

Message

From: Riana [REDACTED] [/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=[REDACTED]]
Sent: 8/30/2023 12:30:02 PM
To: Antigone Davis [REDACTED]@meta.com]; Ravi Sinha [REDACTED]@meta.com]; Arcadiy [REDACTED] [REDACTED]@meta.com]
CC: Ashly Nikkole [REDACTED] [REDACTED]@meta.com]
Subject: Re: [a/c priv] Child safety task force update - August 28, 2023

Hi all – The meeting with Michael Johnson has been scheduled for Friday, September 8. Thank you!

From: Antigone Davis <[REDACTED]@meta.com>
Date: Monday, August 28, 2023 at 7:40 PM
To: Riana [REDACTED] <[REDACTED]@meta.com>, Ravi Sinha <[REDACTED]@meta.com>, Arcadiy [REDACTED] <[REDACTED]@meta.com>
Cc: Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>
Subject: Re: [a/c priv] Child safety task force update - August 28, 2023

Thank you!! I really appreciate the follow up (and all the work done to get us to this point).

It would be great to be briefed by Michael Johnson/his team. I'd appreciate including Ravi as well.

From: Riana [REDACTED] <[REDACTED]@meta.com>
Sent: Tuesday, August 29, 2023 7:23:05 AM
To: Antigone Davis <[REDACTED]@meta.com>; Ravi Sinha <[REDACTED]@meta.com>; Arcadiy [REDACTED] <[REDACTED]@meta.com>
Cc: Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>
Subject: Re: [a/c priv] Child safety task force update - August 28, 2023

1. Previously, multiple FB and IG-specific child safety keyword lists were maintained and used across various search interventions. We now have two lists and streamlined search interventions across FB, IG and soon to be Messenger. You can find more details in Leah's document here: <https://docs.google.com/document/d/1gWt6qKRDMoahmWuowVNg0mze1C1CNqt8GFa8noQPXtA/edit>.
2. You can read more about the second line of defense work [here](#). The org managing this work is new. It's run by Michael Johnson, head of Integrity, Support, Security, and Operations (ISSO) Governance, Risk and Compliance (GRC) and Second Line of Defense (2LOD). Him and his team have offered to brief you if that's something you'd be interested in. Let me know, and I can help make sure it gets set up.
3. Sounds good – Ravi, you're on the chat thread we're going to use moving forward which comms is on too.
4. Arcadiy briefly mentioned this in the meeting on Thursday, but we didn't plan for any of these metrics to be public yet. We're working with comms to scope research on what parents/caregivers want to hear about this integrity work, and then we'll see if we have the right metric to support that.

Happy to answer any other questions,
Riana

From: Antigone Davis <[REDACTED]@meta.com>
Date: Monday, August 28, 2023 at 5:02 PM
To: Riana [REDACTED] <[REDACTED]@meta.com>, Ravi Sinha <[REDACTED]@meta.com>, Arcadiy [REDACTED] <[REDACTED]@meta.com>
Cc: Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>
Subject: Re: [a/c priv] Child safety task force update - August 28, 2023

HIGHLY CONFIDENTIAL (COMPETITOR)
HIGHLY CONFIDENTIAL (ATTORNEY'S EYES ONLY)

META3047MDL-172-00000572
METANMAG-043-00000572
PLTF EX-01874_0001

PLTF EX

PLTF EX-01874

1. Thanks, this looks good to me. How does it compare to what we did before?
2. Where can I read more about "Second Line of Defense" work? Maybe it is just me that does not have the needed context.
3. Thanks, comms will need to be prepared to answer questions on our investment in human investigation teams. I think you know this and are working with them on how to manage response. @Ravi Sinha, let's plan to stay close to this open conversation both what numbers we might share and more importantly, our investment here.
4. These numbers are all good. It was my understanding that we would be working toward some numbers we would share publicly and with some kind of regularity. Where are we on that?
5. Sounds good, do remember, was just trying to understand our numbers for other harm types so I can begin thinking about this for that conversation down the road. On this one, Nick, Joel and Kevin, will want to hear the reply your wrote to me. Fine to wait for the question – but I am getting pinged on the side.

And thanks for coming back so quickly and fully!

From: Riana [REDACTED] <[REDACTED]@meta.com>

Date: Tuesday, August 29, 2023 at 4:38 AM

To: Antigone Davis <[REDACTED]@meta.com>, Ravi Sinha <[REDACTED]@meta.com>, Arcadiy [REDACTED] <[REDACTED]@meta.com>

Cc: Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>

Subject: Re: [a/c priv] Child safety task force update - August 28, 2023

Hi Antigone –

Thanks for reading through this and your response. Arcadiy is out on parental leave for the next month.

We provided more detail in the email we sent you, Molly, Guy, etc. last week because we knew you all would want to click in a bit more! We don't plan on following up with this group unless they come back with questions, but I answered your questions below. Please let me know if you want to discuss any of these further. Thank you!

1. Ops is responsible for this – Diego San Martin and Thantap [REDACTED] are owning. The process document is here: <https://docs.google.com/document/d/1wepsYw9zChUNNOsaw-R6kykPZsOfm7KXV5f2HduAyRs/edit?usp=sharing>. Leah [REDACTED] on your team has been instrumental in putting this together.

2. Second Line of Defense work has been stood up beyond Child Safety. Right now, the GRC team is focused on HEx and Child Safety to start. We ran this by Molly and Guy, and they didn't think this would be news to this group.

3. We had a conversation with Sarah [REDACTED] on i3 about this and while we didn't commit to a specific number of SNDs, they are going to continue through the end of the year. There is an open conversation between i3 and XI around whether we share the # of SNDs publicly or if we just share how i3's findings are incorporated into integrity measures.

4. As part of the Task Force, several metrics were stood up or adopted, both temporarily for the duration of the task force and as a longer term measurement. In the process, several of them went through multiple iterations to ensure Risks they are addressing are accurately captured. With the task force ending, we wanted to close off this goal by summarizing these metrics:

1. Theme 1: Reactive Recall

- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]

2. Theme 3: Recommendations

3. Theme 5: Futureproofing

- a. Risky Actor Connectedness (RAC): We worked through several iterations and metric proposals on how to capture degree to which Bad Actors are successfully connecting with each-other on our platforms, finally landing on a proposal which is currently under leadership consideration.

5. I also saw your comment in the doc re: RoaR that said, "How does this compare to other harm types? What do we consider an acceptable target? How long do we think it might take to get there?" As discussed during the meeting last week, we have just established RoaR as a north star metric. We first need to close out theme 1 and complete the "no regrets" fixes (additional improvements to ranking to better match our ideal design and additional protocol/training improvements on human review) before pursuing changes that might have greater trade-offs, and see how far those no regrets investments get us. Then, we can assess how we're doing on all user reports and XI can escalate the tradeoffs for increasing RoaR. That will probably take us to the end of the year.

From: Antigone Davis <[REDACTED]@meta.com>

Date: Monday, August 28, 2023 at 3:55 PM

To: Ravi Sinha <[REDACTED]@meta.com>, Arcadiy [REDACTED] <[REDACTED]@meta.com>, Riana [REDACTED] <ria[REDACTED]i@meta.com>

Cc: Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>

Subject: FW: [a/c priv] Child safety task force update - August 28, 2023

Hi all (Ravi, don't think you were on this email, looks like an oversight) – first thank you for all your work on this, herculean -- a few things, that the three of you might want to go back and clarify:

1. See first highlighted item below: I think Nick, Joel and even myself will want to how are we ensuring, frequency, and who holds responsibility for this.
2. See second highlighted item below highlighted below: I don't think many on this email will understand what this means (including me). You might want to consider following up on the thread and explaining in lay person terms. What do we mean by second line of defense processes, what are the P1 controls, ill GRC be responsible for constantly monitoring or did they just help build the system, is one person/team accountable for this monitoring? The linked doc doesn't specify either.
3. I also think we need to address investment in SNDs in this area, did the Task Force assess need for greater investment here to support future proofing?
4. What happened to developing child safety metrics (Chirs [REDACTED]'s AI)?

Thanks again!

From: Riana [REDACTED] <[REDACTED]@meta.com>

Date: Monday, August 28, 2023 at 7:35 PM

To: Nick Clegg <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn

<[REDACTED]@meta.com>, Scott Tucker <[REDACTED]@meta.com>, Frank [REDACTED] <[REDACTED]@meta.com>, Victoria [REDACTED] <[REDACTED]@meta.com>, Amit Bhattacharyya <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Arcadiy [REDACTED] <[REDACTED]@meta.com>
Cc: Henry [REDACTED] <[REDACTED]@meta.com>, Diana [REDACTED] <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, Markus [REDACTED] <[REDACTED]@meta.com>, Jonathan [REDACTED] <[REDACTED]@meta.com>, Carl [REDACTED] <[REDACTED]@meta.com>, Nicky [REDACTED] <[REDACTED]@meta.com>, Scott [REDACTED] <[REDACTED]@meta.com>, Victoria [REDACTED] <[REDACTED]@meta.com>, Mary [REDACTED] <[REDACTED]@meta.com>, Joseph [REDACTED] <[REDACTED]@meta.com>, Stephanie [REDACTED] <[REDACTED]@meta.com>, Hayley [REDACTED] <[REDACTED]@meta.com>, Kate Gotimer <[REDACTED]@meta.com>

Subject: [a/c priv] Child safety task force update - August 28, 2023

Hi all,

This will be our final leadership update on the child safety task force. As a reminder, we stood up the task force as a result of the WSJ inquiry and subsequent [story on pedophile activity on Instagram](#). At the time, we expected a series of stories to be published, but we have not seen subsequent stories and are not expecting any at the moment. Of the 4 areas the task force identified based on the reporter's initial inquiry (user reports, recidivism, recommendations, and future proofing), we met exit criteria for three of them. The team will continue working on the fourth area - user reports.

A few highlights of the task force work:

1. We improved our ability to automatically disable bad actors in this space and closed gaps to limit recidivism.
 - a. [REDACTED] with 95% precision which is leading to around ~30k accounts disabled a month.
 - b. We made significant improvements to limit recidivism. For both account linking and device disables, child safety now matches our strictest policies. Today, we will device block for child safety in the same way we device block for terrorism.
2. We have a unified keyword list (currently in the tens of thousands of terms) across apps and interventions (i.e if we block a term such as "pizza" it will then apply across IG, FB and soon to be Messenger). We are ensuring we keep this list up to date using multiple inputs.
3. [REDACTED]
4. The task force also focused on future proofing this work and ways to better anticipate problems in this space. We improved the process for incorporating signals from i3 network disruptions to feed into the work done above (keyword list, filtering etc.). We also worked with the Governance Risk and Compliance (GRC) team to stand up the second line of defense processes by identifying and validating relevant controls. P1 controls testing was completed on Friday with few exceptions. The GRC team aims to share findings from testing within the next 1-2 weeks with relevant stakeholders to kick start tracking of any identified issues.

User reporting is the one theme the task force did not meet its exit criteria on and the subset of the task force working on this theme will continue. As discussed before, a vast majority of child safety content is removed before it is reported by users, so reporting plays a small but important role in our overall risk reduction. Over the course of the task force we were able to update reviewer guidance and better prioritize user reports by incorporating additional classifier signals into prioritization algorithms. However, we are still missing violating content when it is reported to us more often than we would like. Of the violating child safety content that is reported to us we successfully take down 31% on Facebook and 17% on Instagram, and we intend to improve these numbers through additional work on report prioritization, training, etc. before fully winding down theme 1 of the task force.

You can find more details on the work of the task force [here](#). Please let us know if you have any questions and thank you for all the help and support through the duration of this task force!

Thank you,
Riana, Arcadiy and the Child Safety Task Force

On Aug 19, 2023, at 2:00 AM, Arcadiy [REDACTED] <[REDACTED]@meta.com> wrote:

Folks,

Comms continues to hear from Jeff Horwitz. He's now interested in our IG Live detection and enforcement because he says he's witnessed accounts boosting the spread of child safety-related content by co-hosting Lives. [REDACTED]

[REDACTED] While we consider this out of scope for the Task Force, we found a SEV related to this issue ([S356462](#)). We have pulled in the right teams to support and are tracking the outcome of the SEV.

