Message

From: Miki Rothschild [/O=THEFACEBOOK/OU=EXTERNAL (FYDIBOHF25SPDLT)/CN=RECIPIENTS/CN[redacted]
Sent: 12/19/2020 5:10:39 AM
To: Arturo Bejar [redacted]@fb.com]
Subject: Re: Update since we spoke

These are really good point. I think I need to add one more step at the top of the funnel for production, where prevention could include things like nudges/information and norm setting. I love your example around self remediation by improving reporting flow, that's super insightful.

I'll give this funnel another rev after New Years and share with you for thoughts.

Happy holidays!
Miki

Sent from my iPhone

On Dec 18, 2020, at 10:13 AM, Arturo Bejar <[redacted]@fb.com> wrote:

It does, I think it communicates how non-policy violating experiences/policy violating experiences are present across the board, and as we get data on those proportions that will help communicate.

Question - for potential bad experiences - do you mean content that is discernible by us (i.e. we could build a classifier/criteria to detect)? Or does it capture everything?

I think one of the areas with a lot of potential is work that addresses the supply side of bad experiences. This is beyond the current fast detection & removal work around prevalence. I think there are two kinds of content creators, those who would change behavior given respectful feedback, and those who would not, and right now those are mixed together. Finding ways to address the supply side of potential bad experiences seems high leverage given volume, and will give cleaner signals on the content creators that are intentionally generating them.

One area that will probably shift is the definitions of the labels on the left, in particular self-remediation and support. Once we understand better how people behave given a bad experienced we might end up with a definition of support that is centered around the person's bad experience, and Shilpa/S[redacted] have begun using the word resolution, which I includes possible outcomes which include self-remediation, and submitting feedback or a report.

One of the interesting findings from the work we did was around hate speech in India, it had an action rate of 3-4%, and when reviewing the reports we found that there were a lot of things having to do with religion and cricket. After research we found that people in India found those hateful and they did not share our definition of hate. We added options for that market that included 'this insults my religion', 'this insults something important to me', this led to a significant workload reduction & action rate increase in hate queue. We were surprised to find though that people would choose the options and would not complete the flow. We surveyed to figure out what was going on and it turned out people were satisfied and did not want to take any

of the self-resolution actions, what they wanted was to be heard on the issue that they were having.

Arturo

On Dec 18, 2020, at 9:47 AM, Miki Rothschild <[redacted]@fb.com> wrote:

Thank you for sharing and for all your work helping us articulate this gap much better, Arturo! I think this is starting to sink across the organization, which is really exciting.

BTW, I had a chat about it with Guy in our 1:1 where I used the attached funnel to try to illustrate this point. Curious if this resonates?

From: Arturo Bejar <[redacted]@fb.com>
Date: Friday, December 18, 2020 at 8:55 AM
To: Miki Rothschild <[redacted]@fb.com>
Subject: Fwd: Update since we spoke

Hi Miki,

As the year winds down I sent Chris Cox the update below. I'm also planning to touch base with Schrep and A[redacted] I'm also talking to Chris Struhar as I think similar thinking is happening on the new group the formed around this on FBApp. I know this is a long road and I am grateful to get to work with your team, in particular working with Sam and Sayed.

I hope you have a good holiday season, if it is ok when Yoav is on leave I will reach out more.

Arturo

Begin forwarded message:

From: Chris Cox <[redacted]@fb.com>
Subject: Re: Update since we spoke
Date: December 14, 2020 at 8:35:51 AM PST
To: Arturo Bejar <[redacted]@fb.com>

I really like this thought, especially the framing you and D[redacted] propose at the bottom.
I know that Guy's team as well as John H.'s team has been looking hard at the question of how these interactions can be more positive experiences vs. where they are right now. My sense is the most actionable thing would be to pass this feedback along to them. Are you okay if I do that or did you want to send directly?

Thanks for writing this up and especially for the thought behind it. I hope all is well.

Chris

From: Arturo Bejar <[redacted]@fb.com>
Date: Saturday, December 12, 2020 at 4:33 PM
To: Chris Cox <[redacted]@fb.com>
Subject: Update since we spoke

Hi Chris,

I wanted to give you an update since we last spoke. After we spoke I felt I did not do a good job of articulating what I've come to understand, and after conversations with people across the company, and working closely with the IG well-being team, I have a better framing.

Today we don't understand well the relationship between bad experiences or harm from our user's perspective (what TRIPS captures) and prevalence (policy violating content). In IG 20% of DAU witness what they consider to be hate over the last 7 days, 5% report being targets. During that period 70,000 reports are submitted the Hate category in FRX, and only a small fraction of those is actioned. I think this explains one of the sources of negative perception of our products and by our employees.

In the conversations I've had I've found that for many people prevalence equals harm (my guess from what I've seen is that Mark thinks about it his way), and when faced with non-policy violating harmful content our choice is whether to change our policy or not. But what if most content that people experienced as harmful was far from policy violating? What would we build then?

Most people when they enter the reporting flow, do so because they want to help make the community safer/better, but right now that only happens when content violates policy. Prevalence focuses on our rapid detection and removal of policy violating content. Today we have no feedback mechanisms that would help well intentioned creators be better members of our communities. The transparency experiments I've seen have the feedback come from FB, which people don't appreciate, I think that makes sense, it is like when the police knocks on your door. I think there are design/product ways around this.

If we use TRIPS, user experienced harm, as a key metric and most content is far from policy violating, then we need to start thinking of different solutions. I've spoken to different people who have tried to move TRIPS (D[redacted], V[redacted], C[redacted] U[redacted], M[redacted], Jennifer Guadagno, etc), and what I'm proposing they think could make a difference. There are two problems: our user's definitions of [hate|misinformation|etc.] are different from ours, and today we don't understand how people behave given a bad experience (do they just scroll away? hide?). I'm working with the team in IG on two efforts around this, a funnel analysis from TRIPS to prevalence, and a research project on behavior given bad experiences.

The other thing I've noticed in my discussion is that when people get the broader definition of user experienced harm, they get stuck on current set of solutions (block/mute/unfollow/downrank/report). But if you start from the point where most content is not policy violating then the reporting flow becomes about receiving support for your experience independent of the right decision on the content, and we could do things like community classifiers/ranking (I don't want to see provocative content and help others who don't want to see that kind of content).

D[redacted] V[redacted], gave me a great framing where our solutions for this space should meet two criteria:

- Does it make the person feel safer? (i.e. do we improve their individual experience)
- Does it make the community better?

The second one is something I think we've been missing. If you see a friend share something that makes you upset, and you unfollow/unfriend them, that would improve your experience, but it does not make the community better.

There are many examples where our current handling does not work. A couple of weeks ago I heard about a transgender body positive creator receiving rape threats. The threats are not credible so their tool is to block, it might make them safer, but it does not make the community safer, and the person making the threats continues to do so to other people because the worst that can happen is a block from someone they don't know. One question I've been using to focus conversations, if we wanted to create a community free from harassment, what would we build?

If we get really good at finding, understanding, and addressing the unintended harms that people experience in our products I think it would be motivating for employees as well as true to our mission.

If this is too high level I can give more specific examples, the Instagram team has been really great to work with on these things, and I've found people in the FB App team who are thinking about this problem space in this way too.

Let me know if you'd like to talk about this, or I can also keep you posted.

Arturo

<Bad Experiences Funnel.001.jpeg>