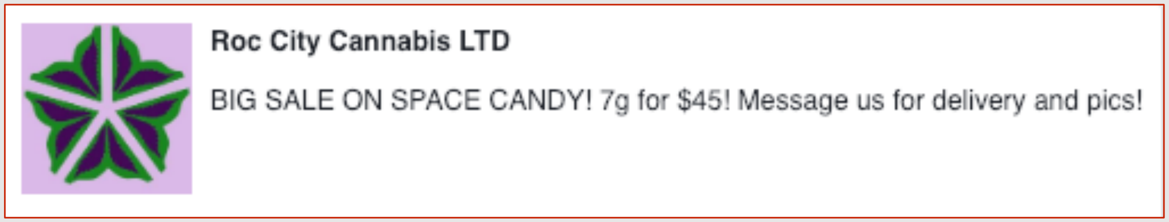
Marijuana and Cannabis Products

Option 1: Enforcement Updates

- Work with Community Operations Training team to understand reviewer challenges, confusion or error, and improve training accordingly
- Improve communication to dispensaries to help them understand our policy (e.g., revise Page appeals messaging to make it clear that marijuana sales are prohibited from commercial and retail pages in addition to peer-to-peer sales)
- Update Operational Guidelines to help prevent over-enforcement on dispensaries that aren’t using platform for sales

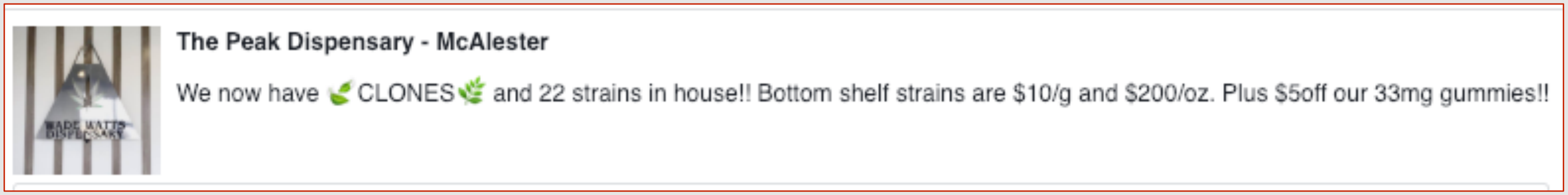
Marijuana and Cannabis Products

Option 1: Examples (not allowed)



Roc City Cannabis LTD

BIG SALE ON SPACE CANDY! 7g for $45! Message us for delivery and pics!



The Peak Dispensary - McAlester

We now have 🌱CLONES🌿 and 22 strains in house!! Bottom shelf strains are $10/g and $200/oz. Plus $5off our 33mg gummies!!

Marijuana and Cannabis Products

Option 2: Allow marijuana sales from brick & mortar stores

Allow the sale, trade, and advertisement of marijuana from brick and mortar stores. But maintain the peer-to-peer prohibition on marijuana sales.

Pros:

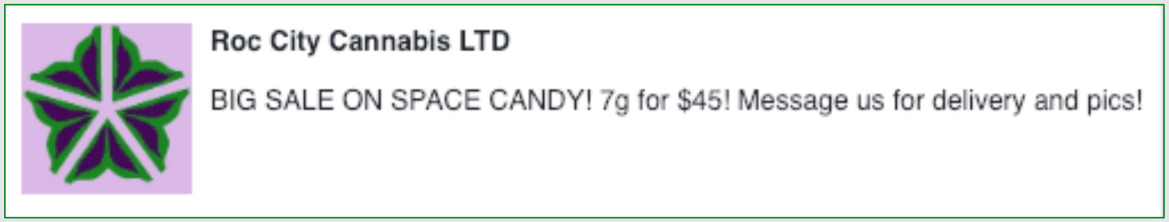
- Allows for businesses in localities that allow the sale of marijuana to operate on the platform

Cons:

- Facebook is not in a position to assess whether a seller complies with local requirements
- Product limitations regarding geographic restrictions of content to only suitable countries
- May be both over and under restrictive, depending on the country

