

Message

From: Faye [REDACTED]/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN [REDACTED]
Sent: 10/31/2023 9:50:49 PM
To: [REDACTED]@meta.com]; Taylor [REDACTED]@meta.com]; Rishi [REDACTED]@meta.com]; Swithin [REDACTED]@meta.com]; Rahul [REDACTED]@meta.com]; Lexi [REDACTED]@meta.com]; Nir [REDACTED]@meta.com]; Gabriel [REDACTED]@meta.com]; Tony [REDACTED]@meta.com]; Sophie [REDACTED]@meta.com]; Mallory [REDACTED]@meta.com]; Alison [REDACTED]@meta.com]; Corri [REDACTED]@meta.com]; Kathy [REDACTED]@meta.com]; Stephanie [REDACTED]@meta.com]; Faye [REDACTED]@meta.com]; Carolina [REDACTED]@meta.com]; Arcadiy [REDACTED]@meta.com]; Malia [REDACTED]@meta.com]; Hudson [REDACTED]@meta.com]; Komal [REDACTED]@meta.com]; Kate [REDACTED]@meta.com]; Rory [REDACTED]@meta.com]; Amit [REDACTED]@meta.com]; Yashaswi [REDACTED]@meta.com]; Victoria [REDACTED]@meta.com]; Andy [REDACTED]@meta.com]; Jimmy [REDACTED]@meta.com]; Kaitlin [REDACTED]@meta.com]; Ashly [REDACTED]@meta.com]; Mike [REDACTED]@meta.com]; Melissa [REDACTED]@meta.com]; Kristina [REDACTED]@meta.com]; Sayed [REDACTED]@meta.com]; Mike [REDACTED]@meta.com]; Richard [REDACTED]@meta.com]; Sarah [REDACTED]@meta.com]; James [REDACTED]@meta.com]; Scott [REDACTED]@meta.com]; Leah [REDACTED]@meta.com]
Subject: Message summary [{"otherUserFbId":null,"threadFbId":6427561907365847}]
Attachments: video-e0033bc9-90c3-4c67-ac9f-2fd4b6a4f03d-1698767061.mp4;
377124407_284350133924228_7896967635058393494_n.png;
370241162_303782345840542_536791761734269577_n.png

Malia [REDACTED] (10/31/2023 01:31:09 PDT):
>@Carolina I'll be curious of the outcome for these breastfeeding groups; is here the right way to follow?

Kaitlin [REDACTED] (10/31/2023 06:39:52 PDT):
><https://www.internalfb.com/cases/4439816>

Kaitlin [REDACTED] (10/31/2023 06:42:18 PDT):
>Looks like current consensus is to remove as an SoP call for ANSA

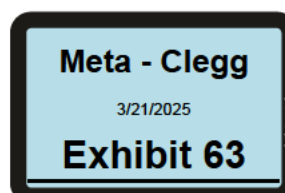
Carolina [REDACTED] (10/31/2023 06:44:12 PDT):
>Yes, you can follow C#4438125. For now, PRESC ran a fanout of the original group's admins and was able to disable a total of 25 groups for ANSA violations, with alignment from CPT.
>
>They are now running inauthenticity checks for the admins.

Andy [REDACTED] (10/31/2023 08:45:19 PDT):
>Flagging this new video Horwitz just sent of the AYSF recommendations for one of the Instagram accounts he's used to do this chid safety research.
>
>Can we please look into this? I'd have assumed at this point the steps we've taken would have accounted for at least most of these recommendations.

Andy [REDACTED] (10/31/2023 08:45:20 PDT):
shared: video-e0033bc9-90c3-4c67-ac9f-2fd4b6a4f03d-1698767061.mp4

Alison [REDACTED] (10/31/2023 08:59:23 PDT):
>Oof - thanks for flagging Andy.
>1. We need to nuke these accounts ASAP. @Carolina [REDACTED] - what is the best way to have ops not only remove the accounts that are being recommended, but also can we do a fanout of the accounts that are following the accounts being recommended here.
>2. MaSCA scores. Can we figure out what Jeff's account's MaSCA score is and why it is being recommended not only seemingly minor content, but what also looks like accounts that *should* have a MaSCA score themselves and be non-rec. @Rishi [REDACTED], is this something your team can prioritize today? It seems like our thresholds for AYSF may not be in the right place?
>3. For #2 - @Carolina [REDACTED], can we ensure ops pulls the list of handles before removing the accounts so we can do that review?

Gabriel [REDACTED] (10/31/2023 09:01:11 PDT):



CONFIDENTIAL

METANMAG-039-00000003



PLTF EX-06012_0001

>If we could get a list of these account IDs I think we can then check what % of them had a MaCSA score and whether there are signals we are missing

Carolina [REDACTED] (10/31/2023 09:01:42 PDT):

>Hey! I'll spin up a case for this right now and will investigate. Gabriel, I'll subscribe you so you can get the IDs.

Alison [REDACTED] (10/31/2023 09:04:15 PDT):

>Thank you @Carolina [REDACTED] - please subscribe me and Andy as well.

