

[IIC on IG] PRI Analysis Quip v1

Executive Summary

The investigation conducted several tests looking at inappropriate interactions with minors on Instagram. We hope the findings can be useful for CO, CI, and Instagram teams who continue to work to prepare for encryption and who are prioritizing the safety of minors on the platform. Tests were conducted looking at the current state of IIC on Instagram and learnings are intended to be used to inform next steps in preparation for end to end encryption.

IIC enforcement currently focuses heavily on detecting *real world harm* (such as grooming, incest, bestiality, and position of trust) and *online harm* (CEI and sexual conversations). These abuses largely occur in messenger. The policy, however, describes IIC more broadly and could also include unsolicited interactions in comments as well as DMs. Furthermore, it is possible that unsolicited contact isn't always sexual in nature and the current approach and actioning of other abuse does not take reporter age into consideration. A more holistic approach to child safety may put us in a better position to protect minors from both unsolicited sexualized interactions from adults as well as unsolicited non-sexual interactions in comments and messages protect minors.

[Updated 6/18] Cause Analysis Summary

	Problem	Root Cause	Recommendation
1	IG recommended a minor through top suggested to an account engaged in groomer-esque behavior.	The recommendation algorithms worked as intended. 1. Product	Establish a whitelist for recommendations on IG.
2		There is not currently a whitelist to determine which accounts can and cannot be recommended in top suggested. 3.a.Protocol	
3	3 accounts engaged in IIC remained live on the platform even after FB received signal that they were engaging in child exploitation.	This was the most relevant option for the rep. 3.a.Protocol	Add escalation path for reps at sc to escalate jobs for IIC.
4		The ops guidelines for this queue call for CEI review and	Update ops guidelines and add more context to include age and photos of reporter and responsible

CONFIDENTIAL
CONFIDENTIAL

META3047MDL-111-00149153
METANMAG-021-00151987
PLTF EX-01730_0001

PLTF EX

PLTF EX-01730

5	Minors reported adults interacting inappropriately in IGDM and we left the accounts up.	there was no CEI. 3.a.Protocol These accounts did not meet the 30% threshold. 3.a. Protocol	so that there is sufficient information to action for IIC. Update account disable threshold IG and explore opportunities to incorporate age in determining severity of actions.
6	Minors have a limited ability to prevent receiving violating unsolicited messages from adults.	The functionality does not exist. 1. Product	Consider adding functionality to turn off requested inbox.
7	Adults interacted with minors inappropriately in comments.	We don't explicitly call out commenting in our IIC policy. 2. Policy	Update IIC policy to include comments as an example and clarify behaviors that are and are not IIC
8		We have not allocated resources to preventing IIC in comments. 3.a. Protocol	Consider proactive and reactive opportunities to address IIC in comments.
9		The default private settings for minors do not limit who can comment. 1. Product	Update default privacy settings for minors to be more secure.

[Updated 6/18] Discussion for PRI CIPCO

1. Do we want to be taking a more holistic approach (surface and behavior) to combating IIC on the platform?
2. If so, how can we ensure that age is taken into account throughout our prevention and detection efforts?
 - a. This includes:
 - i. Default Privacy Settings
 - ii. Recommendations
 - iii. User Messaging
 - iv. Report Routing
 - v. SRT Context
 - vi. Severity of Actions
3. How can we most effectively incorporate different IIC violator personas into our prevention and detection efforts?

Background

Inappropriate Interactions with Children is a complicated abuse area that encapsulates numerous violation types ranging from creepy to criminal (see policy definition in the appendix). The behaviors of victims, abusers, and the interactions between the parties

can vary according to the particular relationship and violation type. Currently, in order to identify something as IIC we have to know the relative **age of the two parties** (adult and minor) and the two must be **engaged in sexualized conversation**. This has been historically challenging to enforce on IG both because we do not collect age of users and the detection of sexualized conversation usually occurs in messages which receive limited review and will no longer be accessible going forward (unless user reported).