We continue to track several issues to address the problem of FB Groups created for the purposes of child sexual exploitation/solicitation/endangerment or groups containing a high volume of Child Safety violating content. [REDACTED]

[REDACTED] We are working with ops and others to scope this.

[REDACTED]
[REDACTED]
[REDACTED]

[REDACTED] We are actively working on resolving these by early next week.

On the Task Force, work continues at an aggressive pace especially as we intend to close things out by next Friday. A high-level update is below, and our final update next week will be a detailed look at how we fared reaching our exit criteria. Please let us know if you have any questions in the meantime.

1. **Theme 1: User Reporting.** As mentioned last week, [REDACTED]
[REDACTED] We are now scoping options for additional improvements and will have a plan with timelines by next week.
2. **Theme 2: Account enforcement.** Child Safety Eng has been working through a backlog of issues discovered during the Theme 2 audit of our enforcement. The team ranked the issues and drew a line on what we would complete as part of the Task Force and what would be scoped for Q4. Of all the issues (detailed below), only 1 is at risk of missing the target end of task force date. If that slips, we will address the aspect of the issue that makes it a high priority and schedule any follow up work in Q4.
 1. IG Findings & Fixes:
 1. **Issues with reporting in app:** [REDACTED]
[REDACTED]
[REDACTED]
 1. **Issues with SRT enforcement flows:**
 1. A small number (<1%) of SRT jobs are failing to be created, which is leading to any violation that would have been detected in the job to go unnoticed.
 2. A small number of jobs (<1%) are failing to render, which in some cases makes the job unworkable and in other cases just negatively impacts precision/recall of the job.
 3. [REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
 1. **Issues with enforcement action applying correctly:**
 1. A small number (<2%) of content deletes result in exceptions, which could indicate the content delete failed.
 2. A small number (<2%) of disables result in exceptions, which could indicate the users are not disabled. This excludes any soft actions.
 3. [REDACTED]
[REDACTED] the action is being logged incorrectly for compliance.
 1. **Logging issues:** [REDACTED]
[REDACTED]
[REDACTED]
 1. **Test tooling issues:**
 1. Some generated IG test users are unable to report profiles. This is a suspected bug with the testing tool rather than user reporting flows.
 2. Messaging flows testing: All generated IG test users are IG messaging interop users, which means some test scenarios can't be run.
 1. FB Findings & Fixes:
 1. **Issues with reporting in app:**

1. FB profiles can't be directly reported for Child Safety concerns. This is possible for Pages and SOAPs but not for plain old Profiles
2. Some report options that are Child Safety related (e.g. Child Abuse) don't create the expected review jobs
1. **Logging issues:** Profile review jobs aren't searchable by reporter, which makes it difficult to test.
3. **Theme 3: Recommendations and discovery.**
 1. We continue to be on track to consolidate our keyword lists that trigger interventions. Implementation is only at risk for Messenger, but if they're not able to complete the work next week, it will be done as a fast follow to the Task Force.
 1. We have implemented recommendation filters for key account and content discovery/recommendation surfaces to reduce non-rec account/content exposure by 50% for content and profiles each across Facebook and IG. The only major surface that remains an issue is People You May Know on Facebook, where we are working through growth metrics trade-offs. We are working on getting to a resolution early next week.
4. **Theme 5: Future proofing & auditability.** Part of this theme was ensuring that the "second line of defense" work that was already happening was in lockstep with our work. That work will continue beyond the Task Force, and we feel confident we've partnered with that team in a sustainable way to ensure we have systems in place to continuously monitor our systems. In addition, we're finalizing our bad actor interaction metric and finalizing the talking points for an update on the outcomes of the task force with the European Commission.

From: Arcadiy [REDACTED] <[REDACTED]@meta.com>

Date: Friday, August 11, 2023 at 23:50

To: Nick Clegg <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>, Scott Tucker <[REDACTED]@meta.com>, Victoria [REDACTED] <[REDACTED]@meta.com>, Frank [REDACTED] <[REDACTED]@meta.com>, Henry [REDACTED] <[REDACTED]@meta.com>, Diana [REDACTED] <[REDACTED]@meta.com>

Cc: Amit Bhattacharyya <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, Markus [REDACTED] <[REDACTED]@meta.com>, Jonathan [REDACTED] <[REDACTED]@meta.com>, Carl [REDACTED] <[REDACTED]@meta.com>, Nicky [REDACTED] <[REDACTED]@meta.com>, Scott [REDACTED] <[REDACTED]@meta.com>, Victoria [REDACTED] <[REDACTED]@meta.com>, Mary [REDACTED] <[REDACTED]@meta.com>, Riana [REDACTED] <[REDACTED]@meta.com>, Joseph [REDACTED] <[REDACTED]@meta.com>, Stephanie [REDACTED] <[REDACTED]@meta.com>, Hayley [REDACTED] <[REDACTED]@meta.com>

Subject: [a/c priv] Child safety task force update - August 11, 2023

Hi folks,

Beyond the updates on the task force itself, we wanted to mention S357962, an escalation from Google about Messenger searches for community chats returning results that suggested they may be used to share CSAM, which put us at risk of Google pulling the Messenger app from the Play Store.

This was an issue we already intended to address as part of the task force and had already mitigated on Facebook and Instagram; bringing the same search term filtering to Messenger was planned work. As a result of the SEV, teams accelerated that effort and shipped search filtering across all Messenger platforms (Android, iOS, web, desktop).

In addition, this escalation and parallel investigations around a list of groups reported to us by the Stanford Internet Observatory turned up gaps in our coverage of Groups specifically; teams have started identifying and addressing issues and we hope to land improvements to these in the coming weeks.

As for the task force, we'll keep this week's update brief. As we get close to the end of the Task Force, we are starting to measure the effectiveness of our work. Below is an overview of what we're seeing:

1. **Theme 1: User Reporting.** While we believe we will better quantify this problem and improve our systems and will hit most of our exit criteria, we believe we will not be able to meaningfully mitigate this risk by the end of the task force.
 1. We launched a measurement for Recall over all Reports (RoarR) as our new north star metric for defensible user reporting enforcement. This captures multiple aspects of recall, including whether we reviewed a report in the first place and whether a rep is likely to action it correctly. Initial results on content RoarR suggested we were at ~24%.
 2. Last week, in our effort to improve Rep Recall, we launched updates across the Child Sexual Exploitation, Abuse and Nudity (CSEAN) policies & protocol to address enforcement issues identified in the audit. Given the cadence of policy/protocol launches and the need to re-train reps, this is our "one shot" at improving recall through training and guideline improvements as part of the task force.
 1. [REDACTED]
 2. We expect there is some ramp time for reps to internalize the latest updates, so we'll repeat this test in the next few weeks. As we find additional issues, we'll look for additional opportunities for improvement to be scoped for H2 roadmaps.
2. **Theme 2: Account enforcement.** We believe this risk will be partially mitigated by the end of the task force.
 1. We have launched a range of new Child Safety account disable triggers leveraging newly developed detection approaches. This week, we expanded actor enforcement [REDACTED]
 2. The initial data shows we're linking and disabling additional accounts/devices:
 1. [REDACTED]
3. **Theme 3: Recommendation and discovery.** We believe this risk will be partially mitigated by the end of the task force.
 1. We are systematically reviewing thousands of keywords as we consolidate our trigger lists for things like search interventions in an effort to snap to a simplified and longer-term defensible list of search filters (and turn off the temporary levers we had put in place at the start of this escalation that blocked an expansive set of search terms). We are now on track to implement the consolidated lists across FB, IG, IGD and Messenger. We'll start testing the implementation early next week.

2. We continue to run tests on FB before implementing recommendation filters for key account and content discovery/recommendation surfaces to reduce non-rec account/content exposure by 50% for content and profiles each.
4. **Theme 5: Future proofing & auditability.** We believe this risk will be partially mitigated by the end of the task force.

1. By next week, legal

PRIV

PRIV

From: Arcadiy [REDACTED] <[REDACTED]@meta.com>

Date: Friday, August 4, 2023 at 23:26

To: Nick Clegg <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>, Scott Tucker <[REDACTED]@meta.com>, Victoria [REDACTED] <[REDACTED]@meta.com>, Frank [REDACTED] <[REDACTED]@meta.com>, Henry [REDACTED] <[REDACTED]@meta.com>, Diana [REDACTED] <[REDACTED]@meta.com>

Cc: Amit Bhattacharyya <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, Markus [REDACTED] <[REDACTED]@meta.com>, Jonathan [REDACTED] <[REDACTED]@meta.com>, Carl [REDACTED] <[REDACTED]@meta.com>, Nicky [REDACTED] <[REDACTED]@meta.com>, Scott [REDACTED] <[REDACTED]@meta.com>, Victoria [REDACTED] <[REDACTED]@meta.com>, Mary [REDACTED] <[REDACTED]@meta.com>, Riana [REDACTED] <[REDACTED]@meta.com>, Joseph [REDACTED] <[REDACTED]@meta.com>, Stephanie [REDACTED] <[REDACTED]@meta.com>

Subject: [a/c priv] Child safety task force update - August 4, 2023

All,

We're about three weeks from the end of the task force, and we remain broadly on track to complete our exit criteria. While there are some items that are behind, we've worked to accelerate these and expect to complete the work within the timeframe of the task force.

We saw two key launches this week, including:

1. Launched updates across the Child Sexual Exploitation, Abuse and Nudity (CSEAN) policies & protocol to address enforcement issues identified in the audit. The updates were informed by analyzing an extensive sample of user reports that were not actioned by scaled review, with a particular focus on providing additional examples and clarification on Solicitation and Sexualization of Children. We also updated guidelines for age determination and addressed gaps in decision trees across multiple review flows. Please find the key changes below and refer to the launch task for further details: [T158555096](#). We will assess the impact of these updates on reviewer recall after re-injection of the golden frame sample and share updates with this group.
2. Lowered the threshold for disabling the highest risk actors based on behavioral signals (i.e. very high MaCSA score) on Facebook from 8 to 6 points. In our initial enforcement, we disabled ~62,000 accounts. We expect to disable ~1,300 accounts daily.

Child Safety Taskforce Update

1. [Theme 1] User Reports:

1. Child Sexual Exploitation, Abuse and Nudity Policy, Guidelines and UDT Updates Launched:

1. Topline Policy Updates

1. Consistently clarified references to real and non-real children across the CSEAN policy and introduced a definition for non-real children to ensure reviewers take action on content involving fictional characters, dolls, AI-generated content, and art where the CSEAN policies apply (KQ A2)
2. Overhauled the solicitation KQ to address enforcement issues and added examples of new trends (KQ C1)
3. Added clarifications to the definitions for near nude imagery of children and sexually suggestive poses (KQ C2 & C4)
4. Refined our definition of sexualized imagery of children under explicit sexualization (KQ D1) and introduced a definition of commonly sexualized body parts (KQ D2)
5. Clarified and removed redundant KQs related to Implicit Sexualization of Children (KQ D3)

1. Key Operational Updates

1. Age Determination: Added guidelines for “definite” vs “age-ambiguous” references to children and provided examples of how different age signals can be used to identify violations
2. Solicitation: Added new “Delete Solicitation” guidelines and relevant decision tree labels to action solicitation content that does not meet escalation thresholds
3. Instagram Profile Review: Updated the review flow to align with the three strike disable threshold for Child Sexual Exploitation violations (as it was set to 13) and added labels for Solicitation and CSE Interest to the decision tree to ensure reviewers are able to take action on these violations
4. Complex Objects Review: Added Solicitation label to the decision tree to help reviewers better identify Solicitation violations (reviewers previously used ‘CSAM’ label to escalate)

2. [Theme 2] Account Enforcement:

1. We lowered the threshold for disabling the highest risk actors based on behavioral signals [REDACTED] from Facebook from [REDACTED], aligning enforcement thresholds with those we use on Instagram. In our initial enforcement, we disabled ~62K accounts. We expect to disable ~1,300 accounts daily.

