

Message

From: Jennifer Newstead [/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN= [REDACTED]]
Sent: 6/1/2023 6:10:51 PM
To: Nick Clegg [REDACTED]@meta.com; Alison [REDACTED] [REDACTED]@meta.com; Guy Rosen [REDACTED]@meta.com; Adam Mosseri [REDACTED]@meta.com; David Ginsberg [REDACTED]@meta.com; Antigone Davis [REDACTED]@meta.com; Joel Kaplan [REDACTED]@meta.com; Tucker Bounds [REDACTED]@meta.com; Lena Pietsch [REDACTED]@meta.com; Andy Stone [REDACTED]@meta.com; Nicole [REDACTED] [REDACTED]@meta.com; Tim [REDACTED] [REDACTED]@meta.com; Molly Cutler [REDACTED]@meta.com; Ravi Sinha [REDACTED]@meta.com; Riana [REDACTED] [REDACTED]@meta.com; Amit Bhattacharyya [REDACTED]@meta.com; Monika Bickert [REDACTED]@meta.com; Mark Zuckerberg [REDACTED]@meta.com
Subject: Re: Heads up on a coming story - A/C Priv

+ Mark for context For the Board discussion

From: Nick Clegg <[REDACTED]@meta.com>
Sent: Thursday, June 1, 2023 3:50 AM
To: Alison [REDACTED] <[REDACTED]@meta.com>; Guy Rosen <[REDACTED]@meta.com>; Adam Mosseri <[REDACTED]@meta.com>; David Ginsberg <[REDACTED]@meta.com>; Antigone Davis <[REDACTED]@meta.com>; Joel Kaplan <[REDACTED]@meta.com>; Tucker Bounds <[REDACTED]@meta.com>; Lena Pietsch <[REDACTED]@meta.com>; Andy Stone <[REDACTED]@meta.com>; Nicole [REDACTED] <[REDACTED]@meta.com>; Tim [REDACTED] <[REDACTED]@meta.com>; Jennifer Newstead <[REDACTED]@meta.com>; Molly Cutler <[REDACTED]@meta.com>; Ravi Sinha <[REDACTED]@meta.com>; Riana [REDACTED] <[REDACTED]@meta.com>; Amit Bhattacharyya <[REDACTED]@meta.com>; Monika Bickert <[REDACTED]@meta.com>
Subject: Re: Heads up on a coming story - A/C Priv

+ Monika

Whilst I'm v grateful for all the work that has gone into this – and the clarity with which our enforcement shortcomings have been identified – it is of course hugely worrying that we find ourselves in this position, and that we are performing so much worse in such a sensitive area than our peers.

How quickly do you think we will be able to demonstrate a material improvement in our enforcement performance as we address the root causes of our current flawed practice (errors with auto-closing and reviewer guidance)? I think it's right to engage openly with Jeff as proposed – we cannot shirk our responsibilities here – but it would be even better if we could show him in fairly short order that we are quickly improving our performance.

Thx

N

From: Alison [REDACTED] <[REDACTED]@meta.com>
Date: Wednesday, 31 May 2023 at 00:28
To: Nick Clegg <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Joel Kaplan

CONFIDENTIAL
CONFIDENTIAL

META3047MDL-239-00010159
METANMAG-082-00010159
PLTF EX-02197_0001

PLTF EX

PLTF EX-02197

<[REDACTED]@meta.com>, Tucker Bounds <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Nicole [REDACTED] <[REDACTED]@meta.com>, Tim [REDACTED] <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Ravi Sinha <[REDACTED]@meta.com>, Riana [REDACTED] <[REDACTED]@meta.com>, Amit Bhattacharyya <[REDACTED]@meta.com>

Subject: Re: Heads up on a coming story - A/C Priv

Hi Nick —

We have made progress investigating Jeff's assertions, and found that despite our **policies** being in generally the correct place and substantial investment and improvements in **proactive detection and enforcement**, Jeff has accurately identified previously unknown/unintended gaps in our **reactive enforcement** when violating content is reported by users. Those gaps are going to enable Jeff to publish an extremely salacious and damaging story about how we fail to take action on the overwhelming majority of clearly problematic content flagged to us, complete with vivid and disturbing images and negative comparisons to our competitors.

