

FILED UNDER SEAL

EXHIBIT 72

From: [REDACTED] [/O=THEFACEBOOK/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=[REDACTED]]
Sent: 8/8/2018 9:51:10 PM
To: [REDACTED] [REDACTED]@fb.com]; [REDACTED] [REDACTED]@fb.com]
CC: [REDACTED] [REDACTED]@fb.com]; [REDACTED] [REDACTED]@fb.com]; [REDACTED] [REDACTED]@fb.com]
Subject: HB IIC HPM 8 August 18

Harmful Behavior IIC HPM - August 8, 2018

CI Mission: Build a safe, inclusive community by minimizing bad experiences for people on Facebook

HB Mission: Prevent behavior leading to physical or emotional harm from happening on Facebook

This is the first Inappropriate Interactions with Children (IIC) focused HPM. We'll be sending one out every 2 weeks! We've previously covered the scams work within the HB wide HPM.

IIC is defined as an adult contacting or attempting to build a relationship with a minor for the purposes of sexual gratification. A full definition of IIC including Tier 1 and Tier 2 IIC can be found [here](#)

Tl;dr in <5 bullets

1. We've established Q3 roadmap goals for IIC: (a) Build high precision detection signals for Tier 1 IIC (real world harm and CEI cases), (b) action accounts across Tier 1 (10k) and Tier 2 (online harm) IIC (1m) and (c) run experiments with broader product teams to limit future IIC violations.
2. Prevention of future IIC: We've launched an experiment in partnership with the Friending team to measure the impact of removing minors from the PYMK of adults suspected of IIC.
3. Understand work: A few key insights - (a) We've identified PYMK as the top victim discovery channel (79%) for IIC suspects. (b) ~30% of disabled IIC adults are recidivist i.e. they come back onto the platform and resume violating behaviour.
4. Tier 1 IIC: We have two promising heuristics for detection of real world harm and CEI.

Finalised Q3 Roadmap

Q3 Detailed Goals

We recently exited Q3 roadmapping and are executing against the following goals:

Area	Goals
Tier 1 IIC	
Measurement	[P1] Stand up lower bound estimate of prevalence
Detection	[P1] Ship heuristic/classifiers with >30% precision
Remediation	[P1] Action 10k Tier 1 (real world harm) accounts via CO manual review, stretch goal - action 50k accounts with additional CO capacity
Review Tooling	[P1] Ship tooling and review tree changes to increase accuracy of Tier 1 harm decisions

XFN	[P2] Ship NCMEC reporting flow for text only solicitation of CEI
Tier 2 IIC	
Measurement	[P2] Ship account level prevalence metric
Detection	[P1] Ship actioning classifier at > 90% precision
Remediation	[P2] Automatically action 1 million accounts i.e. ~20% of estimated prevalence
Remediation	[P2] Ship actioning experiments to measure effectiveness of increasing friction for bad actors
Prevention of future IIC	
Friending team	[P1] Ship experiment with PYMK team to measure impact of removing minors from adult PYMK for adults flagged by IIC classifier (can do this at lower precision)
Messenger team	[P1] Ship experiment with Messenger team to measure impact of preventing exchange of nnp images between adults and minors
Understand work on Family of apps	
IG	[P2] Understand IIC prevalence on IG platform and cross-disable accounts soft linked to FB IIC accounts
Dating	[P3] Support user gating for Dating app via IIC and Adult prediction classifiers

Full details in - [Harmful Behaviour - Q3 Roadmap](#)

Highlights

Shipped

We've launched an experiment in partnership with the Friending team to measure the impact of removing minors from the PYMK of adults flagged by the IIC behavioural classifier

[REDACTED], [REDACTED], [REDACTED], [REDACTED], [REDACTED]

- Context: PYMK is the top victim discovery channel for suspects (see bullet in understand section below)
- Experiment set up: Experiment will be run over a month, with 50% of adults flagged by the IIC classifier being exposed to the treatment (i.e. no minors in PYMK). Full details [here](#)
- Key metrics (full list in [quip](#)): all metrics will be compared for treatment and control groups and analysed by classifier score buckets.
 - Friend requests sent and accepted (minors vs others)
 - New threads started (with minors vs with others)
 - New accounts created by suspected IIC adults
 - Counter metrics: impact on friending (segmented for impact on US friending)

