Message

From: Johanna [redacted]/O=THEFACEBOOK/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=[redacted]
Sent: 9/28/2020 5:13:56 PM
To: Johanna [redacted]@fb.com]; Daniel [redacted]@fb.com]; Vaishnavi J [redacted]@fb.com]; Hitomi [redacted]@fb.com]; John [redacted]@fb.com]; Malia [redacted]@fb.com]; Mariana Gamez [redacted]@fb.com]
Subject: Message summary [{"otherUserFbId":null,"threadFbId":3115552808553662}]
Attachments: 120314681_3859322934090404_2097861768699583109_n.png; 120444773_703808340478779_9131252759469479692_n.png

Jennifer [redacted] (9/28/2020 00:55:30 PDT):
>@silent On 2, I'd be careful of wording because this makes it sound a bit worse than it is. I would say that Child Safety struggles with ensuring correct actions for new/niche surfaces because these rely on the surfaces' engineers understanding and implementing our legal obligations. Child Safety legal obligations is not something known by the general population of the country.
>
>Instagram suffered here because:
>1. Surfaces tend to be newer and in a different codebase
>2. Volumes are very small anyway (in comparison to FB) so it's hard to spot issues.

Jennifer [redacted] (9/28/2020 01:02:57 PDT):
>@silent I would instead here point out that our CEI deletes on IG vs FB are ~21x fewer.

Jennifer [redacted] (9/28/2020 01:03:10 PDT):

shared: 120314681_3859322934090404_2097861768699583109_n.png

Jennifer [redacted] (9/28/2020 01:04:06 PDT):
>@silent Basically our enforcement overall on IG is not as good and prone to errors we have less experience with.

Jennifer [redacted] (9/28/2020 01:08:31 PDT):
>@silent Another stat: We are autoenforcing on 90x fewer IIC on IG than on FB. I'm told this is mostly classifier recall because we don't have more labeled data to improve the classifier.

Jennifer [redacted] (9/28/2020 01:08:44 PDT):

shared: 120444773_703808340478779_9131252759469479692_n.png

John [redacted] (9/28/2020 01:41:22 PDT):
>@silent no additional feedback on my end on top of what Malia has said

Vaishnavi Jayakumar (9/28/2020 08:05:58 PDT):
>Is this still the case with the IIC classifer that's been calibrated for IG?

Jennifer [redacted] (9/28/2020 08:11:16 PDT):
>Yes, it is.

Vaishnavi Jayakumar (9/28/2020 08:11:22 PDT):
>My comments would be that on (1) I think achieving parity with FB is a shifting target - we may have the same classifiers and interventions rolled out on IG but they may not perform as well, to Malia's point above. I'd reframe as achieve parity with FB's *outcomes
>
>On (4) I'm skeptical that it's gaps in policy that are driving our current child safety challenges, unless you're not specifically talking about content policies. There are certainly gaps in our policies (have outlined the few that exist in our child safety landscape) but it seems lower in priority than all the detection, measurement, and enforcement challenges

Jennifer [redacted] (9/28/2020 08:12:15 PDT):
>We additionally have another classifier on FB that looks specifically at text in messages (in addition to our behavioural classifier). I do not think we would be permitted to scan IG Direct messages in this way.

Vaishnavi Jayakumar (9/28/2020 08:13:51 PDT):
>Hmm which one? Both IICT1 and T2 run across Direct

Jennifer [redacted] (9/28/2020 08:14:18 PDT):
>This is a third classifier than you're thinking about .

Jennifer [redacted] (9/28/2020 08:14:30 PDT):
>The IIC T2 classifier for IG looks just at behaviours.

Jennifer [redacted] (9/28/2020 08:14:51 PDT):
>IIC T1 looks at a subsample including text, to find T1 violations

Vaishnavi Jayakumar (9/28/2020 08:14:54 PDT):
>That's the IIC actor classifier no?

Jennifer [redacted] (9/28/2020 08:15:36 PDT):
>On FB we have another classifier that is text based (so not e2ee proof) that makes a lot of autodeletes and scans all minor to adult conversations.

Vaishnavi Jayakumar (9/28/2020 08:16:11 PDT):
>My understanding is there are 3 - actor classifier (also disables for IICT2), IICT2 for DM review, and IICT1 for DM review

Jennifer [redacted] (9/28/2020 08:16:13 PDT):
>https://www.internalfb.com/intern/wiki/Harmful_Behaviour_Engineering/Child_Safety/IIC/Proactive_Enforcement/IIC_Tier_2_Proactive_Detection/#1-borat

Jennifer [redacted] (9/28/2020 08:16:28 PDT):
>there are two IIC T1 classifiers

Jennifer [redacted] (9/28/2020 08:16:45 PDT):
>One for reactive, one for proactive. Both for human DM review.

Jennifer [redacted] (9/28/2020 08:17:11 PDT):
>On FB, two classifiers with no review, just autodelete: the behavioural one, and Borat (message text-based)

Vaishnavi Jayakumar (9/28/2020 08:18:22 PDT):
>Ahh I see! Autodelete for text-based is new to me

Jennifer [redacted] (9/28/2020 08:18:43 PDT):
>Right. But that is a significant chunk of FB autodelets.

Jennifer [redacted] (9/28/2020 08:19:19 PDT):
>We could turn this on for IG with (I think) little effort because text is text, but we'd need permission from IG Direct, and in the past they've been less willing to give this.

Hitomi [redacted] (9/28/2020 08:55:16 PDT):
>On your second point here, is the volume (of CEI?) smaller because of difficulties in detection (since we have surfaces that are newer)?

Jennifer [redacted] (9/28/2020 08:56:21 PDT):
>Yes, our detection is known to be worse.

Hitomi [redacted] 9/28/2020 09:11:54 PDT):
>For the 4th point. Would you suggest removing work on policy from this list? or are there other policy-related opportunities that you would mention here.

Vaishnavi Jayakumar (9/28/2020 10:13:56 PDT):
>I think it's more a matter of prioritisation - I'd move up the understand / sizing work and move down the policy work - there are examples in the note of policy-related opportunities