Content Standards Forum – November 27, 2018

Agenda

1. Recommendation: Designation Reversal for Hate Figures
2. Recommendation: Age-gate Alcohol & Tobacco Sales
3. FYI: Content Escalations to our Law Enforcement Response Team

Recommendation: Designation Reversal for Hate Figures

Issue: We have a robust process through which we designate hateful organizations and individuals as dangerous and remove them from the platform. At the same time, we know there are individuals who have renounced former hateful affiliations and ideologies; however, operating at scale and with limited context, it can be difficult to assess whether an individual has genuinely disavowed previously held beliefs and associations.

Recommendation for consideration: Establish a process for reversal of hate designations, but under a rubric that sets higher bar for reversal.

Status Quo: We don't have a process through which individuals designated as hate figures can be removed from our list of established hate figures.

- Option 1 - Status Quo
- Option 2 - Establish lower threshold for reversal
    - Hate Figure must be alive
    - Hate Figure must issue a public apology and / or express remorse for previous actions
    - Hate Figure must publicly renounce affiliated hate entities
    - Hate Figure may not meet any Level 1 or Level 2 designation signals after the repudiation/apology
    - 1 year waiting period after above criteria is met before reinstatement
    - **Check in:** We will review the Hate Figure six months after being reinstated
    - **Platform Restrictions:** Hate Figure may not access monetization tools (e.g. Ads, fundraising tools) for 1 year
    - Pros -
        - The policy option does not require a hate figure to advocate against hate/hate entities, a factor which could put the hate figure's safety at risk after disavowal
        - Addresses the primary concern: disengagement with hate entities
        - Check-ins and platform restrictions ensure compliance and promote good behavior
    - Cons -

        - Establishing a lower threshold is inconsistent with the rigorous designation process - and accompanying enforcement we have in place for hate entities (ex: Removal of all praise, support, and representation).
        - Potential for abuse / gaming
        - May be viewed as too simplistic
- **Option 3 - Recommendation for consideration - Establish higher threshold for reversal *(differences between Option 2 and 3 are highlighted in bold, below)***
    - Hate Figure must be alive
    - Hate Figure must issue a public apology and / or express remorse for previous actions
    - Hate Figure must publicly renounce affiliated hate entities
    - Established by
        - (a) statement from Hate Figure OR
        - **(b) negative sentiments about the Hate Figure from former affiliates**
    - **Hate figure must advocate against hate/hate entities**
    - **3 year waiting period** between apology and consideration for reinstatement
    - Hate figure may not meet any Level 1, Level 2, **or certain Level 3 designation signals** after the repudiation/apology
    - **Check-in:** We will review the Hate Figure six months after being reinstated
    - **Platform Restrictions:** Hate Figure **may not access** monetization tools
    - Pros -
        - Establishing a higher threshold is consistent with our designation process and accompanying enforcement for designated hate entities (ex: Removal of all praise, support, and representation)
        - Less potential for abuse
        - Check-ins and platform restrictions ensure compliance and promote good behavior
    - Cons -
        - The policy option requires a hate figure to advocate against hate/hate entities, which could create safety risks for the hate figure
        - The criteria could be seen as overly restrictive and too high of a burden
        - The high threshold goes beyond disengagement with hate entities; instead, it is an attempt to change behavior
- External engagement -
    - All stakeholders agreed that we should allow people who disavow hateful beliefs/associations back on the platform at some point
    - With regard to the low threshold, people raised concerns about having to speak out against the hate org they were previously affiliated with
    - Many wanted to set the bar as high as possible. They didn't want any mistakes made in this area
    - Most stakeholders agreed that a waiting period between disavowal and reinstatement was necessary
- Working Groups -
    - We agreed that we did not want a revolving door and that this process was as clear
- Next Steps -

    - Work with Community Operations to develop Operational Guidelines
    - Work with Community Operations to clarify policy to allow praise, support, and representation of a Hate Figure's disavowal during the waiting period
    - Announce Date: November 28, 2018
    - Go Live Date: December 11, 2018
- Discussion -
    - Question: How do we define ethno-state? Have we thought about some way of allowing someone to come to us on how they may be in real world danger if they publicly disavow their former hate org but wants to be back on the platform? What about Isreal/ Palestine content around advocating for an ethno-state?
        - Answer: Advocating for an ethno-state is one of the signals we use to determine whether someone may qualify as a hate figure. This is calling for any state that is limited to just one protected characteristic (e.g. white-only, black-only). It is one of many signals we use for designation so merely advocating for an ethno-state does not mean you will be designated, but to answer your question, this would include calls for a Jewish-only states.
        - Answer: Many stakeholders did flag the safety concern as well so we did keep this in mind.
        - Answer: Also important to note that just like hate org/figure designations, reversal of a designation would have to be approved by a cross-functional internal group, including our public policy team.
    - Question: Could you touch on what you mean by a "check-in"?
        - Answer: Six months after the initial decision to reverse designation, we will review our decision again to make sure the individual in question has continued to adhere to the criteria laid out in our recommendation.
    - Question: If the decision is reversed, what if someone posted praise, support and representation of actions the figure did before disavowal? How would we deal with that at scale?
        - Answer: We are working with Community Operations to operationalize, but praise, support and representation of the figure prior to disavowal would still count as a violation under the updated policy and guidelines.
    - Discussion: Something mentioned earlier that I would like to go back to. The fact that a person was publicly known for hateful acts and met the high standard we maintain for designation makes <u>public</u> disavowal extremely important. Also very unlikely that a person who falls into this camp would be able to privately disavow since they tend to be high profile figures.