We have so far attributed the reactive enforcement gaps that Jeff has surfaced to technical errors that we've opened a SEV to address - reports not reaching reviewers - and to human error, where reviewers incorrectly actioned reports. The team is working on addressing the issues that have surfaced now, while also reflecting on how to ensure that we have the ability to systematically detect issues and reactive enforcement in the future. More detail below. (We have been meeting with Guy and Molly regularly over the past week to review investigation progress and our plans to remediate - please let us know if you would like to meet as well as a group.)

From the comms side, the team has confirmed that Jeff intends to characterize our enforcement as abysmal, call out instances in which he believes we've incorrectly drawn policy lines, and negatively compare our efforts and approach to other platforms. Jeff will emphasize that he found - and continues to find - exploitative child sexualization imagery on our platform, and he says that he intends to publish such imagery in the Journal. The team is still working on specific language but intends to make the following points:

1. Explain that most of what Jeff has identified to us violates our existing policies and should have been removed from the platform, either based on user reports or automation.
2. Acknowledge and own the enforcement failures and offer an off record explanation of what took place to build trust.
3. Highlight specific steps we'll take moving forward based on what we've learned and that are directly responsive to what Jeff raised, including blocking additional hashtags and updating operational guidelines for reviewers.

Jeff's allegations generally do not relate to gaps in our content policies. The content that Jeff is concerned about generally violates our policies, should be removed, and frequently include context that should also lead to actor-level penalties beyond our standard repeat offender policies, such as removing an account immediately. You can see more detail about our policies in the table below. The one area where we think that we should make a change to our policies relates to hashtags. Specifically, we will block terms descriptive of sexual activity or clearly associated with child sexual exploitation or nudity, even if those terms are not violating on their own. We will clarify the existing policies and the change to the policy with respect to hashtags with Jeff.

We have a very high volume of proactive enforcement of our policies - and we have improved this very substantially in the last few years. We proactively action this content via classifier work and strategic network disruptions (SND). For Child Sexual Abuse Material (CSAM), almost all of the content we action (99.9%) is detected proactively before it is reported because we match images against previously banked content. For child sexual solicitation content, which generally requires us to understand both image + caption (examples that Jeff flagged) and cannot be matched to image banks - our proactive rate is still high. Over the last 90 days (as of 23 May), of the 24,621 pieces of content we actioned for Child Sexual Solicitation (CSS), 87% was detected proactively (content was not reported prior to being actioned). As mentioned previously, we are also pulling stats on our SNDs in this space and hope to have a stat we can provide Jeff.

Investigating Jeff's reports surfaced enforcement gaps when content is reported by users - gaps that we need to close given imperfect proactive detection and the severity of the violation type. Of the 428 reports Jeff appears to have made from various test accounts, we found that we correctly actioned only 25%, or 105 reports. Of the 75% of reports (323 reports) that we failed to review, remove and/or apply appropriate account-level consequences, we are conducting a comprehensive root cause analysis and have so far attributed the errors to the following issues:

1. *Errors with auto-closing.* We do not intend to auto-close (without human review) any user report that is tagged as "involving a child" - every report should be reviewed. We have discovered 69 reports where this happened and have

opened a SEV to mitigate and resolve those issues. (There were a number of reports that were properly eligible for auto-closure as they were reported for different violation types or not tagged as involving a child.)

2. *Reviewer guidance.* We found that 104 reports were reviewed by and incorrectly ignored by human reviewers, likely because of the use of coded language that may be unfamiliar to reviewers. We are updating the reviewer guidance and our systems to include the problematic terms so that reviewers can have that context when looking at a piece of content. Given the adversarial nature of the space (as our enforcement becomes effective against one set of codes, bad actors will move to another set), policy will also implement a regular cadence of updating guidance to reviewers.

Policy Table

	A. Exonerating context	B. Ambiguous context	C. Malevolent context	D. Severely malevolent context
1. Violating image	Remove image; strike; disable account at a higher threshold; possibly report to NCMEC Ex. [REDACTED]	Remove image; strike; disable account at a low threshold; possibly report to NCMEC Ex. [REDACTED]	Remove image; disable the account immediately; possibly report to NCMEC Ex. [REDACTED]	Same as C.
2. Non-violating image	Allow Ex. <i>Child in swimsuit</i>	Allow	Remove image; disable account at a low threshold; possibly report to NCMEC. Ex: [REDACTED]	Remove image; disable the account immediately; report to NCMEC if accompanied by solicitation indicators Ex. [REDACTED]

