Message

From: Guy Rosen [/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=[redacted]]
Sent: 10/3/2024 5:35:04 PM
To: Raena [redacted]@meta.com]; Ravi Sinha [[redacted]@meta.com]; Max Eulenstein [[redacted]@meta.com]; Arcadiy [redacted]@meta.com]; Nell McCarthy [[redacted]@meta.com]; Sayed Otaru [redacted]@meta.com]; Naomi Gleit [redacted]@meta.com]; Molly Cutler [[redacted]@meta.com]; Miki Rothschild [[redacted]@meta.com]; Amit Bhattacharyya [[redacted]@meta.com]; Mel [redacted]@meta.com]; Joel Kaplan [[redacted]@meta.com]; Maggie [redacted]@meta.com]; Nino [redacted]@meta.com]; Josh [redacted]@meta.com]; Tessa Lyons-Laing [[redacted]@meta.com]; Don [redacted]@meta.com]; Kevin [redacted]@meta.com]; Stephanie [redacted]@meta.com]; Antigone Davis [redacted]@meta.com]; James Turner [[redacted]@meta.com]; Dustin [redacted]@meta.com]; Allison Hartnett [[redacted]@meta.com]; Lewis [redacted]@meta.com]; Aastha [redacted]@meta.com]; Courtney [redacted]@meta.com]; Curtiss Cobb [[redacted]@meta.com]; Laura [redacted]@meta.com]; Rahul [redacted]@meta.com]; Prashanth [redacted]@meta.com]; Nicole [redacted]@meta.com]; Kathryn [redacted]@meta.com]; Bryce Bartlett [redacted]@meta.com]; Adam Mosseri [redacted]@meta.com]; Brady Lauback [redacted]@meta.com]; Veronika Belokhvostova [[redacted]@meta.com]; Eric [redacted]@meta.com]; Timothy [redacted]@meta.com]
CC: Sarah [redacted]@meta.com]; Phil [redacted]@meta.com]; Santhosh [redacted]@meta.com]; Jon [redacted]@meta.com]; Sophie Vogel [redacted]@meta.com]; Eric [redacted]@meta.com]; Matthew [redacted]@meta.com]; Allison [redacted]@meta.com]
Subject: Re: [Final Readout] Teen Recs Lockdown for IG Teen Accounts Launch

Thank you team! This was such an impressive effort from start to finish on a very complicated area, and I'm really impressed by the dedication and creativity everyone brought to it.

I'm excited to see the seeds we planted here continue into our AAC approach over the long term.

I am hugely grateful for everyone's work here!

- Guy.

From: Raena [redacted]@meta.com>
Sent: Tuesday, October 1, 2024 2:56 PM
To: Guy Rosen <[redacted]@meta.com>; Ravi Sinha <[redacted]@meta.com>; Max Eulenstein [redacted]@meta.com>; Arcadiy [redacted]<[redacted]@meta.com>; Nell McCarthy <[redacted]@meta.com>; Sayed Otaru [redacted]@meta.com>; Naomi Gleit [redacted]@meta.com>; Molly Cutler <[redacted]@meta.com>; Miki Rothschild <[redacted]@meta.com>; Amit Bhattacharyya <[redacted]@meta.com>; Mel [redacted]@meta.com>; Joel Kaplan <[redacted]@meta.com>; Maggie [redacted]@meta.com>; Nino [redacted]@meta.com>; Josh [redacted]@meta.com>; Tessa Lyons-Laing <[redacted]@meta.com>; Don [redacted]@meta.com>; Kevin [redacted]@meta.com>; Stephanie [redacted]@meta.com>; Antigone Davis [redacted]@meta.com>; James Turner <[redacted]@meta.com>; Dustin [redacted]@meta.com>; Allison Hartnett <[redacted]@meta.com>; Lewis [redacted]@meta.com>; Aastha [redacted] <[redacted]@meta.com>; Courtney [redacted]@meta.com>; Curtiss Cobb <[redacted]@meta.com>; Laura [redacted]@meta.com>; Rahul [redacted]@meta.com>; Prashanth [redacted]@meta.com>; Nicole [redacted]@meta.com>; Kathryn [redacted]@meta.com>; Bryce Bartlett [redacted]@meta.com>; Adam Mosseri [redacted]@meta.com>; Brady Lauback <[redacted]@meta.com>; Veronika Belokhvostova <[redacted]@meta.com>; Eric [redacted]<[redacted]@meta.com>; Timothy [redacted]<[redacted]@meta.com>
Cc: Sarah [redacted]@meta.com>; Phil [redacted]@meta.com>; Don [redacted]@meta.com>; Santhosh [redacted]@meta.com>; Jon [redacted]@meta.com>; Sophie Vogel <[redacted]@meta.com>; Eric [redacted]@meta.com>; Matthew [redacted]@meta.com>; Allison [redacted]@meta.com>
Subject: [Final Readout] Teen Recs Lockdown for IG Teen Accounts Launch

