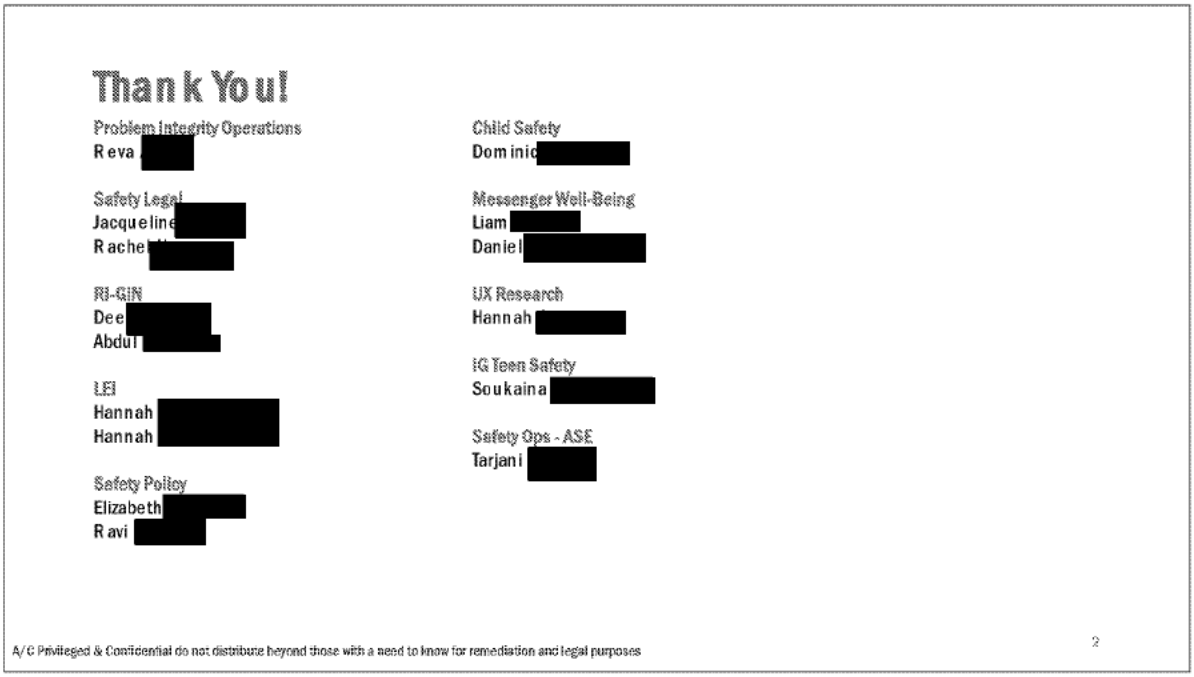
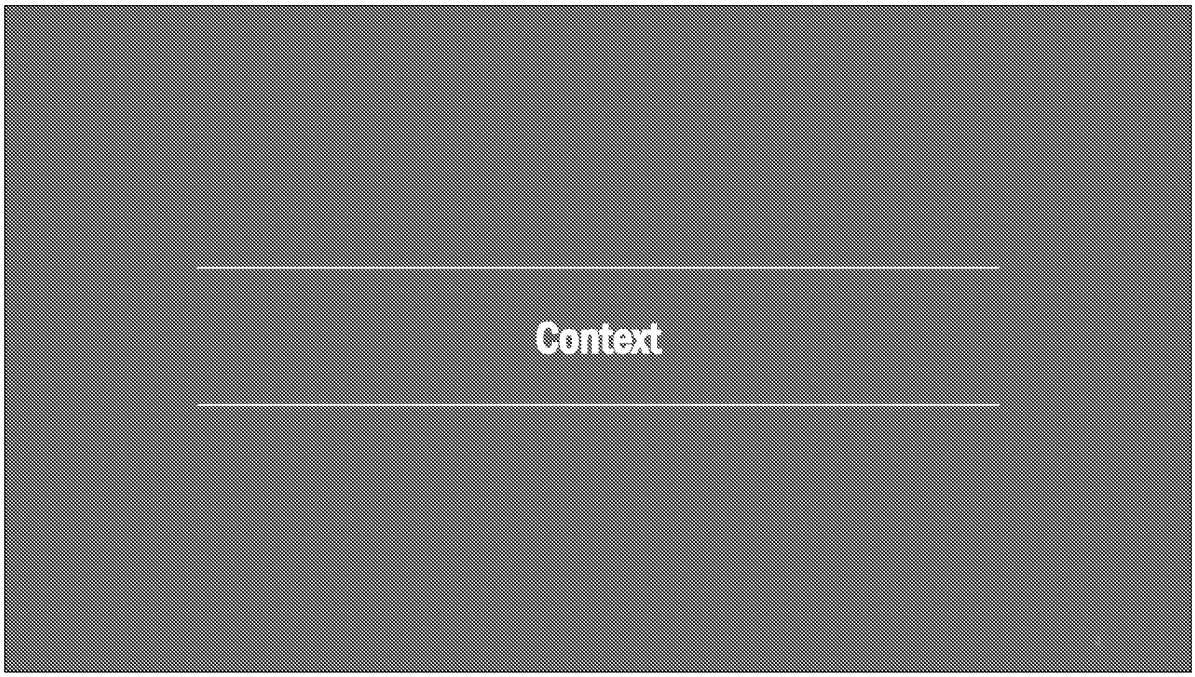
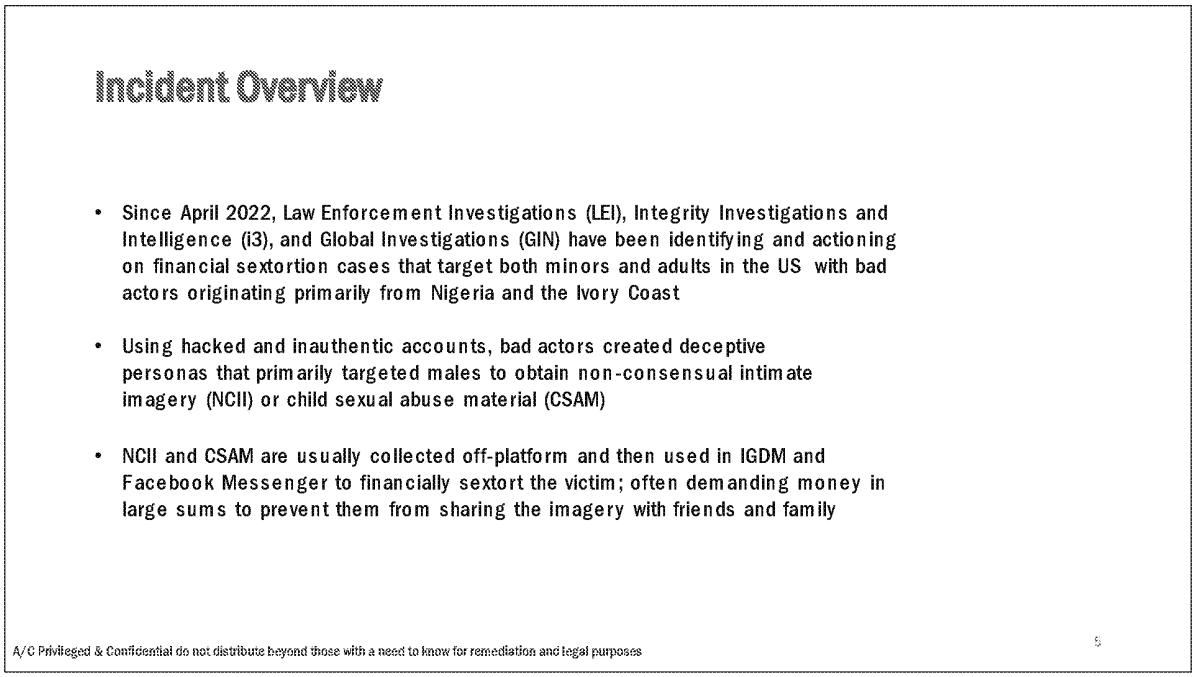
Hello everyone. My name is Taylor [redacted] and I'll be presenting this risk intelligence integrity investigation on financial sextortion in instagram direct message.



