Message

From: Miki Rothschild [/O=THEFACEBOOK/OU=EXTERNAL (FYDIBOHF25SPDLT)/CN=RECIPIENTS/CN=[redacted]]
Sent: 7/14/2020 2:57:05 PM
To: Sayed Otaru[redacted]@fb.com]
CC: Yoav Shapira [redacted]@fb.com]
Subject: Re: IIC on Instagram

Super helpful. If MWB is taking this on I don't think you need to reconcile. I just wanted to know in case because the HPM said they're still figuring out (probably older version). I'll forward it to you and you can ask Sam[redacted] to add you to the distribution if you'd like to get future ones.

Sent from my iPhone

On Jul 14, 2020, at 7:48 AM, Sayed Otaru <[redacted]@fb.com> wrote:

Apologies for my relative lull in correspondence here, I am pursuing one last bit of data from DS to complete the picture.

Last Wednesday, Ed confirmed that the MWB's Support & Inclusion team will take on H2 2020 goals to lower IIC prevalence on IGD. Chris [redacted] is also on the same page about this. Can you please link me to the CI HPM? I would be happy to help them reconcile the commitment from MWB. Separately, I will have periodic check-ins with Ed through H2 to understand progress/relationship building.

There are still some outstanding questions to Malia (Child Safety DS) on the finalization of the specific measurement and I believe that may inform that target metric for Ed. I have a follow up with Andrew [redacted] later today to understand where the DS teams aligned on the final numbers.

Here are the specific H2 2020 goals the MWB have undertaken:

1. **[P0, Q3] Proactive Warnings for IIC on IGD:** On Messenger, Proactive Warnings for IIC generate 1 million blocks of suspected IIC groomers per week. We will bring this feature to IGD to lower IIC prevalence, with support for both in-network and x-network threads.

2. **[P1, Q3 + Q4] Reduce prevalence of IIC on IGD by adapting measures on Facebook and Messenger:** Recent prevalence metrics indicate that Tier 2 IIC is 3x more prevalent on IGD than Messenger; on IGD, 500k non-adult accounts receive Tier 2 IIC every week. We will audit the measures in place on Facebook/Messenger and adapt high-impact measures for IGD. For example, we will consider restricting flagged IIC groomers from sending message requests to non-adult accounts.

From: Miki Rothschild <[redacted]@fb.com>
Sent: Tuesday, July 14, 2020 12:31 AM
To: Sayed Otaru <[redacted]@fb.com>; Yoav Shapira <[redacted]@fb.com>
Subject: Re: IIC on Instagram

Hey @Sayed, curious what's the latest on this one? On CI HPM it said they're looking to partner with either Msgr Integrity or IGWB, but not sure where we landed?

From: Sayed Otaru <[redacted]@fb.com>
Date: Monday, June 29, 2020 at 12:56 PM
To: Miki Rothschild <[redacted]@fb.com>, Yoav Shapira <[redacted]@fb.com>
Subject: Re: IIC on Instagram

Makes sense, Miki - thanks for your suggestion. I will keep you apprised if there are any noteworthy challenges following Ed's response and our subsequent discussions on the main thread.

Regards,
Sayed.

From: Miki Rothschild <[redacted]@fb.com>
Sent: Sunday, June 28, 2020 9:06 AM
To: Yoav Shapira <[redacted]@fb.com>
Cc: Sayed Otaru <[redacted]@fb.com>
Subject: Re: IIC on Instagram

Thanks for the heads up, Sayed!

I agree with Yoav that this is currently a non goal but also agree these numbers are quite alarming (500k per week). I think we want to be clear about the working model for who needs to address it — my sense is CI should provide the signal and a surface team, in this case growth (?), would apply the surface level action. So it's not actually a ton of work on Well-being (some coordination, perhaps adding the signal to WBSP?). So my suggestion is:

1) vet the numbers and solution
2) break out owners according to working model
3) discuss with the related teams and if they can't prioritize escalate decision to IG WB, Msgr WB and I2 leads.

