IGTS: Protect Minors in Comments - Product Brief

LAMA: [HYPERLINK "https://www.internalfb.com/intern/launch_manager/projects/2294330190711548/privacy/248707480304844/Assessment" \h]

Background

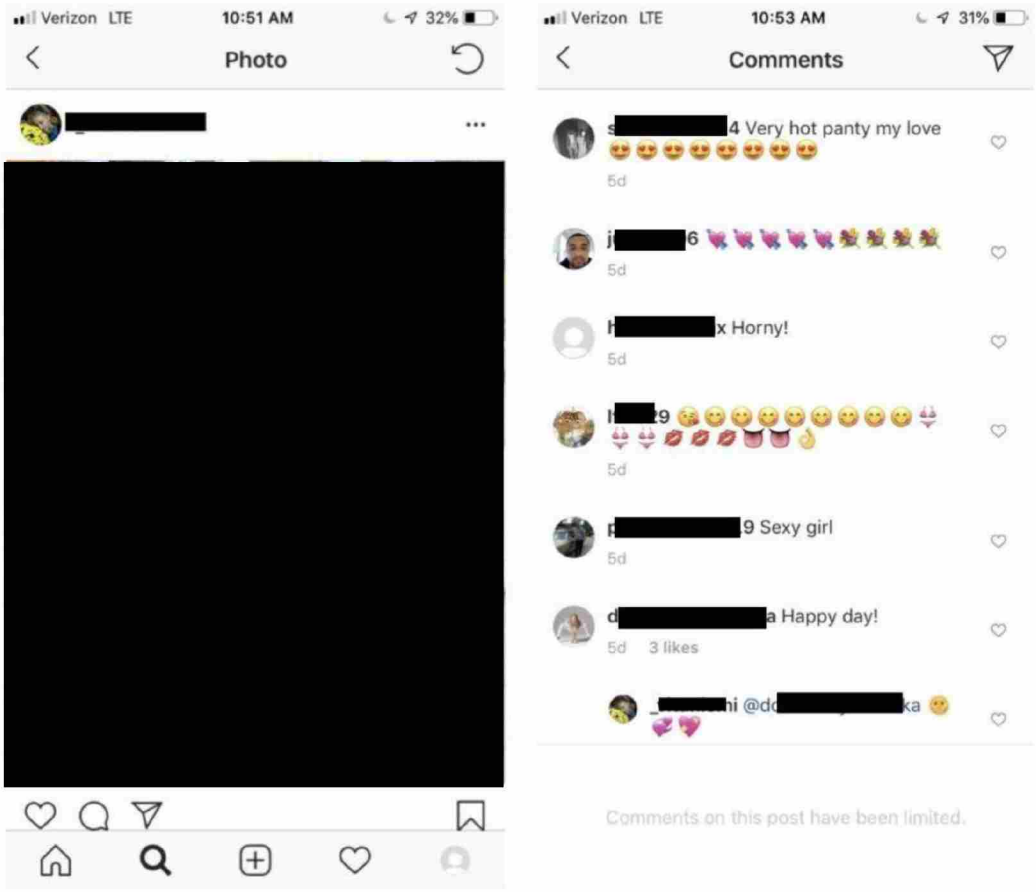
- **Goal of this product intervention:** Protect Minors in the Comments environment from adult "potential violators" who could present harm to them and/or inhibit their freedom to freely express themselves on IG.
- **Why we need to solve this problem now:**
  - Inappropriate interactions with Children (IIC) Tier 2 **manifested 38 times more on Instagram than on Facebook** based on a CIPCO study in August 2020 ([HYPERLINK "https://fb.workplace.com/groups/135709743752034/permalink/624917758164561/" \h]) or ~200k victims/day. When normalized for the number of teens, the relative prevalence on Instagram compared to Facebook is about 5x.
  - The National Center on Sexual Exploitation labelled Instagram as **"Predators' Paradise for Sex Trafficking, Child Abuse, and Pedophilia"** in November 2019 ([HYPERLINK "https://www.youtube.com/watch?v=NpkHYVYvb8s" \h]).
  - Out of 11 social apps, Instagram was ranked in the **top three least safest apps** by both parents and teens based on an [HYPERLINK "https://docs.google.com/presentation/d/13SZLEnpC04bdnTvqzjrNWccZ_YroNjWAJbL1lie3p7s/edit" \l "slide=id.ga7bfc63f48_0_325" \h] by the Messenger Rooms integrity team in December 2020.
  - When analyzing the most common instances of harmful contact with minors, **we found that comments drove X% of total harmful contact.** ([HYPERLINK "mailto:[redacted]@fb.com" \h] please add verbiage per your analytical framework)