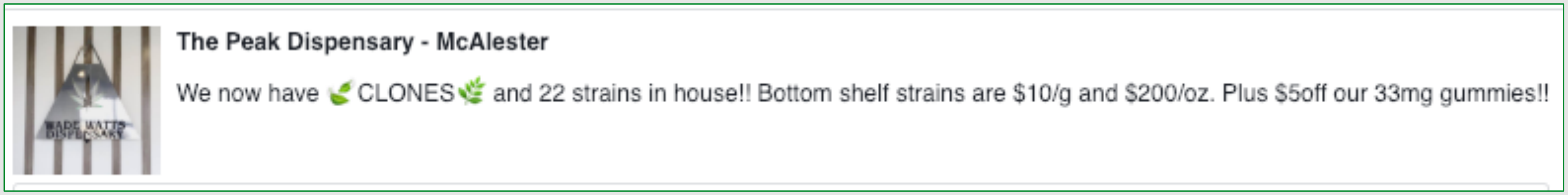
Marijuana and Cannabis products

Option 2: Examples (allowed under this proposal)



Roc City Cannabis LTD

BIG SALE ON SPACE CANDY! 7g for $45! Message us for delivery and pics!



The Peak Dispensary - McAlester

We now have 🌱CLONES🌿 and 22 strains in house!! Bottom shelf strains are $10/g and $200/oz. Plus $5off our 33mg gummies!!

Marijuana and Cannabis products

Option 3: Allow medical marijuana from brick & mortar stores

Allow the sale, trade, and advertisement of medical marijuana in brick and mortar stores and online retailers. Peer-to-peer sales and recreational sales from brick and mortar stores will be prohibited.

Pros:

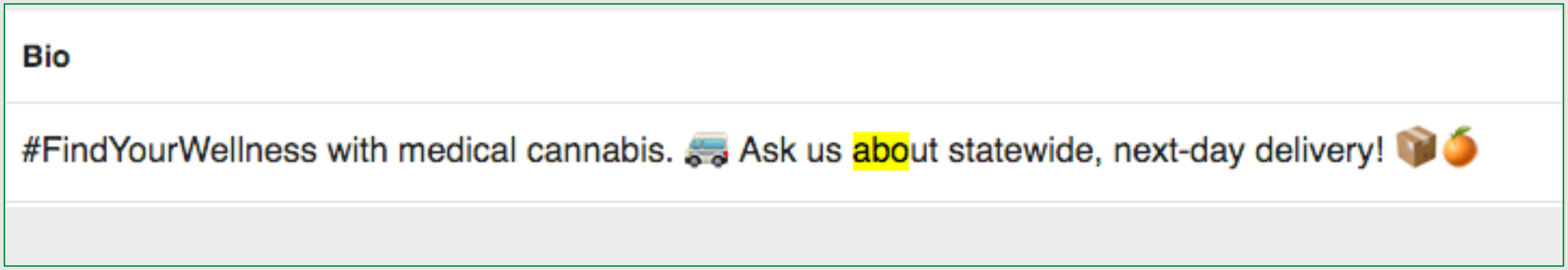
- Aligns with growing movement in some areas to legalize marijuana for medicinal purposes
- Aligns with goals of connecting people with businesses they find relevant
- May increase access to emerging potential medical applications

Cons:

- Product limitations regarding geographic restrictions of content to only suitable countries
- Will both be over and under restrictive, depending on country
- Facebook is not positioned to know the full range of a business's activities nor can we assess whether a seller complies with local requirements

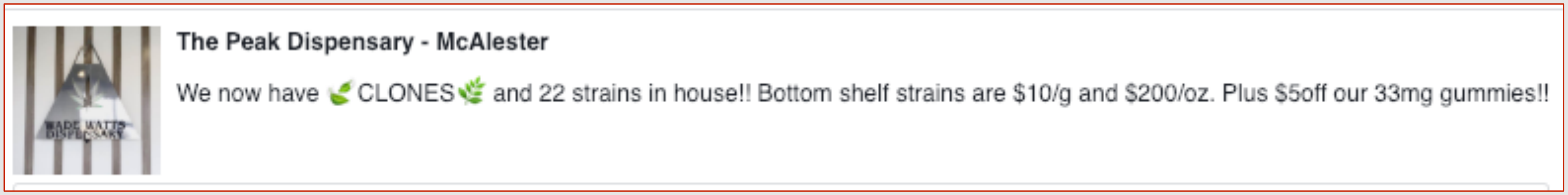
Marijuana and Cannabis products

Option 3: Examples



Bio

#FindYourWellness with medical cannabis. 🚐 Ask us about statewide, next-day delivery! 📦🍊



The Peak Dispensary - McAlester

We now have 🍃CLONES🌿 and 22 strains in house!! Bottom shelf strains are $10/g and $200/oz. Plus $5off our 33mg gummies!!