3. [Theme 3] Recs and Discovery:

1. We’ve made significant progress in consolidating the keywords lists used to regulate hashtags, typeahead, search terms and search interventions on Facebook and Instagram. The team developed a clustering system to review more than 80K unique search terms to ensure high enforcement accuracy, while minimizing strain on operations and policy resources. We are now working on the plan to maintain this list on a regular basis - this will require input from multiple teams including safety policy and i3.

4. [Theme 5] Future Proofing & Auditability:

1. As part of this theme, we committed to executing network disruptions and manual risk investigations. In June 2023 the Stanford Internet Observatory (SIO) escalated to Meta 168,575 FB groups suspected to be engaging in Child Sexual Exploitation (CSE) activity.
 1. i3 Child Safety reviewed a sample (374) of these FB groups and identified:
 1. 30% (111 FB groups) were confirmed as True Positives (i.e CS violating groups) which we define as FB groups created for the purposes of CSE

or FB groups containing a high volume of CS violating content such that it meets the threshold for disabling in line with CS policy. Groups identified as CS violating were deleted except for those identified as candidates for future i3 SNDs.

2. 70% (263 FB groups) were confirmed as False Positives (i.e non-CS violating)
 1. 44% (115 FB groups) of the False Positives were ANSA, ASE and SSPX violating.
 1. 56% (148 FB groups) of the False Positives were non-violating groups.
1. Next steps:
 1. From the sampled data i3 identified 17 groups that were created for the purposes of CSAM solicitation and child sexualization which will form the basis of a future Strategic Network Disruption (SND).
 2. i3 created two other prioritized sampling groups (1,158 Groups) which were enqueued to CSAM Escalation queues. We'll follow up with a breakdown of the enforcement actions.

From: Arcadiy [REDACTED] <[REDACTED]@meta.com>

Date: Friday, July 28, 2023 at 22:20

To: Nick Clegg <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>, Scott Tucker <[REDACTED]@meta.com>, Victoria [REDACTED] <[REDACTED]@meta.com>, Frank [REDACTED] <[REDACTED]@meta.com>, Henry [REDACTED] <[REDACTED]@meta.com>

Cc: Amit Bhattacharyya <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, Markus [REDACTED] <[REDACTED]@meta.com>, Jonathan [REDACTED] <[REDACTED]@meta.com>, Carl [REDACTED] <[REDACTED]@meta.com>, Nicky [REDACTED] <[REDACTED]@meta.com>, Scott [REDACTED] <[REDACTED]@meta.com>, Victoria [REDACTED] <[REDACTED]@meta.com>, Mary [REDACTED] <[REDACTED]@meta.com>, Riana [REDACTED] <[REDACTED]@meta.com>, Joseph [REDACTED] <[REDACTED]@meta.com>, Stephanie [REDACTED] <[REDACTED]@meta.com>

Subject: [a/c priv] Child safety task force update - July 28, 2023

Hi folks,

The Comms team has been in regular contact with Jeff Horwitz, and there is still no clarity on when he plans to run the next stories in his series. Our current understanding is:

1. The Project Gamma story is on hold. He didn't feel like he had the sourcing to run it.
2. Jeff seems to be closest to publishing a piece on device level verification. As a reminder, this story will focus on how Apple and Google have managed to stay quiet as support for device level

verification has grown in the industry. Antigone spoke to him this week about how device level verification is a good solve for issues apps face with underage enforcement.

3. He continues to flag Facebook Groups which we're reviewing as he shares, but we don't have any new details on the angle of the story. It will be tied to Stanford's report on FB Groups and pedophilic networks.
4. Jeff is also working on a story about our move to end-to-end encryption.

This week, the Child Safety Task Force completed launching multiple interventions intended to reduce the ability of risky actors to discover other risky actors across our platforms, filtering discoverability surfaces based on MaCSA scores:

1. Recommendation filtering for AYM, Reels, Explore, Feed surfaces: Experiment results demonstrated significant reduction on MaCSA <> MaCSA engagement **-11.50% follows, -6.68% threads, -11.85% comments, -32.85% likes**. We observed regressions on guardrail metrics which were below materiality threshold.
2. Search filtering for typeahead, and showing in search results user only after exact name search. Experiment results showed reduction on MaCSA <> MaCSA engagement **-2.07% threads, -0.9% (NSS) comments, -0.39% (NSS) follows**. We observed regressions on search metrics that we reviewed with the Disco Safety team and we will run a backtest to monitor them.
3. Launched FB Group, Page and Event recommendation filters based on high percentage of MaCSA scores among admins or members. Over last week we completed the launch to 100% of the recommendation filters for groups or pages with indicators of child safety risks. (brief, launch post)

A fulsome update on the Child Safety Taskforce work is below.

Child Safety Taskforce Update

1. [Theme 1] User Reports:
 1. Impact of "lightweight" reporting experience: In January, we started to experiment with a new way to report Facebook content that does not require users to specify why they are reporting that content; instead, we automatically apply tags using ML. Between 21 Apr and 21 May, test users (lightweight experience) submitted 24% more reports compared to control, and the number of CSE violating reports increased by 90%. For reports with [REDACTED], test users submitted 114% more reports compared to control, but 43% fewer reports were found to be violating for CSE. [REDACTED]
[REDACTED] It's clear that the lightweight reporting flow enables more reports to be submitted, and more CSE violating reports to be found. However this implies [REDACTED]
[REDACTED] We're discussing next steps.
 1. Tooling, guidelines, training: We continue to address technical and Guided Review Tree (GRT) fixes. Two trainings were delivered this week, and we got approval to make an Emergency Policy update on Child Sexual Exploitation, Abuse and Nudity Policy & Protocol Clarifications (T158555096). This is set to launch internally and externally via CS update on Aug. 2.
2. [Theme 2] Account Enforcement:

1. Performance of CSE Violating IG Profile Classifier: Since the launch on June 27, the CSE-violating IG profile classifier detected and disabled 1704 unique IG Profiles for Child Safety Violations and demonstrated a 14.5% action rate based on human review.
1. Lessons Learned
 1. Reviewers are incorrectly assessing user profile age - Child Safety Ops is collaborating with Core Policy to provide more robust guidance on age assessment within Policy / Ops Guidelines in the update scheduled for August 2.
 1. Reviewers are over-enforcing for CSAM Solicitation due to false assumptions of nature of content being shared via external links, reviewers tend to err that content may be sexualized content. CS Ops to re-calibrate CSAM Solicitation KQs with review teams.
1. Some profiles enqueued by the classifier are accurately being assessed as violating our Adult Sexual Exploitation, Sexual Solicitation and Prostitution or Adult Nudity & Sexual Activity policies, but current tooling does not allow these profiles to be taken down. We are exploring improvements to ensure we're capturing Adult Sexual Exploitation / Sexual Solicitation & Prostitution violations.
3. [Theme 3] Recs and Discovery:
 1. IG MaCSA <> MaCSA discovery intervention launched across multiple surfaces. See details above.
4. [Theme 5] Future Proofing & Auditability:
 1. We are making progress on multiple fronts. Most notably:
 1. Legal [REDACTED] PRIV
 1. [REDACTED] PRIV
 1. We have developed and are finalizing a runbook for learning from network disruptions / investigations and incorporating them into existing mitigations. We're also evaluating goals on conducting and publicizing additional strategic network disruptions as a component of our external narrative around child safety.

From: Arcadiy [REDACTED] <[REDACTED]@meta.com>

Date: Friday, July 21, 2023 at 21:03

To: Nick Clegg <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>, Scott Tucker <[REDACTED]@meta.com>, Victoria [REDACTED] <[REDACTED]@meta.com>, Frank [REDACTED] <[REDACTED]@meta.com>, Henry [REDACTED] <[REDACTED]@meta.com>

Cc: Amit Bhattacharyya <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, Markus [REDACTED] <[REDACTED]@meta.com>, Jonathan [REDACTED] <[REDACTED]@meta.com>, Carl [REDACTED] <[REDACTED]@meta.com>, Nicky [REDACTED] <[REDACTED]@meta.com>, Scott [REDACTED] <[REDACTED]@meta.com>, Victoria [REDACTED] <[REDACTED]@meta.com>, Mary [REDACTED] <[REDACTED]@meta.com>, Riana [REDACTED]

<[REDACTED]@meta.com>, Joseph [REDACTED] <[REDACTED]@meta.com>, Stephanie [REDACTED]
<[REDACTED]@meta.com>

Subject: [a/c priv] Child safety task force - mid-task force check-in (July 21, 2023)

Hi all,

As of last Friday, we are halfway through the Child Safety Task Force.

As you know, we have been prioritizing four key areas, three of which map to key allegations in the Wall Street Journal article. The fourth is focused on long term sustainability: ensuring that the issues we address stay addressed. For each of these areas, we'll share a brief summary of the improvements we've made, expected progress for the rest of the task force and any key insights we've gleaned from our understand work.

Our goal remains to wind down the task force by Friday, August 25, and we are on track to do so. This aligns with the timing to deliver an update to the European Commission before our DSA effective deadline.

Looking beyond the immediate task force, we've identified several important questions about our long term efforts in this space we'll want to evaluate in the coming weeks. We will share a quick preview of some of these questions at the end of this summary.

Please let us know if you have any questions and thank you for all the help and support during the first half of the task force!

1. Some user reports have not been accurately handled by Meta (theme 1)

The key concern had to do with reports of potential child safety issues not properly being addressed by Meta. By the end of the task force we plan to improve user report ranking and rep guidelines, align user report messaging and have metrics to measure this over time. **However, we anticipate that our user report recall will continue to be challenging at the conclusion of the task force (see future considerations section).**

1. Launches and wins

1. [REDACTED]
[REDACTED]
2. Shipped updates to improve report ranking for Child Safety by incorporating additional classifier signals into prioritization algorithms. The percentage of jobs reviewed that were found to violate for Child Sexual Exploitation increased by 67% on FB and 166% on IG.

2. Still to come

1. **Early August:** Roll out to reps policy and protocol updates to address known gaps and provide clear examples with coded language, emojis, etc., in order to improve human review recall.
2. **Early August:** Launch aligned responses to users who report problematic content across Facebook and Instagram.

3. **End of July (for content)/end of August (for profiles):** Establish a baseline for our newly-aligned user report recall north star metric (Recall over all Reports or "Roar"), which will allow us to monitor for issues with reports going forward.
4. **Q4:** Roll out additional improvements to user report ranking to better identify Child Safety jobs and ensure they're appropriately prioritized (see insights below).
5. **Q4:** Broader human review improvement to escalation pathways and review tooling in content and profile reviews (improving Roar and efficiency).

2. Some bad actors were not being enforced on (theme 2)

The allegations suggested that bad actors were not being appropriately enforced on (including through more severe interventions like disabling linked accounts and devices), and that we weren't successfully taking any action on certain types of problematic actors, enabling risky adults to connect with one another on our platforms. By the end of the task force, we expect to have significantly expanded the mechanisms by which we disable violating accounts (and other complex entities) on our platform.

1. Launches and wins

1. Completed an enforcement audit to identify technical issues impacting account disables being executed. Following the audit, found that >98% of our end to end account level enforcement works correctly, with <2% of errors and associated logging or testing gaps that make it challenging to validate the remaining enforcement scenarios
2. Launched multiple new mechanisms for disabling Instagram accounts, including [REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
3. Improved logging to better understand the volume and reliability of account actions like device disables, which we previously had little clarity on.

2. Still to come

1. **End of July:** Complete shipping updates to disable additional linked accounts and increase the coverage of device disables for Child Safety.
2. **End of July/early August:** Extend the text matching bank approach to disable Facebook groups and profiles with violating subject line elements.