From: Nick Clegg <[REDACTED]@meta.com>

Date: Tuesday, May 23, 2023 at 3:35 AM

To: Alison [REDACTED] <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, Tucker Bounds <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Nicole [REDACTED] <[REDACTED]@meta.com>, Tim [REDACTED] <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Ravi Sinha <[REDACTED]@meta.com>, Riana [REDACTED] <[REDACTED]@meta.com>, Amit

Bhattacharyya <[REDACTED]@meta.com>

Subject: Re: Heads up on a coming story - A/C Priv

Thanks for the update – there's clearly a lot of review work in flight to chase down the various aspects of this, so looking forward to the complete assessment later in the week. Then we can decide on next steps – both internally and comms.

N

From: Alison [REDACTED] <[REDACTED]@meta.com>

Date: Tuesday, 23 May 2023 at 01:22

To: Guy Rosen <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Nick Clegg <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, Tucker Bounds <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Nicole [REDACTED] <[REDACTED]@meta.com>, Tim [REDACTED] <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Ravi Sinha <[REDACTED]@meta.com>, Riana [REDACTED] <[REDACTED]@meta.com>, Amit Bhattacharyya <[REDACTED]@meta.com>

Subject: Re: Heads up on a coming story - A/C Priv

Hi Nick-

Below please find an update on work in progress to understand and respond to Jeff's reporting thus far. We don't expect him to publish imminently and believe that the allegations that he made to you over the weekend relate to a series of stories, rather than a single piece. We will send a more complete assessment and comms plan for the series later in the week. From the email sent to you and conversations with comms, Jeff is focused on two main allegations:

1. Our policies allow harmful content and harmful actors to remain on the platform.
2. Our enforcement is poor because we do not review user reports.

Allegation 1: Our policies allow harmful content and repeat offenders to remain on the platform

Response: Most of the content Jeff flagged does, in fact, violate our policies/ protocols and reflects enforcement rather than policy gaps. Most of the violation types are "one and done" and result in removal from the platform after a single violation and a report to NCMEC if it reaches the legal threshold. The potential gaps identified so far include:

1. Possible recidivism: We fan out and should be disabling other accounts across our apps likely to be owned by the same person. Product is working to confirm whether this is an automated or manual process. But even when we do this, our understanding is that [REDACTED]
2. Hashtags: Hashtags are found via user reports or surfaced in ongoing ops sweeps that we are not doing consistently. We are going to expand coverage of those sweeps and improve our ability to more effectively catch different versions of keywords ((i.e., child p0rn" and "ch1ld porn") that change quickly due to the adversarial nature of this group.

Allegation 2: Our enforcement is poor because we don't review user reports.

Response: Jeff is primarily focused on lack of enforcement for user reports he and the safety activists he is in touch with have made. We're looking at all of the reports that Jeff sent in to understand what happened in each instance. We are also determining what we can share about proactive enforcement in this space, which is a significant and missing piece of Jeff's reporting. Specifically for proactive enforcement:

1. We rely on classifiers to surface and prioritize review of likely violating content. Reviewing and removing content in response to user reports is a small fraction of our enforcement activity in this space.
2. Our enforcement approach across Child Sexual Exploitation content in public is to prioritize content based on probable severity and [REDACTED]. This helps us reduce exposure to benign users, but leaves a potential gap for concentrations of users proactively seeking out [REDACTED].
3. We are organizing all of the network disruptions we have done on CSAM content and pulling together a stat to use.

Below is more detail on current work that is being done to address these allegations and the gaps we've identified so far. Please let me know if you have any questions.

Thanks,
Alison

Action Item Deep Dive:

1. Specific charge: The company's action rate on user reports is low.
 - a. AI: Determine our proactive vs. reactive work in this space. Product to pull our enforcement stats for the child sexual solicitation violation type.
2. Specific charge: The company is failing to review most user reports of underage sexual exploitation on the platform.
 - a. AI: I3 to review reports from content flagged by the reporter to understand whether it was reviewed/what action was taken.
3. Specific charge: The reporting flow doesn't offer an option to report child endangerment, only explicit underage nudity.
 - a. AI: Safety policy to provide detail on the testing we've done around the options we offer to report this type of content.
4. Specific charge: IG doesn't monitor hashtags or emojis
 - a. AI's:

Content specifically sent in reporter's email has been reviewed.