In order to continue to protect younger users it is imperative that we continue to improve proactive detection, upstream prevention, and up-sell and correctly action user reports for this abuse area. Due to existing work in this area across numerous CIPCO teams, this investigation focused largely on **discoverability** or upstream prevention and **reactive reporting flows** or user reports.

Testing

Discoverability

- **Hypothesis:** We may be facilitating possible groomers finding young people through following, hashtags, explore, and search.
- **Evidence:**
 - Recommending groomers follow minors
 - Overall, 7% of IG recommendations are related to cases where we recommended teens to non-teens.
 - This is compared to 26% of people we recommended to Groomers were Teens: which is way above the IG teen-non_teen recommendations threshold.
 - This could be due to different factors, as indicated in the recommendation sources below, one of which could be the normal followers graph, and other include, similar hashtag interest, facebook friends etc.
 - 22% of the teens recommended to Groomers, received a follow request from Groomers when the recommendation was shown to them. In other words, groomers clicked on the follow button 22% of the time after we recommended a teen to them.
- **Tests:**
 - Test 1: Make test account, follow a bunch of grooming-type hashtags, see if platform recommends children as people to follow.
 - i.e. sexyteens, hotteen, sugardaddy, gayboysex, boy_and_boy, sexyteengirls
 - Test 2: Make test account, follow a bunch of hashtags, see if platform recommends other creepy hashtags
 - Test 3: Make test account, follow a bunch of creepy hashtags, see what gets recommended in search.

- Test 4: Make test account, follow several teen/grooming accounts, see if platform recommends other teens to follow.
- Test 5: Make test account, follow several teen/grooming accounts, see if platform recommends questionable hashtags to follow
- Test 6: Make test account, follow several teen/grooming accounts, see what gets recommended in search
- **Results:**
 - We are facilitating adults finding minors and minor-focused content through recommendations and explore.
 - Tests 1-3
 - Hashtags on their own did not result in much content being served to us. In other words, following hashtags did not significantly alter search, explore, our feed, or produce other similar hashtag recommendations.
 - Some hashtags like #sugardaddy and #gayteen, however, would result in teen photos appearing in stories.
 - While a significant amount of minor-focused content may not be recommended to adults if they follow hashtags, a dedicated user would be able to surface photos of minors with very little effort and comment or send a DM to those accounts if they choose.
 - The primary way we were able to surface content was by using hashtags and seeking out photos of minors that were sexualized in some way. We then used that account to fanout and look at other followers to see what hashtags they used and what accounts they followed.
 - Users that follow minors' accounts will be recommended similar accounts as well as photos of teens in explore.
 - Tests 4-6
 - Following teen modeling accounts, borderline CEI accounts, and minors resulted in our being served a lot of similar content in explore.
 - Numerous accounts were dedicated to borderline CEI and aggregated teen photos of teens in bikinis, for example. Many of those account photos had a significant number of comments sexualizing the photo even if the photo itself was not violating.
- **Problems:**
 - Adults are easily able to find minors to inappropriately interact with on IG.
- **Root Cause(s):**
 - IG algorithm does not take in to account violations when serving content.
 - Sexualized teen accounts exist on the platform because it is not clear that they violate (age is hard to determine) and we aren't set up to enforce against them at scale.

- We don't take age into account when deciding what to recommend to someone.

- **Risk Area:** Harmful content

	Problem	Intermediate Cause	Root Cause	Rec
1	IG recommended a minor through top suggested to an account engaged in groomer-esque behavior.	The account followed and searched for sexualized minor accounts.	The recommendation algorithms worked as intended. 1. Product	Estab for recom on IG
2			There is not currently a whitelist to determine which accounts can and cannot be recommended in top suggested. 3.a.Protocol	