3. Recommendations, search and hashtags contained terms and results related to alleged exchange of CSAM off platform (theme 3)

The key allegations were that it was easy to find problematic, child safety-related content on our platforms through search and other discovery mechanisms, and that once you did, recommendations would help surface other similar contents and accounts. By the end of the task force we will restrict potentially risky accounts from being able to engage with one another, reduce child safety related content on recommendations surfaces, and strengthen & consolidate existing filters for search, hashtags etc. to be based on consolidated and maintained lists.

1. Launches and wins
 1. Assessed whether the content identified in the escalation is covered by existing policies and concluded that we do not need to create a new non-recommendable Child Safety policy; rather, we just need to better enforce already existing policies.
 2. Audited the full set of recommendations surfaces & scenarios to identify gaps around Child Safety, enabling us to define the places we need to ensure enforcement exists.
 3. Aligned on new recommendation and search filtering policies for bad actor<>minor and bad actor<>bad actor.
2. Still to come
 1. **End of July:** Implement recommendation filters for groups and pages on Facebook based on Child Safety signals.
 2. **End of July:** Implement additional keyword-based interventions for search, discovery and hashtag surfaces across FB and IG, leveraging a newly centralised and governed keyword solution
 3. **End of July/early August:** Implement recommendation filters for key account and content discovery/recommendation surfaces (including restrictions for bad actor:bad actor discoverability).
 4. **End of August:** Turn off temporary, overly-expansive hashtag/search restrictions and transition to a consolidated, well-maintained list of risky terms.

4. Process improvements: future-proofing and auditability (theme 5)

While it's important to fix the immediate issues, we also identified gaps in our ability to ensure these issues stay fixed over time as our systems change and as the problem space evolves.

PRIV

PRIV

1. Launches and wins
 1. Completed a strategic network disruption that took down over 14,000 Instagram accounts and identified 46,000 accounts suitable for custom messaging. Insights identified as part of this disruption are being used as an input into other themes (e.g. reviewer training, keyword lists).
 2. Created requirements for the (parallel) end-to-end testing task force to ensure we can monitor key scenarios on an ongoing basis.
2. Still to come
 1. **End of August:** Create an update for the European Commission on the progress made as part of the task force.
 2. **End of August:** Finalize processes/runbooks for intake of insights and intelligence from network disruptions into scaled enforcement.
 3. **End of August:** Operational second line of defense coverage for the Child Safety space

Key insights

1. **Rep recall measured:** Analysis of a golden set of known-violating Child Sexual Exploitation profiles/content suggested that reps have ~22% recall on Instagram profiles and ~42% recall on Instagram content, meaning of all the known violating content reviewed by reps as part of this golden set, they would have taken down only ~1 in 5 profiles and ~2 in 5 pieces of content. We are hopeful that our protocol, training and review improvements will move the needle on these metrics.
2. **Significant opportunities to improve review flows identified in shadowing:** Task force members visited multiple review sites to shadow reps who are reviewing content and profiles for Child Safety and identified a number of issues, from flows missing important context to confusion about the purpose of certain review queues. We're following up on ownership of these issues and working to address them.
3. **It's difficult to test user report ranking:** A significant challenge we face in improving our ranking approach is that it's hard to measure the effect of these changes on known Child Safety jobs. We are coordinating with the end-to-end testing task force to create an approach to measure and validate the improvements we're trying to make.
4. **Discovery and Recommendation surfaces are highly fragmented:** Ownership of solutions and measurement is distributed broadly across teams, and approaches/tools vary between Accounts and Content as well as Recommendations and Search and between Facebook and Instagram. The Task Force has had to work across teams, solutions and metrics to mitigate the same holistic risk relating to unconnected user discovery. Though advancements have been made, it will be challenging to maintain unified policies and mitigations beyond the task force while this organizational and tooling fragmentation persists.
5. **Device blocking has limited effectiveness in keeping bad actors off the platform:** [REDACTED]

[REDACTED]

Device Blocking is highly nuanced, inconsistently applied and has limited logging, so understanding to date has been limited and increased understanding is driving improvements we're making.

Future considerations

One of the most important takeaways we have from the task force so far is that while we'll significantly mitigate risks in the Child Safety space, we won't be able to fully mitigate all harms, preempt questions from regulators or prevent "the next article" from being written: the outcomes of the task force will allow us to tell a much better story, but there may be more we need to consider to keep reducing harm and build defensibility in this space. We are not yet ready to present options around these questions for discussion, but these are some of the things we're starting to think about and expect to raise for further discussion by the conclusion of the task force.

1. **Profile-level enforcement lags content and message-thread based enforcement**, with unclear escalation paths from reactive user reporting on FB and multiple challenges identified with the IG scaled enforcement reviews. Profile-level enforcement cannot easily support appeals, precision measurement, or the requirements for a profile-level RoaR metric. Beyond the taskforce, we expect to be limited in our ability to measure and improve Profile-level RoaR for Child Safety and other VTs without further investment in FB and IG profile enforcement—work that is currently unowned by any product team. What is the right support/ownership model for these efforts?

2.

PRIV

3. While this was not an explicit focus of the task force, regulators and activists continue to raise questions around **ways for us to quantify the risks to children on our platform**. Do we need to attempt to create an externally-explainable metric that reflects child to bad actor interactions ("how likely is my child to interact with a potentially problematic person on Meta platforms?")
4. We don't have a great way to answer the question of how easy it is for bad actors to find one another on our platform, and while we're both shipping interventions and exploring metrics, **we expect motivated users to continue to be able to find each other**. How do we know we're doing "well enough" on bad actor <-> bad actor discovery/engagement and is there an external narrative we can build around this?

From: Arcadiy [REDACTED] <[REDACTED]@meta.com>

Date: Saturday, July 15, 2023 at 00:11

To: Nick Clegg <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>, Scott Tucker <[REDACTED]@meta.com>, Victoria [REDACTED] <[REDACTED]@meta.com>, Frank [REDACTED] <[REDACTED]@meta.com>, Henry [REDACTED] <[REDACTED]@meta.com>

Cc: Amit Bhattacharyya <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, Markus [REDACTED] <[REDACTED]@meta.com>, Jonathan [REDACTED] <[REDACTED]@meta.com>, Carl [REDACTED] <[REDACTED]@meta.com>, Nicky [REDACTED] <[REDACTED]@meta.com>, Scott [REDACTED] <[REDACTED]@meta.com>, Victoria [REDACTED] <[REDACTED]@meta.com>, Mary [REDACTED] <[REDACTED]@meta.com>, Riana [REDACTED] <[REDACTED]@meta.com>, Joseph [REDACTED] <[REDACTED]@meta.com>, Stephanie [REDACTED] <[REDACTED]@meta.com>

Subject: [a/c priv] Child safety task force update - July 14, 2023

Folks,

As of today, we are halfway through the planned timeline for the Child Safety Task Force. By the end of next week, we will share a mid-point check in, including a summary of changes to our systems to date, work planned for the next six weeks and early thoughts on post-task force efforts.

Before we jump into an update on the task force progress, the comms team spoke with Jeff Horwitz today, and our current understanding is:

1. Jeff will write a short story about Project Gamma first. He accepts we hadn't planned on launching something with this project - but that we were in very early explorations. He pressed on whether we would continue to explore such an offering in light of his story on pedophilic networks on IG, and how we're assessing these decisions in light of the Task Force's findings. He wants to understand why the Gamma project came to a close, and where we stand now on a u13 offering on IG. Comms is going to develop a recommendation on how to respond - our deadline is mid-next week.

1. We launched Text Matching Service (TMS) banks aiming to proactively detect usage of inherently violating Child Safety keywords and phrases within IG Subject Elements (Name, Username, Bio), covering child sexual solicitation, child sexualization and uncategorized child sexual exploitation. The intent was to increase enforcement against profiles containing this kind of violating text. Within the first 48 hours of the launch, we enqueued ~5.5K jobs for IG profile review queue, of which 18.7% of the jobs were deemed to be violating per Child Safety policy. We enrolled ~1.3K IG profiles to IG UFAC for high severity violation types.
1. [Theme 3] Recs and Discovery:
 1. We are working to directly tackle the volume of Child Safety non-rec exposures, and have updated the exit criteria for this theme to reflect a goal to reduce non-rec account/content exposure by 50% for each of content and profiles. This comes as a result of scoping the current state of said exposures across content and accounts, based on the Taxonomy agreed within the Theme. We believe that [REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED] surface on FB is PYMK (>95%) while on IG it is Search (58%)
1. [Theme 5] Future Proofing & Auditability:
 1. We are on track with our second line of defense work and are coordinating independent testing of the prioritized set of 178 controls (Child Safety and Human Safety).

From: Arcadiy [REDACTED] <[REDACTED]@meta.com>

Date: Friday, July 7, 2023 at 23:24

To: Nick Clegg <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>, Scott Tucker <[REDACTED]@meta.com>, Victoria [REDACTED] <[REDACTED]@meta.com>, Frank [REDACTED] <[REDACTED]@meta.com>

Cc: Amit Bhattacharyya <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, Markus [REDACTED] <[REDACTED]@meta.com>, Jonathan [REDACTED] <[REDACTED]@meta.com>, Carl [REDACTED] <[REDACTED]@meta.com>, Nicky [REDACTED] <[REDACTED]@meta.com>, Scott [REDACTED] <[REDACTED]@meta.com>, Victoria [REDACTED] <[REDACTED]@meta.com>, Mary [REDACTED] <[REDACTED]@meta.com>, Riana [REDACTED] <[REDACTED]@meta.com>

Subject: [a/c priv] Child safety task force update - July 7, 2023

All,

As shared earlier in the week, Jeff Horwitz confirmed that the next stories in his series focused on youth safety will be on Project Gamma. The earliest he'll write about Project Gamma is next week, but we believe it's possible that timeline slips.

On the Project Gamma story, comms pushed back on his allegation that we're unpausing IG Kids and clarified the nature of the project as a competitive and industry analysis. Horwitz does not dispute this, but believes it merits writing about because the fact we gave it a code name means we're significantly invested.

Horwitz acknowledges there's asymmetry in the market given that YouTube and TikTok have u13 products, and he intends to spend some time digging into their offerings.

We know that Horwitz reached out to Larry Magid from Connect Safely. Larry plans to tell Horwitz that he believes a well-designed, safe u13 product is a good thing to train young people on social media before they're "unleashed in the wilds."

This change in the series has not impacted the work of the Task Force. This week teams were focused on executing, and we anticipate a series of impactful launches next week. An overview of the work done this week is below.

Thank you,
Arcadiy, Riana, and the TF

Workstreams update

1. Proactive Detection/Sweeps
 1. [GRO Targeted Searches & Enforcement] GRO-W Regional teams removed another ~500 complex objects for child safety policy violations this week, taking the total number of proactively detected and removed assets to ~5.5k. GRO-W FTCs have also now reviewed ~4.5k complex objects, and have removed ~1k for violating child safety policies. Action rates for complex objects have dropped significantly across all markets currently being supported.
2. Child Safety Taskforce
 1. [Theme 1] User Reports:
 1. The team has drafted Policy updates (Implementation Standards/Known Questions) as well as Operational Guidelines and submitted them for XFN review. We are on track to send these updates to Policy leadership by July 17, and we will schedule a launch date and reviewer training shortly after that. The goal is to have these changes fully implemented by mid-August.
 1. This team has also added three exit criteria:
 1. Queue Management & Ranking: pIV for CS jobs is sufficiently high (meets high pri DCI cutoff) and maps to IV in a way that is extensible to IGPR + Complex Entities.
 2. Review Quality: Reviewers are easily able to Escalate CSAM/Solicitation jobs across all surfaces (coverage gaps across IGPR, Groups, and Pages are filled), and successfully escalate >0 jobs across each to demonstrate it's working as expected.
 3. User Messaging for Reporting & Takedowns: FB and IG have user experience parity across situations (Human Review, ML Review, Content is Taken Down), including messaging for User Removed the Post and Report Appeals.