Before we begin, I want to thank all of our cross-functional partners that we met with as we could not have completed this investigation without all your help and insights.

Now, here is the agenda for the presentation.


Context

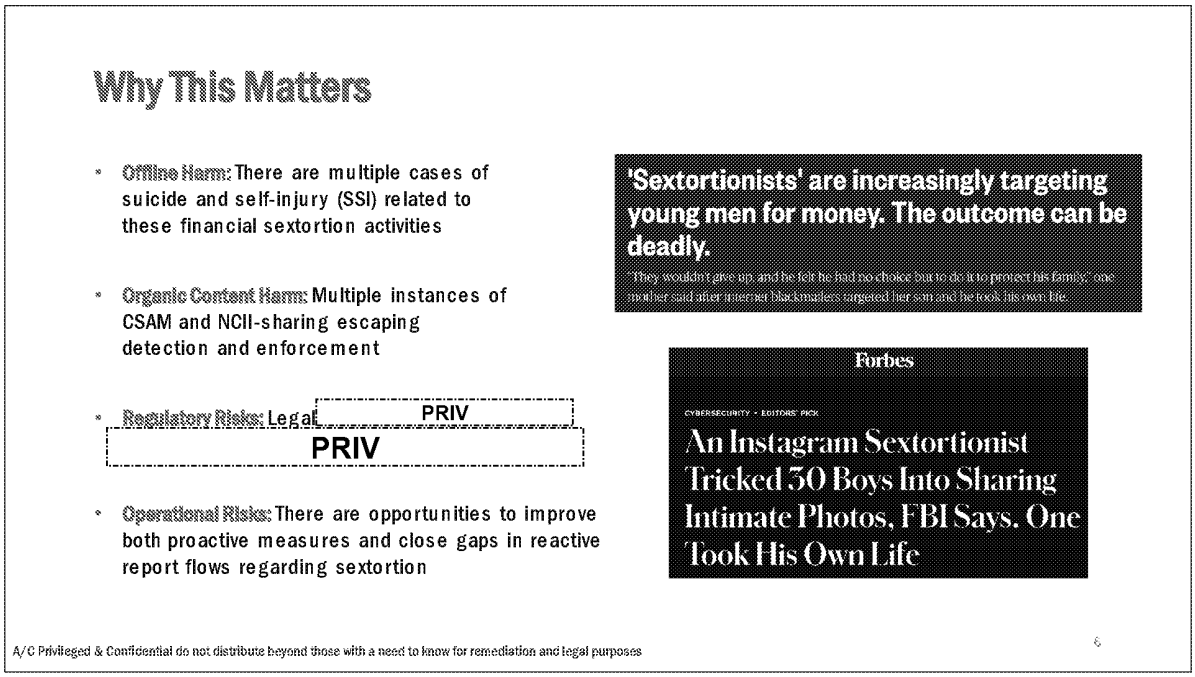


Now, let's talk through the overview of this case.

Earlier this year, Law Enforcement Investigations, i3, and GIN began identifying and actioning on financial sextortion cases that were targeting both minors and adults in the US. The majority of these cases were coming from bad actors located within Nigeria and the Ivory Coast.

Using hacked and inauthentic accounts, these bad actors were creating deceptive personas that would target users and manipulate them into sending non-consensual intimate imagery, NCII, or CSAM if the victim was a minor, and use that to sextort the user. For the purpose of this investigation, the deceptive persona activity is out of scope for this investigation and will focus on the sextortion activity.

In a lot of cases, the actual NCII or CSAM was collected off-platform in places like Snapchat. After the bad actor obtained the imagery, the conversation would move back to our private surfaces, where they would demand large sums of money. If the victim did not comply immediately, they would threaten to share the imagery with close friends and family.



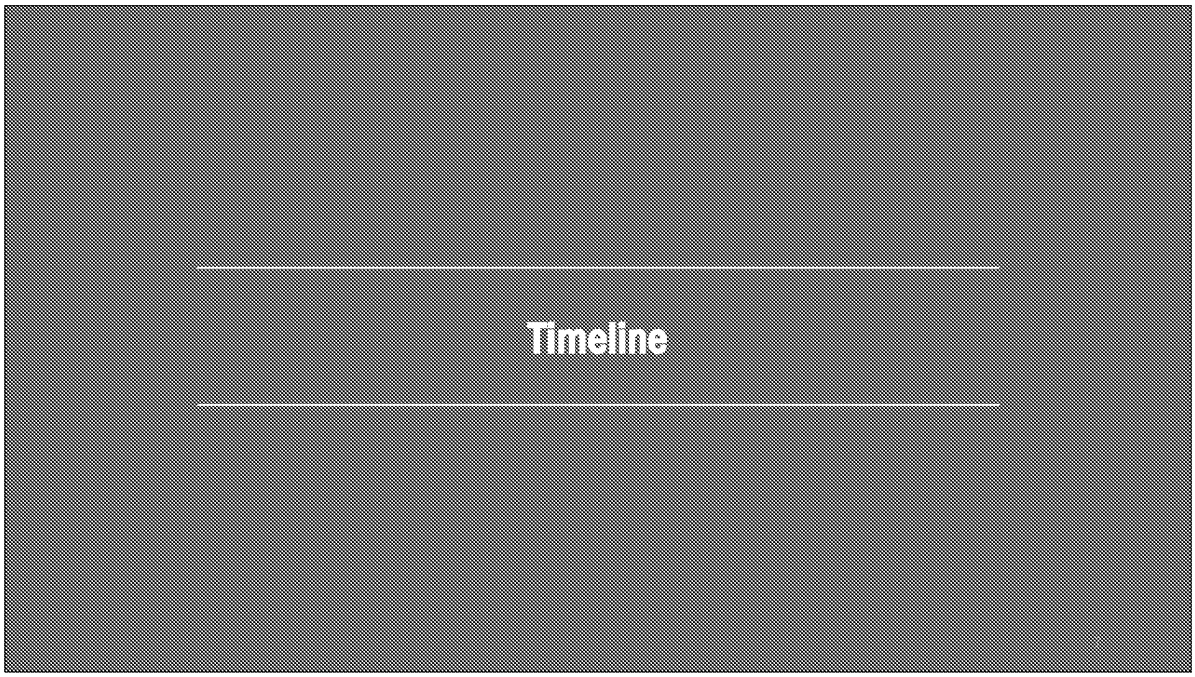
So, why does this matter and why did we investigate this issue.

First off, there are multiple confirmed cases of suicide in the US related to these instances of sextortion which has garnered public and media attention, as evident from these two media headlines.