a/c priv - Mel, [redacted]

Hi everyone,

Sharing a final update from our Teen Recs lockdown to recap our progress over the lockdown, our current status, and next steps below.

First - a huge THANK YOU to the working teams here who sprinted leading up to our Teen Accounts launch! Through the lockdown, we updated multiple policy lines, shipped the Pristine and Recommended in Limited Volume solutions, established a survey to understand parent sentiment on content we know People Disagree on, and worked across the company to improve the overall content we're recommending to teens. Our policy harmonization efforts **reduced age inappropriate content prevalence by 13.5%, and initial data from Pristine sources shows we are slightly above our weekly 0.2% non-rec prevalence goal,** so we will continue to monitor and assess the need for any Temporary Risk Response Measures (TRRMs).

Our work on Teen Recs and Age Appropriate Content (AAC) more broadly for teens continues, but we have shifted out of Lockdown and continue to focus on execution. Going forward, execution will be tracked through standard Execution Tracking and Review Forums led by Youth Integrity as a part of WS2.2: Age Appropriate Content (STO: Arcadiy [redacted] QB: Don [redacted]) – upcoming milestones as a part of this:

1. This week, we've kicked off a vTeam to partner across the company on establishing a Strategy, North Star & measurement, & Plan for Teen Recs/AAC work overall (Driver: Eric [redacted] (Central Youth) with vTeam from XI, IG, FB, Central Youth, & Policy).
2. Next week (week of 10/7), we will have the review on long-term policy options for Teen Recs/AAC (Driver: Nell)
3. October 24th, we will publish the Age Appropriate Content Transparency Center Page (Drivers: Sophie Vogel, Phil [redacted]

Lockdown Summary & Outcomes

WS1.1 Extreme Clarity Nomenclature + Root Cause Analysis (Owner: Molly Cutler)
The goal of this sub-workstream was to ratify revised canonical nomenclature & visuals + complete Root Cause Analysis of WSJ/Reels

- **Status**: Complete - deck here
- Teams are actively using this nomenclature and framework, and in our longer-term North Star discussions we are looking for opportunities to simplify.

WS1.2 Pristine Sources on IG for Nido (Owner: Prashanth)
The goal of this sub-workstream was to launch a version of Pristine Sources for New Users with a stated age of 13-17 by September 17th.

- **How Pristine Sources works**: We take well performing content and heavily filter it with content and actor integrity signals, in addition to applying personalized ranking signals that include integrity demotions. This process limits the recommendations that teens can see to around 44% of our total available inventory, and we are actively working on follow-ups to safely bring in more inventory to improve the new teen user experience while meeting our safety goal.
- **Status**: Launched pristine pool on 8/30, current overall weekly prevalence is above goal at 0.22%.
- **What we shipped for 9/17**
    - An improved pristine pool that retains ~44% inventory and improves the engagement tradeoff for stated age "new teens"
    - A TRRM filter for "Risky Behavior" content using the Risky Behaviors classifier in ESR ranking.
    - Note: We aligned with Tessa not to ship a graduation experience for stated age "new teens" because the production pristine pool already has 0.4% non-rec prevalence for "new teens" from 1000 impressions to the proposed 2500 graduation cap. We will have this mitigation as a TRRM that can be turned on if needed.