1. [Theme 2] Account Enforcement:
 1. We're already seeing broader impact than we anticipated of the CSE Violating Profile classifier that launched last week. The Discovery Safety XFN said they've saved efforts because of this work.
 1. We will be shipping updates to [REDACTED].
 1. The team is preparing to launch a [REDACTED]. After this launch, the team will work towards launching the TMS bank for FB Groups and Profiles.
1. [Theme 3] Recs and Discovery:
 1. The team is making progress on how we're going to consolidate keyword lists and maintain them regularly. This includes a plan to [REDACTED] while ensuring we'll continue to capture violating content.
1. [Theme 5] Future Proofing & Auditability:

1.	PRIV	legal	PRIV
PRIV			

From: Arcadiy [REDACTED] <[REDACTED]@meta.com>

Date: Friday, June 30, 2023 at 23:54

To: Nick Clegg <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>, Scott Tucker <[REDACTED]@meta.com>, Victoria [REDACTED] <[REDACTED]@meta.com>, Frank [REDACTED] <[REDACTED]@meta.com>

Cc: Amit Bhattacharyya <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, Markus [REDACTED] <[REDACTED]@meta.com>, Jonathan [REDACTED] <[REDACTED]@meta.com>, Carl [REDACTED] <[REDACTED]@meta.com>, Nicky [REDACTED] <[REDACTED]@meta.com>, Scott [REDACTED] <[REDACTED]@meta.com>, Victoria [REDACTED] <[REDACTED]@meta.com>, Mary [REDACTED] <[REDACTED]@meta.com>, Riana [REDACTED] <ria[REDACTED]i@meta.com>

Subject: [a/c priv] Child safety task force update - June 30, 2023

Folks,

The task force made significant progress this week. We closed out most of the 'Understand' work, and as a result of the audits, there are only a limited number of additional exit criteria. This means we

confirmed our understanding of the issues at the start and are more informed on how to resolve them. We've begun executing and have started to see the impact of the work. Our next major milestone is the midpoint of the task force: July 14.

Highlights include:

1. [Theme 2] Account Enforcement: [REDACTED]
[REDACTED]
[REDACTED] Since the launch on June 27, we have actioned ~100 profiles from SRT. We'll continue to explore ways to improve the recall and precision of the model.
2. [Theme 3] Recs & Discovery: We explored whether we might need a new Borderline Child Safety policy in order to capture the kinds of content and entities we saw in the current escalation. Policy, product and ops reviewed a meaningful sample of content and concluded we **do not** need a new policy as we expect the content to be covered by existing policies and our primary issues are enforcement gaps on recommendation surfaces. This means our in-progress efforts to improve enforcement (both human and automated) may be able to mitigate many of the risks and we do not need to embark on the long process of creating new policy, protocol and enforcement.
3. [Theme 5] i3 conducted two Strategic Network Disruptions (SNDs) focused on FB Groups - Op Caldo de Pollo 2.0 and Op Journal.
 1. Op Caldo de Pollo 2.0 resulted in 2,053 FB accounts getting disabled and 34,722 FB accounts receiving custom deterrence messaging. This messaging is sent to users that are considered passive observers. If passive observers are in any subsequent investigations, they are then disabled.
 1. Op Journal resulted in 6,160 FB accounts getting disabled and 47,208 accounts receiving custom messaging.
 1. i3 also completed their review of the groups surfaced with admins/users with high MaCSA scores. 222 of 237 groups reviewed were adult porn related, scam related, or benign groups that were not violating from a child safety POV. 12 were found to be violating (posted some content related to CSAM trading) and 3 were created for the sole purpose of trading CSAM. All 15 of these groups were enforced on.

A detailed look at all the workstreams is below. Please let us know if you have any questions!

Arcadiy, Riana and the rest of the CSTF

External

1. Policymaker outreach
 1. US: We received a bipartisan letter yesterday from members of the Senate Judiciary. It has received attention from the WSJ and Punchbowl, thus far. Comms is referring press who inquire back to what was shared with the Journal for the underlying story.

Workstreams update

1. Proactive Detection/Sweeps

1. [Regions Targeted Searches] Total number of complex objects removed by Regions FTEs (LATAM, Brazil, NORAM, MENA, India, and APAC) = 5069. GRO-W FTC enforcement efforts have lead to the removal of an additional c.700 complex objects.
2. [Paid Risk Scoping] A GRO XFN investigated the risk associated with inorganic/paid entities relating to the distribution of CSAM material. Investigations focussing on ads linked to assets removed in response to the WSJ enquiry, and a broader look at paid CSAM risk potential, both resulted in extremely low risk/low prevalence of this type of CSAM distribution / solicitation content. GRO will continue due-diligence detection exercises, and is working to ensure discoverability interventions are also applied on Commerce surfaces where relevant
3. [Discoverability] GRO is partnering with DI-Ops and other Discoverability & Search stakeholders to support an audit of TRRMs launched in response to the WSJ enquiry. We're also engaging teams with D&S expertise to support additional scoping and mitigation of risks in languages other than English.

2. Child Safety Taskforce

1. [Theme 1] User Reports:

1. The team audited the IG Child Safety reporting flow for organic content, so we know exactly what messaging is delivered upon reporting content and profiles on IG and after the content has been assessed by our systems. The biggest takeaway is that we have an opportunity to align messaging across surfaces.

2. [Theme 2] Account Enforcement:

1. [REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED] More findings [here](#).

3. [Theme 3] Recs and Discovery:

1. The team did an audit of Recommendation and Discovery surfaces and found a number of issues that will be addressed in the near-term. The biggest opportunity is to consolidate keyword lists - we currently use >7 different lists of terms to trigger different interventions. This fragmentation of keyword lists across apps makes it difficult to maintain a consistent or robust keyword-based filtering approach.
2. Policy with product and ops evaluated the need for a new Borderline Child Safety Policy to capture content not covered by existing components of the Community Standards or our other Borderline Policies in order to make this content Non-Rec (No); and whether we need to update our entity Non-Rec Policy to capture "borderline" child safety entities as Non-Rec (No). After evaluating samples of "borderline" content, we determined that this content should fall within our existing Community Standards Child Sexualization policy.

Core Policy will instead update Known Questions and OGs to ensure this policy is understood by reviewers.

4. [Theme 5] Future Proofing & Auditability:

1. As part of this theme, we are tracking the need to provide an update on the task force to Commissioner Thierry Breton and the DSA European Commission before the end of August.
2. i3 conducted two Strategic Network Disruptions (SNDs) focused on FB Groups - Op Caldo de Pollo 2.0 and Op Journal.
3. i3 completed their review of the groups surfaced with admins/users with high MaCSA scores. 222 of 237 groups reviewed were adult porn related, scam related, or benign groups that were not violating from a child safety POV. 12 were found to be violating (posted some content related to CSAM trading) and 3 were created for the sole purpose of trading CSAM. All 15 of these groups were enforced on.

From: Riana [REDACTED] <[REDACTED]@meta.com>

Date: Tuesday, June 27, 2023 at 12:17

To: Nick Clegg <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Arcadiy [REDACTED] <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>, Scott Tucker <[REDACTED]@meta.com>

Cc: Amit Bhattacharyya <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Abby [REDACTED] <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, MC [REDACTED] <[REDACTED]@meta.com>, Markus [REDACTED] <[REDACTED]@meta.com>, Frank [REDACTED] <[REDACTED]@meta.com>, Mary [REDACTED] <[REDACTED]@meta.com>

Subject: Re: [a/c priv] Child safety task force update - June 23, 2023

Hi Nick,

Yes, we're tracking providing the DSA EC team with more information about the task force before the end of August.

Thank you,
Riana

From: Nick Clegg <[REDACTED]@meta.com>

Date: Tuesday, June 27, 2023 at 1:08 PM

To: Jennifer Newstead <[REDACTED]@meta.com>, Arcadiy [REDACTED] <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>, Scott Tucker <[REDACTED]@meta.com>
Cc: Amit Bhattacharyya <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, Riana [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Abby [REDACTED] <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, MC [REDACTED] <[REDACTED]@meta.com>, Markus [REDACTED] <[REDACTED]@meta.com>
Subject: Re: [a/c priv] Child safety task force update - June 23, 2023

Regardless of where we net out on our public facing statements on this work, Mark has committed to share the outcome of our work with Breton and his team privately over the Summer. Assume that is being incorporated into our work here?

Thx

N

From: Jennifer Newstead <[REDACTED]@meta.com>
Date: Tuesday, June 27, 2023 at 7:50 AM
To: Arcadiy [REDACTED] <[REDACTED]@meta.com>, Nick Clegg <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>, Scott Tucker <[REDACTED]@meta.com>
Cc: Amit Bhattacharyya <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, Riana [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Abby [REDACTED] <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, MC [REDACTED] <[REDACTED]@meta.com>, Markus [REDACTED] <[REDACTED]@meta.com>
Subject: Re: [a/c priv] Child safety task force update - June 23, 2023

PRIVILEGED

Hi everyone.

PRIV

Jen

From: Arcadiy [REDACTED] <[REDACTED]@meta.com>
Sent: Monday, June 26, 2023 1:32 PM
To: Nick Clegg <[REDACTED]@meta.com>; Joel Kaplan <[REDACTED]@meta.com>; David Ginsberg <[REDACTED]@meta.com>; Jennifer Newstead <[REDACTED]@meta.com>; Monika Bickert <[REDACTED]@meta.com>; Guy Rosen <[REDACTED]@meta.com>; Tom Alison <[REDACTED]@meta.com>; Adam Mosseri <[REDACTED]@meta.com>; Kevin Martin <[REDACTED]@meta.com>; Nicola Mendelsohn <[REDACTED]@meta.com>
Cc: Amit Bhattacharyya <[REDACTED]@meta.com>; Antigone Davis <[REDACTED]@meta.com>; Alison [REDACTED] <[REDACTED]@meta.com>; Chris Sonderby <[REDACTED]@meta.com>; James [REDACTED] <[REDACTED]@meta.com>; Joey Simhon <[REDACTED]@meta.com>; Molly Cutler <[REDACTED]@meta.com>; Anna [REDACTED] <[REDACTED]@meta.com>; Riana [REDACTED] <[REDACTED]@meta.com>; Sayed Otaru <[REDACTED]@meta.com>; Liz [REDACTED] <[REDACTED]@meta.com>; Abby [REDACTED] <[REDACTED]@meta.com>; Andrea Saul <[REDACTED]@meta.com>; Andy Stone <[REDACTED]@meta.com>; Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>; Kaitlin [REDACTED] <[REDACTED]@meta.com>; Krish [REDACTED] <[REDACTED]@meta.com>; Lena Pietsch <[REDACTED]@meta.com>; Gabriel [REDACTED] <[REDACTED]@meta.com>; MC [REDACTED] <[REDACTED]@meta.com>; Markus [REDACTED] <[REDACTED]@meta.com>
Subject: Re: [a/c priv] Child safety task force update - June 23, 2023

Hi Nick,

Thanks for your question.

Yes, we're expecting to be able to publish an update on the task force at its conclusion. The timeline for the task force aligns well with the DSA effectiveness date (8/25). We are planning to work with comms, legal and policy on a public facing update.

We're also exploring an opportunity for an interim check-in to be shared with a TBD audience (e.g. advertisers who have explicitly reached out in response to the WSJ escalation) around late July.

However, it's important to note that we aren't going to be able to mitigate 100% of risks within the task force (or even beyond). We'll need to be able to explain and defend the lines we've drawn and the progress we've made, and to lay out the future work we plan to do in this space.

Thanks!

Arcadiy

From: Nick Clegg <[REDACTED]@meta.com>
Date: Monday, June 26, 2023 at 11:14
To: Arcadiy [REDACTED] <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>
Cc: Amit Bhattacharyya <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, Riana [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Abby [REDACTED] <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, MC [REDACTED] <[REDACTED]@meta.com>, Markus [REDACTED] <[REDACTED]@meta.com>
Subject: Re: [a/c priv] Child safety task force update - June 23, 2023

+ Markus

Thanks for the update on this important work.

One additional note: in the meeting with Commissioner Breton and Mark on Friday, Breton made it clear: a) that he views this work as falling squarely within the terms of the DSA and; b) that if we are in a position to complete our remedial work here before the DSA is formally active (August) we should set this out publicly.

