Content Standards Forum - November 13, 2018

As Mark mentioned in his note on content and governance issues, our Content Policy team runs a meeting to discuss potential changes to our policies, what we refer to as our Community Standards, every two weeks. Teams represented at the meeting generally include: Legal, Safety Policy, Counterterrorism specialists, Cybersecurity Policy, Community Operations, Public Policy (including regional public policy from around the world), Communications, product teams, Diversity, U.S. State and Federal Policy, and Government and Politics from around the world.

There are two types of presentations given at the meeting by the Content Policy team's subject matter experts: what's known as a 'heads up' and a policy "recommendation".

A 'heads up' is a short presentation that the Content Policy team uses to introduce an issue they plan to work through with internal and external input. In some cases, a heads up is prompted by input the team receives from external stakeholders; in others, changes are necessitated to account for the way language is used or because the Community Operations team has identified a gap in existing policy.

Upon presenting a heads up at the Content Standards Forum, the Content Policy subject matter experts convene internal and external working groups to inform the policy development process. They also analyze data and examples to better understand the ways that people speak on the platform, and to help inform how abuse may manifest itself.

Based on the input received and the research completed, the subject matter experts will consider several options for updating or changing a policy. It's these options, among which is the team's policy recommendation, that ultimately make their way back to the Content Standards Forum for discussion.

Today, for the first time ever, we're sharing the minutes from the Content Standards Forum held on November 13. The meeting agenda included one policy recommendation and three heads ups.

Today's Agenda

1) Recommendation: New Coordinated Inauthentic Engagement and Spam Update
2) Heads Up: Criminal Entities Designation Signals
3) Heads-Up: Non-Exploitative Child Nudity
4) News Feed FYI: Off-Platform Content in News Feed Ranking

Notes

1) Recommendation: New Coordinated Inauthentic Engagement and Spam Update

Overview

- Issue: Authenticity is the cornerstone of our community. As such, we prohibit the spread of commercial spam under our Spam Policy. We also prohibit coordinated inauthenticity. Recently, new forms of abusive behavior and deceptive engagement have highlighted the need for us to re-examine the scope of our policies. However, behavioral enforcement is difficult to scale and prone to false-positives. Therefore, we need to be careful that enforcement is not over broad or arbitrary because we do not want to remove genuine political speech.
- Summary to Date
    - Held several internal meetings and one cross-functional working group
    - Consulted with six external experts
- Recommendation for Discussion:
    - Broaden the scope of our Spam and Coordinated Inauthentic Behavior (CIB) policies to capture false and inauthentic engagement across content types.

Status Quo Policy

At present, spam and coordinated inauthentic behavior are covered by a number of policies, including:

- Spam Policy
    - Attempting to gain distribution for commercial speech for financial gains.
    - Attempting to gather sensitive information.

- Misrepresentation Policy
    - Users are not allowed to maintain multiple profiles.
    - Users are not allowed to create inauthentic profiles.
    - We remove accounts that participate in, or claim to engage in, CIB.
- Review Guidelines for Things Like Pages or Groups
    - We consider multiple elements such as the title, description, and posts within a Page or Group to determine whether or not it should be unpublished
    - We review administrators for authenticity

Examples of Violating Content per Status Quo Policy:

- Alternative for Sweden (Sweden) — we took down this Page and an associated Group after an investigation revealed that the organization was covertly coordinating re-shares of content and posts to mislead people into believing that the content was more prevalent than was actually the case.
- FollowNow (Brazil) — similarly, we took down this Page and an associated Group upon learning that the admins were inauthentic and were using the Page to post political content, misleading followers into believing that certain elections issues were receiving more support than others.

Options for Consideration

Option 1 (Recommendation): Broaden the scope of existing policies by removing commercial limitation on Spam and expanding CIB to CIE to capture manipulative and deceptive engagement practices.

Under this approach, we would remove the commercial limitation on our Spam policy and expand our Coordinated Inauthentic Behavior (CIB) policy by creating a complementary Coordinated Inauthentic Engagement (CIE) policy to capture deceptive and manipulative engagement. This will require that we develop specific guidelines for tiered enforcement under CIE, and that we review and adjust the action on Pages and Groups to align with the severity of malicious signals that are new or evolving.