Sent from my iPhone

> On Jun 26, 2020, at 4:18 PM, Yoav Shapira <[redacted]@fb.com> wrote:
>
> Child Safety is explicitly called out as a non-goal in our H2 plans. So if we can do something here, cool. But if we can do nothing at all, that's fine too.
>
> ---
>
> **From:** Sayed Otaru <[redacted]@fb.com>
> **Date:** Friday, June 26, 2020 at 16:17
> **To:** Miki Rothschild <[redacted]@fb.com>, Yoav Shapira <[redacted]@fb.com>
> **Subject:** Re: IIC on Instagram
>
> + Yoav (FYI)
>
> ---

From: Sayed Otaru <[redacted]@fb.com>
Sent: Friday, June 26, 2020 9:08 AM
To: Miki Rothschild <[redacted]@fb.com>
Subject: Fw: IIC on Instagram

FYI on the conversation around IIC thus far.

TL;DR occurrence of IIC on IGDM is ~500k (victims) a day in English markets - Child Safety DS needs to finalize the number per IG nuances but anticipate the number to remain in/around the ~500k/day. CI team are not able to take a goal here in H2 2020 and Messenger Well-being (Ed [redacted] is the PM) may be able to take a goal. We'll check in next week.

Regards,
Sayed.

From: Edward [redacted]@fb.com>
Sent: Thursday, June 25, 2020 4:45 AM
To: Sayed Otaru <[redacted]@fb.com>; Malia [redacted]@fb.com>; Chris [redacted]@fb.com>
Cc: Dhiraj [redacted]@fb.com>; Andrew [redacted]@fb.com>; Nicola [redacted]@fb.com>; Akum [redacted]@fb.com>
Subject: Re: IIC on Instagram

Hi all,

Lots to consider in this thread and still working through implications for our roadmaps. I think it is fair to assume we will be tackling this to some degree in Q3, most likely through rollout of Proactive Warnings to IGD . If we have confidence in this metric, we should publish it and hold ourselves accountable to it. I'm not expecting to make much movement towards solution #1 in Q3, given the org priority to ship interop, but Q4 will hopefully give us some bandwidth to do this (we'll use Q3 to scope out the right approach).

Will circle back start of next week once our roadmaps for Messenger Well-being are in a more final state.

Thanks,
Ed

From: Sayed Otaru [redacted]@fb.com>
Date: Thursday, 25 June 2020 at 07:49
To: Malia [redacted]@fb.com>, "Chris [redacted]@fb.com>
Cc: Dhiraj [redacted]@fb.com>, Andrew [redacted]@fb.com>, Nicola [redacted]@fb.com>, Edward [redacted]@fb.com>, Akum [redacted]@fb.com>
Subject: Re: IIC on Instagram

Thanks for the clarification, Chris and Malia.

Chris, how did your roadmap review with CIFM go and are there any related noteworthy discussions or action items from that review?

Ed, can you please let us know if your team is able to support the work in H2 2020 to help mitigate this issue?

Regards,
Sayed.

From: Malia [redacted]@fb.com>
Sent: Wednesday, June 24, 2020 3:10 AM
To: Chris [redacted]@fb.com>; Sayed Otaru <[redacted]@fb.com>
Cc: Dhiraj [redacted]@fb.com>; Andrew [redacted]@fb.com>; Nicola [redacted]@fb.com>; Edward [redacted]@fb.com>; Akum [redacted]@fb.com>
Subject: Re: IIC on Instagram

Just to add a few things onto what Chris said.

A. Small correction to your understanding of the metric. **The metric is ~500k victims per DAY in English markets only** (so far; we'll add additional markets as more capacity comes online). So we expect the true situation is worse.

B. However, this is a preliminary metric for a few reasons. **Although this same method has been carefully vetted for FB, there are a couple of differences we want to ensure aren't affecting it.**

1. We want to tighten up any differences in defining "sex talk" between our prevalence labelers (PDO) and reps that enforce IIC in IG Direct normally (CO).
2. Age estimation is better in FB than in IG.

To test the effects of both of these possibilities, we're sampling a number of our initial IG prevalence labels to send through our normal CO enforcement stream. From this we'll be able to RCA any discrepancies and make corrections. In addition, you're probably aware that a new "IG adult" classifier is due to ship soon that we hope will be more precise than current predictions.

Nevertheless, the initial metric values match the intuition of our CO and investigator SMEs, who have observed higher incidence in IG anecdotally. We think mitigation is really important here.