Are there plans to do b)?

N

From: Arcadiy [REDACTED] <[REDACTED]@meta.com>
Date: Friday, June 23, 2023 at 11:48 PM
To: Nick Clegg <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>
Cc: Amit Bhattacharyya <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, James [REDACTED]

<[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Molly Cutler
<[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, Riana [REDACTED]
<[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Liz [REDACTED]
<[REDACTED]@meta.com>, Abby [REDACTED] <[REDACTED]@meta.com>, Andrea Saul
<[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Ashly Nikkole [REDACTED]
<[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>,
Lena Pietsch <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, MC [REDACTED]
<[REDACTED]@meta.com>

Subject: [a/c priv] Child safety task force update - June 23, 2023

Folks,

Sharing another weekly update on the Child Safety Task Force and related escalation.

We have not received any new information from Wall Street Journal Reporter Jeff Horwitz. External outreach from policymakers and advertisers has slowed, though we continue to work on responses and our approach for future articles in the series.

The Child Safety task force hit its first milestone this week with several teams closing out their 'Understand' exit criteria and publishing their audits internally. Highlights of the work are below, including the [launch](#) of MaCSA (Malicious Child Safety Actor) heuristic-driven disables on Instagram which disabled ~20K users on rollout. Full details of audit readouts are available in the [Workplace group](#).

Please reach out with any questions.

Thank you,
Arcadiy, Riana & the CSTF

Workstreams update

1. Proactive Detection/Sweeps
 1. [GRO-W Targeted Enforcement] GRO-W FTCs have reviewed c.1.2k complex objects, and have removed c.200. GRO-W has also removed c.150 simple objects, and will continue to drive the majority of GRO's steady-state enforcement efforts in [REDACTED]
[REDACTED]
 2. [Regions Targeted Searches] GRO Regions teams are currently supporting proactive detection and enforcement across the [REDACTED]
[REDACTED]
As of 6/23, GRO teams have disabled c.4k complex objects as a result of these proactive detection efforts. This is in addition to the c.6k assets removed by PResc as part of the response efforts.
 3. [Discoverability] GRO is working with DI-Ops to share all the FBIDs of the content removed by GRO. DI-Ops will run an analysis to find related hashtags and search terms that would have caused these assets to appear on discoverability surfaces. We believe this will lead to more impactful search interventions.
1. Child Safety Taskforce
 1. [Theme 1] User Reports:

The team audited the ranking and prioritization of Child Safety-related reports in our systems and found:

1. The overall system is working "by design" in that Child Safety user reports go into our central ranking systems, have standard ranking formulas applied to them, and are routed to the correct queues.

2. However, [REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
3. Also, our predicted "Integrity Value" for Child Safety jobs is lower than expected (even if we accurately predict the violation type), leading to fewer jobs getting prioritized than we would expect. The team is investigating the causes.

The team audited the FB Organic Reporting experience for Child Safety and found:

1. The average time it takes to send an outcome message for child safety and other related reports (like sexual exploitation and pornography) is 4 days. We are exploring ways to reduce that timeline.
2. People think we're ignoring their reports, so we're looking at ways to highlight how we use technology for harm detection and reduction in response to user reports.
3. For Child Safety, all reports should generate an SRT job. We found that around 30% of FB profile reports did not result in job creation. This could be due to deduping or logging issues, but we are continuing to investigate.

This team also conducted moderation site visits and found:

1. There is room to improve the feedback loop between Content Moderation Sites, Markets teams, Child Safety Ops / Policy team to improve guidance regarding emerging trends amongst malicious Child Safety Actors to bypass our current enforcement protocol.
2. There are Child Safety Ops Guidelines and reviewer error gaps. We are developing a plan to provide clearer guidance to reps on CSAM escalation criteria. We will also review and audit operational guidelines and KQs training and trigger a refresher training across Content Moderation teams.

1. [Theme 2] Account Enforcement:

This week, we globally launched MaCSA IG disables. We used the CHILD_SEXUALIZATION violation type and disabled ~20K users on rollout. We expect to disable ~8k users per month. Precision is measured at 94%, after being calibrated via I3 investigator review, with ongoing sampling at low volumes to tune signals. Appeals will not be available immediately; we will work on appeals as one of the H2 deliverables after the taskforce exit criteria.

We've made progress on understanding how our hard-linked propagation efforts impact recidivism. [REDACTED]
[REDACTED]
[REDACTED]

[REDACTED] We are working with policy to understand the device experience once a device is blocked, and then we'll move towards fixing the issues identified during our tests.

1. [Theme 3] Recs and Discovery:

We have made progress on auditing existing Child Safety related hashtag and search enforcement to ensure behavior matches expectations, and we are beginning work on the NonRec policies we want to take across surfaces.

The team aligned on having two lists provisionally 'high' and 'medium' confidence to consolidate all keyword-based interventions. The next step is to propose a solution for

updating the lists on an ongoing basis and get a sample of the keywords to ratify with safety and app leads.

1. [Theme 5] Future Proofing & Auditability: The end to end testing taskforce has been working towards a proposal that will be shared with leadership next week on proposed strategy, goals, success criteria, execution approach, etc.

From: Arcadiy [REDACTED] <[REDACTED]@meta.com>

Date: Friday, June 16, 2023 at 16:54

To: Nick Clegg <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>

Cc: Amit Bhattacharyya <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, Riana [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>

Subject: [a/c priv] Child safety task force update - June 16, 2023

Folks,

Sharing our weekly update on the Child Safety Task Force. With the ongoing efforts to address immediate escalations, we have a bit of overlap on comms, so passing along the latest (daily) overall escalation update below as well.

Within the task force, this week we transitioned from planning to execution, with a particular focus on completing our understand work and audits. We've pasted the full Friday update just below this note, but a few key things to call out from the course of the week:

1. We've analyzed our performance on user reports and concluded that content reports are not very likely to get prioritized, but are likely to get accurately actioned if they actually get worked, while profile reports have the opposite behavior: they're less likely to be actioned correctly, but we work more of them. We'll work on both sides of this issue in the next several weeks.
2. We launched updated models for ranking reports for human review, incorporating them into our overall ranking function (multi-armed bandit). It's a bit early to see clear results from these changes, however.
3. We've aligned with policy on which violation types should result in device blocking, soft link propagations, and feature limits. Those policies are in shadow mode, and we should have data to share early next week.

The task force team is on track to complete understand work/audits by the end of next week. We will share a more detailed read-out of what we've learned at that time.

Full task force update below:

1. [Theme 1] User Reports: We have an initial understanding of child safety signals used in ranking today putting us on track to complete the user reports audit by June 23. We also have site visits scheduled in Bogota and Lisbon next week to get a deeper understanding of what tooling, guidelines, and training need to be adjusted.
2. [Theme 2] Account Enforcement: The team is making good progress on understanding recidivism & hard-linked propagation. We conducted an audit of all propagation across violation types. With policy, we aligned on which violation types should result in device blocking, soft link

propagations, and feature limits. Those policies are in shadow mode, and we should have data to share early next week.

3. [Theme 3] Recs and Discovery: The team put together a proposal for policy with existing Child Safety related hashtag and search enforcement to ensure behavior matches expectations. The proposal also has suggestions on how we keep lists updated on an ongoing basis. The team should have a recommendation on how to move forward by the end of next week.
4. [Theme 5] Future Proofing & Auditability: With the exit criteria fully finalized, we were able to assign owners to each of them.

Thanks,

Arcadiy

From: Riana [REDACTED] <[REDACTED]@meta.com>

Date: Friday, June 16, 2023 at 15:20

To: Guy Rosen <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Amit Bhattacharyya <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Nissa [REDACTED] <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>

Cc: Arcadiy [REDACTED] <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Audrey [REDACTED] <[REDACTED]@meta.com>, Abby [REDACTED] <[REDACTED]@meta.com>, Richard [REDACTED] <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Jimmy [REDACTED] <[REDACTED]@meta.com>, MC [REDACTED] <[REDACTED]@meta.com>, Chris [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>, Ashly Nikkole [REDACTED] <[REDACTED]@meta.com>

Subject: Re: WSJ CSAM Series EOD Update: 6/16 (A/C Priv)

Sorry for the second email – the update below now includes GRO's latest numbers. Please let us know if you have any questions.

From: Riana [REDACTED] <[REDACTED]@meta.com>

Date: Friday, June 16, 2023 at 6:04 PM

To: Guy Rosen <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Amit Bhattacharyya <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Nissa [REDACTED] <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>

Cc: Arcadiy [REDACTED] <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Audrey [REDACTED] <[REDACTED]@meta.com>, Abby [REDACTED] <[REDACTED]@meta.com>, Richard [REDACTED] <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>

<[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Liz [REDACTED]
<[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Jimmy [REDACTED]
<[REDACTED]@meta.com>, MC [REDACTED] <[REDACTED]@meta.com>, Chris [REDACTED]
<[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>,
Alison [REDACTED] <[REDACTED]@meta.com>

Subject: WSJ CSAM Series EOD Update: 6/16 (A/C Priv)

Hi all -

Please find today's update on the WSJ Escalation and the Child Safety Task Force. Please reach out with any questions.

Thank you,
Riana

External

1. Policymaker outreach

US: On June 15, we received a letter from Rep. Buck regarding the WSJ's recent reporting. The letter requests by the end of the month information regarding the child safety task force and the company's remediation efforts.

1. Comms

No updates

1. Advertiser

1. We have closed the loop on all open client questions and most recently responded to Home Depot, Dentsu, and Ralph Lauren.

Antigone had a call with Procter & Gamble this morning, and while this could change, the sales team doesn't believe they'll pause.

Now that advertiser interest has tapered off for this article, the focus now is on preparing the brand approach for the future WSJ articles in the series.

Workstreams update

1. Proactive Detection/Sweeps

1. [GRO-W Targeted Enforcement] Global Operations Learning Organization (GOLO) delivered training to FTC reps today. Workflows go live in English, Spanish, and Portuguese on Monday, June 19th.

2. [Regions Targeted Searches] GRO is engaging additional Regions support, with 10 additional markets / languages to be covered from Monday, June 19th: Iberia (Spain and Portugal), DACH, Filipino, French, Maghreb, Thai, Vietnamese, Indonesian, and Hindi (India).

1. As of 6/16, GRO teams have disabled over 3.6K assets. The majority of removals have taken place on IG, with 1.2K IG users disabled. Teams have also taken down 682 Pages, 959 FB Groups and 685 FB Users across the markets currently participating in proactive detection efforts (North America, LATAM, Brazil and Iraq).

1. Child Safety Taskforce

1. [Theme 1] User Reports: We have an initial understanding of child safety signals used in ranking today putting us on track to complete the user reports audit by June 23. We also have site visits scheduled in Bogota and Lisbon next week to get a deeper understanding of what tooling, guidelines, and training need to be adjusted.

2. [Theme 2] Account Enforcement: The team is making good progress on understanding recidivism & hard-linked propagation. We conducted an audit of all propagation across violation types. With policy, we aligned on which violation types should result in device

blocking, soft link propagations, and feature limits. Those policies are in shadow mode, and we should have data to share early next week.

3. [Theme 3] Recs and Discovery: The team put together a proposal for policy with existing Child Safety related hashtag and search enforcement to ensure behavior matches expectations. The proposal also has suggestions on how we keep lists updated on an ongoing basis. The team should have a recommendation on how to move forward by the end of next week.
4. [Theme 5] Future Proofing & Auditability: With the exit criteria fully finalized, we were able to assign owners to each of them.

From: Riana [REDACTED] <[REDACTED]@meta.com>

Date: Thursday, June 15, 2023 at 8:06 PM

To: Guy Rosen <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Amit Bhattacharyya <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Nissa [REDACTED] <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>

Cc: Arcadiy [REDACTED] <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Audrey [REDACTED] <[REDACTED]@meta.com>, Abby [REDACTED] <[REDACTED]@meta.com>, Richard [REDACTED] <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Jimmy [REDACTED] <[REDACTED]@meta.com>, MC [REDACTED] <[REDACTED]@meta.com>, Chris [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>

Subject: WSJ CSAM Series EOD Update: 6/15 (A/C Priv)

Hi all -

Please find today's update on the WSJ Escalation and the Child Safety Task Force. Please reach out with any questions.

