Tennessee v. Meta
Case No. 23-1364-IV
PX-00020

Child Safety State of Play *Interim update - September 2020*

Note: Updates have been ported over to Child Safety State of Play (9/24). Please highlight any additions and tag Karuna [redacted] in comments.

July document Child Safety - State of Play (7/20)

Public Policy landscape

Europe

- FB EMEA leadership sees a concerted effort to go after tech in Europe and Facebook's business model specifically.
  - Google and Microsoft all considered to have too much power in the European and Global marketplace.
  - Surveillance narrative (as expounded by Zuboff) broadly accepted in Europe and US tech has not been able to mount an effective counter narrative.
- Safety, due to the consensus around the importance of protecting children in the digital space, being used successfully in Europe, by the EU and Member States, through regulation, in part, to reassert control of Europe's digital destiny.
- Our decision to encrypt Messenger is seen as symptomatic of a pivot to privacy, at the cost of children's safety. That US notion of data privacy, is not the same in Europe and therefore exacerbates the situation.
  - The European Union Internet Forum for example, consciously excluded Facebook from its draft technical paper on encryption.
- A number of separate stakeholder groups (predominantly privacy, safety and security but including some of our competitors) are mounting well-coordinated attacks, leveraging the regulatory momentum both within countries (member states) and across the European Union.
  - Our competitors (Microsoft in particular and Google to a lesser degree) are accentuating that approach and along with so-called experts, seek a EU legislative response during Q2 2021, that could be detrimental to our plans and in Microsoft's case, have little impact.

[redacted]

[redacted]

[redacted]

AMET Policy and Regulatory Context

There are neutral to negative sentiments generally around Facebook in the region with concerns over our platforms impact on misinformation, hate speech, privacy and its effect on young users in the region. Our market domination in the continent is also seen as largely problematic by policy makers who are churning out social media, tax and competition legislation to deal with this.

Child safety concerns are growing in the region especially after the increased use of digital platforms by children because of school closures due to covid-19. Access issues and digital literacy form the majority of concerns as the level of online safety awareness is very low and younger users are exposed to inappropriate content. Facebook's efforts to address this have come under scrutiny from CSO partners discontent with our digital literacy investments in the region. Our competitors, Google are investing heavily on child safety in the region; Be Internet awesome and they recently announced a 1 million USD fund to support safety NGO's in Africa.

[redacted]

APAC Trends

[redacted]

- **Dedicated child safety regulators with increasing powers.** Australia and Fiji already have an esafety commissioner, the model (which Australia continues to spruik) is gaining some traction among CSOs and legislators in Taiwan and to some extent, the Philippines, at present. Various government entities in South Korea are also attempting to stake claim over online child safety (and NCII). While updated regulation and government coordination in these countries could be a welcome improvement, we are seeing multiple proposals, largely in conflict with US law, to expand our reporting obligations and data sharing to local LE and gov and - in the case of South Korea - require us to proactively detect/prevent uploads of content the gov deems violating without the ability to verify what the content is (i.e ingest CSAM/NCII hashes from gov not just other industry/NGOs).
- **Safety by Design.** eSafety's **SbD** approach continues to gain traction with an increasingly global audience (collaboration with VCs and Alex Stamos in the US, multi-lateral/INGOs such as UNICEF and ICTU and seeding approach with APAC/South Asian countries). Phase 2 moves beyond principles to best practice recommendations on internal processes, integrity measures and settings to protect minors. While these are still voluntary, the proposed AU Online Safety Act already points to SBD as the basis for minimum requirements/expectations of intermediaries, so could form the basis of legal obligations in future and will continue to have visibility.
- **Safety/Privacy tension increasingly apparent**: secrecy of communications provisions in Japan and Korea are calling into question the legality of our scanning for harmful content while other regulatory proposals in both countries require either collecting and sharing of more user information with local authorities/civil complainants or increasing efforts in proactive detection of violating illegal content. (mirrors tension in EU with ePrivacy directive?)
- **Push for age-verification increasing.** In South Korea and Japan current proposals requiring we verify age through mobile phone number verification. Philippines Dep of Ed issuing advisories for u13s to use Messenger for home-based learning leading to clear exposure to us having millions of u13s on the platform, which we aren't currently detecting/enforcing against.
- **CEI/NCII prevalence risks fueling throttling or unhelpful data regulation:** Pakistan (X platforms) already banned due to "immoral content" and FB child safety issues used to fuel Citizen Online Protection Rules/control of speech.
- **Livestreaming abuse and coordination with 5 eyes.** Local momentum on public calls by senior gov (Dutton etc) is linked particularly to narrative around live-stream abuse between ANZ and PH. ANZ also in lock-step in response to our e2ee intentions and engagements, AU leading coordinated engagement with UK/US.