From: Chris [redacted]@fb.com>
Sent: Wednesday, June 24, 2020 2:07 AM
To: Sayed Otaru [redacted]@fb.com>
Cc: Dhiraj [redacted]@fb.com>; Andrew [redacted]@fb.com>; Nicola [redacted]@fb.com>; Edward [redacted]@fb.com>; Akum [redacted]@fb.com>; Malia [redacted]@fb.com>
Subject: Re: IIC on Instagram

Thanks Sayed, copying Ed (Messenger Well-Being PM), Akum (Child Safety EM) Malia (Child Safety DS), and Nicola (I2 eng lead)

To provide a little more information and a couple of clarifications:

1) Details of the measures we took to reduce IIC Tier 2 on Messenger can be found here, total prevalence reduction was ~23%. We apply these friending restrictions based on a set of encryption-resilient heuristics (adults reaching out to minors and getting NFX), rather than a classifier. This has been a *major* win for Child Safety, and an important step for us demonstrating to external stakeholders than we can mitigate inbox-based abuse by preventing it at source, which is a key narrative ahead of E2EE.

2) In H2 2020, the focus for the CIPCO Child Safety team will be on a narrow set of extremely high-priority problems (roadmap here). In particular, I want to call your attention to the follow *non*-goal: "*we will hold the line on IIC T2 prevalence on IG & Messenger, and respond to any spikes, but will not be goaling on prevalence reduction*". This line was written before we had a clear idea of the scale of the problem on IG, which we now know to be approximately 3x the size of the Messenger problem.

3) We are reviewing our roadmap with CIFM today, who will likely agree that the Child Safety goals are correctly stack ranked, but will want to see a clear plan and ownership for IGDM-based IIC T2 - the number is worryingly high. We don't need to give them crystal-clear execution details, but I'd like to let them know the thoughts of this collective group.

4) Having discussed briefly with Ed yesterday, it looks like something his team way be able to help with (@Ed please chime in). In my opinion, we could take a substantial dent out of prevalence by replicating the principles which worked so effectively on Messenger (a couple if which you have described) below.

From: Sayed Otaru <[redacted]@fb.com>
Sent: Wednesday, June 24, 2020 5:48 AM
To: Chris [redacted]@fb.com>
Cc: Dhiraj [redacted]@fb.com>; Andrew [redacted]@fb.com>
Subject: IIC on Instagram

Hi Chris,

Note: *I have CC'ed Dhiraj and Andrew - who head up Engineering and DS support respectively for IG Well-being Safety.*

I had conversations with Dhiraj and Andrew, as well as another with Miki. We would like to clarify the 'Inappropriate Interactions with Children' (IIC) problem you described to me earlier today and understand some more details around the possible paths forward. Here is a recap of our discussion as I understand, please correct any mischaracterizations that I may have below:

Problem Summary
Messenger implements a classifier that predicts adult accounts that engage in IIC with underage accounts. Messenger subsequently restricts those flagged adult accounts from being able to make friend requests to all non-adult accounts. Messenger does not de-platform flagged adult accounts and believes the restriction is a sufficient deterrent. Instagram (IG) has historically not been able to classify non-adult accounts, until recently. The CI team has concluded a study (the numbers have not yet been formalized) and currently show that **500,000 IG underage accounts receive IIC on a weekly basis**. The IG prevalence number is 3x Messenger's.

Why Now?
The CI roadmap review is this Thursday, Jun 25th. This recently obtained data understanding the size of the problem on IG is likely to be discussed with leadership and we (CI & IG) should have a plan on who is accountable for driving this problem down and by when.

Potential Solutions
1. IG could replicate the user experience from FB/Messenger and restrict classified adult IG accounts from making 'follow requests' to underage IG accounts.
2. IG can think through an alternate product experience to address this problem.

Resourcing Challenge
1. *Option #1* CI will not be able to prioritize taking on this work on IG in H2 2020 due to other higher priorities.
2. Option #2 Chris [redacted] and team may possibly be able to take on this work in H2 2020.
3. *Option #3* IG would like to:
    1. 1) First better understand the sizing of this problem on IG,
    2. 2) Review the potential product experience options for protecting non-adult accounts,
    3. 3) Understand what pieces of work CI would own for any for the eventual solution and what pieces IG would need to own.
    4. 4) Prioritize against our H2 2020 plans (with US 2020 elections related work being our top-most priority).

In the event that neither CI, IG, or Ed's team is able to prioritize remedying this experience on IG in H2 2020, we may need to elevate the discussion to our respective leadership teams to align on tradeoffs/resourcing.

Regards,
Sayed.