IGWB to confirm ground truth on hashtag enforcement.

Ops to provide detail on sweep cadence/how the process works.

5. Specific charge: IG's implementation guidelines seem surprisingly tolerant on near/implied nudity of kids, more so than peer platforms (TikTok and YouTube).
 - a. AI: Policy/ops to confirm that reviewers see more/have more context than what the reporter suggests at scale.
6. Specific charge: Instagram is applying strikes to accounts that post their "menus" of underage sex content, but does not remove these accounts
 - a. AI's

Since 2021, we block devices and fan out on any account that is owned by the same person and we do it cross-app. Product to confirm whether this is automated or manual.

Legal/SR

PRIV

7. Specific charge: The frequency with which users deploy multiple linked accounts, combined with Meta's account-level vs. user/device level penalty policies, undercuts the effectiveness of IG enforcement.
 - a. AI: We do propagate enforcements but recall is low (~19.39%) for all enforcements. Comms and policy to discuss if we want to acknowledge this is an area of improvement that we are moving fast to respond in this area.
8. Specific charge: Instagram does not prevent its users from visiting even hashtags that it knows are frequently used to post child exploitation imagery.
 - a. AI: IG product policy and safety ops to determine if we can expedite changing interstitial language particularly the "see results anyway" language.
9. Specific charge: The reporter has heard from activists that they're being recommended violating accounts in Accounts You Should Follow (AYSF) after seeking out violating content to report.
 - a. AI: IGWB to provide what the intended behavior for AYSF is and whether there are any known gaps we should expedite.

From: Guy Rosen <[REDACTED]@meta.com>

Date: Monday, May 22, 2023 at 1:43 PM

To: Adam Mosseri <[REDACTED]@meta.com>, David Ginsberg <[REDACTED]@meta.com>, Nick Clegg

<[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, Tucker Bounds <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Nicole [REDACTED] <[REDACTED]@meta.com>, Tim [REDACTED] <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Ravi Sinha <[REDACTED]@meta.com>, Riana [REDACTED] <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>, Amit Bhattacharyya <[REDACTED]@meta.com>

Subject: Re: Heads up on a coming story - A/C Priv

My very unheroic answer is that I have learned not to have a directional top line take, but rather to do the root-cause analysis and understand specifically where we have policy gaps, protocol gaps, enforcement gaps or where people may disagree with where we're at. I know the working team is running these down (Adam, Sayed from your team plugged in) – so far I think I've seen a mix, but I want to run these to ground.

Also +Amit who is plugged into the working team from my side.

From: Adam Mosseri <[REDACTED]@meta.com>
Sent: Monday, May 22, 2023 9:15 AM
To: David Ginsberg <[REDACTED]@meta.com>; Nick Clegg <[REDACTED]@meta.com>; Antigone Davis <[REDACTED]@meta.com>; Joel Kaplan <[REDACTED]@meta.com>; Tucker Bounds <[REDACTED]@meta.com>; Lena Pietsch <[REDACTED]@meta.com>; Andy Stone <[REDACTED]@meta.com>; Nicole [REDACTED] <[REDACTED]@meta.com>; Tim [REDACTED] <[REDACTED]@meta.com>; Guy Rosen <[REDACTED]@meta.com>; Jennifer Newstead <[REDACTED]@meta.com>; Molly Cutler <[REDACTED]@meta.com>; Ravi Sinha <[REDACTED]@meta.com>; Riana [REDACTED] <[REDACTED]@meta.com>; Alison [REDACTED] <[REDACTED]@meta.com>
Subject: Re: Heads up on a coming story - A/C Priv

If Jeff's take is even directionally accurate that's super concerning. Guy, I know your team is already digging into this, but it would be great to get your take on whether we do think we're systemically behind or if Jeff's take is based on a small number of anecdotes? Also, I want to be clear that we're happy to redirect resources from our safety teams on Instagram as well if that would be helpful.

