

From: Vaishnavi J [/O=THEFACEBOOK/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=REDACTED]
Sent: 9/24/2020 7:23:15 PM
To: Sarah [REDACTED] [REDACTED]@fb.com]; Karuna [REDACTED] [REDACTED]@fb.com]
CC: Antigone Davis [REDACTED]@fb.com]
Subject: Re: Child Safety State of Play - Timely/Action needed

Hi Sarah - on 1) This is a policy gap. We do not have policies in place to account for this behaviour at scale. An empty account is one that has no posts or content, but just comments on other people's posts.

Right now, IGPR only reviews reported profiles for violations in their subject elements like bio, photo, username, and doesn't review the actual content on the profile (except in some very limited queues). CO is looking to bring back holistic review of IG profiles in 2021 beginning with our most severe violations.

Hope this helps!

From: Sarah [REDACTED] <[REDACTED]@fb.com>
Sent: Thursday, September 24, 2020 12:12:23 PM
To: Vaishnavi J <[REDACTED]@fb.com>; Karuna [REDACTED] <[REDACTED]@fb.com>
Cc: Antigone Davis <[REDACTED]@fb.com>
Subject: Re: Child Safety State of Play - Timely/Action needed

I've started a Google doc where I'm tracking changes as I revise. Vaishnavi, I tried to incorporate your comments below but a few things to clarify:

1. We do not enforce against certain concerning behaviours at scale across FB / IG:

- Empty accounts that primarily leave explicitly sexualising comments on pictures of minors - currently, upon escalation where we confirm the account-holder is regularly posting sexualizing comments on images of minors, the account will be disabled.
- Empty accounts that primarily leave implicitly sexualising comments on pictures of minors or that only engage in behavioural actions e.g. likes, saves of images of minors

Can we cover with: "While our minor sexualization policies will be completed this half, our AI tooling is not mature enough to proactively detect this content at scale which often occurs in a highly contextualized manner." Or are you pointing out a policy gap? Not sure what you mean by "empty accounts"?

And when you say "holistic IG profile review in 2021", what do you mean? I added the below for proliferation of our FRX tag "Involves a Child."

Enabling users to report child sex abuse anywhere they see it	We recently shipped a user report feedback tag, "Involves a Child" on IG Profiles. However, we need to continue closing existing gaps in reporting coverage and launch this feedback tag where it is missing across the platform, including FB Groups and IG comments.
---	--

From: Vaishnavi J <[REDACTED]@fb.com>
Sent: Thursday, September 24, 2020 10:27 AM
To: Karuna [REDACTED] <[REDACTED]@fb.com>; Sarah [REDACTED] <[REDACTED]@fb.com>



Cc: Antigone Davis <[REDACTED]@fb.com>

Subject: Re: Child Safety State of Play - Timely/Action needed

Hi Karuna,

Thanks for sharing - suggested edits from me are below.

Thanks,
Vaishnavi

CHILD EXPLOITATION

Under "ready / ongoing"

Sexualisation of minors on IG: Ongoing - Expanded minor sexualisation policy in October, holistic IG profile review in 2021, profile-level reporting option for CEI that is routed for holistic review by Child Safety teams, Resources for people who seem to be seeking out CEI, Hiding minors' context lines (e.g. additional info about location) in cross-app threads when an unconnected adult messages a minor,

Messenger Safety Tips on IG: Ongoing - Messenger Safety Tips on track to launch on IG in October

Under "not ready" - move up to "somewhat ready"

Block reach of unconnected adults to IG minors: Unconnected IG adults cannot initiate messaging with FB minors, and we are working to prioritise work with product to prevent unconnected FB adults with a high IIC score from being able to reach out to IG minors.