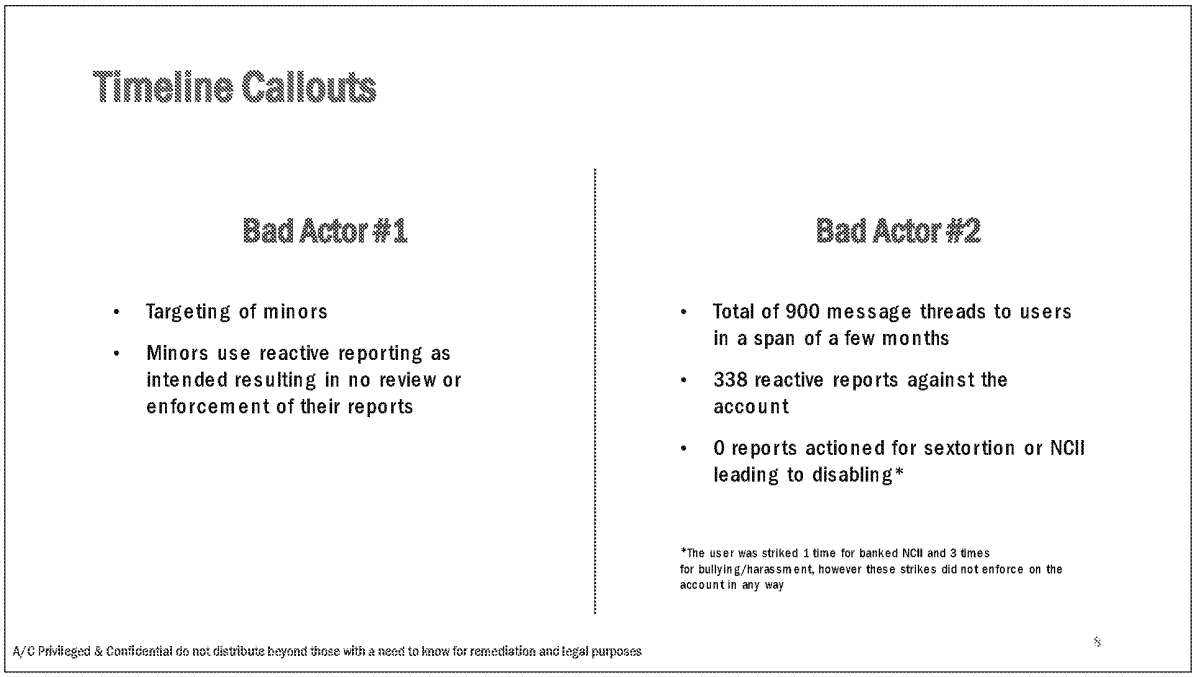
We have multiple instances of both NCII and CSAM sharing occurring on our platforms that escape detection and enforcement creating organic content harm.

PRIV lega PRIV

Finally, there are opportunities to improve both our proactive measures and close gaps in reactive report flows regarding sextortion, especially as we prioritize growth and user voice in the company moving forward.



Now, we will walkthrough two bad actor timelines to give a clear picture of a small portion of the events that took place. I'd like to start with a few callouts regarding these bad actors. Both timelines will reflect bad actor sampling that we performed in this investigation. These two bad actors are sampled from a strategic network disruption performed by LEI. Both bad actors we chose were disabled for confirmed financial sextortion through that SND. I'd like to reiterate the great work by here LEI, i3, and GIN in identifying and disabling a large network of bad actors.



As we jump into the timeline portion, we have decided to create 2 separate timelines to illustrate different problems across our bad actor types.

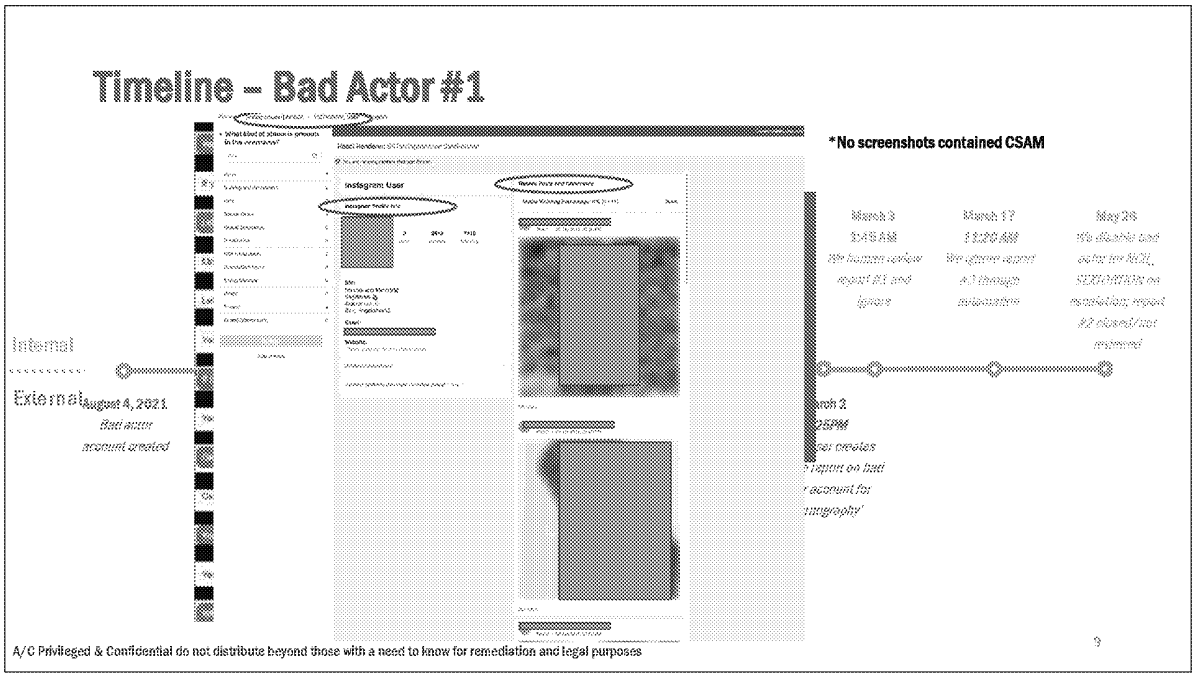
Bad actor number one, you will see, knowingly targeted minors after asking their age.

Subsequently, we see these minor users using our reactive reporting as intended, however resulting in no review or enforcement of these reports

Taking a look at bad actor number 2 will give you a picture of the amount of impact a single bad actor can have when we do not address gaps in our reactive review flows. This bad actor had a total of 900 message threads within a few months

With 338 reactive reports against both the account and message threads.

Many of these reports contained signals of sextortion. However, you'll see none of them were actioned for sextortion or NCII leading to disabling. We will call out the user was striked for bullying and harassment and banked NCII, but these strikes did not enforce on that bad actor in any way. Now we can jump into the first timeline.



(click) Starting with a common theme, the bad actor creates the account months before there is any activity.

After 7 months, the messaging and targeting of the minor user begins.

You can see the typical messaging through this screenshot, requesting a picture exchange and asking for their Snapchat account.

(click again) We see signals of sextortion only hours after the messaging began.

Here is an example of the signals of sextortion we saw. In addition to the typical discussion, signals included messages from the bad actor demanding an amount of money and threatening they will share pictures with friends, with the minor repeatedly stating they don't have enough money and the don't want them to share the imagery.

(click again) 8 minutes later, the minor user creates a reactive report against the bad actor's account.

I also want to talk through the reporting flow for users and our reviewers. When there is inappropriate activity taking place on our platforms, the victim can submit a report on the user or the message thread. When the report is on the user, our reviewers are unable to see the message thread, even if that is where the potentially violating content is taking place, as seen on this screenshot. However, if the review is on the message thread itself, our reviewers can view the thread.