From: David Ginsberg <[REDACTED]@meta.com>
Date: Saturday, May 20, 2023 at 23:30
To: Nick Clegg <[REDACTED]@meta.com>, Antigone Davis <[REDACTED]@meta.com>, Joel Kaplan <[REDACTED]@meta.com>, Tucker Bounds <[REDACTED]@meta.com>, Lena Pietsch <[REDACTED]@meta.com>, Andy Stone <[REDACTED]@meta.com>, Adam Mosseri <[REDACTED]@meta.com>, Nicole [REDACTED] <[REDACTED]@meta.com>, Tim [REDACTED] <[REDACTED]@meta.com>, Guy Rosen <[REDACTED]@meta.com>, Jennifer Newstead <[REDACTED]@meta.com>, Molly Cutler <[REDACTED]@meta.com>, Ravi Sinha <[REDACTED]@meta.com>, Riana [REDACTED] <[REDACTED]@meta.com>, Alison [REDACTED] <[REDACTED]@meta.com>
Subject: Re: Heads up on a coming story - A/C Priv

Thread forked so I, nick, wanted to make sure you saw Andy's response earlier also.

We do not have a firm sense of timing from Jeff yet.

Get [Outlook for Android](#)

From: Nick Clegg <[REDACTED]@meta.com>
Sent: Saturday, May 20, 2023 6:29:42 PM
To: Antigone Davis <[REDACTED]@meta.com>; Joel Kaplan <[REDACTED]@meta.com>; David Ginsberg <[REDACTED]@meta.com>; Tucker Bounds <[REDACTED]@meta.com>; Lena Pietsch <[REDACTED]@meta.com>; Andy Stone <[REDACTED]@meta.com>; Adam Mosseri <[REDACTED]@meta.com>; Nicole [REDACTED] <[REDACTED]@meta.com>; Tim [REDACTED] <[REDACTED]@meta.com>; Guy Rosen <[REDACTED]@meta.com>; Jennifer Newstead <[REDACTED]@meta.com>; Molly Cutler <[REDACTED]@meta.com>; Ravi Sinha <[REDACTED]@meta.com>; Riana [REDACTED] <[REDACTED]@meta.com>; Alison [REDACTED] <[REDACTED]@meta.com>
Subject: Re: Heads up on a coming story - A/C Priv

Thanks - yes an update asap on Mon would be good so we can decide next steps etc.

When is Jeff H planning to publish?

N

Sent from Outlook for iOS

From: Antigone Davis <[REDACTED]@meta.com>
Sent: Saturday, May 20, 2023 11:28:48 PM
To: Nick Clegg <[REDACTED]@meta.com>; Joel Kaplan <[REDACTED]@meta.com>; David Ginsberg <[REDACTED]@meta.com>; Tucker Bounds <[REDACTED]@meta.com>; Lena Pietsch <[REDACTED]@meta.com>; Andy Stone <[REDACTED]@meta.com>; Adam Mosseri <[REDACTED]@meta.com>; Nicole [REDACTED] <[REDACTED]@meta.com>; Tim [REDACTED] <[REDACTED]@meta.com>; Guy Rosen <[REDACTED]@meta.com>; Jennifer Newstead <[REDACTED]@meta.com>; Molly Cutler <[REDACTED]@meta.com>; Ravi Sinha <[REDACTED]@meta.com>; Riana [REDACTED] <[REDACTED]@meta.com>; Alison [REDACTED] <[REDACTED]@meta.com>
Subject: Re: Heads up on a coming story - A/C Priv

+ Ravi Sinha

From: Antigone Davis <[REDACTED]@meta.com>
Sent: Saturday, May 20, 2023 6:24 PM
To: Nick Clegg <[REDACTED]@meta.com>; Joel Kaplan <[REDACTED]@meta.com>; David Ginsberg <[REDACTED]@meta.com>; Tucker Bounds <[REDACTED]@meta.com>; Lena Pietsch <[REDACTED]@meta.com>; Andy Stone <[REDACTED]@meta.com>; Adam Mosseri <[REDACTED]@meta.com>; Nicole [REDACTED] <[REDACTED]@meta.com>; Tim [REDACTED] <[REDACTED]@meta.com>; Guy Rosen <[REDACTED]@meta.com>; Jennifer Newstead <[REDACTED]@meta.com>; Molly Cutler <[REDACTED]@meta.com>
Subject: Re: Heads up on a coming story - A/C Priv

Hi Nick,

The allegations are serious, and the team has been working both on a response and a comprehensive review to identify any gaps/potential deficiencies etc. and close them. They also have an update to leadership in the works. My understanding is that they are/were planning to send it on Monday or Tuesday. I believe they are waiting on a few additional insights. I will see if they can expedite the update.

