From: Arturo Bejar[/O=THEFACEBOOK/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=[redacted]]

Sent: Wed 10/6/2021 4:46:10 AM (UTC)

To: Yoav Shapira[redacted]@fb.com]; Miki Rothschild[redacted]@fb.com]; Samir [redacted]@fb.com]

Subject: Fwd: Gap in our understanding of harm and bad experiences

Attachment: BEEF by Age.webp

Hi Samir, Yoav, Miki,
After seeing Zuck's post tonight, I decided it was time to send him the note below. Not sure what will come from it but figured it to be the time. I'll keep you posted if there are any developments.

Arturo

Begin forwarded message:

From: Arturo Bejar <[redacted]@fb.com>
Subject: Gap in our understanding of harm and bad experiences
Date: October 5, 2021 at 9:37:59 PM PDT
To: Mark Zuckerberg <[redacted]@fb.com>
Cc: Sheryl Sandberg <[redacted]@fb.com>, Chris Cox <[redacted]@fb.com>, Adam Mosseri <[redacted]@fb.com>, Mark Zuckerberg <[redacted]@fb.com>

Dear Mark,
I saw the note you shared today after the testimony, and I wanted to bring to your attention what I believe is a critical gap in how we as a company approach harm, and how the people we serve experience it. I've raised this to Chris, Sheryl, and Adam in the last couple of weeks.

I want to start by saying that my personal experience, and what I believe, is that you and m-team care deeply about everyone we serve, and my goal in sending this is to be of service to that. It's been 2 years since I've been back part-time.

51% of Instagram users say 'yes' to having had a bad or harmful experience in the last 7 days. Out of those 1% of report and of those 2% have the content taken down (i.e. 0.02%). The numbers are probably similar on Facebook.

Two weeks ago my daughter[redacted] and an experimenting creator on Instagram, made a post about cars, and someone commented 'Get back to the kitchen.' It was deeply upsetting to her. At the same time the comment is far from being policy violating, and our tools of blocking or deleting mean that this person will go to other profiles and continue to spread misogyny. I don't think policy/reporting or having more content review are the solutions.

There is detailed data about what people experience in TRIPS, a statistically significant survey. We ran a more detailed survey, I've attached the full age breakdown below, but here are some key numbers (these questions are in the last 7 days):

21.8% of 13-15 year olds said they were the target of bullying.
39.4% of 13-15 year olds said they experienced negative comparison.
24.4% of 13-15 year old responded said they received unwanted advances.

Why does someone think it is ok to post 'get back to the kitchen' or harass someone? I believe it is because it doesn't violate policy, and other than deleting or blocking, there is no feature that helps people know that kind of behavior is not ok. Another example, is unsolicited penis pictures [redacted] has received those from boys too since the age of 14, and her tool is to block them. I asked her why boys keep doing that? She said if the only thing that happens is they get blocked, why wouldn't they?

Why the gap between Prevalence and TRIPS? Today we don't don't know what % of content people experience as misinformation, harassment, or racism is policy violating. We have done great work in driving down prevalence, and there will always be more to do, but what if policy based solutions only cover a single digit percentage of what is harming people?

Policy is necessary when the content is unambiguously inappropriate, yet it has many limitations. It trails behavior, the interventions are heavy and risk over-enforcement and getting the border line right is extraordinarily difficult. Policy enforcement is analogous to the police, it is necessary to prevent crime, but it is not what makes a space feel safe.

What makes a workplace, or a school, or a dinner table feel safe is social norms.

If someone goes around telling women to 'get back to the kitchen', and the only thing that happens is their content is deleted or they get blocked, don't we run the risk of normalizing bad behavior? How are people to learn to be members of a safe and supportive community without visible interventions that help set the social norms for the environment? I believe social norms also protect speech.

At dinner tonight [redacted] said: my car videos are getting 100,000 views, it's natural that I'm going to get a lot of hate with that. Is it? Why is it acceptable for someone to harass someone on their surface? The most powerful solution for the integrity and safety space is to affect the supply of bad experiences via the actors creating them.

I might be wrong about my assessment, and welcome feedback about any effort or data that I'm missing. I believe that it is important to get the following efforts well-funded and prioritized:

- What is the content that is causing bad experiences for our users? How intense is the experience?
- What % of that content is policy violating? (i.e. how much of TRIPS is driven by content other than what drives Prevalence?)
- What are visible product solutions that make the community better over time? e.g. actor feedback, comment covers, pinned comments, etc.

The solutions we create I believe should have the following properties:

- The person who has the negative experience should feel heard, you don't 'perceive' racism or harassment, you experience it, and you are the source of truth for that. The feedback flow should not be just about filing a report, but about understanding the experience the person is having so we can give them the right solution.
- We should empower creators, communities, and Instagram, in setting the social norms for the spaces they are a part of.
- Where appropriate we should give feedback to actors, in the belief that they are acting with good intention and might have caused unintentional harm. There can be a range of interventions that start with 'nudges' that assume positive intention. This will allow us to separate the people who would behave differently given feedback, from the ones who are intentionally causing harm. We can then approach people who are intentionally malicious with the integrity tools.

If you would like I can give more details or specifics on this. I am appealing to you because I believe that working this way will require a culture shift. I know that everyone in m-team team deeply cares about the people we serve, and the communities we are trying to nurture, and I believe that this work will be of service to that.

Arturo