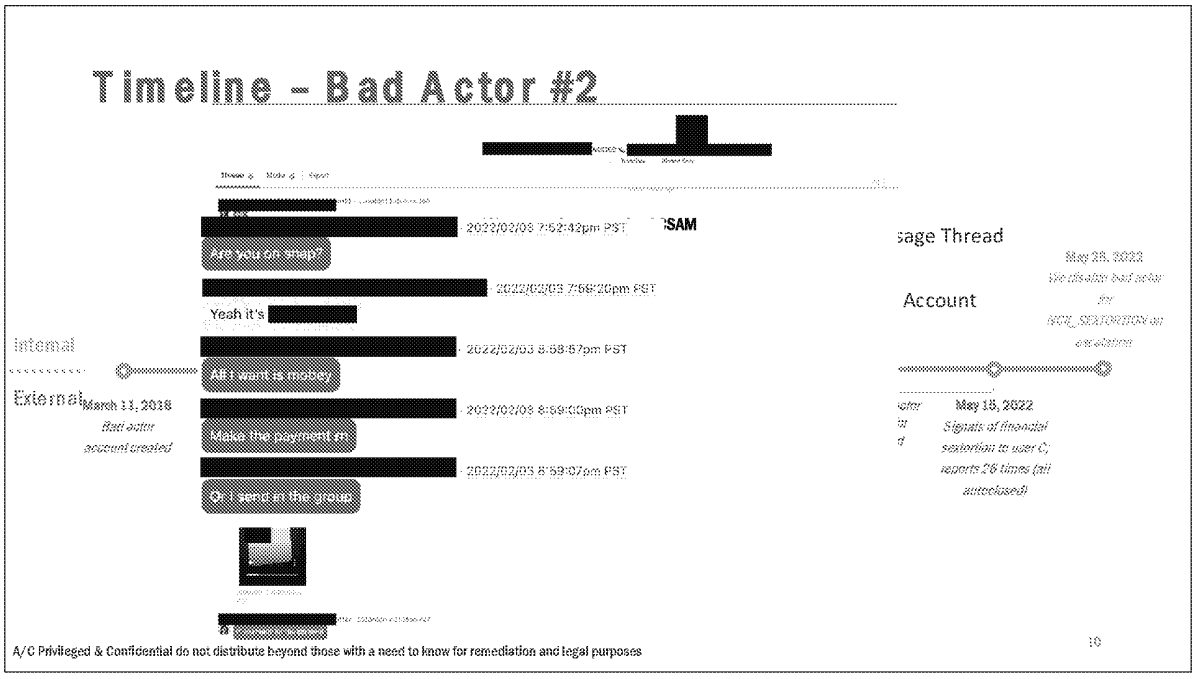
(click again) 13 minutes later, the minor user creates another reactive report but this time on the message thread, which would be the appropriate way to report this harm as reviewers could view the thread.

Just over 30 minutes later, the minor user creates another report on the bad actor's account.

Now within our reporting flow, the first report on the user account for cyberbullying is ignored by human review hours after receiving the report. This outcome was expected as there was no violating content on the user's account.

14 days later, we ignore the third report on the user's account through automation. This was again the expected decision as the activity was taking place in the message thread.

The second report from our user was correctly placed on the message thread and never reviewed by either automation or human review. Therefore, we did not catch the signals of minor sextortion that were present. We will review this and the potential mitigations further in the cause map. We finally disable this bad actor for NCII_SEXTORTION on escalation through project doubtfire performed by LEI.



Now moving on to the second bad actor. As a reminder, we want to focus on the shear number of reports against the bad actor and the lack of human review on any sextortion-related report.

Similar to the first timeline, this bad actor creates an account in 2018, over 3 years before any activity is recorded. In 2022, we see the actors begin messaging multiple male users. I want to reiterate we will show the experience of only three users here, however many more were targeted.

Beginning on February 3rd of 2022, we begin seeing signals of financial sextortion to user A.

Here is a screenshot example of the signals of sextortion present in these threads. As you can see, the conversation moves off-platform immediately and just one hour later, returns demanding money.

Another common method our bad actors used ncluded obtaining imagery off-platform, returning to IGDM and creating a group titled "User X nudes" containing a video with the user's snapchat page and their nude imagery. (click twice)

Between February 3rd and February 27th, this one bad actor was reported a total 179 times including 123 reported correctly on the messenger thread.

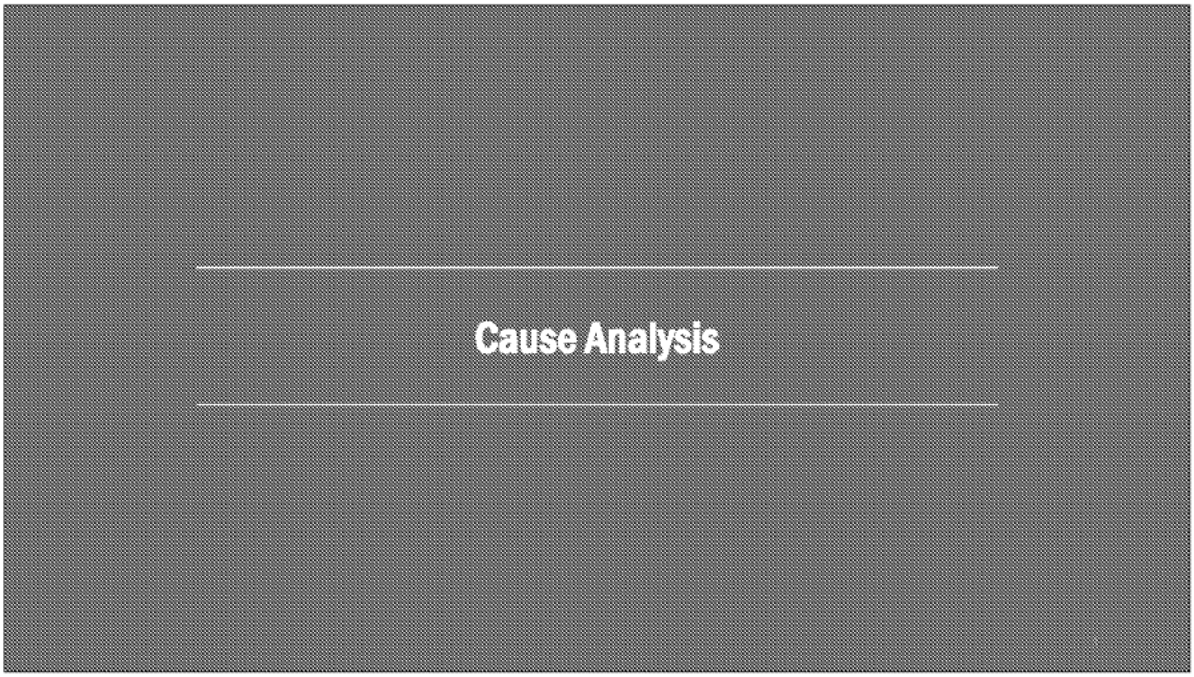
Moving on another month to March 25th, our bad actor was reported 43 additional times. As we did not conduct exhaustive review of all 900 message threads from this bad actor, we reviewed a handful of one on one and group chats during this timeframe and found 8 threads with signals of financial sextortion. Additional group names we identified in the bad actor's threads indicate this occurred many more times.

On March 4th, our next user, user B, reported this activity 18 times in the message thread and none were actioned for sextortion or NCII.

Through May 15th, the bad actor activity was reported another 117 times. Our final user in this timeline, user C, reported this sextortion message activity 26 times and all were autoclosed with no action. We do want to mention again the bad actor did receive 3 strikes for b/h and one for banked NCII during this time period, but these had no affect on the use of the bad actor account throughout this process.