Reactive Reporting - IGDM

- **Hypothesis:** We may be missing IIC content on IG that is reported through reactive reporting.
- **Evidence:** Teens report rate for DMs on IG is at 51% of which 40% is related to IIC cases.
 - 20% of All DM reports are related to IIC cases (where the reporter is a teen and the sender is a non-teen)
- **Test:** Pull a list of 100 recent IG DMs that have been escalated for child exploitation (there is no option to report for IIC). Go through and manually label them to see if any are IIC.
- **Results:**
 - Labeled 50 jobs that were marked for child exploitation, marked those where IIC was taking place
 - 6% Confirmed as IIC
 - 20% Labeled as Possible IIC
- **Problem:**
 - We are missing IIC in at-scale review of IGDMs.
 - Minors are subjected to IIC in comments.
- **Root Cause(s):**
 - Reporting process is confusing.
 - Jobs cannot be escalated for IIC currently within SRT.
 - It is difficult for reps to determine age of reporter and responsible.
 - There is not a clear way for minors to report IIC on IGDM jobs.
 - Age is not a factor in how we triage reporting.
 - There are limited resources currently allocated to prevention efforts.
- **Risk Area:** Harmful Content

	Problem Statement	Proximate Cause	Intermediate Cause	Root Cause
--	-------------------	-----------------	--------------------	------------

1	3 accounts engaged in IIC remained live on the platform even after FB received signal that they were engaging in child exploitation.		They were escalated for child exploitation.	This was the most option for the rep. 3.a.Protocol
2				The ops guidelines queue call for CEI and there was no CEI. 3.a.Protocol

	Problem	Proximate Cause	Intermediate Cause	Root Cause
1	Minors reported adults interacting inappropriately in IGDM and we left the accounts up.		We only remove accounts for violating DMs if 30% of their DMs violate.	These accounts meet the 30% threshold. 3.a.

	Problem	Proximate Cause	Intermediate Cause	Root Cause
1	Minors have a limited ability to prevent receiving violating unsolicited messages from adults.			The functionality exist. 1. Product

Reactive Reporting - Comments

- **Hypothesis:** We may be missing IIC content on IG that is reported through reactive reporting.
- **Evidence:** In the course of conducting discoverability tests we found numerous incidences of sexualization of teens through comments in addition to observing other types of possible IIC in comments i.e. bullying.
- **Test:** Observe comments on teen accounts and determine how and if we are actioning comments for IIC.
- **Results:**
 - IIC is happening in comments.
 - In one month there were approximately 8 million comments reported on IT, 7.5% of those comments were a teen reporting a comment from an adult
 - Nudity, 18% of all comments deleted were a teen reporting a non-teen

- Hate speech, 52% of all comments deleted were a teen reporting a non-teen
- Harassment/bullying, 55% of all comments deleted were a teen reporting a non-teen
- Violence, 36% of all comments deleted were a teen reporting a non-teen
- **Problem:**
- **Adults interacted with minors inappropriately in comments.****Root Cause(s):**
 - We don't explicitly call out commenting in our IIC policy which has contributed to not building out protocol and enforcement of IIC in comments.
 - We have assumed that comments can be dealt with through self-remediation.
 - We have been focused on more severe cases that can be referred to law enforcement.
- **Risk Area:** Harmful content and reputation
 - As a result of IIC in comments and the ability to DM minors they don't know, Instagram could be seen by many parents, media outlets, and governments as an unsafe platform for minors.
 - Furthermore, IIC in comments is highly visible and opens us up to further scrutiny and questions about why we are not doing more to protect minors.
 - The platform continues to receive negative international media coverage for its' handling of inappropriate interactions with minors as well as more serious cases of grooming.
 - Lack of enforcement and prevention may normalize IIC and in particular adult sexualization of minors.
 - Moreover, it may embolden adults to engage in behavior that they otherwise wouldn't.
 - Sexualization of minors, as a particular form of IIC, in comments may lead to more serious forms of online and offline harm.