<u>Recommendation: Age-gate Alcohol & Tobacco Sales</u>

<u>Issue:</u> We do not want alcohol and tobacco products to be sold or promoted to minors on our platforms. We also recognize that many countries restrict the online promotion or sale of tobacco and alcohol products. However, if we ban all promotion and sale of these products, licensed retailers and brands who have voluntarily age-gated their content would be adversely and

significantly affected. This is also not currently addressed clearly in our policies.

Recommendation for consideration: Age-gate to 18+ all sale and promotion of alcohol, tobacco and e-cigarettes in organic posts.

Status Quo: Currently, our policies stipulate that the sale and promotion of alcohol, tobacco and e-cigarettes cannot be targeted at minors. And our commerce policies do not allow the sale of this content at all.

- Option 1 - **(Recommendation for consideration): Age-gate to 18+ all sale and promotion of alcohol, tobacco and e-cigarettes in organic posts** *(FB Profiles, IG Profiles, Groups)*
    - Pros -
        - Allows legitimate brands to advertise and sell their products
        - Restricts exposure to the demographic that is legally permitted to purchase those products
    - Cons -
        - Leaves scope for minors to access the content if they haven't provided Facebook with their real age
        - Falls short of legal requirements in certain jurisdictions that ban tobacco sales
        - Legal age for alcohol consumption varies between jurisdictions, so a single age-gate won't align with legal requirements in every jurisdiction
- Option 2 - **Ban all sale or promotion of alcohol, tobacco and e-cigarettes** *(FB Profiles, IG Profiles, Groups)*
    - Pros -
        - Comprehensively addresses the issue of alcohol and tobacco sales to minors
        - Aligns with legal regimes in many countries that ban all online tobacco promotion and sales
    - Cons -
        - Legitimate businesses cannot use Facebook to build their brand or advertise
        - Over-enforces on alcohol promotion and sale relative to local laws in most jurisdictions
- Option 3 - **Age-gate commercial posting of alcohol; Ban commercial posting of tobacco and e-cigarettes** *(FB Profiles, IG Profiles, Groups)*
    - Pros -
        - Aligns with legal regimes in many countries that ban online tobacco and e-cigarette sales and promotion but allow alcohol sales and advertising to adults
    - Cons -
        - Distinguishes between alcohol and tobacco on grounds that may not be explicable
        - Prevents legitimate tobacco and e-cigarette businesses from using the platform to advertise and build their brand

- External Engagement -
    - Safety Team and External Engagement worked together on this
    - Experts were divided on the path forward, with the biggest issue being exposure to minors.
    - The experts suggested that we should take whatever action would delay exposure to the products, brands and overall lifestyle as long as possible because that was the key issue.
    - Experts that supported the complete ban noted that our age-gating capabilities were not strong/reliable enough and until we strengthened those capabilities, they couldn't trust our age-gating to properly protect minors from this content.
    - On the other hand, groups, particularly in the UK, felt like our recommendation to age-gate was in line with the regulations in place in their country, noting that:
        - These products were in fact legal so a complete ban might be too far reaching, and
        - Partnering with using a third party verifier could help to enhance our age-gating capabilities and outcomes.
    - Overall, all of the groups noted that any action forward, enhancing our policies in this area, would be a win for the protection of minors on the platform.
- Next Steps -
    - Partner with our engineering teams on technical dependencies regarding the Age Gating capabilities
    - Announcement and launch: January 2019
- Discussion -
    - Question: Are ads being age-gated? Or the Hennessey Page for example?
        - Answer: If it is an ad it is subject to our Commerce Policies. But right now the Hennessey Page with organic posts would not be.
        - Answer: We do have a policy about sale to minors but this is about enforcement
        - Answer: A lot of this is already banned in ads anyway so this is about organic content.
    - Question: So there is no policy on this at all for organic post?
        - Answer: There isn't one. Any organic post selling alcohol stays up unless we have a legal takedown request so we are trying to cover that gap
        - Discussion: Here, someone on the legal team noted that we do get these requests and this does happen.
        - Follow-up: That was what I was wondering especially because many countries in MENA ban the sale of alcohol completely.
    - Question: If a celebrity is promoting an alcohol brand or smoking an e-cig as part of a promotional deal, would this fall within policy?
        - Answer: Yes, we would consider this promotion and this type of content would be age-gated - so anyone under 18 would not see it. However, currently, people cannot age-gate their own content - we would age-gate the post specifically.
    - Question: Is it worth postponing this recommendation and enforcement until we get confirmation that the tooling would be available?

        - Answer: We don't recommend postponing the update because we believe there's risk in allowing this type of content.
        - Follow-up: Let me know if I can connect you to someone on the tooling team as well and we can reconnect after the meeting.

FYI: Content Escalations to Law Enforcement Response Team

- Issue: We have initiated a project to map content areas that result in an escalation to our Law Enforcement Response Team, and analyze the policies and operations related to them (ex: grooming, suicide, credible threats, etc.) in order to determine if we are executing effectively and to assess whether we need more resources in certain areas. This mapping will be done with Community Operations.
- Next Steps -
    - Partner with Law Enforcement Response Team, Law Enforcement Outreach, Safety Policy and Operations, and, among others, the team that works on Dangerous Orgs and Individuals
    - Report back with any recommendations