Product Strategy

- **Our Mission:** "Make Instagram the safest online community for teens and a place where they feel free to express themselves."
- **Strategy:** By restricting potential violators from interacting with minors in public comments, we are cutting off a primary gateway to (a) harmful messaging that

could lead to grooming and sexual exploitation, and (b) unwanted/"creepy" follows where the bad actor can research the minor's habits, gain access to the minors's social graph, and comment on the minor's posts/stories. Similar to having a sex offender in a neighborhood in the real world, we simply don't want these individuals reaching out to our children (or vice versa). Even if these follows don't result in direct IIC, they create an environment of fear and harm where minors don't feel free to express themselves.

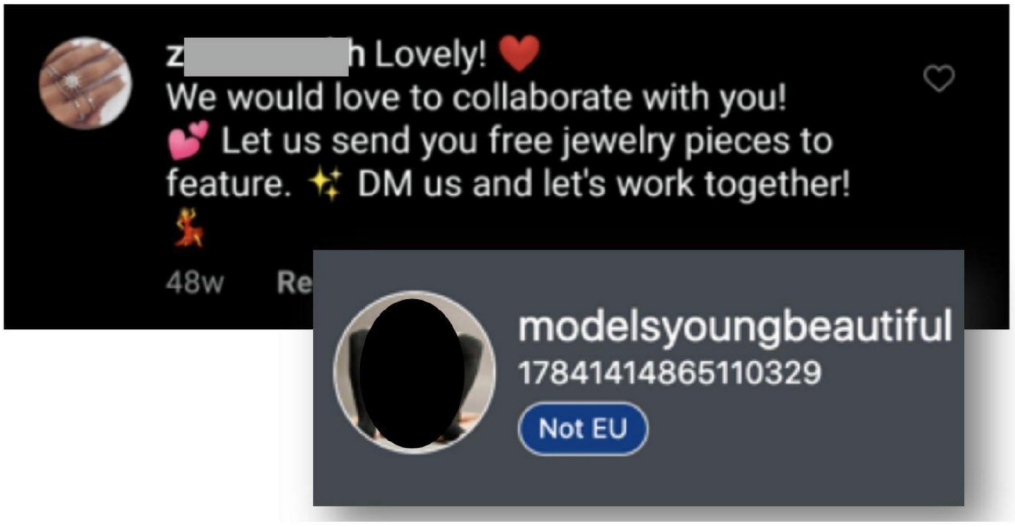
- **Commenting in public spaces often leads to:**
  - **"Creepy" comments**: An opportunity for potential violators to to make "creepy" comments that deter minors from feeling free to express themselves (see example below from our [HYPERLINK "https://urldefense.proofpoint.com/v2/url?u=https-3A__www.dropbox.com_l_scl_AAC1jDIVv-5FfrIHN0yMJE80tBRZDCquSwW8A&d=DwMFaQ&c=5VD0RTtNlTh3ycd41b3MUw&r=h2jf4KJObPmVvP-vLQjqqg&m=TORinNzlOH-fhQjk-oGcuALXJlxXxpvVQezZb78_5Gg&s=aKsqfWCL8qukbSWR-79IdoxjGLayNGuWqT0jf6QFHJQ&e=" \h] in Q4 2020).