	Problem	Proximate Cause	Intermediate Cause	Root Cause
1	Adults interacted with minors inappropriately in comments.	We do not have protocols and enforcement to action against IIC in comments.	There is not widespread clarity that comments can violate for IIC.	We don't explicitly commenting in our Policy
2				We have not allocated resources to preventing IIC in comments. 3.a. Problem
3		Privacy settings are often not preventing inappropriate	Minors are not using privacy settings to prevent adults they are not	The default privacy settings for minors do not limit comments. 1. Problem

		interactions in comments.	connected with from commenting on their media	
--	--	---------------------------	---	--

PPPEP Breakdown

Root Cause	% of total
1. Product	33%
2. Policy	11%
3. Protocol	56%
<i>a. Process is missing or incorrect</i>	
<i>b. Documentation is missing or incorrect</i>	
<i>c. Education was not provided</i>	
4. Enforcement	
<i>a. Problem was detected or user reported</i>	
<i>i. Report was not looked at</i>	
<i>ii. We looked at the report but tools were bad</i>	
<i>iii. We looked at the report but rep made a mistake</i>	
<i>b. Problem was not detected or user reported</i>	
<i>i. Conscious decision to have lower precision/recall tradeoff</i>	
<i>ii. Unintended issue</i>	
5. People Disagree	

Appendix Background



Namrata Bostrom is with Christopher Palow.
May 13 at 11:14 AM

SCA



Guy Rosen is with Namrata Bostrom and 4 others.
May 13 at 11:02 AM

SCA

✖ 2. Inappropriate interactions with children (IIC) such as:

1. Initiating unsolicited contact with minors (for example, private messages between stranger adults and minors)
2. Soliciting, displaying, sharing, or viewing imagery of nude, **sexualized**, or sexual activity with minors
3. Arranging real-world sexual encounters or obtaining sexual material from a minor directly.
4. Displaying nudity to minors

- Minors are subjected to IIC in comments.

- Below are screenshots of comments where reporter was a minor and responsible was an adult. Report reasons below include hate speech, bullying/harassment, and nudity.

instagram_comment_text "Whipping the black students in my class for refusing to pick cotton"

Text

instagram_comment_text "@futuremodel_ eww yuck you scary uglyy"

Text

instagram_comment_text "Bend ya over and pound tat sweet juicy pussy with my fat thick ten inch cock balls slapping off your ass cheeks deep babe"

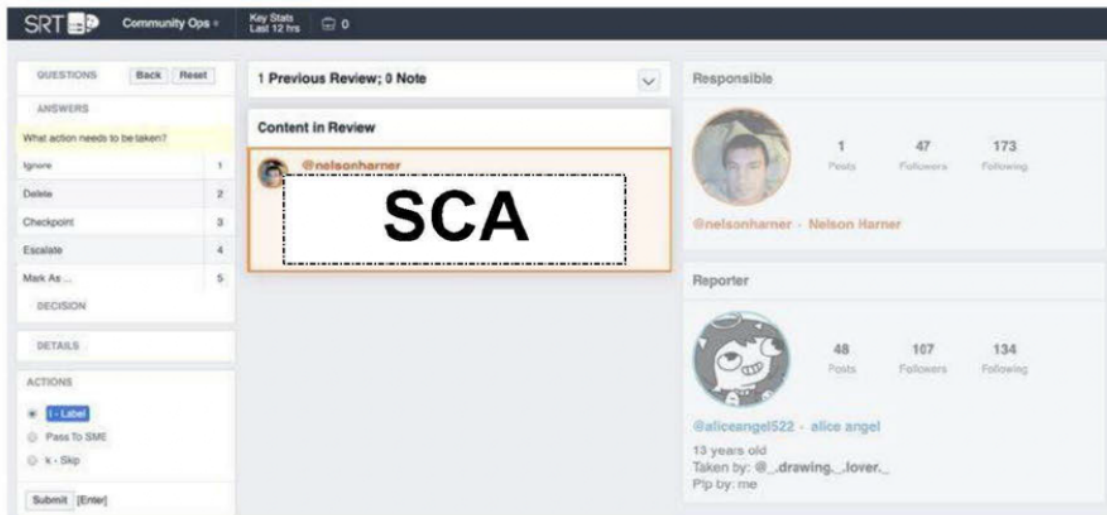
Text

- Minors are also sexualized in comments which may also qualify as IIC under the letter of the policy.

