Message

From: Cliona [redacted] [redacted]@fbworkmail.com]
Sent: 9/25/2020 9:32:25 AM
To: Child Safety Working Group - Interaction Integrity [post+2096976820549581@mail.workplace.com]
BCC: Sarah [redacted] [redacted]@fb.mail.onmicrosoft.com]
Subject: [Child Safety Working Group - Interaction Integrity] [24 Sept] MWB x Child Safety Weekly Sync

[redacted]

Cliona [redacted] posted in Child Safety Working Group - Interaction Integrity
September 25 at 2:32 AM

[24 Sept] MWB x Child Safety Weekly Sync

MEETING GOAL

Align and execute across MWB, Child Safety and IGWB on mitigations and their measurement to decrease prevalence of Tier 2 IIC on IGD

TL;DR

- The outcome of CEP approval for IGD is expected today.
- There was agreement to hold off on officially setting the baseline and goal until we have good data on the age estimates– current age estimates are resulting in false positives and Malia is working on improving the labelling process to minimise them.
- The 38x number should not be used to compare messenger to IGD prevalence because of differences in the populations. Malia is doing some due diligence on the metric and when it's more comparable to Messenger the number will be rolled out more broadly.
- The initial approach to identifying potential IIC violators will be to use the behavioural classifier but it should be noted that this will have a really small impact and we need to follow with the NFX heuristic as soon as possible. Pouya will do NFX sizing this week with a goal of implementing this approach more quickly.
- Rahul has started scoping the first mitigation (prohibit potential violators from sending messages to minors). The implementation will not be invisible to suspects as it follows reachability guidelines. Product brief: https://docs.google.com/document/d/1ZpozZE-mogPlak_K82r-6CbmiYQIKekfLSU9zD8Sd6c/edit
- Age verification via ID is a possibility to reduce false positives, but it would be difficult on Instagram because (1) many minors have no ID and (2) Instagram accounts don't have true identities and people could pass with IDs that didn't belong to them. Follow up is to look at conclusions from research on this topic.

- Discoverability analysis is still in the early stages. So far, analysis shows that 55% of positively labelled suspects sent message requests (were not following the minor) but this is very similar to the rate of people who were not involved in IIC.

NOTES

1. UPDATE ON CLASSIFIER ESTIMATED PREVALENCE (CEP) APPROVAL FOR IGD

- PXFN took place on Monday Sept 20th and there was strong support for it; we're expecting to hear from legal and leadership about approval by EOD.

2. **Update on adding 'age question' to labelling flow [MA]**

- Context: we've launched IIC Tier 2 measurements, but the age estimate is very challenging. Investigators have examined a sample of labelled data and have said there are age false-positives – most are on the dividing line between minor and adults. Identifying age in this group is challenging on IG compared to FB where the estimate is much more mature
- Malia is working on mitigating this and tightening up the measurement – for ex exploring whether additional widgets for context should be added to those jobs.
- The existing age prediction model predicts teen vs non teen and this doesn't help because 'teen' includes 18/19 year old adults, while the policy specifies contact between adult and child. Even some level of granularity would be good here.

2a Related: Update on setting goal and baseline

- We want to hold off on officially setting the baseline and goal until we have good data on the age estimates. We had tentatively agreed on a 25% goal but we should wait to publicise that until we have better data.
- Also, CW pointed out that we shouldn't use the initial 38x number to compare Messenger to IGD prevalence; we don't want to directly compare with FB side because we're not comparing the same population in each group (different age populations etc).
- Malia is going to do some due diligence to make the metric more comparable to the FB side, and will also be working on the filters and the age measurement. When we're comfortable with the changes it will be rolled out more broadly.

3. IIC mitigations: Agree on first approach to identifying potential IIC violators

- We discussed the proposal (ie start by using the behavioural classifier developed by child safety, beginning at 60% precision (same classifier as proactive warnings). Over time, we will try different precision thresholds and heuristics.)
- There was a discussion about the low expected impact of classifier-only in identifying suspects and there was a strong recommendation to follow FB approach from the start (uses a lower classifier threshold and also includes NFX signals from minors) since taking away the message button is not a hard actions and won't be noticeable.
- However, when we talked about the implementation it was noted that this will NOT be an invisible action to suspects since they will see the message about the message not being sent. The implementation is being done this way because of requirements of the

new reachability controls in Messenger/IGD. Because of this we have to evaluate the NFX actions false positive rates.

- Follow up is to do a sizing for FPs from NFX actions this week (Pouya)
- There was a question about having the NFX data labelled – this would need to be discussed with CO partners; it may be difficult to get labelling at the moment because of WFH.
- In the meantime, Rahul will start by implementing the classifier approach. We will follow as soon as possible with the NFX heuristics.

4. IIC mitigations: update on implementation of first mitigation (Prohibit potential violators from sending message requests to minors)

- Rahul has started to scope but is currently off sick
- We don't have PXFN approval yet but LaMa is in assessment. Vaishnavi said that she can help with PXFN if there's any need there
- Product brief: https://docs.google.com/document/d/1ZpozZE-mogPlak_K82r-6CbmiYQIKekfLSU9zD8Sd6c/edit

5. Discussion on potential UFAC checkpoint for age verification

- Context: False Positive impact of the mitigation will be asymmetric - if an adult is caught it won't be much impact but if we stop a minor from messaging friends it will have a big impact.
- Suggestion: we should gate the soft actions behind a UFAC, where ID is uploaded and AI can decide whether the ID is fake.
- There was another comment that we could also apply this to the labelled positives
- Counter-argument 1: Children don't usually have ID. Although this is mainly in the US where teens often have driving licences, this mainly applies to 16+ (varies by state).
- Counter-argument 2: n IG it's really hard to verify ID. On IG you can be anything, so people would only have to find any fake DI under the age of 18 and they'd be fine
- There was a note that other services like Snapchat and Tiktok don't do this either because lying is so easy.
- Trevor mentioned that there was lots of research on this topic for the election and we should look at their conclusions (link: https://www.internalfb.com/intern/research/p/ovrq8oiz0l/)

6. Update on discoverability analysis so far

- So far, Cliona has been trying to estimate how successful we might be in removing messaging ability by looking at how often messages were sent before follow (ie. Message requests that will be intercepted with intervention)
- In 55% of prevalence-labeled cases, responsible messaged before follow --> message request was needed (15% target sent first message, but some of those were followed first by responsible) – very similar % in the negatively labelled cases.

View The Post on Workplace.

Click here to see what they posted.

Like Comment Share

View on Workplace

This message was sent to [redacted]

If you don't want to receive these emails from Workplace in the future, please **unsubscribe**.

Facebook, Inc., Attention: Community Support, 1 Facebook Way, Menlo Park, CA 94025