Under "not ready" - move up to "ready"

IIC classifier use on IG: IIC classifier has now been calibrated for IGD

Add to "Not ready"

1. We do not enforce against certain concerning behaviours at scale across FB / IG:

- Empty accounts that primarily leave explicitly sexualising comments on pictures of minors - currently, upon escalation where we confirm the account-holder is regularly posting sexualizing comments on images of minors, the account will be disabled.
- Empty accounts that primarily leave implicitly sexualising comments on pictures of minors or that only engage in behavioural actions e.g. likes, saves of images of minors

AGE ASSURANCE

For age assurance, under "Not ready", "Age data on IG" - recommend revising to:

Age models do not work as well on IG because IG only started to collect age data for new sign ups in December 2019. Thus, we currently only have stated age for ~67% of IG users. Age collection on IG has previously been criticized as being for the purpose of targeting ads, and intrusive to users who do not necessarily need to bring their authenticated selves to the platform. Relying on stated age on similar apps in the industry also have also resulted in age lying and misrepresentations, and fines for collecting children's personal information. We recommend engaging with IG leadership on building better age affinity models to predict age of our users, and prioritise age model maturity on IG.

From: Karuna [REDACTED] <[REDACTED]@fb.com>
Sent: 24 September 2020 2:02 AM
To: Sarah [REDACTED] <[REDACTED]@fb.com>; Vaishnavi J <[REDACTED]@fb.com>
Cc: Antigone Davis <[REDACTED]@fb.com>
Subject: Child Safety State of Play - Timely/Action needed

Hi Sarah and Vaishnavi, thanks for all the work on the interim update to the child safety state of play doc. Can you please review the recommendations section for child exploitation and age assurance (pasted below) and send any updates to me in your morning.

Thanks!
Karuna

Recommendation - Child Exploitation

Internal readiness to rise to the challenge. Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged, suggestions on how we might kick start conversations on this work.

Ahead of Interop, work is in progress to block unconnected adults on FB and IG from initiating messages with minors on FB, however work is not planned to prevent FB and IG adults from initiating messages to IG minors. Limited country tests and data analyses have shown that [REDACTED] can demonstrably reduce CEI sharing but product leadership has been unwilling to prioritize these mitigations. We are currently escalating the [REDACTED] conversation up through Monika to meet with MLT. We are also getting ready to meet with Jay [REDACTED] to discuss how best to present [REDACTED] to MLT. We need leadership support to have product prioritize this work because it will allow us to make a much stronger case with our external critics and stakeholders that we are doing good work on this front. We have progress on blocking predicted groomers from using Disappearing Mode but need CO support to apply the same protection in IG. Work has been completed on basic safety mitigations for Rooms but work still needs to be done to enable in-call reporting with evidence capture in both Rooms and RTC (aka Video Chat). RTC, which is the predecessor to Rooms, is not yet in parity with Rooms (i.e. feature gating high-risk users, user education/self-remediation interventions) and will require stronger leadership buy-in to accomplish even bare minimum protections this half. We continue to advocate for strong discoverability and reachability controls in anticipation of Interop, including further restrictions to PYMK and GYSJ. [REDACTED] is currently ongoing but we need more resources for this in addition to engineering and review resources to develop an operational sex trafficking classifier. Please see the chart below for detailed summary of the specific related work and its status.

Area	Additional notes
Ready; work is complete / ongoing and on track	
Sexualisation of minors on IG	Ongoing: Expanded minor sexualisation policy in H2; holistic IG profile review in 2021
Rooms	Completed: Safety mitigations include 1) feature gating high-risk users (NCMEC reported and predicted groomers) 2) proactive detection of CEI associated Room names 3) we do not recommend Rooms to unconnected adults if they contain minors (and vice versa) 4) we prevent unconnected adults from inviting minors directly into Rooms 5) we block Room links from comments in a minors' post if the commenter is an unconnected adult 6) we block Room creation by Groups with membership of 1 and photos, and 7) we also log and retain identifying metadata on all users, including anonymous joiners for 90 days
NCMEC numbers contextualization	Ongoing: prioritized by Child Safety DS
Block reach of unconnected adults to FB minors	Ongoing: Just as unconnected FB adults cannot currently initiate messaging with FB minors, IGD adults will not be