- Antigone

From: Nick Clegg <[REDACTED]@meta.com>
Sent: Saturday, May 20, 2023 5:45 PM
To: Joel Kaplan <[REDACTED]@meta.com>; David Ginsberg <[REDACTED]@meta.com>; Tucker Bounds <[REDACTED]@meta.com>; Lena Pietsch <[REDACTED]@meta.com>; Andy Stone <[REDACTED]@meta.com>; Adam Mosseri <[REDACTED]@meta.com>; Antigone Davis <[REDACTED]@meta.com>; Nicole [REDACTED] <[REDACTED]@meta.com>; Tim [REDACTED] <[REDACTED]@meta.com>; Guy Rosen <[REDACTED]@meta.com>; Jennifer Newstead <[REDACTED]@meta.com>; Molly Cutler <[REDACTED]@meta.com>
Subject: Fwd: Heads up on a coming story - A/C Priv

David, Joel, Guy - see below. Obviously these are serious assertions - I know we have been working on responses, but it would be good to get an update on the central allegation that we are out of step with other industry players here. Is there a process underway? N

Sent from Outlook for iOS

From: Horwitz, Jeffrey <jeff.horwitz@wsj.com>
Sent: Saturday, May 20, 2023 9:13 pm

To: Nick Clegg <[REDACTED]@meta.com>
Subject: Heads up on a coming story

This Message Is From an External Sender

Hello Nick --

I don't imagine this has risen to your folks' attention, but there's an Instagram child safety-focused project I'm working on that I wanted to tell you about. It's already in the able hands of Andy Stone and Stephanie in comms, so everything's in fine shape in terms of seeking official response and getting specific problem points resolved. But given the limited mandate of comms, I figured I might try to raise a few things that I've found which would be relevant on the underlying policy front.

There's no need to get back to me if you're confident everything is being handled/escalated properly. You've got my word that this email is not a weasely gotcha move so I can print that "*Nick Clegg didn't respond to an email about the problem*" or some crap like that. However, if it's of interest to you, I'm happy just to walk through the below in detail and answer any questions you might have.

The broad overview of the problem is that there seems to be a gap between Instagram's child safety measures and some rivals. Pretty much all of the below came as a surprise to me.

On an Ops level, I'm comfortable saying empirically that the company is failing to review most user reports of underage sexual exploitation on the platform. This isn't isn't a brief capacity shortfall. When reported to IG, posts and accounts featuring near-naked minors which explicitly state they sell underage sex content often receive a response stating that Instagram doesn't have the resources to review them. (I understand that the realities of content moderation require triage of user reports, but I had always assumed that this was an area where the company would generally close out the queue.) Also, the User Reporting Flow doesn't offer an option to report child endangerment, only explicit underage nudity.

A related ops problem is that, having seen hundreds of reports filed by child safety activist types over the last year, when Instagram does review the accounts of users clearly stating they sell underage sex content, it generally does not take them down. I'm still figuring out what the numerator and denominator are, but if IG removes 20% of the blatant stuff reported by users I'd be surprised.

There's also been a notable failure to monitor hashtags that are extremely well known among Instagram's pedophile community. Meta has blocked a number of these (#pedowhore) that have been flagged by me and Stanford researchers already, but there are still plenty left, such as #MNSFWContent. (As of today it returns 1,700 overwhelmingly inappropriate posts from users purporting to sell novel underage sex imagery). These hashtags are disabled on TikTok, YouTube, Twitter or Snap, suggesting that Instagram isn't doing some basic tracking of adversarial behavior.

There's also a problem with emojis. There are probably a half dozen that feature in what could only be considered an extremely lazy effort at pedophile code, but 🍆 and 🍑 are glaring. If Ops can't turn up 5,000 bad accounts with high precision via a search for users who combine those two symbols/words in their handles or bios, I'd be shocked. (In case you're wondering why the map, please consider that some pedophiles would like to be known as "minor attracted persons.")

These are operational issues, however -- not your part of the org. The stuff I wanted to raise with you is more policy-focused. Here's a list of a few things that I've come up with. It's possible that comms will push back on one or two of them in a way that's compelling, so please consider this stuff provisional.