Carolina [REDACTED] (10/31/2023 09:05:44 PDT):

>Will do!

Alison [REDACTED] (10/31/2023 09:05:54 PDT):

>Can we please specifically ask for whoever is reviewing to list out the IDs somewhere in the case so they can be pulled easily.

Alison [REDACTED] (10/31/2023 09:05:55 PDT):

>Thanks!

Kristina [REDACTED] (10/31/2023 09:08:10 PDT):

>i3 Child Safety identified several accounts belong to Jeff. Last we checked, they didn't have MaCSA scores.

>

>#shareadoc comment

https://docs.google.com/spreadsheets/d/1H4W5mnJGEyHo16HIsO6wg78dzWdsRssGRYSdDpL_12s/edit#

Alison [REDACTED] (10/31/2023 09:13:37 PDT):

>@Gabriel [REDACTED] - can we take a look here while ops is pulling the accounts from the video for why his accounts don't have a MaCSA score given how often he is searching for/viewing this type of content.

Malia [REDACTED] (10/31/2023 09:18:17 PDT):

>There are no MaCSA signals for views.

Malia [REDACTED] (10/31/2023 09:18:46 PDT):

>There are MaCSA signals for likes and saves of content. And there are signals for searches *if* the search terms are in the high-precision list.

Malia [REDACTED] (10/31/2023 09:19:03 PDT):

>We have not been able to include the wider set from the taskforce in this list as the precision is varying.

Gabriel [REDACTED] (10/31/2023 09:19:50 PDT):

>Yes - @Kristina [REDACTED] and @Malia [REDACTED] - for these accounts

>1/ should the MaCSA signals be triggered? is there some gap in what we are detecting?

>2/ are there any additional human-driven signals on looking at these accounts which, if automated, would have led to these accounts being considered risky (1 or 2 points)

Alison [REDACTED] (10/31/2023 09:21:24 PDT):

>+1 my point above was based on what I assume the behavior of his accounts are (viewing/searching) - we should confirm first if the MaCSA signals should be triggered here (per @Gabriel [REDACTED]'s point).

Carolina [REDACTED] (10/31/2023 09:28:22 PDT):

>Also, just flagging that I'm using my test user account to surface the accounts, and also started being recommended some myself.

Carolina [REDACTED] (10/31/2023 09:28:32 PDT):

shared: 377124407_284350133924228_7896967635058393494_n.png

shared: 370241162_303782345840542_536791761734269577_n.png

Carolina [REDACTED] (10/31/2023 09:28:47 PDT):

>I'm adding these to the investigation and will take them down as well

Alison [REDACTED] (10/31/2023 09:30:16 PDT):

>Can you share the case here when you have it? Separately confirming that @Malia [REDACTED] @Kristina [REDACTED] - you are looking into Gabe's ask above?

Kristina [REDACTED] (10/31/2023 09:32:12 PDT):

>Yes - I can look into this

Swithin [REDACTED] (10/31/2023 09:33:04 PDT):

>Shouldn't these accounts also be auto enqueued for review based on account names esp pizzalink, kidseller? @Diego

Carolina [REDACTED] (10/31/2023 09:34:35 PDT):

>C#4447847

PLTF EX-06012_0002

Alison [REDACTED] (10/31/2023 09:35:58 PDT):

>Great - IDs are in the case @Gabriel [REDACTED]

Melissa [REDACTED] (10/31/2023 09:43:30 PDT):

>Also looks like a handful do have reactive user reports on the profile, but are sitting in the queues/pending

Alison [REDACTED] (10/31/2023 09:43:53 PDT):

>Sorry - what profile?

Melissa [REDACTED] (10/31/2023 09:44:15 PDT):

>a handful of those in the case

Diego [REDACTED] (10/31/2023 09:47:27 PDT):

>We have variations of this. Would be helpful to see what are the new names that are being used and we can add them to the TMS banks for enqueueing or auto actioning.

Alison [REDACTED] (10/31/2023 09:48:14 PDT):

>Coming back to this:

>1. @Carolina [REDACTED] - when we have it, please provide an update on the accounts that have been actioned/result of the fanout.

>2. @Malia [REDACTED] @Kristina [REDACTED] @Gabriel [REDACTED] - put a doc here so we can centralize across the reporters accounts + IDs from the video

https://docs.google.com/document/d/1eZa0GCrc1m_SyUyrNF1Vtv4r910dNQoKdSoq6brEvNs/edit#shareadoc

Alison [REDACTED] (10/31/2023 09:49:19 PDT):

>Great - can you please review them from the case shared above and add new content to the TMS banks if permissible.

Carolina [REDACTED] (10/31/2023 09:51:28 PDT):

>Diego - I'll include these insights in the case once I'm done with the investigation, but based on initial fan-outs, I'm seeing a lot of keyword obfuscation (i.e. "megha" or "megga" instead of "mega", and instead of "seller", using "sellar" or "zshallar") But also more obvious versions like, "pizza mega sellers" or "trusted cp seller".