Understand work

PYMK is the top victim discovery channel for IIC suspects [REDACTED]

- Context: Identifying the key channels through which IIC suspects find minors will help prioritise product changes that reduce future IIC violations on the platform
- Results: 79% of new conversations between IIC suspects and minors were conversations between friends. (full details in [post](#))
 - 88% of friending was initiated by the adult

- o PYMK was the top friending source accounting for 79% of the friend pairs
 - o Profile was the second most common source accounting for 18% of the friend pairs.
- Next steps: we've launched an experiment with the friending team to measure the impact of removing minors from the PYMK of adults flagged by the IIC classifier (covered above)

Recidivism rates among disabled IIC accounts is high at ~30% [REDACTED]

- Context: Analysis is based on 4.5K IIC accounts disabled between 4th to 10th June, with recidivism measured as of 9th July through device/DATR sharing or accounts identified by GUARDS, SUMA (details in [quip](#))
- Results: ~42% disabled accounts created new accounts. Of these ~60% were labelled for IIC violations in the manual review. An additional ~11% were check pointed for other violations. Overall a lower bound 30% were recidivist.
- Next Steps: To establish an upper bound by tracking accounts over a longer time period, and exploring UIG in addition to DATR /device for identifying recidivist accounts.

Progress

Tier 1 IIC: we've identified two promising heuristics based on analysis of previously disabled tier 1 accounts and are working towards generalising the approach for Tier 1 (CEI and real world harm) detection

1. **Heuristic 1: Long two-way conversations between adults and minors that are sexual in nature [REDACTED]**
 1. Heuristic description: Long, two-way conversations between adults and minors (>1000 messages exchanged), with multiple sexual terms used. This was based on the insight that Tier 1 cases of harm tend to have very long message threads with minors containing sexual conversations.
 2. Early results: ~19% of accounts were labelled as tier 1 IIC in a one off labelling exercise. A significant portion of the negative labels were age liars (i.e. minors pretending to be adults).
 3. Next steps are to 1) test lower thresholds for conversation length, 2) develop text classifiers to replace the current bag of words approach and 3) improve filtering for age liars.
2. **Heuristic 2: High scoring NNP images sent by minors to adults [REDACTED]**
 1. Heuristic description: high scoring NNP2 images sent by female minors with a unique photo dna hash to screen out viral memes. Data on disabled groomers revealed that minors were more likely to have sent imagery in cases that get escalated as tier 1.
 2. Early results: ~35% of accounts were labelled as containing CEI in a one off labelling exercise. A deeper look at the 'ignores' revealed that images had been deleted, and were therefore not seen by the reviewers.
 1. Caveat: this data is based on results for the Algerian market, where the heuristic was applied as a test case.
 3. Next steps are to 1) refine the heuristic and 2) tackle the deleted images problem to improve accuracy. 3) We're also setting up a test to measure the results on a global sample.

Review Tooling: the IIC team was in Dublin last week to meet market SMEs and shadow their workflow to identify gaps in tooling and improvements in the review process. The emerging insights are listed below.

More details can be found [here](#) [REDACTED]

Emerging insights

1. Thread ranking is critical to the guided review process as reps rely heavily on the top 3 ranked threads to make their decision. Our current ranking is not optimised and is likely leading to false negatives.
2. Verifying the age of the participants in the conversation is hard and time consuming, and simple UI improvements could make this easier and would increase reviewer efficiency and accuracy.
3. We need to dig deeper to understand whether the current CEI reporting guidelines in the queue err on the side of over-reporting to NCMEC.

Thank you to [REDACTED] for rallying the troops and all the Market SMEs that joined us last week!

Team and Me

Last few weeks have been very hectic with road-mapping, so it was nice to get a quiet weekend to spend with my son and watch him take his first few steps! :D