- **The start of the grooming process**: The initial point of contact where adult potential violators and minors begin a relationship that leads to grooming and exploitation (see example below from our [HYPERLINK "https://urldefense.proofpoint.com/v2/url?u=https-3A__www.dropbox.com_l_scl_AAC1jDIVv-5FfrIHN0yMJE80tBRZDCquSwW8A&d=DwMFaQ&c=5VD0RTtNlTh3ycd41b3MUw&r=h2jf4KJObPmVvP-vLQjqqg&m=TORinNzlOH-fhQjk-oGcuALXJIxXxpvVQezZb78_5Gg&s=aKsqfWCL8qukbSWR-79IdoxjGLayNGuWqT0jf6QFHJQ&e=" \h] in Q4 2020).

Fake model scouts tricking UK girls into sharing explicit photos

Leading model agencies say calls from worried parents are increasing as scammers use social media to exploit children



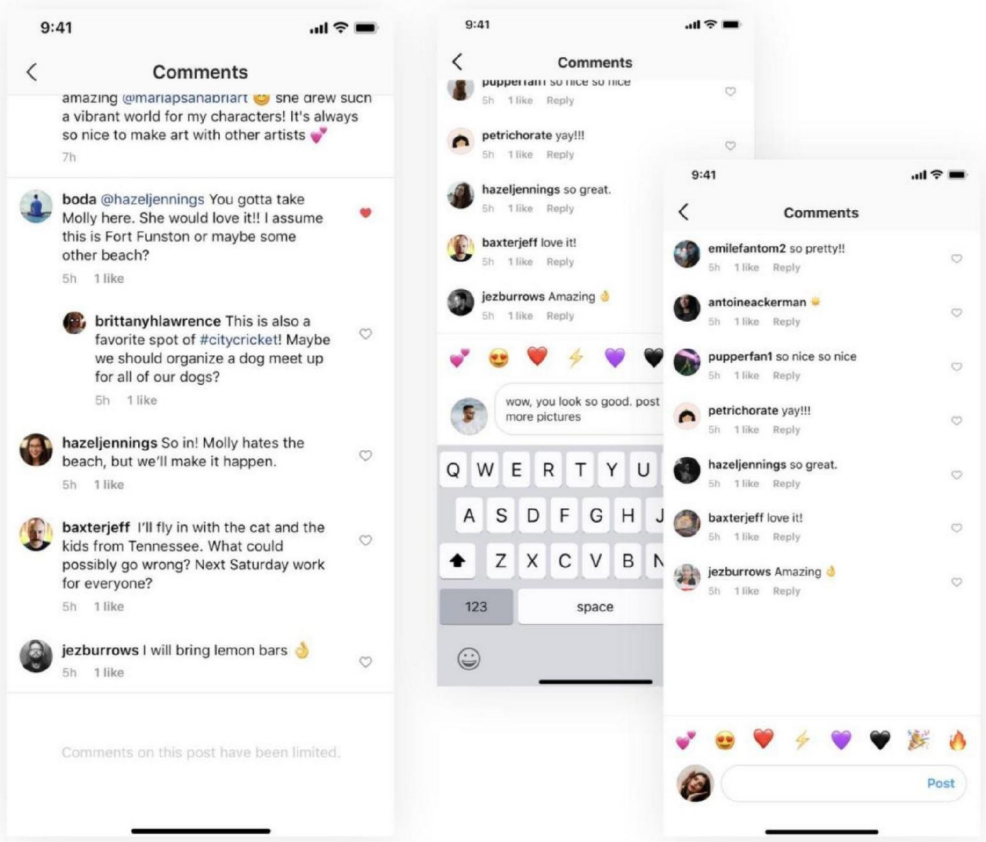
- **Patrolling for targets**: An opportunity for potential violators to find targets by patrolling through the comments on a public post. In the example below, the potential violator scans the comments for a post by a public figure with a large teenage female fan base and begins to make contact either through on surface interaction (e.g. reply) and/or by visiting their account where they can follow, DM, and/or conduct research into images/video, habits, interests, social graph, etc...


justinbieber
970,220 likes
justinbieber : @rorykramer
View all 3,826 comments
6 days ago


Comments
yasminmalak93 I love you so much Justin
6d 6 likes Reply
isabellaavegaa
6d 6 likes Reply
haileydrenters
6d 11 likes Reply
bonillaricci
6d 5 likes Reply
bizzraxhl i love u
6d 10 likes Reply
eylen_20
6d 3 likes Reply
rosy.roselovers Follow us for free roses
6d 4 likes Reply
marlene.camposl
6d 5 likes Reply
keiryeridania