- **Metrics:**
  - We established the goal of non-rec prevalence <.2%.
  - We launched Pristine (very low risk) recommendations to mitigate the P0/launch blocking issue of ensuring new teens are not recommended sexually suggestive content. Our goal was 0.2% non-rec prevalence and we are currently at 0.22% weekly non-rec prevalence. We are continuing to monitor and assess the need for any Temporary Risk Response Measures (TRRMs).
- **Current state & Next Steps:**
  - Our weekly overall non-rec prevalence is slightly above goal at 0.22%, driven by prevalence on Tab (0.3%). We are seeing increases in Restricted Goods and Services (RGS) prevalence on the tab.
  - As a first step, we have backfilled the pristine pool content with scores from RGS classifiers so that the signal can be used in ranking.
  - We have also instituted a daily review of top reported and non-recommended content. Our observations are that the increase in prevalence is a mixture of both true and false positives. True positives include videos such as people smoking, false positives include movie clips featuring guns (fiddler). RGS prevalence is higher than before, but maybe not to the extent indicated by metrics. (Historically, regulated goods and services is the non-rec topic with the most labeler error and policy ambiguity).
  - Our integrity systems have also had some issues over the last week with multiple SEVs (1,2), which could be contributing to increase in non-rec prevalence.
  - We are not inclined to intervene yet with TRRMs, but are monitoring the rise closely. We expect the lingering effects of the Sevs to persist into early next week.
- This work will now be led by Prashanth [redacted], taking over from Jon [redacted].

WS1.3 Recommended in Limited Volume for IG (Owner: Prashanth)
The goal of this sub-workstream was to launch diversification for New Users with a stated age of 13-17 by September 17th.

- **How Rec in Limited Volume works:** To mitigate inappropriate content that remains in the Pristine pool, we'll introduce diversity requirements so new teens are less likely to see a high concentration of content about a given topic. We'll do this in a topic-agnostic way first, but could also do it for specific topics (e.g., body change/judgment) if/when a policy framework exists.
- **Status:** At low risk
- **Metrics**: We established a goal of 1% Reels Concentration of Exposure (CoE) DAU, and 2% Explore CoE DAU. Labeling based measurement indicates we have met our goals. *Note: CoE is currently defined as # of new teens with high exposure to sexual content, >10% classifier-based sexual or body-prominent content.
- **Current State & Next Steps:**
  - From weekly labeled data, CoE on reels chaining is between 0-0.1%, and between 0.1-0.5% on the Explore grid (details).
  - Classifier based measurement reads 1.02% for reels chaining, and 1.86% for Explore grid.
- This work will now be led by Prashanth [redacted], taking over from Jon [redacted]

WS1.4 Short-Term Integrity Levers to Improve Content & Account Enforcement for Teen Recs (Owner: Arcadiy, Phil)
The goal of this sub-workstream was to test levers to reduce youth non-rec and violating prevalence overall and to review non-rec lines across multiple policy areas that are high-impact to youth to determine if they can be strengthened and enforced on by September 17th.

- **Context**: We explored levers to reduce harm for all teens by experimenting with additional applications of existing classifiers and LLMs, improving detection & enforcement on profile links, age-gating content based on actor signals, and expanding applications of human review.
- **Updates Made over Lockdown:**
  - **New Policies**
    - **Improved Policy and Enforcement to Reduce Sexually Suggestive Teen Non-Rec Prevalence:** Policy approved using Brand Safety sexual and suggestive content policy as a short-term lever to increase U18 non-rec enforcement on sexually suggestive content.
    - We ended up finding an even more performant solution with our refreshed IG Non-rec Sexually Suggestive classifier, which was introduced/refreshed for all teens (not just cold start, where it