Rishi [REDACTED] (10/31/2023 09:53:37 PDT):

>Gabriel/Malia: for the accounts that are being surfaced here, should they be caught by MacSA? Or are there other CSE-I profile level classifiers that can catch these?

Malia [REDACTED] (10/31/2023 09:57:29 PDT):

>We'll take a look and assess. I'm assuming your question here is on the recommended profiles and not JH's.

Rishi [REDACTED] (10/31/2023 09:57:53 PDT):

>yes

Rishi [REDACTED] (10/31/2023 09:58:04 PDT):

>Could we use <https://fb.workplace.com/notes/311970791222605> for soft actions?

Malia [REDACTED] (10/31/2023 09:58:28 PDT):

>Is this not live?

Rishi [REDACTED] (10/31/2023 09:59:07 PDT):

>Reading the note, it seems like we're using it only to enqueue/UFAc

Rishi [REDACTED] (10/31/2023 09:59:57 PDT):

>Hypothetically, we could use lower thresholds to filter out of recs

Malia [REDACTED] (10/31/2023 10:00:09 PDT):

>Ah yes, ok.

Gabriel [REDACTED] (10/31/2023 10:00:56 PDT):

>Yeah i think that's right. We didn't look into this for taskforce but might be an option; think we could probably get a precision number based on what's getting enqueued

Carolina [REDACTED] (10/31/2023 10:07:16 PDT):

>@Malia [REDACTED] /@Gabriel [REDACTED] on the case the accounts that were recommended to Jeff are labeled as "JH AYSF", you'll notice the case now has about 172 IG accounts, the bulk of which I've found using keyword combinations on front-end and conducting fanouts to find device-linked IG accounts.

Rishi [REDACTED] (10/31/2023 10:08:08 PDT):

>ok lmk.

>

>Could we also use rough heuristics for eg, "[REDACTED] [REDACTED] [REDACTED]" and we filter these out of recs.

Gabriel [REDACTED] (10/31/2023 10:09:40 PDT):

PLTF EX-06012_0003

>worth a sidebar: I think this is how we should be doing nonrec more generally; will need a broader conversation incl policy and upstream teams who work on X-Ray image features

Rishi [REDACTED] (10/31/2023 10:10:56 PDT):
>ok let me spin up a side chat

Amit [REDACTED] (10/31/2023 10:13:22 PDT):
>i think this is a promising direction. are we going to expand on this line of thinking and actually eliminate borderline seeds from the rec selection process?

Amit [REDACTED] (10/31/2023 10:13:39 PDT):
>is that something we are also discussing

Swithin [REDACTED] (10/31/2023 10:15:10 PDT):
>Some of these do have a ton of stories. We should definitely remove them from seed sourcing. I'm also seeing some UFAC activity on the accounts and am unclear about what those mean

Sayed [REDACTED] (10/31/2023 10:15:32 PDT):
>Can we also find out what the stated/predicted ages are for these recommended accounts? It can help inform some of how we are thinking about AYMF for teens.

Amit [REDACTED] (10/31/2023 10:20:49 PDT):
>so is this change to seed sourcing live now or we need to do it?

Rishi [REDACTED] (10/31/2023 10:24:42 PDT):
>It's not live. Once we align on which signals can/should be used, we should probably do both: (a) stop seed sourcing and (b) do a last pass ranking integrity filter.
>
>(a) will take time as it gets deep into the AYMF stack
>(b) can be a quicker change

Gabriel [REDACTED] (10/31/2023 10:49:13 PDT):
>Folks, it might be helpful to step back a little on IG recs;
>
>1/ We have shipped a bunch of things already - we can validate they are working as intended
>
>2/ More broadly, AYMF and Content filtering are still not where we would want them to be. There are multiple levers we could pull but all would take some focus and both eng+policy
>
>3/ business outcomes needed and timelines are currently unclear; it's not obvious to me that pivoting back into war room mode across IGWB and XI teams and shipping some hotfixes now would either address the Jeff story issue (given timelines of this week) or set up teams for longer term success - but I would like to get some SR steer on this cc @Alison [REDACTED]

Carolina [REDACTED] (10/31/2023 14:43:46 PDT):
>Hey all, following up with the latest enforcement figures from the GRO investigation on the AYSF case (C#4447847).
>
>In addition to the 12 accounts featured in Jeff's video, and the 204 IG accounts we disabled from preliminary fan-outs, GRO reviewed 1.6K users that were device linked to the original disabled accounts. We conducted subject element reviews (profile photo, username, and bio), and disabled 987 additional IG accounts for CSAM trade.
>
>In total, the investigation resulted in 1203 IG account takedowns. The majority of the users were based in India and engaging in trading/selling of telegram links.
>
>As a next step, we're going to collect the common keywords we spotted and will share with Diego and Jonny to inform further TMS banking efforts.
>
>Let us know if you have any questions or have any additional IDs for review.
>
>cc. @Andy @Alison

Diego [REDACTED] (10/31/2023 14:50:49 PDT):
>Thank you, Carolina!