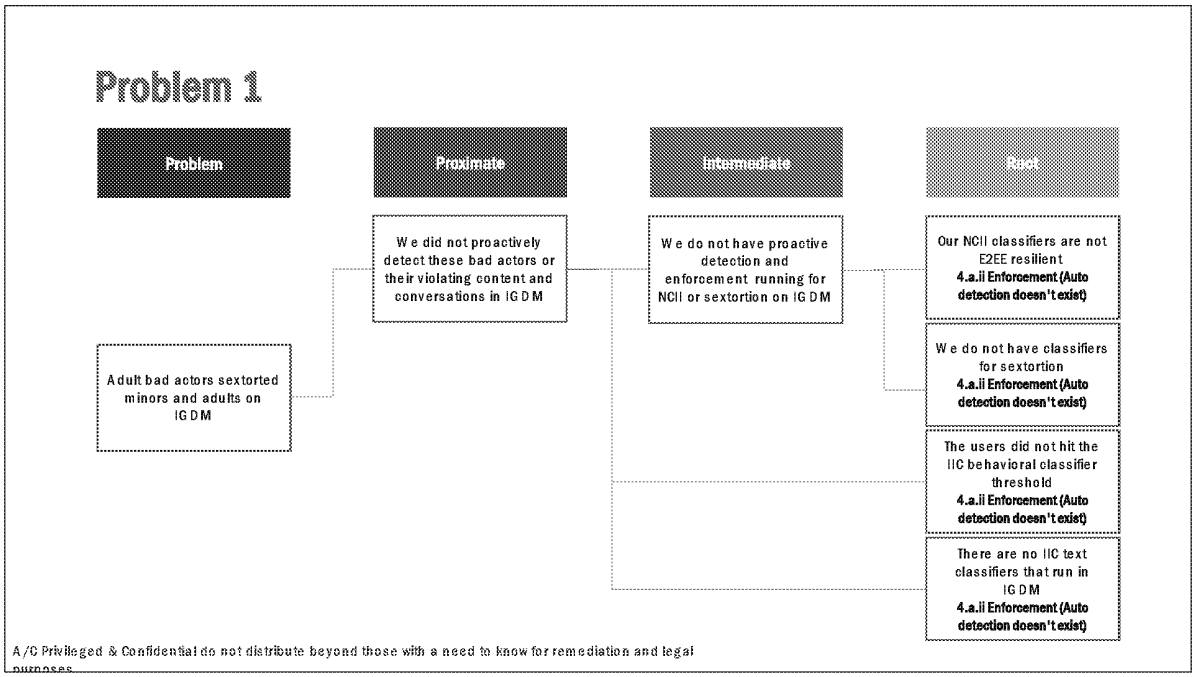
And then finally on May 26th, the bad actors is disabled on an escalation for NCII_SEXTORTION through LEI's SND.

Just to reiterate, this particular bad actor was reported 328 times with 210 were reported within messenger threads, 128 were against the account and none of them actioned or enforced for sextortion or NCII.



Before we walk through our cause analysis, we would like to mention there are currently strategic plans in place to help solve for many of the root causes we will talk through. Our partner Reva[redacted] will talk through many of these at the conclusion of this presentation.

Now for this RCA, we continued the focus on the sample of 2 bad actors.



(click again) Our first problem statement is that adult bad actors sextorted minors and adults on IGDM.

The first proximate for this problem is that we, as in Meta, did not proactively detect these bad actors or their violating content and conversations in IGDM

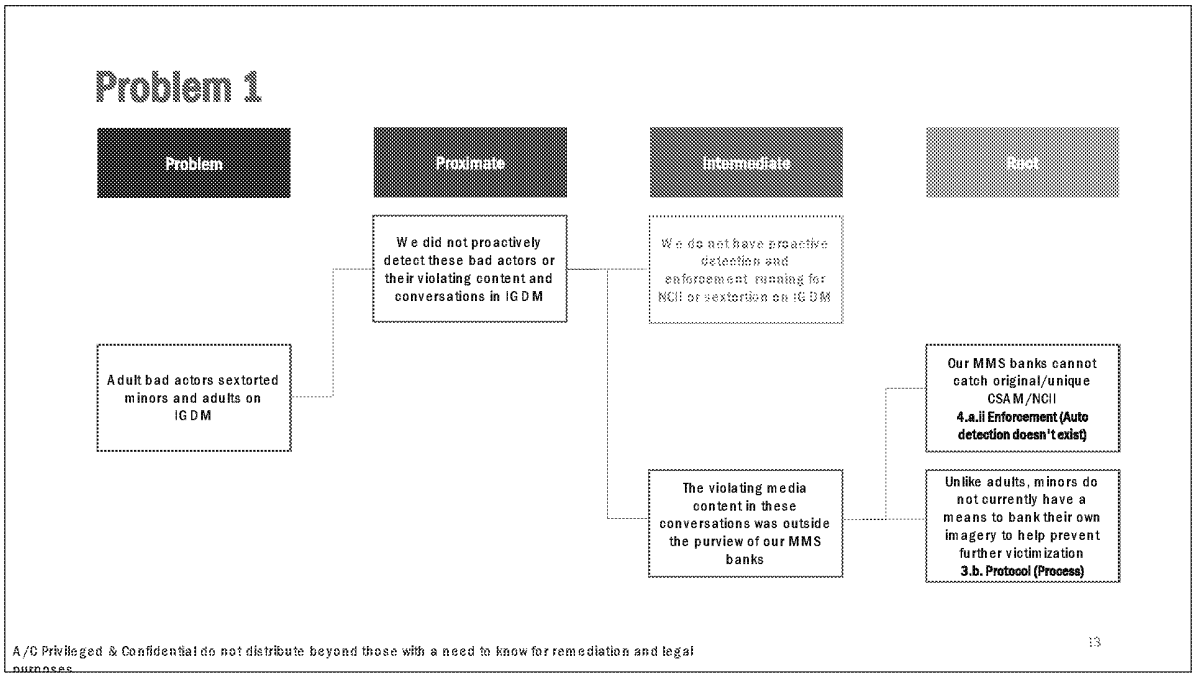
And that is because we do not have proactive detection and enforcement running for NCII or sextortion in IGDM

We don't have this proactive detection because our NCII classifiers are not end to end encryption resilient, meaning they can't be active once the company is required to be E2EE.

We also do not have any classifiers running for sextortion specifically.

The third root cause for this proximate is that the users did not hit the IIC, or inappropriate interactions with children, behavioral classifier threshold.

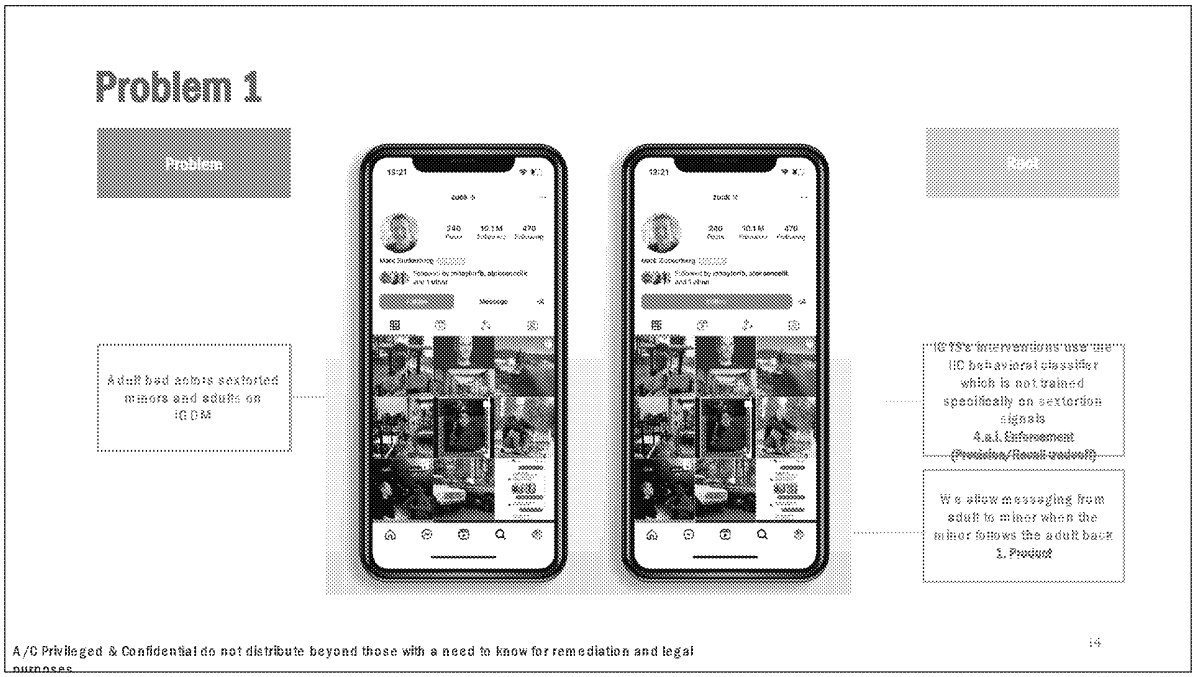
With the final root cause being there are no IIC text classifiers that run in IGDM. Now, there are various archetypes in classifiers that are currently launched or planned such as UnIS and MaCSA. Each of these may contribute to partial mitigation of these risks, however, we do not have any solutions that completely cover sextortion. Again, Reva will go into more detail regarding these mitigations in a bit.