- Pros
  - Captures newest forms of suspicious and malicious behavior
  - Easier for users to understand
  - Allows for more scaled enforcement
  - Aligns policy and protocol so that enforcement is more consistent
- Cons
  - Increased risk of over-enforcement
  - Risk of Spam becoming a catch-all policy

Option 2: Maintain the status quo and keep the scope of the CIB and Spam policies narrow

- Pros
  - Less risk of over-enforcement as behavioral enforcement is more complex to implement at scale

- Cons
  - Leaves considerable policy gap that information operations actors can take advantage of
  - Enables malicious financially-motivated spam behavior
  - Not easy for users to understand

Internal Cross-Functional Working Group

- We engaged with teams across Public Policy, Community Operations, Communications, Legal, Community Integrity, and Risk and Response.

External Outreach

- We engaged with experts, among them academics and think tanks, who have studied the spread of online spam and the behaviors of people who run online scams.
- The feedback we received is outlined here:
  - Some said to only focus on commercial spam since it is far more difficult to keep up with, identify, and define human inauthentic behavior.

    - Some experts noted that past hesitation to include political spam in a regulatory definition of spam has allowed non-commercial spam to grow. As such, Facebook should work to capture both commercial and non-commercial spam.
    - Others advocated for an industry wide definition for spam and suggested that we help align the industry on terminology.
    - There was general consensus that the average Facebook user doesn't want to see or engage with deceptive content on the platform and we should work to tackle this issue.
- In expanding the Spam policy beyond commercial motivations while also creating a CIE policy to capture deceptive and manipulative content, we are balancing external perspectives that called for ensuring this abusive behavior is in scope of our policies, while looking to ensure Spam does not become a catch-all policy.

Timeline/Next Steps

1. Convene smaller working groups to:
    1. Develop guidelines for CIE enforcement based on the severity of a violation
    2. Develop precise policy language and add CIE as a policy section within the Community Standards
    3. Refine existing spam policy language and review scaled operational guidelines for Community Operations
2. Update policy language and launch internally and externally

Questions/Discussion

- *Question:* Can either of you touch upon the recent India escalation?
- *Answer:* One of the ways that we catch spam is through machine learning classifiers that have been trained to recognize and remove spam from our platforms. In the India escalation, a civic spam classifier disabled Indian journalists on the basis of a behavior — several of them were re-sharing their own content many times over from domains that had been identified as hosting, what appeared to be, low quality content or affiliated with ad farms. By applying our standard spam classifiers to India, we failed to capture what the Internet looks like in India and how people use our platforms. Coordinated Inauthentic Engagement (CIE) needs manual review and its efficacy is contingent on understanding local Internet culture. Within CIB, we simply enforce by removing content or not. In CIE, we want to come up with a tiered enforcement to help people reform their behavior before we take them off the platform.
- *Question:* Has there been any thought about being proactive in our communications about our work here? People aren't trying to be deceptive but trying to take advantage of site functionality.
- *Answer:* Like all of our policies, this update will be made public in our Community Standards, and we'll work with comms to establish language that gives people more guidance on what they can or can't post and helps them better understand CIB and CIE. We also want to build more in product education so people know why they may have violated our policies and what they can do to adjust their behavior.
- *Comment:* These behaviors change quickly and keep evolving. We need to keep an eye on how abuse changes and incorporate that into enforcement.

- *Question:* I'm focused on the secret coordination in a secret group that was seen in Sweden. What is the difference if, for example, a political party does a secret conference call offline with social media managers to coordinate how to spread content online?
- *Answer:* The behavior wasn't necessarily a secret in Sweden. The answer here may not be outright removal of content. We might consider a feature limit or in product warning about too much sharing. What we want to work out in the future is what are the set of behaviors we want to fix or inform. The goal is to make this policy public so we will work with comms and others to determine how to present this within the Community Standards.