Specific Product Implementations

- **Initiative #1: Prevent "potential violators" from commenting on public posts by teens (and vice versa).** Currently, "potential violators" can find a minor's profile (if public) and comment on their posts. We will prevent "potential violators" from commenting using one of the two design treatment below and vice versa (Still TBD - depends on ease of implementation and XFN feedback):
    - 1a) Block Comment Design treatment #1 (mock on left): We will block potential violators from posting with the message "Comments on this post have been limited" (or similar language). This is more transparent than treatment #2 (autohide), but there is a risk that the potential violator will realize their account classification. However, this concern is mitigated as there are multiple reasons why comments can be blocked for users including: restrictions made by the target on who can comment, restrictions on banned keyword/phrases, over engagement by the user (likes, comments, follows, unfollows), multiple duplicate comments, etc... Therefore specific language in app on why this feature is gated is not necessary. However, some additional language in the help center may be justified.
    - 1b) Autohide Design treatment #2 (mock on right): We will auto-hide potential violators' comments from everyone else (including post owner). This is less transparent than option #1 above, but less risky. The potential violator will just believe that his comment was ignored.



Block likely violators

Auto-hide likely violators' comments

Designs provided by [HYPERLINK "mailto:katherineliu@fb.com" \h]

- **Initiative #2 - Remove or hide minor comments on a 3rd party public post when a potential violator is viewing (and vice versa)**
  - Currently, "potential violators" can see comments from minors on a third party's post. This allows potential violators to scan the comments for potential targets. The potential violator can then reach out via a reply comment, a tag/mention, follow, DM, etc.. or the violator can simply begin to research the target's images/videos, interests, habits, and social graph.
  - If the viewer is a potential violator, we will remove or hide minors' comments from view. We will also make this bi-directional as grooming often happens in both directions.
- **How are we defining "potential violators" and "minors"?**
  - For "potential violators":
    - We will be using a set of heuristics as a proxy for "potential violators" based on research from Central Integrity and child safety

experts, and will be an account that has one OR more of the following characteristics:

- User is defined as a non-teen (adult) per the Adult Classifier Model **AND**
  - User has at least 2 Blocks taken against them in the last 28 days **OR**
  - User has at least 2 NFX taken against them in the last 28 days **OR**
  - User has a Child Sexual Exploitation violation within past 90 days
- This definition will not apply to an individual account that is verified or x-checked (as users with large fandoms are more potential to receive blocks/NFX without causing harm)

- In future iterations
  - We will be using the **IG IIC Classifier with appropriate precision scores** in accordance with CI and Child Safety guidance. However, due to [HYPERLINK "https://fb.workplace.com/groups/135709743752034/permalink/724978978158438/" \h], the classifier will not be available for use for several months. See FAQs for more details around the use of the IIC classifier.
  - We will consider including Adults (SUMA) who have multiple accounts with one account disabled for CSE, as a safeguard against recidivism

- For "minors: We will be leveraging the age-model developed and maintained by the Central Product Platform Team. See FAQs for details.

Measuring success:

- Goal Metric: We will measure the impact on overall **IIC Tier 2 prevalence.**
- Main Tracking Metric: We will measure the **reduction of contacts** between "potential violators" and minors across all entry points, surfaces, and/or behavioral patterns.
- Other tracking metrics: We will measure the **reduction of Follows and Blocks/NFX** (negative feedback from adult:minor interactions). We will also explore **sentiment** metrics to see if minors feel more free to express themselves.
- TBD - details of how to setup experiment, time range, etc.. [HYPERLINK "mailto:[redacted]@fb.com" \h]

Milestones:

Both of these initiatives are P0 initiatives expected to launch in Q1 and Q2 2021 in their respective countries and form factors.