(Click again) Our second intermediate for this proximate is that the violating content in these conversations was outside the purview of our MMS banks.

And that's because our MMS banks cannot catch original or unique CSAM or NCII, which is the expected process for these banks.

And unlike adults, minors do not currently have a means to bank their own imagery to help prevent further victimization. There are plans from the safety policy team to support NCMEC in the development of a platform for minors and guardians to proactively hash intimate images and videos of minors and share the hashes with participating companies in order to prevent uploads of CSAM in H2 of this year. We also recommend ensuring users are aware of these updated measures.



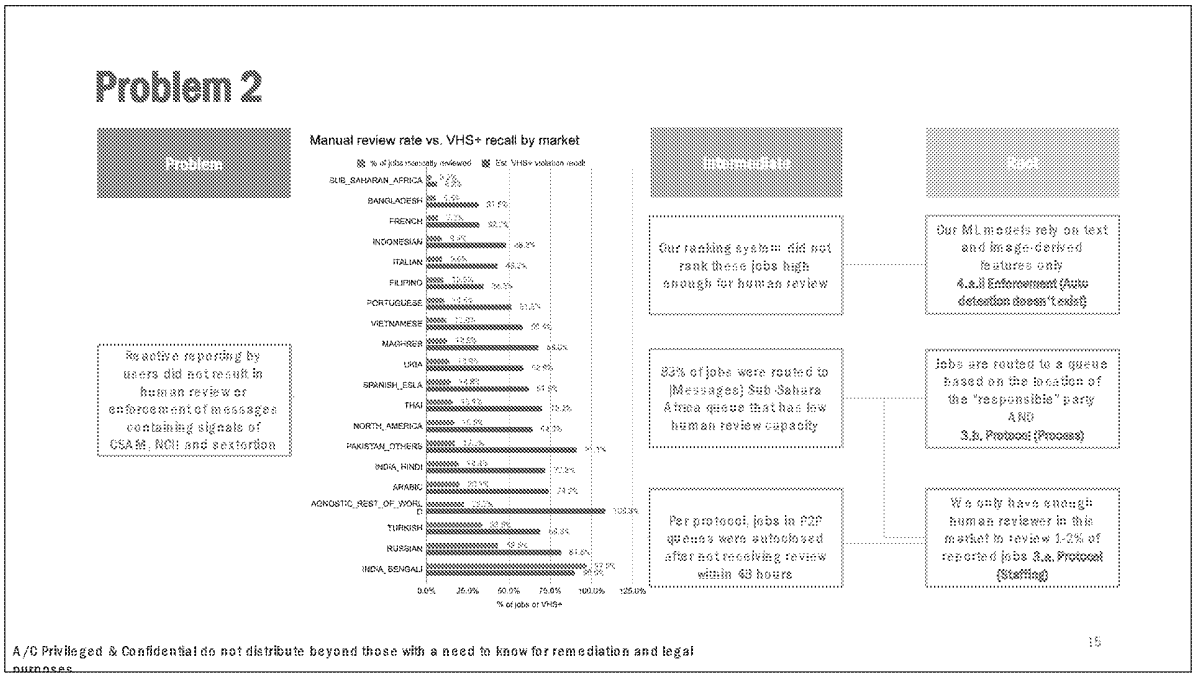
(click again) Now moving onto our second proximate for this problem. These adult bad actors were able to message minors in IGDM.

As the minor-specific product interventions did not prevent these interactions.

And that's because IG teen safety's interventions use the IIC behavioral classifier which is not trained specifically on sextortion signals.

And we also allowed messaging from adult to minor when the minor follows the adult back. This is because not all adult to minor interactions are risky. However, the child safety team has implemented additional frictions that help prevent risky adults messaging minors, effectively hiding the primary messaging button from risky adults.

Here is an example of that new friction with the message button hidden on the right. We tested this functionality with one bad actor to see if their interactions would have hit the risky-adult-to-minor and it did not appear to meet the threshold so it is unclear how much this mitigation would impact these specific cases.



And that wraps up our first problem so moving on to problem number 2. The reactive reporting by users, including multiple reports from individual users, did not result in human review or enforcement of messages containing signals of CSAM, NCII, and sextortion.

Our first proximate is that 0% of these IGDM reports against these bad actors were given human review.

We did this because our ranking system did not rank these jobs high enough for human review.

One root cause for this intermediate is that our ML models rely on text and image-derived features only. Videos are features that are currently not in the model which is what one of our bad actors relied on in their sextortion activity. The messenger well-being team has a H2 goal to introduce video-based features from reported conversations into our ML-based automation pipeline. It's also important to note that although there are additional root causes, due to data retention policies, we were unable to fully understand why these specific jobs were ranked too low for human review.

Also, 83% of these jobs were routed to the sub-Sahara Africa queue that has low human review capacity

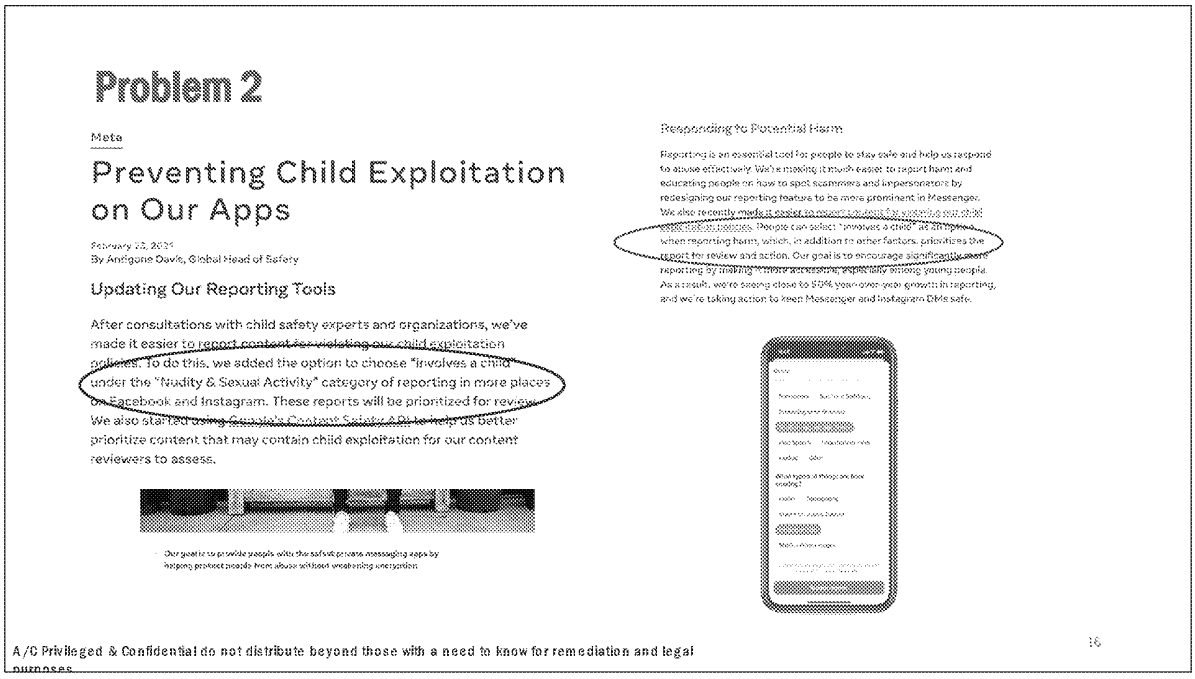
This is because our jobs are routed to a queue based on the location of the responsible party, in this case Nigeria.