Immediate Product Vulnerabilities

Product

- **Facebook**
  - **Inability to auto-action using our proactive CEI classifiers:** Our CEI video and photo proactive classifiers do not perform well enough to auto-action content because we don't meet required hard action (delete) or soft action (downrank/quarantine/friction) thresholds. This has implications for company priority products and the future of their child safety measures:
    - Rooms: Product continues to push for growth and plans to allow Rooms creation directly from WA and IG without FB log-in. They are also currently testing relaxing initial integrity measures, such as requiring an account age of 30 days to see if this leads to growth. **If we want to avoid having Zoom's child safety problems, we need to prioritize filling the following product and enforcement gaps:**
      - We hope to land alignment amongst policy, legal, and product leadership for an in-call evidence capture solution in H1 2021 but in order to review and enforce against our policies quickly and at scale, we'll need improved CEI proactive video classifiers that can detect child nudity;
      - We are also discussing with product the possibility of moving some of our riskiest users and communities onto public spaces — whether via public Rooms (subject

to proactive harmful content scanning much as we do for FB Live) or public Messenger threads (forked off from e2ee MSGR infra) that contain more than 1K users. Any such solutions will require improved CEI proactive image and video detection.

- Before we move to e2ee, we should **prioritize reducing CEI VPV's (views) and preventing bad actor connections:** Data analysis has shown that large message threads may harbor a higher concentration of bad users and harmful content shares. If we can drive down our NCMEC report numbers prior to our move to e2ee with demonstrable reduction in CEI shares, we could be better equipped to push back against calls for exceptional access. As product expands into "Community Chats" (chats for >1K users) and seeks to raise the ceiling on thread sizes as we move to Cross-App Communications, we need to invest in data science and product intervention experiments to understand what solutions best achieve these goals.
- **Immature minor sexualization classifier:** Implicit Minor sexualization policies may launch this half but our AI tooling is not mature enough to proactively detect complex objects (where a page/group/profile shows a collection of borderline to benign images of minors accompanied by minor sexualizing comments/captions) because our page/group/profile-level detection has not been set up yet. In addition, we've been detecting and auto-actioning explicit minor sexualizing comments after launching a policy to enforce against explicit minor sexualization last half. However, our detection only runs on comment text, which can lead to false positives if the media being commented on is benign/borderline, or false negatives if we do not detect explicit minor sexualizing captions attached to a photo/video featuring a minor.
- **Provide users with meaningful ways to report ephemeral content:**
  - Vanish Mode (VM) and Unsend: If we are to stay within our commitment to ensure users can report harms even when we're e2ee, we must provide users with a meaningful way to report.
    - Our current safety mitigations for VM include (1) Gating suspected groomers from the product experience altogether; (2) allowing users to report within 1 hour (for Messenger) and 14 days (for IG, though this is likely to change to achieve parity with Messenger) after the message is seen; (3) educating users about their ability to report vanish mode messages; (4) VM is only available in existing threads.
      - Educating users about their ability to report is likely to be a low return on investment without extending the 1 hour retention period of VM messages. Recent data analysis shows that only 57% of grooming reports occurred within the hour of the first reported message sent. For CEI, it's even worse — only 46% of CEI reactive reports were sent within the hour. However, the data also shows that after the first hour, the number of reports drop off dramatically (2-5% made within each subsequent hour). We should further explore upselling reporting so and provide sufficient CO review capacity to work the queues in a timely manner.
    - Unsend is also (1) reportable within 14 days of unsend; and (2) accompanied by user education regarding their ability to report within 14 days of unsend.
- **Close enforcement gaps revealed by NCMEC reports data analysis:**
  - Child Safety data science work is ongoing to contextualize NCMEC report numbers. One of the key enforcement gaps that have emerged is what to do with CEI recipients. **Currently, we neither detect nor enforce against CEI recipients.** However, recent data analysis has shown that for only 38% of recipients of NCMEC-reported content was this the first content they received while 29% had received 5 or more NCMEC-reported content (1.8M accounts). When compared against the fact that only 7% (58K accounts) of CEI uploaders, whom we do proactively detect/report/disable, have uploaded 5 or more pieces of NCMEC-reported content, we see that our current enforcement regime is capturing an incomplete picture and perhaps incorrectly over-enforcing against uploaders and/or under-enforcing against recipients. Interestingly, Asian and MENA dominate both uploaders and receivers and only 3% are from the U.S.
  - Based on this ongoing work, **we should refine the way we currently gate or feature block NCMEC reported users from products and product experiences.** Currently, only Messenger Kids and Rooms are using a filter to block NCMEC reported users who have shared viral or severe meme CEI content but have not yet been disabled from our platform. This is a stop-gap measure to protect our most vulnerable users and risky products but in the near future, we need to a) revise our enforcement policy against CEI recipients and include malicious CEI recipients in these product gating filters as well as b) refine our ability to target malicious NCMEC reported uploaders to make consistent product gating decisions across FB and IG.
- **COVID review constraints:**
  - CO review capacity for CEI is at ~30%; CO Safety suspects this will continue to be true for many more months to come for large pools of queues (125k hr/wk across all teams) with <20% projected supply for the near future.
  - With the exception of MMS and CEI content classifiers, we're not currently running classifiers