	able to initiate messages with FB minors (ETA ahead of Interop global launch)
Messenger Safety Tips	Ongoing: Launch Messenger Safety Tips on IGD (ETA: Q3); Cross-network threads (ETA: November; prior to Interop global launch)
Tincan and Disappearing Mode carve-out	Ongoing: Use the IIC actor level classifier to block our riskiest users from using Tincan and Disappearing Mode (the successor to Tincan), which is a high-risk feature for child safety
Somewhat ready; work is ongoing but more resources are needed	
Interop: discoverability and reachability	Needs more resources: Strong discoverability and reachability controls, including further restrictions to PYMK and GYSJ
RTC and Rooms	On-going: In-call reporting with evidence capture work; RTC has committed to bringing themselves into parity with Rooms safety mitigations, which at minimum includes feature gating high-risk users. <i>In discussion:</i> user education/self-remediation interventions (i.e. safety notices, upsell reporting/blocking)
██████ and ██████████	Ongoing: ██████████ ██████████ for ████████ Needs more resources: Child Safety ██████████ ██████████
X-industry video hash sharing	Needs more resources: Initial conversations underway to develop universal video hash sharing in collaboration with key partners like Thorn who have done early work and our opensourcing of PDQ-TMKF in 2019.
██████████ ██████████ ██████████ ██████████	██████████ ██████████ ██████████ can demonstrably reduce sharing of CEI (a recent test showed that ██████████ reduced our NCMEC Cybertip report numbers by 7% in test countries) and could help us push back on calls for backdoor legislation. We are currently escalating the ██████████ conversation up through Monika to meet with MLT and could use your support. We are also getting ready to meet with Jay ██████████ to discuss how best to present ██████████ ██████████ to MLT.
Not ready; work has not been prioritised and more resources are needed	
Aggregated content on IG surfaces	Build classifiers to detect and take action on problematic aggregated content on surfaces like Explore, hashtag pages, Reels
Block reach of unconnected adults to IG minors	Currently, on MSGR, we do not deliver messages initiated by unconnected adults to minors. In IGD, we do not have this protection in place. For Interop, product has committed to preserving this rule in MSGR, but not for cross-network messages sent by FB adults to IGD minors. Nor will they protect IGD minors from receiving messages from unconnected IGD adults. In light of the fact that grooming on IG occurs most likely ~2x as much as FB, their main reason not to build this protection for IGD minors is less than compelling: we do not want to block FB parents and older relatives from being able to reach out to their IGD children/relatives. They claim this is one of the main value props for Interop and a big growth bet.

IIC classifier use on IG	Product and Policy have aligned that using the IIC classifier is the best available safety mitigation for minors in Disappearing Mode, planned for global launch in November. However, the actor-level groomer classifier is not yet operational on IG, which means that we cannot use it to block groomers from accessing high-risk features like Disappearing Mode. We need about a week or two of CO review capacity in order to calibrate the classifier for IG use. This is low hanging fruit and fully supported by Child Safety CI and IG Policy. We just need help getting CO to prioritize this.
Upsell reporting for Disappearing Mode	Disappearing Mode messages sent in IGD are retained for 14 days and in MSGR are retained for 1 hour. The bulk of user reports for IIC (57%) and CEI (46%) in MSGR and IGD are sent within the first hour. After the first hour, the number of reports drop off dramatically (2-5% made within each subsequent hour). However, product does not have the CO resource capacity to support reporting upsells within Disappearing Mode to educate the user that even ephemeral messages can be reported and that they have a limited window of time in which to do so. Once our transparency report is released, governments will seize upon spikes in our NCMEC reporting numbers during COVID-19 to fuel calls for exceptional access legislation. We should make clear to product and CO that during a time when we have increased CEI activity on our platform, we should prioritize reporting upsells for Disappearing Mode before global launch.
Sex trafficking classifier	We have one engineer dedicated to this work and we have no data science support to do some basic understand work to lower the strikes threshold for prostitution and sexual solicitation strikes (currently set at 17x, which is too high). This area is severely under-resourced given our SESTA-FOSTA liability.