- Countries:
  - Initiative #1 Prevent commenting on minor posts (P0 priority/Q1 Timing): This will only launch in the US. As described in [HYPERLINK "https://www.internalfb.com/intern/wiki/Core_Dimensions/Age_Prediction/Age_Modeling_FAQ/" \l "q-where-are-age-models-a" \h] the age model in IG is fully operational only in the US as of now for "high risk" initiatives where the gating of product features are transparent to the user.
  - Initiative #2 Remove or hide minor comments on a third party post (P0 priority/Q2 Timing): This will launch in the US, AU, FR, JP, and UK. Per the age modeling team (see FAQs), low risk initiatives mainly related to "demoting or down-ranking content/suggestions" is available for AU, FR, JP, and UK.
- Form Factors:
  - We will launch initially across iOS and Android as this comprises XX% of our total harmful connections.
  - We will subsequently roll this out to other form factors including Web and Lite, which make up Y% of total harmful connections.

Intended Learnings:

- Impact of "comments" on total reduction of contact events between "potential violators" and minors: In Q1, the IG Teen Safety DS team will build an analytical framework for tracking all contact events between potential violators and minors across entry points, surfaces, and/or behavioral patterns. While our data shows that "comments" is a primary method for "potential violators" to reach minors, we're interested to see if & how they adapt once the restriction is in place. Will cutting off one outlet for contact lead to increases in other avenues? This will give us understanding on subsequent product interventions.
- Impact of "Followers" on IIC Prevalence: While we expect IIC prevalence to decline when we restrict upstream "comments" from occurring, we are unsure of (a) how significant the impact will be and (b) when the benefits will manifest (as we're unclear of how long the "grooming" process would normally take before IIC events occur). These data points will help us to better understand the problem space, set realistic expectations/goals for the future, and make adjustments to our product roadmap.
- Recidivism and Actor Accountability: Similar to bullet #1 above, we would like to understand if recidivism continues with bad actors emerging with new/alternate accounts to follow minors and/or whether bad actors can re-emerge once they are no longer classified as "potential violators".

Handshakes:

- Surface Team: Had initial discussions with [HYPERLINK "mailto:[redacted]@fb.com" \h] (PM on Relevance Integrity Team) on strategy, requirements, and prioritization.
- CI Product/Eng Team: Working with [HYPERLINK "mailto:[redacted]@fb.com" \h] and/or [HYPERLINK "mailto:[redacted]@fb.com" \h] on the use of the IG IIC Classifier to determine potential violators and [HYPERLINK "mailto:[redacted]@fb.com" \h] and [HYPERLINK "mailto:[redacted]@fb.com" \h] on use of heuristics.
- Child Safety Policy Team: Aligning with [HYPERLINK "mailto:[redacted]@fb.com" \h] and [HYPERLINK "mailto:[redacted]@fb.com" \h] on guidelines for using IG IIC Classifier thresholds for gating product features consistently across Facebook, Inc.
- Centrals Product Platforms Team: Aligned with [HYPERLINK "mailto:[redacted]@fb.com" \h] on use of age models (see FAQs) and communicating needs for increased country coverage in the future.
- Signals Foundation Team: TBD - [HYPERLINK "mailto:[redacted]@fb.com" \h]
- Privacy XFN: TBD

GTM Plan:

- Per discussion with PMM and Policy/Comms teams, (a) our policy team is planning to give an IG Teen Safety update to external stakeholders in January 2021 and (b) our GTM team will make announcements on [HYPERLINK "https://www.saferinternetday.org/" \h] on February 9, 2021. We are working with the teams to provide a high level narrative on our vision and efforts in H1 and beyond, but will not communicate any specific actions we are taking during these events.
- As we roll out these initiatives, we will work closely with our PMM, Policy, Comms, and Legal team to highlight any storylines that build our legitimacy as a top online platform for child safety.
- TBD - Do we need to add additional language in the help center that blocks could happen due to integrity issues in addition to the other reasons currently in place (especially if we pursue initiative #1 option a).

FAQs

- **Why are we limiting this product intervention to public spaces only?** Our intention is to reduce the likelihood that potential violators (a) Make "Creepy" comments that prevent minors to freely express themselves, (b) contact minors that leads to grooming, and (c) Patrol comments from minor's on third party posts

for potential targets. We believe that these three use cases happen mainly in public spaces (i.e. when the "potential violator" and minor are not connected). This also alleviates potential backlash where there are false positives between close contacts (e.g. when parent can not comment on their child's post).