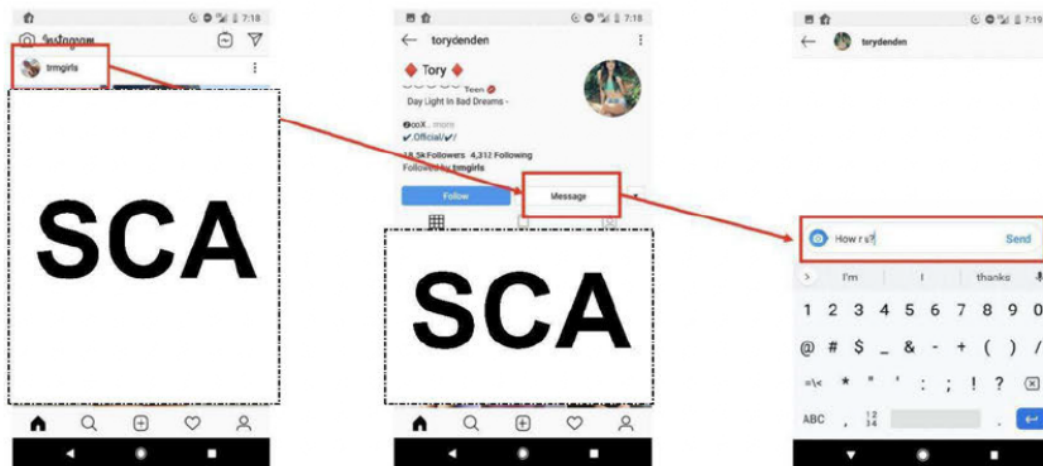
- We are missing reports of grooming and inappropriate interactions with children in DMs at scale.

User reporting process

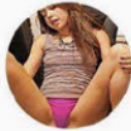




- Adults are easily able to DM minors that they find on Instagram.



- "Sexy teen" accounts are on the platform.



slavgirl

Follow

15 posts 15.4k followers 81 following

♥ Katya

BestGirl 🇺🇸

POSTS

TAGGED

SCA



joven16citas





wow_hot_teens



1,245 likes

View all 32 comments

May 7

Discoverability Observation Summary - [DRAFT] Personas

This table represents a culmination of observation through conducting discoverability testing, note on [[HYPERLINK "https://fb.workplace.com/notes/leah-treiman/understanding-predatory-behavior-upstream/199636287446261/"](https://fb.workplace.com/notes/leah-treiman/understanding-predatory-behavior-upstream/199636287446261/)], and some preliminary data work on groomers.

	A	B	C	D
1		[Draft] IIC Adult Personas/Scale		
2		Interested	Exploratory	Want to interact online
3	Motivation	Not motivated	Somewhat motivated	More motivated
4	Focused on age	Age agnostic	More motivated to seek out photos of minors	Looking for minors
5	IG use	Looking for sexual content in general	May DM a few minors they gravitate towards	Sexual comments including dmme
6		Might use the platform to post their own photos	Largely consumers of photos of minors	May send photos
7		May use IG regularly for activities not involving minors	May come across borderline CEI accounts including accounts dedicated to posting 'sexy teens'	May solicit photos
8			May follow several to many borderline CEI accounts	May DM many minors
9				May engage in sextortion
10				
11				
12				
13	Search	May be heavy consumers of porn	May follow/seek out accounts that involve minor CEI? may look	May use hashtags to look at photos of minor