Marijuana and Cannabis products

Next Steps

- Work with Community Operations to further refine Operational Guidelines to maximize alignment with policy intent
- Coordinate with Community Operations to ensure that streamlined training is delivered to drive increased accuracy
- Coordinate with Communications team and update our appeals messaging to improve clarity and people’s understanding of our policy

HEADS-UP:
Manipulated Media

Manipulated Media

Overview

Issue: When independent fact-checkers identify misinformation, we reduce the distribution of the content and inform users of the fact checkers' warning. People are concerned that this approach may not be sufficient to counter the persuasive effect of synthetic media (e.g., deepfakes); however, removing all "manipulated media" would lead to difficult decisions around what is actually "manipulated" and may also lead to removal of political speech and satire.

Goal: Develop a manipulated media-specific policy that goes beyond the status quo covered by our misinformation policy.

Manipulated Media

Status Quo

- Through partnerships with third-party fact-checkers, we reduce the distribution of misinformation and inform people when something has been labeled as false.

- Manipulated media, including text, image, audio, or video, are eligible for fact-checking.

- Many of our partners have expertise in visual verification techniques, such as reverse image searching and analyzing image metadata. However, internal research indicates that fact-checkers find detection difficult.

- Investing more in technical detection, including through AI, will be crucial.

Manipulated Media

Examples

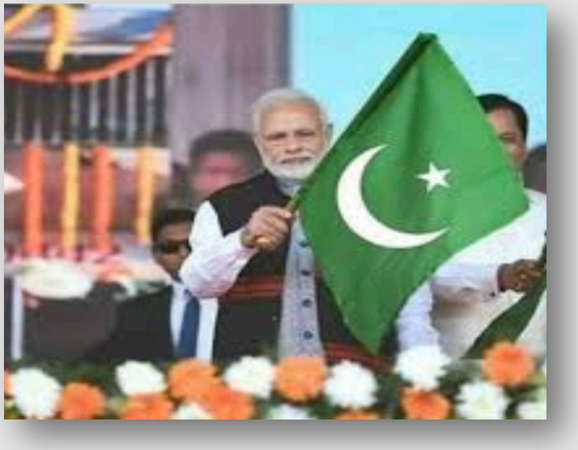
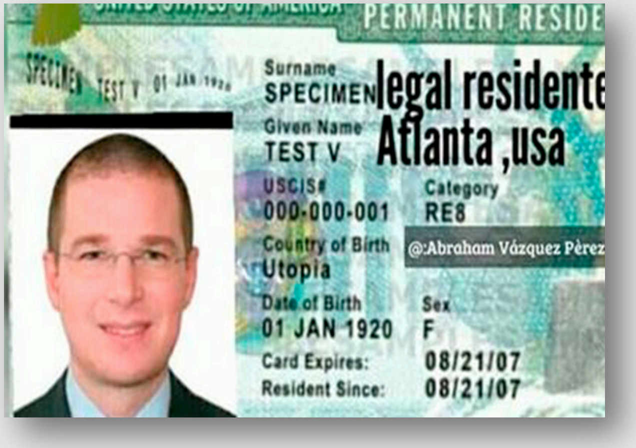


Photo of Indian PM Modi edited to suggest he was waving a Pakistani flag.



PERMANENT RESIDE

SPECIMEN TEST V 01 JAN 1920

Surname
SPECIMEN legal residente

Given Name
TEST V Atlanta ,usa

USCIS#
000-000-001

Category
RE8

Country of Birth
Utopia

@Abraham Vázquez Pérez

Date of Birth
01 JAN 1920

Sex
F

Card Expires: 08/21/07

Resident Since: 08/21/07

Photo of Mexican politician Ricardo Anaya photoshopped onto a template of a U.S. green card suggesting that he is a resident of Atlanta, Georgia.

Video of CNN reporter Jim Acosta's interaction with White House staff generated debate about whether footage had been intentionally edited to deceive, or merely changed when it was converted from a video file to a GIF.

Manipulated Media

Questions

- What kinds of manipulated media pose the greatest threats, and do those threats vary by country/region or time period (e.g., election cycles, national crisis moment)?

- Are there opportunities for cross-industry collaboration, and partnerships critical to addressing the threats posed by the most sophisticated actors?

- Can we introduce product solutions that would help (i.e., new "inform" treatments such as labeling, content removal)?

- How should we treat out-of-context/recontextualized media? (e.g., Onion article claiming North Korea's Kim Jong Un "sexiest man alive" understood to be satire by some readers but taken literally by Chinese media.)

Manipulated Media

Next Steps

- Continue working with AI research to understand advances in detection of manipulated media

- Continue working with Product to come up with the universe of product solutions that might be applicable in this sphere.

facebook