Thanks everyone for the discussion. Unless there are objections to the policy recommendation presented, we will consider this passed. For people interested in follow up work, please get in touch with the subject matter experts who presented today.

2) Heads Up: Criminal Entities Designation Signals

Overview

- Issue: We do not allow praise, support or representation of Criminal Organizations, their leaders or prominent members on our platform. We currently use a set of signals to determine if an organization should be designated as a Criminal Organization. Our signals need to objectively account for current trends to ensure we are keeping bad actors off the platform. At the same time, we need to maintain the ability to evaluate organizations quickly and consistently. Given these challenges, we want to review and update our standards for designation.
- Goals
    - Examine our designation signals to ensure they rely on concrete indicators that are clear, consistent, and explicable.

Status Quo

- How do we designate an organization?
    - Potentially violating organizations are flagged to Content Policy, by internal & external partners.
    - Content Policy reviews the organization against a list of signals to determine whether or not the organization should be designated.
    - If Content Policy recommends designating the organization, the team will reach out to an internal cross-functional group, including country managers, who weigh in on the implications of the designation.
- Designation signals
    - Definition: Criminal Organizations often identify by a name, symbol, colors, hand gestures or related indicia, and they perform or threaten to perform criminal activities.
    - Indicators include, but aren't limited to:
        - Organization showcases or promotes criminal exploits on the platform;
        - Leaders and prominent members convicted of homicide or other violence, drug trafficking, arms trafficking, human trafficking, smuggling,

prostitution, racketeering, illegal gambling, CEI production and distribution, bid rigging, extortion, kidnappings, money laundering, financial fraud;
        - Leaders and prominent members intimidate and threaten anyone with violence or other criminal activities, including other Criminal Organizations.
- We remove all praise, support, and representation once designated.

Key Questions

- How should each signal be weighted?
- Are there additional signals that should be considered?
- Are there existing signals that should not be included as part of the designation process?
- How can we move to a more formulaic structure to ensure consistent outcomes?

Next Steps

- Convene cross-functional working group to consider options
- Consult with a diverse group of academic and law enforcement experts to better inform our development of the respective indicators.

3) Heads-Up: Non-Exploitative Child Nudity

Overview

- Issue: We know that people share imagery that depicts nude or semi-nude children with good intentions. However, we often remove this imagery because of the potential for misuse or abuse, which may feel like censorship to many people. We would like to explore the possibility of allowing more non-sexualized imagery of children between 4 and 18 years old without compromising the safety of children.
- Goals
    - Convene a working group to discuss a more nuanced age classification system that allows some non-sexualized child nudity.

Status Quo

- Child nudity restrictions are applied to content that depicts real people <18:
    - Babies (0-1.5 y.o.) - almost no nudity restrictions
    - Toddlers (1.5-4 y.o.) - very few nudity restrictions (no genitalia)
    - Minors (5-<18 y.o.) - more restricted, stricter than adult nudity policy (no genitalia, breasts, or butts)
- Imagery with possible indicators of child sexual exploitation where age of child is questionable is always escalated and subject to a strict zero tolerance policy.

Examples

- In the past, we have made newsworthy exceptions to leave up historical images and to document mass atrocities such as the Holocaust or famine in Yemen or the Terror of War image.
- We also also leave up images of toddlers and babies (0-4 y.o.) in the nude because over the years, we found that these images are less likely to be exploited.

Next Steps

- Convene a working group to discuss policy options and the possibility of creating a stand-alone policy section for non-sexualized child nudity.
- Partner closely with Safety policy to ensure we are aligned on any changes to the policy.
- Engage with external experts for evaluation and input.