Recommendation – Age Assurance

Internal readiness to rise to the challenge. Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged, suggestions on how we might kick start conversations on this work.

In March it was decided that we would not build a classifier to detect minors under 13 years old. Instead, we would push industry partners for device/OS level age verification and management solutions. Since then, Central Privacy has stood up a small team on youth privacy but a PM is yet to be hired. They have offered to prioritize a few low business cost/high policy impact projects for H2. Identity Integrity has also identified a PM who would own age verification but they will not prioritize this work until we have clear legal obligations to do so. We need leadership support to push product to act on knowledge we already have, i.e. clear our backlogs, disable verified u13s, build an u18 checkpoint to review minors who have been reported on 18+ products like FB Dating. Against this backdrop, our enforcement is not consistent across all apps for a single known person. In addition, age models perform inconsistently across different markets. We recommend that Policy leadership stack rank and prioritize our most sensitive policy driven markets and make clear to product that we should operationalize those markets first. We also need leadership to support the basic principle that if we're using a signal to predict age for business purposes, it should be used to enforce on age. For example, we need to use age prediction to gate minors from alcohol ads and other high-risk features where we currently

rely solely on stated age. Age models also do not work as well on IG because IG only started to collect age data in December 2019 and we currently only have stated age for ~67% of IG users. We need leadership support to convince IG leadership to collect age for the remaining users for whom we have no stated age. We also would like leadership sign off to conduct an audit of special protections for minors and then bring in a trusted expert/group of experts to review our default settings and protections for minors, identify improvements (e.g. GYSJ and PYMK restrictions), and implement them.

Area	Additional notes
Somewhat ready; work is ongoing but more resources are needed	
Age Management	An 18-month long escalation that ended with a Cross-Family Review on 3/3/20 with Nick's input resulted in a decision not to build a classifier to detect minors under 13 years old. Instead, leadership decided to invest in policy work to push industry partners for device/OS level age verification and management solutions. In the meantime, our basic COPPA compliance is at risk when product does not prioritize checkpointing and disabling u13s. We also do not yet cross-enforce against u13s discovered on IG who have hard-linked FB accounts. In addition, we do not yet have product support or resources for an u18 checkpoint where we can efficiently manage u18 users who are reported or detected on 18+ products like FB Dating, Marketplace, Mentorship, or as under-age parents on Messenger Kids. Identity Integrity, which owns underage checkpoints, has no plans to invest in age management in H2 and will instead prioritize impersonation work. Central Privacy has decided to stand up a small team on youth privacy but they have not yet hired a PM and have offered to prioritize a few low business cost/high policy impact projects for H2. If Identity Integrity cannot prioritize age because of U.S. GE 2020, our best opportunity might be with Central Privacy's willingness to help and we could push for immediate hire of a PM to make some progress on this work in H2.
Global use of age models	The best performing age model is not yet globally operational, which leads to a patchwork of integrity measures across the company that leverage age. We need more review resources to establish ground truth data on age in priority markets. We should start by engaging with Core Dimensions (the product team who owns the age models) on what markets we've identified and stack ranked as a Policy org to prioritize with their available H2 resources.
Not ready; work has not been prioritised and more resources are needed	
Age data on IG	Age models do not work as well on IG because IG only started to collect age data in December 2019. In addition, Adam Mosseri will not support upselling age collection for existing users because he sees it as intrusive data collection. Thus, we currently only have stated age for ~67% of IG users. We recommend engaging with IG leadership on collecting age for the remaning ~33% and prioritizing age model maturity on IG.
