**From:** Tomas [redacted]/O=THEFACEBOOK/OU=EXTERNAL (FYDIBOHF25SPDLT)/CN=RECIPIENTS/CN=[redacted]
**Sent:** 6/9/2023 8:59:18 PM
**To:** Tomas [redacted]@meta.com]; Melissa [redacted]@meta.com]
**Subject:** Message summary [{"otherUserFbId":100021609290728,"threadFbId":null}]

Melissa [redacted] (6/09/2023 11:22:31 PDT):
>so disappointing but none of it is surprising

Melissa [redacted] (6/09/2023 11:22:44 PDT):
>I'm not surprised we have bad hashtags. our search safety team is a few employees who look at a list of popular hashtags every week to decide which to filter out :( it's super low tech

Melissa [redacted] (6/09/2023 11:22:49 PDT):
>We also do very little to make sure handles are appropriate which is so creepy and problematic

Melissa [redacted] (6/09/2023 11:23:01 PDT):
>Damn, Stamos finally figured out the most effective path to influencing Zuck

Melissa [redacted] (6/09/2023 11:23:06 PDT):
>it's a really well written article, really frames and contextualizes the issue quite well

Melissa [redacted] (6/09/2023 11:23:24 PDT):
>The fact that we recommended the pedo account that activist viewed once to all her followers is really disappointing

Melissa [redacted] (6/09/2023 11:23:29 PDT):
>Content moderators not enforcing the rules is such a big problem. Could be the biggest problem on right now

Melissa [redacted] (6/09/2023 11:23:37 PDT):
>It's a really big problem that you view one media and then we start recommending many like it. We gotta solve the rabbit hole issue

Melissa [redacted] (6/09/2023 11:24:06 PDT):
>I feel like this should be mandatory reading for all of Relevance on how recommendation systems can go. It's really fair and well written by people who clearly understand the issues well

Tomas [redacted] (6/09/2023 11:25:22 PDT):
>dang

Tomas [redacted] (6/09/2023 11:25:45 PDT):
>I was not aware that this was an issue

Tomas [redacted] (6/09/2023 11:26:08 PDT):
>this = the WSJ article?

Melissa [redacted] (6/09/2023 11:26:31 PDT):
>yes

Tomas [redacted] (6/09/2023 11:27:44 PDT):
>can you tell me more about this?

Melissa [redacted] (6/09/2023 11:28:29 PDT):
>we think rater recall could be as low as 50%, and the non-rec labelers only know English... they label 0 media/week as hate speech, bullying, or suicide and self injury even though we know the prevalence is higher. and they frequently get easier topics wrong that don't require language comprehension, like disturbing content [redacted]

Melissa [redacted] (6/09/2023 11:29:41 PDT):
>this is a more recent check on rater recall that paints a rosier picture, but Rong strongly disagrees with the methodology

Tomas [redacted] (6/09/2023 11:34:50 PDT):
>man

Tomas [redacted] (6/09/2023 11:35:24 PDT):
>You'd think we JUST started working on integrity with issues like this one

Tomas [redacted] (6/09/2023 11:35:41 PDT):
>Clearly this is hard, I don't want to dismiss the good work from the integrity teams

Tomas [redacted] (6/09/2023 11:35:55 PDT):

>But it feels like we're so behind where we should be

Melissa [redacted] (6/09/2023 13:49:31 PDT):
>I mean you see the chaos of the way products are shipped at Meta

Melissa [redacted] (6/09/2023 13:49:49 PDT):
>we don't do comprehensive, holistic overviews often enough

Melissa [redacted] (6/09/2023 13:50:12 PDT):
>we don't hold enough people responsible for metric monitoring. we only staff teams when we want to build, not maintain

Melissa [redacted] (6/09/2023 13:51:34 PDT):
>the central integrity org is pivoting to focus on "voice" aka false positives and overenforcement. their other big new focus is "integrity value": trying to justify how integrity work adds value to the business. as a result, they've stopped investing in SWOP, which is our Meta metric for content violations!

Melissa [redacted] (6/09/2023 13:52:26 PDT):
>integrity is an adversarial space. there's millions of people who will profit from finding ways to exploit it, so they're always trying. product teams change the way things work all the time and don't move in lockstep with integrity

Melissa [redacted] (6/09/2023 13:53:25 PDT):
>and there's just SO MANY TYPES OF HARM. it's hard to think of all of them for everything always. just because there's no harm on an issue type or surface today doesn't mean there won't be harm on hashtags tomorrow, so someone makes the call that it's not a priority and then boom, things evolve

Melissa [redacted] 6/09/2023 13:59:18 PDT):
>and non-rec is owned and operated by the small team on RI, so we don't benefit from Meta's best practices on making sure raters actually label the violating content as violating