And we also only have enough human review in this market to review 1-2% of reported jobs.

Here's a screenshot of that manual review rate versus very high severity plus recall by market, noting the low review rate from SSA at the top.

(click again) Our final intermediate is jobs in p2p queues were autoclosed after not being reviewed within 48 hours. This has recently been changed to jobs remaining open for 10 days.

This intermediate has the same root cause as the previous, not having enough human review in this market.



(click again) The next proximate is even when users report using VHS+ tags, there is no guarantee this will prioritize the report such that it will be reviewed. I'd like to pause here and signpost a delta between our public commitment to reviewing this harm and what users here are experiencing.

In the last year, we released at least two newsroom articles explaining our commitment to users making it easier to report harm for minors and the commitment that we will prioritize these reports for review. This is a screenshot of one article from February of 2021 highlighting that

And then this one from December of 2021, highlighting the same thing.

(click again) As we dig into the root causes of this delta, we can continue with the intermediate, which is that our current tags do not contribute greatly to rank in our queues

Because the data shows user-selected tags do not give us a good idea of the harm taking place. In our upcoming discussion, we will use this intermediate cause to tee up one of our discussion questions regarding how we need to approach this problem in light of this root cause.

And finally, we deprecated the blackmail contact form that formerly allowed users to specifically report sextortion

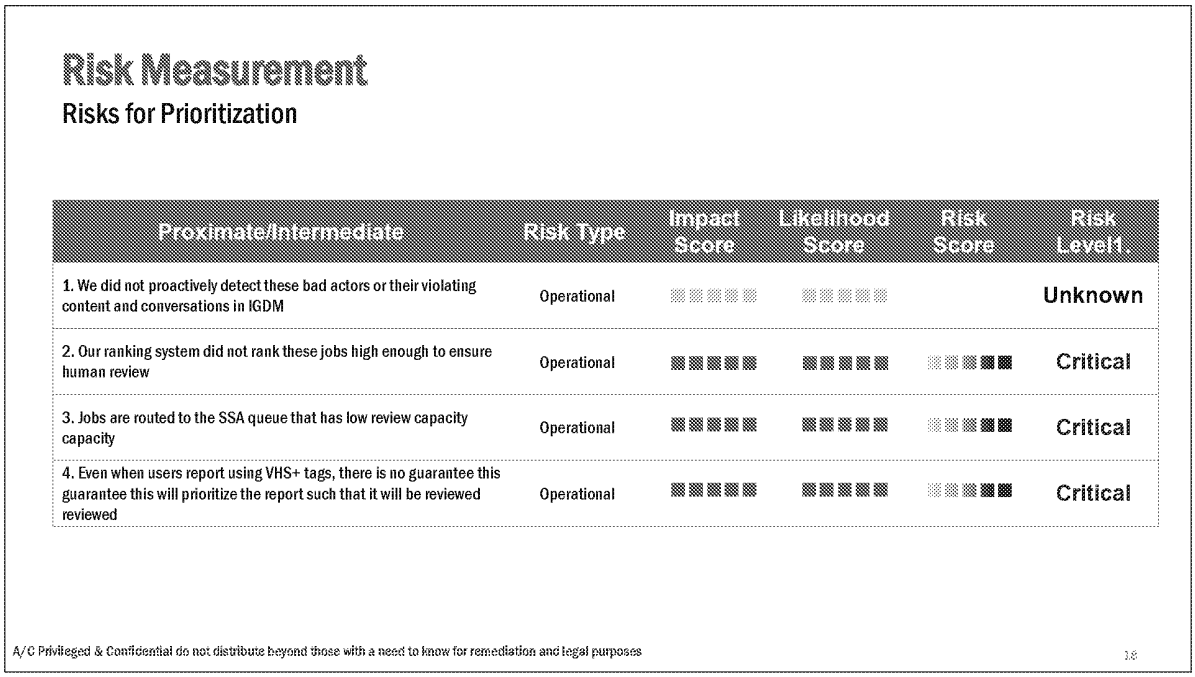
And we did that because of a lack of reviewer bandwidth. In summary, problem 2 displays several root causes that you'll see shorty contribute to a critical risk score. Messenger well being has road mapped items that could potentially update these report flows which will be called out in the resolutions slide at the end. However, we will focus on creative solutions for this problem statement in our discussion.

PPPEP Breakdown

Root Cause	% of Total
1. Product	**8% (1)**
2. Policy	**0% (0)**
3. Protocol	**31% (4)**
a. Staffing	2
b. Process	2
c. Documentation of process	
d. Education about process	
4. Enforcement	**61% (8)**
a. Issue was not detected or reported	
i. Automated detection precision/recall tradeoff	3
ii. Automated detection doesn't exist	5
iii. Other issue	
b. Issue was detected/reported, but	
i. Not looked at in time	
ii. Automated action was incorrect	
iii. Issue with internal tool or product	
iv. Human review error	
5. People Disagree	**0% (0)**

17

Alright here's our PPPEP breakdown, and you'll see majority of our breakdown is Enforcement.



Risk Measurement

Risks for Prioritization

Proximate/Intermediate	Risk Type	Impact Score	Likelihood Score	Risk Score	Risk Level
1. We did not proactively detect these bad actors or their violating content and conversations in IGDM	Operational				Unknown
2. Our ranking system did not rank these jobs high enough to ensure human review	Operational				Critical
3. Jobs are routed to the SSA queue that has low review capacity capacity	Operational				Critical
4. Even when users report using VHS+ tags, there is no guarantee this guarantee this will prioritize the report such that it will be reviewed reviewed	Operational				Critical

A/C Privileged & Confidential do not distribute beyond those with a need to know for remediation and legal purposes

18

Here we want to callout our risk scoring and how that contributes to the prioritization of the risks we identified. In regard to "We did not proactively detect these bad actors or their violating content and conversations in IGDM" this has been given a risk level of unknown as there is currently no prevalence metric available to measure the likelihood. Therefore, we will prioritize this problem alongside our critical risks to amplify the need for this metric. But please note it is currently being built of by the messenger well-being team. Additionally, our latter three causes all come in at critical risk levels. We want to highlight the importance of tackling these issues collectively rather than disparate issues as the way we enable our users to report could have strong downstream affects on our remaining two causes regarding how we rank jobs and staff these queues.

Now, Reva will walk through current and future interventions to tackle sextortion.

In H2 '22, teams drove multiple interventions that directly or indirectly tackle sextortion.