Thank you,
Riana

External

1. Policymaker outreach

US: We did the briefing for bipartisan Energy and Commerce Committee staff for a little over an hour. The briefing team (led by Antigone) did very well explaining the concerns raised in the article, putting them in context, and discussing our existing and task force remediation efforts. Questions were about how long the system was down, why researchers find this stuff and we can't, what sorts of predictive signaling we have, staff size for safety before and after layoffs, whether we had an audit of our child safety practices by an internal or external party (like a law firm) between 2017-2022 or since the article, and exactly how long the system was closing cases before we found the glitch. The staff are going to meet with other parties involved in the article, and we expect more follow up after that.

1. Comms

The House Energy and Commerce Committee's statement was picked up by the [Washington Post](#), which also included part of our statement.

Tomorrow, Alex Heath at The Verge will publish leaked information about Project Salsa.

We'll follow his piece and launch the product. We believe press could mention the WSJ story in their pieces.

1. Advertiser

As of today, all advertisers who paused their campaigns as a result of this article have resumed spending.

Some client calls and messaging continues (with Nike, P&G, Ralph Lauren, Dentsu and Home Depot), but the volume of net-new inbound has tapered off.

Workstreams update

1. Proactive Detection/Sweeps

[GRO-W Targeted Enforcement] Custom SRT Queues have been built. Global Operations Learning Organization (GOLO) is delivering training to FTC reps on Friday, June 16th.

Workflows go live in English, Spanish, and Portuguese on Monday, June 19th.

[Fanouts] Several hundred recidivist entities and hard-linked accounts have been reviewed by PResc. Device blocking has been applied to explicit CSAM violating accounts only, not when other less severe violation types (e.g., child sexualization) are present.

[Regions Targeted Searches] GRO is engaging additional Regions support, with 10 additional markets / languages to be covered from Monday, June 19th: Iberia (Spain and Portugal), DACH, Filipino, French, Maghreb, Thai, Vietnamese, Indonesian, and Hindi (India).

1. As of 6/15, GRO teams have disabled over 3.2K assets. The majority of removals have taken place on IG, with 1.1K IG users disabled. Teams have also taken down 680 Pages, 865 FB Groups and 585 FB Users across the markets currently participating in proactive detection efforts (North America, LATAM, Brazil and Iraq).

[Brand Safety] Escalations workstream aligned with Brand Safety PMMs and are sharing relevant data to enable required monetization interventions.

[TMS Blocklisting] No new updates - we're still targeting activation the week of June 19th, and eng is helping to build the changes we need to launch.

1. Child Safety Taskforce

[Theme 1] User Reports: We discovered a bug in the ranking classifier shipped this week meaning it has not yet impacted queue prioritization. A fix is in progress. We also aligned on an approach for tactical baseline metrics using sampled ground truth data with 'involves a child'.

[Theme 2] Account Enforcement: The proactive disable classifier shows high recall on previously identified samples of self-generating CSAM accounts. The team kicked off a privacy review. MaCSA IG disables UFAC troubleshooting continues, but we are making progress on unblocking precision at 94% via samples at the 5-point threshold for MaCSA.

[Theme 3] Recs and Discovery: The team documented current non-rec implementation per entity type for Child Safety. Given many unknowns on how non-rec is leveraged across surfaces, the audit step may now be shifting into documenting all relevant surfaces and sequencing the non-rec interventions.

[Theme 4] Metrics and Logging: This theme's exit criteria has been redistributed across themes (1) for reactive recall measurement, (2) for systems audit and logging, and (3) for non-rec measurement.

[Theme 5] Future Proofing & Auditability: This team is meeting tomorrow to transition in execution mode now that exit criteria has been finalized.

From: Riana [REDACTED] <[REDACTED]@meta.com>

Date: Wednesday, June 14, 2023 at 8:06 PM

To: Guy Rosen <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Amit Bhattacharyya <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Nissa [REDACTED] <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>

Cc: Arcadiy [REDACTED] <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Audrey [REDACTED] <[REDACTED]@meta.com>, Abby [REDACTED] <[REDACTED]@meta.com>, Richard [REDACTED] <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Jimmy [REDACTED] <[REDACTED]@meta.com>, MC [REDACTED] <[REDACTED]@meta.com>, Chris [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>

Subject: WSJ CSAM Series EOD Update: 6/14 (A/C Priv)

A/C Priv

Hi all -

Below please find today's update on the WSJ Escalation and the Child Safety Task Force. Please reach out with any questions.

Thank you,
Riana

External

1. Policymaker outreach US

1. The House Energy and Commerce Committee Chair Cathy McMorris Rodgers (R-WA) and Committee Ranking Member Frank Pallone, Jr. (D-NJ) issued a joint statement today after last week's WSJ story.
2. We are prepped and ready for the bipartisan briefing for House Energy and Commerce committee staff on Thursday (Antigone will lead).

1. Comms

- a. As of this writing, comms has not received any inquiries on the McMorris Rodgers-Pallone statement. If needed, comms will confirm we're briefing the House Energy and Commerce committee, and off the record will push back on how McMorris Rodgers and Pallone frame the issue by pointing to the Journal's original reporting.

1. Advertiser

Today we shared responses to specific advertiser questions with Mondelez, P&G, Nike, Mailchimp, and Ralph Lauren.

We received just one net new inbound today from Dentsu.

Ralph Lauren continues to be paused. Hondo, Volvo and Toyota have all unpaused, and we expect Ford to unpause within the next few days.

Workstreams update

1. Proactive Detection/Sweeps

[Block List] Technical hurdles have pushed the anticipated launch date back until the week of June 18th. Roadblocks primarily due to automated NCMEC reporting actions not

working as expected. Child Safety Eng and Ops teams are sprinting to analyze and debug errors.

[Targeted Enforcement] As of 6/14, GRO teams have disabled over 2.8K assets. The majority of removals have taken place on IG, with over 1K IG users disabled. Teams have also taken down 530 Pages, 787 FB Groups and 530 FB Users across North America, LATAM, Brazil and MENA.

1. Child Safety Taskforce

[Theme 1] User Reports: On Profiles and Content, we've found that 1) Content jobs are much less likely to be prioritized to human review, but once they are reviewed are more likely to be reviewed correctly and 2) The inverse is true for Profiles - most are reviewed but reviewer error rate is much higher. This suggests that the big levers on prioritization are on Content, but for Profile we need to focus much more on human review decisions, tooling, training etc.

[Theme 2] Account Enforcement: [REDACTED]

[Theme 3] Recs and Discovery: We've created a taxonomy of different types of recommendation, actor, target and entity types. Policy and product aligned on desired non-rec filtering policy which we will want to implement across all relevant recs surfaces as part of the taskforce (or plan anything that needs to go into H2). We are also documenting current nonrec policy and implementation per entity to document the current detection signals used to drive implementation today.

[Theme 4] Metrics and Logging: We are going to move these exit criteria to themes 1-3 since those themes will have to execute against the work.

[Theme 5] Future Proofing & Auditability: We have aligned on exit criteria with a consideration to legal-driven DSA compliance needs and are moving into execution mode.

From: Riana [REDACTED] <[REDACTED]@meta.com>

Date: Tuesday, June 13, 2023 at 8:13 PM

To: Guy Rosen <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Amit Bhattacharyya <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Nissa [REDACTED] <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>, Nicola Mendelsohn <[REDACTED]@meta.com>

Cc: Arcadiy [REDACTED] <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Audrey [REDACTED] <[REDACTED]@meta.com>, Abby [REDACTED] <[REDACTED]@meta.com>, Richard [REDACTED] <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Jimmy [REDACTED] <[REDACTED]@meta.com>, MC [REDACTED] <[REDACTED]@meta.com>, Chris [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>

Subject: WSJ CSAM Series EOD Update: 6/13 (A/C Priv)

Hi all -

Below please find today's update on the WSJ escalation and the Child Safety Task Force. As we await details on the next stories in the series, the product work has moved under the appropriate Task Force themes. We'll accelerate product work as needed if we find that the next story is focused on issue(s) we have yet to address. Please reach out with any questions.

Thank you,
Riana

External

1. Comms

We don't expect to get additional details on the next stories in the series until at least next week.

1. Policymaker outreach
US

We are still expecting a bipartisan press statement from the House Energy and Commerce Committee soon, but it has not been released yet.

Senator Cruz sent us a letter Monday night regarding CSAM and recommendation algorithms. The letter requests responses by June 26. PRIV

PRIV

On Thursday, we've scheduled a bipartisan briefing for House Energy and Commerce committee staff (Antigone will lead).

1. Advertiser

We continue to get advertiser inbound and follow ups about this issue, including from Procter & Gamble, Mail Chimp, Ralph Lauren and Mondelez.

We have offered calls with Antigone to P&G and Nike since both companies have a high volume of questions and concerns on this issue.

On Wednesday, we are briefing Global Business Group (GBG) VPs on this issue who are attending Cannes next week and have advertiser meetings.

Workstreams update

1. Proactive Detection/Sweeps

[Block List] We should be ready to launch the block list on Thursday which will disallow new and remove existing (with a 30-day lookback) IG Accounts FB Pages, Groups, and Profiles using a set of keywords.

[Targeted Enforcement] GRO Detection efforts have so far resulted in 1,800 enforcement actions against IG Accounts, FB Groups, FB Pages, and FB Profiles, in English, Spanish, Portuguese, and Arabic. Efforts are ongoing.

1. Child Safety Taskforce

[Theme 1] User Reports: We are auditing the queue management and ranking to establish ground truth on how child safety jobs are being currently prioritized. This will include understanding which signals are used. We are also establishing root causes for reviewer error on profile and content queues. This will help us determine how to address any technical fixes in review tooling, updating guidelines with policy clarifications and deliver training. As part of this work, we're requesting site visits with vendors in Bogota and Mumbai.

[Theme 2] Account Enforcement: As with User Reports, we're auditing existing Child Safety account enforcement to ensure behavior matches expectations. We are reviewing hard-linked propagation coverage for Child Safety violations, including device blocking coverage, reliability and logging.

[Theme 3] Recs and Discovery: We're working to understand expected behavior for hashtags and search interventions across Facebook and Instagram. We are also creating a taxonomy of different types of recommendation, actor, target and entity type. We plan to align on the recommendation policies with policy teams per category.

- [Theme 4] Metrics and Logging: We are auditing Child Sexual Exploitation enforcement and logging issues and plan to deliver a canonical dashboard for child safety enforcement by violation type for users and content, including enforcement actions.
- [Theme 5] Future Proof and Auditability: We're still aligning on the final exit criteria for this theme because we are making sure the work is complementary to legal-driven DSA compliance needs. We expect to have this complete by Friday.

From: Riana [REDACTED] <[REDACTED]@meta.com>

Date: Monday, June 12, 2023 at 7:37 PM

To: Guy Rosen <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Amit Bhattacharyya <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Nissa [REDACTED] <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>

Cc: Arcadiy [REDACTED] <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Audrey [REDACTED] <[REDACTED]@meta.com>, Abby [REDACTED] <[REDACTED]@meta.com>, Richard [REDACTED] <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Jimmy [REDACTED] <[REDACTED]@meta.com>, MC [REDACTED] <[REDACTED]@meta.com>, Chris [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>

Subject: WSJ CSAM Series EOD Update: 6/12 (A/C Priv)

All -

Below please find today's update on the WSJ Series and the work the XFN is doing. One flag outside of the work we've been doing here is that comms and policy raised that the "cyclone" emoji (🌀), which is quite similar to the Threads logo, is known for being associated with pedophilia. In light of the WSJ escalation, the IG team sprinted on a decision to leadership and made the call to hold the launch and redesign the icon.