			for links off platform and use IG to find dropbox and Patreon accounts?	
1 4		May not know the accounts with minors are minors	May or may not know have insight in to their attraction as a bad behavior	May seek out particular types of photos/accounts i.e. minors displaying attention seeking behaviors, vulnerable youth, etc
1 5		If early 20s or late teens may not see/understand attraction in young people to be harmful (wording?)	Situational - may be attracted to minors and are able to follow/consume sexualized content because it is widely available online and on IG. May be aware of social norms	
1 6	Possible Characteristics	May be young themselves 17+	May receive spammy messages from adult porn accounts	May follow borderline CEI accounts
1 7		May be more likely to present true identity	Might use the platform to post their own photos	May follow CEI accounts
1 8				May be posing as a photographer/legitimate party to get minors to engage/trust them and send photos
1 9	Possible Company Approach	Reduce discoverability of content	Reduce discoverability	
2 0		Taking advantage of content, increasing risk and discoverability of content will mitigate likelihood they follow or engage with minors	Opportunistic, but may not engage in behaviors if perceive higher risk	Opportunistic, but may not engage in behavior if perceive higher risk
2 1		May not exhibit enough obvious behavior to warrant pushing a warning	Potential groomer interstitial	Potential groomer interstitial

2			Panopticon may be effective to prevent further searches, violations, or escalating behaviors	Panopticon may be effective to prevent further searches and violate
2				
3				
2	Examples TBC	hameedullah4866		nigel.cooke.794
4				
2	Minor likelihood to report (cultural dependencies, offline support network, awareness of self, relationship with abuser)	Unlikely	Likely	Conditional
5				

For “Want to Interact in Person” or accounts actioned for grooming we found that:

- Only 105 groomers used hashtags when posting photos. Also, since 6% of the total photos posted by Groomers contained at least 1 hashtag, it seems some groomers are either posting more photos than others or some groomers use more hashtags than others.
 - This might indicate that hashtags are not a common way to surface grooming behavior.
- Also, the median of number of photos posted during the last 3 months by users is 7. With the highest number of photos created/posted being 578.
 - This might indicate that groomers are not very interactive i.e. posting and interacting with their followers through home.

Risk Areas

- Harmful content
 - IIC as it is defined in the letter of the policy is occurring on the platform on more surfaces than we are currently looking at.
 - Minors can be subjected to IIC when they use IG.
 - Lack of enforcement and prevention may normalize adult sexualization of minors.
 - Moreover, it may embolden adults to engage in behavior that they otherwise wouldn't.
 - In addition, sexualization of minors in comments may lead to more serious forms of online and offline harm.
- Reputation
 - As a result of adult sexualization of minors and IIC through comments and the ability to DM minors they don't know, Instagram is seen by many parents, media outlets, and governments as an unsafe platform for minors.

- The platform continues to receive negative international media coverage for its' handling of inappropriate interactions with minors as well as more serious cases of grooming.

Proactive Risk Investigation Specs

- Investigation Master Task: T43728516
- Incident review primary investigators: Luke [REDACTED] & Jourdan [REDACTED]
- Incident review analyst: Yosra [REDACTED]
- PRI presentation date: June 20th
- Contributors: Akum [REDACTED] Chris [REDACTED] Namrata [REDACTED] Karina Newton, Colin [REDACTED] Irem [REDACTED] Vainius [REDACTED] Bernat [REDACTED] Chelsea [REDACTED] Swithin [REDACTED] Zach [REDACTED] Liz [REDACTED] Taylor [REDACTED] Michelle [REDACTED] Gabriela [REDACTED], Matt [REDACTED] Arielle [REDACTED] Nick [REDACTED] Nikhil [REDACTED] Tyler [REDACTED] Aditi [REDACTED] Bettina [REDACTED] Shauna [REDACTED] Antigone Davis, Anna [REDACTED] Jeremy [REDACTED] Tammy [REDACTED] Aditya [REDACTED] Aparajitha [REDACTED] [REDACTED] Christine [REDACTED] Bertram [REDACTED] Brian [REDACTED] Sarah [REDACTED]