Questions/Discussion

- *Comment:* You should make sure to touch base with Community Integrity counterparts on child nudity because they are already thinking about how to train our technology to more effectively find inappropriate content. Community Operations currently errs on the side of caution in escalating child nudity to the Law Enforcement team out of fear of possible child exploitation. This will free up much needed time for our CO colleagues if we can get this right.
- *Question:* In terms of censorship, can you quantify the number of complaints? Are we being proactive on this or is it a reaction to complaints? We seem to be in a good place right now with our newsworthy guidelines. Do we need to change the whole policy?
- *Answer:* People do regularly report this content. In cases of genuine innocent child nudity, it seems unfair to those people sharing the content. We know it's not violating but we don't want to open it up at scale so we have newsworthy guidelines. But your questions are a good call out and we can work on getting the numbers for complaints.
- *Comment:* The Anne Frank Center has complained to us in the past about this. In addition, it's still difficult for Community Operations to enforce all exceptions. It's better to have at scale guidance. A few years ago, there was a segment on Today Show about how we were calling mothers and their children pornographic. So this is definitely an issue.
- *Comment:* Our default is child safety but perhaps we can balance the details better.
- *Question:* On the recent child safety announcement we did within the Community Standards Enforcement Report, we said explicitly that we take down innocent photos too. How will this align with the report?
- *Answer:* It might be that we end up with status quo even after we go through this full process but we have heard from different sources that perhaps we should open this up and want to at least go through an investigation.

4) News Feed FYI: Off-Platform Content in News Feed Ranking

Overview

- Issue: Research finds that people are more satisfied with their experience on Facebook than with their experience on landing pages they navigate to from Facebook. Accordingly, News Feed ranking incorporates assessments of off-platform content, such as evaluating landing pages against our clickbait and ad farm guidelines, in order to demote posts that include links to these pages. News Feed is considering expanding this assessment based on two key considerations:
    - Evaluating landing pages for violations of our Community Standards
    - Evaluating landing pages based on the content on the *entire domain* associated with that page (rather than just the landing page content)

- Goals
    - Work with the product teams and cross-functional stakeholders to identify the challenges this expansion may raise and to mitigate risks appropriately.

Key Components of the Proposal

- Demote links to landing pages that violate our Community Standards:
    - Automatically evaluate landing pages against our Community Standards;
    - Apply gradually as technology is developed;
    - Thresholds and weights TBD.
- Domain-Level Evaluations
    - For links posted to Facebook, automatically evaluate the entire domain associated with that URL (rather than just the landing page itself);
    - Hold domains dedicated to certain types of content (*e.g.*, Your Money or Your Life) to a higher standard than other domains.

Risks and Questions for Consideration

- Perception that we are imposing our Community Standards on the internet at large:
    - Do people want this and would they benefit from it?
    - How do we balance the benefit for users against the possible blowback from news partners, regulators, and others?
- Domain-level assessments also raise questions of fairness and efficacy:
    - Do we want to downrank a link to an informative and authoritative article simply because the article appears on a site that may have low-quality or violating content?
    - Would such assessments unfairly impact certain publishers?
    - Should we, and if so how do we, account for different regional norms?

Next Steps

- Research to better understand user preferences.
- Collect additional feedback from Public Policy and Product Policy teams.

- Review specific ranking changes through the News Feed Core XFN ("Eat Your Veggies") review process.
- Please send questions directly to the News Feed leads so that we can consolidate feedback for the full News Feed team.

Questions/Discussion

- *Question:* How much will we communicate about this? We usually don't want people to game the Feed but this change might hit their pocketbook so my instinct is to be open about this.
- *Answer:* Comms is aware and we will be proactive about communicating this. Inside Feed is one channel we can use to do this. The only way this can happen at the domain level is with full transparency and product is aware we can't do this unless we have that.
- *Question:* Will this evaluation apply to user generated content on the third party domain?
- *Answer:* This wouldn't apply to all content on the third party domain.
- *Question:* What about comments?
- *Answer:* We see supplementary content differently from core content so this wouldn't apply or would at least be weighted differently.
- *Question:* Would this apply to misinformation on an off platform domain?
- *Answer:* Yes, but we don't know how much we should look into that. It'll be a big lift for our third party fact checking partners to look at an entire domain.
- *Question:* Have we given any thought to how publishers might destroy each other's reach if they start to comment against each other on each other's domains?
- *Answer:* This is one of the key issues among others. Folks on our teams have considered some this proposal and are thinking about the appropriate technology to figure out what is an authentic vs. inauthentic comment.