Please reach out with any questions.

Thank you,
Riana

External

1. Comms

Jeff Horwitz redoubled his outreach to current and former employees this weekend. From what we've seen so far, he's pressing on whether:

The issues we've been managing as part of this escalation were well known to people internally, and that those people's warnings weren't heeded.

Given recent layoffs, the company has the resources to make progress on the pledges it made in light of last week's story.

Attention was being paid to content recommendations. He specifically raised concern about Reels recommendations.

1. Policymaker outreach

US

We anticipate a bipartisan press release this evening or tomorrow from the House Energy and Commerce committee criticizing IG and mentioning the briefings and other meetings they intend to do (likely with the researchers involved). We have a bipartisan briefing for committee staff scheduled for Thursday morning (Antigone will lead). We also believe that Chair Rodgers will put out her own blog discussing the issue and her intention to address it.

Senator Cruz put out another series of critical tweets and said Meta needs to be more responsive to congressional oversight requests regarding recommendation algorithms, like his.

The South Carolina AG reached out to ask for a briefing.

1. Advertiser

1. As expected, we continue to receive inbound and follow up from clients (including Nike, Google, Nestle, PepsiCo) — although at a slightly slower pace. Questions include the scope of the reporting issues, our ability to apply remediations globally, our brand safety and suitability controls, and the details of our task force.

Antigone had a call with Nestle today which went well and assuaged their concerns for now.

We received approval on a new stat that speaks to the amount of ad impressions that ran next to potentially violating accounts. This will help us scale our response vs. continuing to run individual ad account analysis.

Workstreams update

1. Immediate Product Interventions

1. [Groups] We are filtering Groups with 20% of members with 0.3 MaCSA scores (this is the same threshold we already use on IG and PYMK from GYSJ). Ops has started to review the filtered Groups. So far, GRO's Detection Pillar has surfaced around 30 minor sexualization Groups on FB which have all been removed.
2. [Search] We've landed all the short term search interventions on Facebook. These should also be implemented on IG Direct and Messenger by the end of the week.
3. [Distribution Surfaces on FB] FB Integrity started the work to adopt the MaCSA signal into distribution surfaces and Search. The adoption will follow a business-as-usual approach, and is anticipated to take about a month to launch.
4. [Appeals] Late Friday, we discovered that some accounts that we were disabling as part of this escalation were able to appeal and be reinstated. We believe this is happening because the accounts are being manually disabled via Centra and the job is created with no context, just the profile information, which is not explicitly violating for Child Safety. In the short term, we've done sweeps of the actioned accounts and have re-disabled any that were able to move through the appeal flow. We have moved the diagnosis of this issue to the task force.

1. Proactive Detection/Sweeps

1. [Block List] We are preparing to add hundreds of terms (e.g., [REDACTED] [REDACTED]) to our text matching banks, both disallowing new and removing existing (with a 30-day lookback) IG Accounts FB Pages, Groups, and Profiles. We are testing the banks to ensure there will be minimal False Positives, and coordinating human review of lower-confidence matching assets.
2. [Targeted Enforcement] GRO's Detection pillar surfaced a further 400 IG accounts and 30 FB Groups that violated our child sexual solicitation policies. All assets have been removed from their respective platforms. GRO is exploring further potential for automated and / or bulk enforcement of violating entities across English, Spanish, Portuguese, and Arabic markets.

1. Child Safety Taskforce

1. Exit criteria for the task force was shared with leads on Friday evening. We are collecting feedback and may tweak these criteria a bit in response, but what we've pulled together paints a comprehensive picture of where we expect to focus. [The full details of our exit criteria can be reviewed here.](#)
2. The team has transitioned into execution mode, and we have set an initial target of June 23 for the "Understand" category of exit criteria. We anticipate that upon completion of these items we will update the exit criteria to better reflect the gaps we find in the audits of our systems.

From: Alison [REDACTED] <[REDACTED]@meta.com>

Date: Friday, June 9, 2023 at 6:59 PM

To: Guy Rosen <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Amit Bhattacharyya <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Nissa [REDACTED] <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>

Cc: Riana [REDACTED] <[REDACTED]@meta.com>, Arcadiy [REDACTED] <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Audrey [REDACTED] <[REDACTED]@meta.com>, Abby [REDACTED] <[REDACTED]@meta.com>, Richard [REDACTED] <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Sayed Oturu <[REDACTED]@meta.com>, Jimmy [REDACTED] <[REDACTED]@meta.com>, MC [REDACTED] <[REDACTED]@meta.com>, Chris [REDACTED] <[REDACTED]@meta.com>, Krish [REDACTED] <[REDACTED]@meta.com>

Subject: WSJ CSAM Series EOD Update: 6/9 (A/C Priv)

A/C Priv

Hi all-

Below find today's update on external reaction + outreach and work that the XFN is doing across workstreams. Please let me know if you have any questions

Thanks and have a great weekend,

[REDACTED]

External

1. Comms

1. The cycle appears to have concluded, barring any new developments that could reignite it.
2. We do not have any new information or detail from the reporter on his next stories but anticipate reconnecting with him next week to learn more.

1. Policymaker outreach

1. US

The House Energy and Commerce Committee minority staff has asked for a briefing next week.

1. Global: India

A consultant associated with the Ministry of Electronics and Information Technology (MeitY) reached out with a request from the Additional Secretary (who is also overseeing the Prajwala CEI litigation) for a response.

1. Advertisers

1. Call with Antigone and advertisers today went well and we got good feedback, especially from agencies. However, there was confusion around what the WSJ article was specifically about (bad actors finding each other vs. prevalence of CSAM on our platforms).
2. To help clarify, we sent a detailed follow up email and are sharing it with internal sales teams for use in the case of future advertiser inbound.
3. Hershey unpaused advertising, but Honda and Ford remain paused, and LVMH paused following the call. There is a risk that the pause expands to other advertisers and becomes public, re-igniting the press cycle.
4. We expect continued inbound and follow ups on this issue from clients over the weekend and into next week.

Workstreams Update

1. Immediate product interventions
 1. [Groups] We will filter any Group in which 20% of members have above a 0.3 MaCSA score (this is the same threshold we already use on IG and PYMK. from GYSJ). This should be implemented by this evening. We are also taking the list of filtered Groups and putting them through ops review to determine if they should be removed.
 2. [Search] IG Direct and Messenger: We are looking at replicating the same search interventions we've implemented on IG and FB on IG Direct and Messenger. We are also exploring sweeps of broadcast and business channels.
1. Proactive detection/sweeps
 1. [Block List] On Monday, we will add hundreds of terms (e.g. [REDACTED] [REDACTED] to our text matching banks, both disallowing new and removing existing (with a 30-day lookback) IG Accounts FB Pages, Groups, and Profiles. We are establishing both high-confidence banks that auto-enforce on matches, and lower-confidence banks that route matches for human review.
 2. [Bulk Actioning] In addition to the blocklist, we bulk actioned ~ 650 IG accounts with [REDACTED] and variations of in the Name and Bio. To limit over enforcement, we excluded XChecked Accounts, and Accounts with large follower counts.
 3. [Targeted Enforcement] GRO has established multiple additional detection mechanisms to support continuous enforcement of violating assets across both IG and FB, in both English and Spanish.
1. Child safety taskforce
 1. Proposed exit criteria for task force will be circulated this evening.

From: Alison [REDACTED] <[REDACTED]@meta.com>

Date: Thursday, June 8, 2023 at 7:55 PM

To: Guy Rosen <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Tom Alison <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Monika Bickert <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Amit Bhattacharyya <[REDACTED]@meta.com>, Joey Simhon <[REDACTED]@meta.com>, Andrea Saul <[REDACTED]@meta.com>, Chris Sonderby <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Nissa [REDACTED] <[REDACTED]@meta.com>, Kevin Martin <[REDACTED]@meta.com>

Cc: Riana [REDACTED] <[REDACTED]@meta.com>, Arcadiy [REDACTED] <[REDACTED]@meta.com>, Anna [REDACTED] <[REDACTED]@meta.com>, James [REDACTED] <[REDACTED]@meta.com>, Audrey [REDACTED] <[REDACTED]@meta.com>, Abby [REDACTED] <[REDACTED]@meta.com>, Richard [REDACTED] <[REDACTED]@meta.com>, Gabriel [REDACTED] <[REDACTED]@meta.com>, Kaitlin [REDACTED] <[REDACTED]@meta.com>, Liz [REDACTED] <[REDACTED]@meta.com>, Sayed Otaru <[REDACTED]@meta.com>, Jimmy [REDACTED] <[REDACTED]@meta.com>, MC [REDACTED] <[REDACTED]@meta.com>,

Chris [REDACTED] <[REDACTED]@meta.com>

Subject: WSJ CSAM Series EOD Update: 6/8 (A/C Priv)

A/C Priv

Hi all -

We are going to start providing daily updates to this group on the WSJ series. Updating on external reaction + outreach and work that the XFN is doing across workstreams. Please flag if you have any questions and Chris and James, [REDACTED]

PRIV

Thanks,

[REDACTED]

External

1. Comms

The cycle appears to be winding down, barring any new developments that could reignite it. It is no longer on Techmeme and most new coverage in the US has been in conservative or lower-tier outlets. There has been some coverage on cable TV with Jeff Horwitz and Alex Stamos conducting a handful of interviews, but nothing on broadcast. In Europe, coverage has included mention of Commissioner Breton's tweet about his upcoming meeting with Mark on June 23.

Comms is in daily informal communication with Jeff, who has indicated he will likely do stories (not imminently) on a) similar activity/content on Facebook, b) brand safety implications c) impact of end-to-end encryption

1. US Policymaker Outreach

We have not received official outreach from the Hill, but expect this to change. We are specifically tracking:

Today, Senator Graham said in an unrelated hearing that they should do a hearing on this topic. Chair of the House Energy and Commerce committee has not officially reached out, but we believe her staff will soon, in the form of a letter or even demanding we testify at a hearing.

We have had limited outreach from AGs but also expect letters to come soon.

1. Advertisers

According to the Client Council, this is some of the worst advertiser reaction to an integrity issue we have seen since the 2020 advertiser boycott.

Global Business Group (GBG) has received dozens of inquiries from North America clients and global agencies including Horizon Media, Nike, PMG, Publicis, Kroger, Disney, Airbnb, Fidelity, Ralph Lauren, Google, Home Depot, Inspire Brands, Honda, WPP, Hershey, GroupM. We have set up a call on Friday at 11AM ET for these partners with Antigone (~80 RSVPs so far)

Hershey's has paused advertising spend, we have arranged a standalone call with Antigone and product to try and alleviate their concerns.

Workstreams Update

1. FB interventions: We are taking a number of immediate actions on FB search/discoverability and are sprinting on additional product ideas for Groups.

[Search] Removed "see results anyway" click through on FB interstitial.

[Search] Temporarily return no hashtag or search results for a very broad set of non-violating terms i.e underage, teen model (same list for what we did on IG).

[Groups] Team is sprinting on a proposal to include Malicious child safety actor (MaCSA) scores into filtering risky Groups from GYSJ (i.e if X% of the Group has a certain score, it is filtered from the surface)

[Brand Safety] Added the CEI signal to our FUSS non monetizable signal to prevent ads from being shown near borderline CSAM

1. Proactive detection/sweeps

We are conducting on-platform sweeps across IG and FB of accounts/complex objects soliciting and distributing CSAM. This includes keyword pipelines and conducting fanouts. We are also starting to assess associated inorganic risk potential.

Finalizing a proposal to use the text matching service tool to prevent entity creation based on keyword strings (e.g., we'd prevent IG Accounts, FB Groups, Pages, and AP+. with "MegalinksCP" as their name). We can also implement this retroactively, which would remove any complex object that matches the text string.

1. Child safety task force

We are finalizing exit criteria and execution plans for the medium and longer term investments for the task force.

We launched an update to our content human review ranking to more directly incorporate Child Sexualization signals, which we expect will result in more user reports for this type of content getting reviewed. We will monitor the results over the next 3-4 days.