outperformed Brand Safety classifier) – and we reduced Teen Non-Rec Sexually Suggestive prevalence by est. 6.71%.
        - We are now working on long-term codification and implementation of the U18 non-rec content policy.
- **Leveraging and Expanding Existing Policies**
    - **Expanded Age-Gating Policy and Enforcement:** Policy approved specific high-leverage policy lines for enhanced age-gating, most significantly a series of policy lines within Adult Nudity and Sexual Activity.
        - These Policy approvals were launched into scaled human review actioning on July 31, and we launched automation via Similarity Detection on August 9.
        - These changes were measured to reduce age-inappropriate prevalence by 3.5% on IG (as well as by 4.7% on FB).
    - **Leveraging Borderline Labeling to age-gate viral teen content:** We created the "Borderline DCI" system to translate classifier labeling of borderline content (e.g. ANSA, RGS) into direct age-gating enforcement actions on the platform.
        - This was launched on August 2 and we reduced age-inappropriate prevalence on IG by 9.0% (and 4.7% on FB).
    - **Expanded and Improved Non-Rec Enforcement via Ads Policies:** Policy reviewed and approved the use of 17 Advertising Standards (12 for all users, 5 for U18 users) to make organic content non-recommendable.
        - This launched on IG on July 26 (and FB on August 23) and reduced age-inappropriate prevalence by 1.0% on IG (and 0.7% on FB). Some of the impact was already captured by expanding age-gating policy lines, which launched earlier.
    - **Removing Additional ANSA Content:** Policy updated the Adult Nudity and Sexual Activity policy to treat digital imagery of nudity as policy violating, which means this content is now removed for all users when detected. This was launched into human review on July 18.
- **Next Steps:**
    - **AAPC**: Policy, Product, Ops, and a wide XFN are currently engaged in "Age-Appropriate Policy Consolidation" (AAPC) to increase age-appropriate policy coverage across key areas and to better leverage our human review workforce to enforce on those policies.
        - Policy development is scheduled to complete by December 2024, and launches will be in February and April 2025.
        - AAPC has its own SteerCo to track progress.
    - **Predicted Age Enforcement:** The IG Youth Appropriate Experiences team rolled back Predicted Age age-gating (which roughly doubles the number of users whose content we enforce on) because of an interaction issue with Youth Ads (S447544). The IG team will address and plan to roll-out again in mid-October.
    - **U18 Non-Rec Actor Enforcement:** The team ran an experiment with stronger U18 non-rec Actor enforcements based on extending the content strike enforcement window, but this solution had inefficient enforcement and was not launched. The team is reevaluating additional experiment launch candidates for Q4'24.

WS1.5 Long-term Options for All Teens

- **Long-term Policy & Enforcement Options (Owner: Nell)**
    - **Update & Next Steps:**
        - Policy shared options for updates for the longer-term approach to Teen Rec Policy & Enforcement, inclusive of a Blue-Sky proposal, with the working team last week on 9/26 and we're updating the proposal based on feedback.
        - We'll review with Joel/Guy/Naomi the week of 10/7 for a Decision (being scheduled)
- **Parent Content Surveys (Owner: Bryce Bartlett, Kathryn [redacted])**
    - *The goal of this workstream is to develop a classifier we can publicly discuss that uses parent preferences for teen content as a signal for teen content experiences.*
    - **Status**: On track to complete first pilot survey in December and start building a prototype classifier [detailed timeline]
    - **Update and Next Steps:**
        - Finalized vendor selection for survey and aligned on pilot wave markets: US, UK, France, Italy, and Spain
        - Finalized main survey question and drafted initial questionnaire, including variables for diversity of parents; worked with vendors for pass-through survey approach

        - Developed content sampling approaches for survey (all content is public and sampled by teen VPVs)
        - **Flags**: (1) working through legal flag re utilizing U18 creator content; (2) working through vendor contract escalation [neither flag currently puts timeline at risk]
- **Age-Appropriate Content Transparency Center Page for week of 10/21 (Owner: Sophie/Phil)**
    - *The goal of this new workstream is to ship a new page in our Transparency Center in w/o 10/21 (to align with external recommendations reset moment) explaining our approach to delivering age appropriate content to teens.*
    - **Status**: On Track
    - **Update & Next Steps:**
        - Leadership provided feedback on outline suggestions, preferring a hybrid option. Comms + Policy will start drafting the content and share a V1 for review on 10/4.
        - Comms + Policy to work with Transparency Center team re: design options.
- **Longer-term Teen Recs/AAC North Star Strategy & Plan for Meta (Owner: Eric [redacted])**
    - **Update & Next Steps:**
        - Per the highlight above, we've kicked off the vTeam to partner across the company on establishing a Strategy, North Star & measurement, & Plan for Teen Recs/AAC work overall
        - The VTeam includes leads from XI, IG, FB, Central Youth, & Policy.
        - We will aim to land the Strategy & Plan in Q4.

Thank you again to the cross for the hard work on tight timelines here team!
Raena