- **In the future how will we use the IG IIC Classifier? ?** We will use the Instagram IIC (Inappropriate Interactions with Children) Classifier [HYPERLINK "https://fb.workplace.com/groups/135709743752034/permalink/552239588765712/" \h] and [HYPERLINK "https://docs.google.com/document/d/1BN43J1FscB1rpUg4Ng08NFZVQiMY4ZyMgEeyR5a50cY/edit" \h]. Precision Thresholds: A 90% precision threshold would be the most conservative (typically used for auto-disabling user accounts) and 30% would be the most aggressive (typically used for adjusting our suggestion algorithms). Example of prior pXFN approval here: [HYPERLINK "https://www.internalfb.com/intern/launch_manager/projects/345161196132632/privacy/1059436120933902/Finalization" \h]
- **How are we defining "minors"**? Minors are defined as users who are 17 years of age and younger. We will rely on the age-modeling developed and maintained by the Central Product Platform team. See details [HYPERLINK "https://www.internalfb.com/intern/wiki/Core_Dimensions/Age_Prediction/Age_Modeling_FAQ/" \l "q-where-are-age-models-a" \h]. Example of prior pXFN approval here: [HYPERLINK "https://www.internalfb.com/intern/launch_manager/projects/398424690890150/privacy/1312372485581307/Consultation" \h]
- **Why don't we launch this globally?** Unfortunately, our impact is limited by the maturity of our age-models. As described above and in [HYPERLINK "https://www.internalfb.com/intern/wiki/Core_Dimensions/Age_Prediction/Age_Modeling_FAQ/" \l "q-where-are-age-models-a" \h] the age model in IG is fully operational only in the US as of now. However, low risk initiatives mainly related to "demoting or down-ranking content/suggestions" is available for AU, FR, JP, and UK.
- **Will users (either minors or potential violators) know that we are launching these initiatives? Do we need to proactively answer any questions in-app?** The initiatives in this brief are highly targeted at restricting "potential violators" from following minors, and vice versa, without any action needed by the minor. Similar to other similar product interventions on xxx and yyy, we do not plan to explicitly communicate these changes to the actors or the wider community in the app. TBD - talk to comms/policy/legal PRIV

PRIV

- **What is considered a CSE violation?** CSE includes profiles that have been flagged for:
  - CEI - Child Exploitation Imagery
  - Child Abuse Graphic Violence
  - Child Grooming

- Child Abuse Non Sexual
- Minors Sexualization
- Child Sexual Exploitation Uncategorized (Other)

Key Data and Research Links

- [HYPERLINK "https://fb.quip.com/9HcCAoY5dKd2" \h]
- [HYPERLINK "https://fb.quip.com/0RGkAmDoRoGX" \h]
- [HYPERLINK "https://docs.google.com/document/d/1qxutBnVxxt_ggf4rwPQ1EQcrlLWPrWuc-Yvl5yMt1qg/edit" \h]
- [HYPERLINK "https://docs.google.com/document/d/1BN43J1FscB1rpUg4Ng08NFZVQiMY4ZyMgEeyR5a50cY/edit?usp=sharing" \h]
- [HYPERLINK "https://fb.quip.com/DZLFAibkIxPQ" \h]
- [HYPERLINK "https://fb.workplace.com/groups/135709743752034/permalink/552239588765712/" \h]
- [HYPERLINK "https://fb.workplace.com/groups/135709743752034/permalink/623622301627440/" \h]
- [HYPERLINK "https://www.internalfb.com/intern/wiki/Core_Dimensions/Age_Prediction/Age_Modeling_FAQ/" \l "q-where-are-age-models-a" \h]
- [HYPERLINK "https://fb.quip.com/xT1uAIxF7K9n" \h]
- [HYPERLINK "https://fb.quip.com/ahdDA6XMnGGl" \h]
- [HYPERLINK "https://fb.quip.com/u6HqAD7DmR6p" \h]
- [HYPERLINK "https://fb.quip.com/B8rmAdunuwoU" \h]