1. IG's implementation guidelines seem surprisingly tolerant on near/implied nudity of kids, more so than peer platforms (TikTok and YouTube). The specific definitions seems to permit images in which

--- a child between the age of ~5 and 18 is wearing "at least one item of clothing between their neck and knees."
--- 2. AND in which there is "no visible genitalia or female nipples"

According to the implementation standards I've seen, this means that a partially nude child whose genitals are covered by a "digital" or "real life" object is acceptable content. So a 12 year old girl/boy wearing only a T-shirt and whose crotch is covered by a teddy bear or heart emoji is permissible, barring extenuating circumstances.

Although the detail of this implementation standard suggests intentionality, I have doubts as to whether it was considered in relation to the privacy-limited structures of outsourced content moderation. For example, it's my sense that a reviewer looking at the above-referenced post showing a partially nude kid would not necessarily see if their handle was "TeenSlut4U" or the account bio said it was "selling content." While this might seem like a gross edge case hypothetical, I assure you it's not -- just take a look at the posts in that "minor not safe for work" hashtag linked to above.

2. Instagram is applying strikes to accounts that post their "menus" of underage sex content. I've gotten the sense that, post the Haugen leaks, IG stopped allowing users to accrue three strikes for posting "child exploitation content" before account deletion. Something which still remains the case, however, is that the platform doesn't remove accounts that have been caught posting "menus" of sex content for sale. This is, at most, a strike. Even in the adult context, this would seem like a lenient interpretation of the rules -- but these are minors.

3. The frequency with which users deploy multiple linked accounts, combined with Meta's Account-Level vs User/device Level penalty policies, undercuts the effectiveness of iG enforcement. It is extremely common for accounts serving Instagram's pedophile community to have multiple linked backups. The 14 year old content seller @Kittyxoxo1 will regularly list @kittyxoxo2 and @kittyxoxo3 in its bio, encouraging an underage girl's fans to follow the redundant, dormant accounts. In the event that Kittyxoxo1 gets taken down, Kittyxoxo2 immediately resumes posting to the thousands of accounts following the girl. This isn't a rare tactic for underage content sellers -- it's standard practice. The result is that Instagram will, over the course of the year, remove multiple accounts belonging to each underage content seller while doing zero damage to the actual network of underage content seller accounts. I could understand why the above is something that the Ops people wouldn't want to stress to Policy, or maybe even think about -- it's depressing.

4. Instagram does not prevent its users from visiting even hashtags that it knows are frequently used to post child exploitation imagery. IG is in the middle of cleaning up some of the worst offenders, but I've had a hard time understanding what purpose the second option of the below popup serves. If the content will likely surface child sex abuse, why would the platform want to allow it to be viewed?

Q #pedo

These results may contain images of child sexual abuse

Child sexual abuse or viewing sexual imagery of children can lead to imprisonment and other severe personal consequences. This abuse causes extreme harm to children and searching and viewing such materials adds to that harm. To get confidential help or learn how to report any content as inappropriate, visit our Help Center.

Get resources

See results anyway

The above brought to mind Antigone Davis's comment to the Guardian a few weeks back that "we've been able to identify the kinds of searches that people do when they're searching for trafficking content, so that when people search for that, we will pop up with information to divert them or to let them know that what they're doing is illegal activity".

I very much understand why offering resources and reminding a user that their search queries are noticeable is a good policy. I have had a hard time understanding, however, why after informing people they're about to open a box of child porn IG would want to let them go ahead and open it.

The above items are of course part of the stuff I expect to write about, and I've raised it with comms. Perhaps this is all known territory where the company has considered the tradeoffs between safety and voice and determined that the current policies strike the right balance. But I'm writing based on my hunch that the current state of play on the platform might come as a surprise.

If that hunch is correct, I'm happy to talk about this stuff in detail before or after publication. Something that the academics and activists who I've consulted on this reporting project have instilled in me is how even half-assed adversarial techniques can exploit soft spots in policy. There's a community of people out there who've studied this network externally with some success, and I'd be glad to share what I've learned from them.

While a large part of my job is devoted to finding and reporting on problems, I'd be very glad if reporting on these topics got more difficult down the road. I may send a version of this note to Adam Mosseri as well, given that the community devoted to this specific area of problem content seems heavily focused on Instagram.

Sincerely,
Jeff

Jeff Horwitz
Reporter, SF Bureau
Wall Street Journal
[REDACTED] cell