Internal Product Interventions

Workstream	Description	Team	Theme	Status	Adult	Minor	FB	IG
1 Baseline & reduce IIC risky thread prevalence	• Build & validate a heuristics-based definition for risky adult-minor interactions • Enable product interventions, including IG message button removal for suspicious adults seeking to connect with minors	Child Safety & UnIS	Stand up prevalence measurement & prevent harm	Complete	✗	✓	Partial[1]	✓
2 Reduce connectivity between suspicious adults & teens	• On MSGR: launch Composer Block preventing suspicious adults & teens from messaging each other • On IGD: launch Composer Block preventing minors from messaging unconnected high-risk adults	MWB Teen Safety & UnIS	Prevent harm	In Progress	✗	✓	✓	✓
3 Improve enforcement accuracy of T1 IIC & minor sextortion user reports	• Increase E2E escalation recall & improve enforcement outcomes for T1 IIC and minor sextortion user reports, including disable & manual NCMEC report volumes • Expected to hit target of 80% E2E recall by EOH	Child Safety	Improve enforcement	In Progress	✗	✓	✓	✓
4 Review more VHS+ & more Child Safety reports	• Review more harm through automation, incl. releasing IIC classifier V4 for MSGR & new auto-enforcing IIC classifier for IGD • Improve ranking & review for VHS+ reports, incl. extending 48hr auto-close rule to 10 days	MWB Reporting & Review (R2)	Improve enforcement	In Progress	Partial[2]	✓	✓	✓
5 Ship Safety Notices for sextortion	• Prompt teens to report accounts after they block someone and send them safety notices with information on how to navigate inappropriate messages from adults	MWB Teen Safety	Provide user support	Complete[3]	✗	✓	✗	✓
6 Expand coverage of sextortion Sigma policies over users from suspicious countries	• Improve precision by overhauling existing Sigma policies wrongly enforcing against 6K accounts/day • Activate/launch Sigma policies to detect sextortion accounts in key locales (including India, Ivory Coast, Indonesia, Yemen, Philippines)	Adult Sexual Exploitation	Improve precision & detection recall	In Progress	✓	✗	✓	✓

1 Prevalence measurement is available for both FB & IG, but product interventions are currently limited to IG due to blockers in launching planned Messenger composer block intervention.
2 VHS+ ranking improvement includes adult sextortion but the extent of coverage is currently unknown.
3 MWB Teen Safety is planning to de-prioritize further investment in Safety Notices in H1 '23 due to modest impact of the launch & new priorities for next half.

In H2 '22, teams drove multiple interventions that directly or indirectly tackle sextortion.

Understand | Investigation & External Partnerships

	Workstream	Description	Team	Theme	Status	Adult	Minor	FB	IG
7	Support NCMEC in launching their "Take It Down" Portal	• Support NCMEC in launching their "Take It Down" portal, to empower teens to prevent the distribution of their self-generated CSAM • Expected launch by EOY or early H1 '23	NCMEC with Safety Policy support	Provide user support	In Progress	✓[4]	✓	✓	✓
8	Supported education & prevention campaign for teens in partnership with Thorn	• Support Thorn in creating educational materials that reduce shame and stigma surrounding intimate images • Support Thorn in empowering teens to seek help and take back control if they've shared intimate images or are experiencing sextortion • Disseminate campaign through influencer activation and IG campaign	Thorn with Safety Policy support	Raise awareness & provide user support	In Progress	✗	✓	✓	✓
9	Conduct a Strategic Network Disruption of financial sextortion actors	• Known as Project Doubtfire, discovered actors engaging in financial sextortion using Snapchat handles to move victims off-platform and entice them to produce nudge imagery • Disrupted 511 accounts for sextortion, reported 88 users to NCMEC, & escalated 25 Snapchat usernames to Snap for redress • Launched non-overlook Sigma rules to auto-disable ~262K sextortion accounts based on behaviors & profile picture reuse	LEI, LEAP, & i3	Improve enforcement & prevent harm	Complete	✓	✓	Partial[5]	✓
10	Identify E2EE-resilient signals for detecting minor sextortion accounts	• Conduct a large scale review of minor sextortion offender accounts & develop a signal grid • Provide recommendations to improve recall of Malicious Actor detection heuristics over [illegible] accounts	i3 Child Safety	Understand trends	Complete	✓	✓	✓	✓
11	Conduct UX research on teen experiences of NCII & sextortion	• Map out user stories for teen victims of NCII sharing and sextortion • Provide recommendations for teen education, prevention, and remediation tools	UXR Teen Safety	Understand trends	Complete	✗	✓	✓	✓

4 This launch does not cover adults; however, StopNCII.org is already operational to support adult victims of NCII abuse.
5 90% of enforcement volumes from Sigma rule disables were from IG accounts.

Teams across Integrity, Messenger, & IG have H1 '23 work planned on sextortion.

Planned interventions

	Workstream	Description	Team	Theme	Status	H1 '23 RM	Adult	Minor	FB	IG
1	Build topline metric for sextortion prevalence	• Build a heuristics-based sextortion definition that passes the UniS quality scorecard checks • Enable prevention of discovery, connection & messaging between risky actors & teens	MWB Teen Safety	Stand up prevalence measurement	In Progress	✓	✓	✓	✓	✓
2	Improve recall of Malicious Actor detection heuristics over IIC & minor sextortion accounts	• Implement signals to increase coverage of Child Safety account score over IIC and minor sextortion actors • Product teams leveraging IIC classifier for product interventions are migrating to current account score by EOY '22	Child Safety	Improve detection recall & prevent harm	Not Started	✓	×	✓	✓	✓
3	Build support for victims of IIC and sextortion	• Share links to help resources, upsell reporting, and connect confirmed victims of minor sextortion with loved ones	IG Teen Safety	Provide user support	Not Started	✓	×	✓	✓	✓
4	Improve enforcement of minor sextortion user reports using Deceptive Persona account score	• Leverage DP ML classifier score to auto-enforce against reported accounts & improve job ranking in MSGR queues to boost escalation probability	Deceptive Persona, MWB R2, & Child Safety	Improve enforcement	In Progress	×	×	✓	✓	×
5	Improve allocation of human review resources	• Navigate to IV to facilitate human review resource allocation & unblock requests for additional human review budget	MWB R2	Improve enforcement	Not Started	✓	✓	✓	✓	✓
6	Lead efforts with external partners to combat sextortion	• Partner with key stakeholders to lead external efforts to build awareness, drive prevention, and incorporate efforts into broader dialogue on child safety	Safety Policy	Raise awareness & prevent harm	Not Started	✓	✓	✓	✓	✓

Discussion: What opportunities should we pursue to further mitigate critical & unknown risks identified by this investigation?

	Question	Risk	Potential Opportunities
1	When minors report sextortion on our platforms, how can we ensure that we handle those reports effectively?	2-4	1. Increase human review capacity in under-staffed marketized queues where these reports are prevalent (e.g., Sub-Saharan Africa) 2. Migrate from market- to language-based routing & increase agnostic review coverage to ensure that reports are not missed 3. Improve FRX tagging for users to reduce discrepancies between user feedback & report ranking and ensure that minors' voices are heard
2	How can we measure the impact of planned mitigations when there is currently no sextortion prevalence metric? What teams need to be involved?	All	1. Launch a x-org measurement horizontal with Integrity data scientists to align on the right set of proxy metrics 2. On proactive detection: estimate recall of eligible classifiers, including MaCSA heuristics, UnIS RTP, Deceptive Persona ML classifier, and sextortion Sigma policies, over confirmed true-positive user reports of sextortion
3	What can we do to improve proactive detection of minor sextortion behavior on IGD?	1	1. Partner with XI Misrepresentation to launch a Deceptive Persona classifier for IG 2. Leverage H2 '22 Understand work by i3 Child Safety & UXR Teen Safety to improve recall of existing classifiers over minor sextortion accounts