nor reviews on threads with >2 users in them during COVID b/c of capacity constraints and b/c these have very low action rates with very time consuming review processes

- Groomer detection remains on for English on both IGD and MSGR and Spanish on FB during COVID. There is no Tier 1 IIC (which detects sextortion or plans to meet up in real life for sex) detection running for French, German, Indian, or Portuguese markets. We are slowly getting back French, German, Portuguese but Indian offices are still offline as they keep having outbreaks. We're also trying to expand to more Tier 1 markets but there's so little review capacity that in order to do so, we'll have to make a tradeoff elsewhere.

- **Nudity & Pornography (N&P) review reductions:** Due to the lack of graphic content review capacity across Community Operations, there is a plan to reduce N&P's review capacity by ~80%. Child Safety teams rely heavily on N&P queues to catch novel CEI content. Short term, this will reduce volumes of escalations, reduce novel CEI banking and increase banking latency. Longer term, the lifting of graphic content WFH restrictions will help to address this.
- **Need more investment in proactive detection of Minor Sex Trafficking:** Need a proactive classifier that allows us target minor sex trafficking and distinguish it from adult prostitution activity. Also need further definition of what constitutes trafficking and where our policies draw the line between it and prostitution.
- **Classifier global resilience:** Need a comprehensive audit to assess whether our child safety classifiers are globally resilient and do not operate with bias. There is ongoing Child Safety engineer work to update our text classifiers to more easily internationalize and integrate new languages for proactive grooming detection but we should do the same global work even for our language-agnostic, image-based classifiers.
- **Improve our text classifiers that detect inappropriate interactions with children**: Child Safety CI is working to develop a text model across both FB and IG that will capture conversations with minors where someone solicits images from them. Proactively detecting this type of conversations within posts and comments would be incredibly helpful. In addition, there is ongoing work to improve our existing text classifiers that detect grooming to achieve two goals: 1) help us land automation of hard actions (disable/report) against groomers who are making plans to meet minors in real life for sex and 2) eliminate the need for new engineering when adding new markets because these new text models will perform well even in languages for which we have no training data.
- **Improve ability for users to report child safety issues anywhere on our platform.** Child Safety CI has continued to charge forward on their goal to ensure that anywhere someone could see or experience child sexual abuse on our platform, they could tell us about it. As part of that effort, we recently shipped a user report feedback tag, "Involves a Child," on IG Profiles. This was met by significant praise from stakeholders who had petitioned for this improvement. However, we need to continue closing existing gaps in reporting coverage and launch the "Involves a Child" feedback tag where it is missing across the platform, for example, in FB Groups and IG comments.

- **IG**
  - **Reduce prevalence of grooming on IG:**
    - The daily number of minor victims having sex talk conversations in IGDM with adults are approximately 38x (though we should note that there are 7x as many minors on IG as on FB, and at least 7x as many victims) more than on FB Messenger.
      - Child Safety CI, Messenger Wellbeing, and IG Wellbeing collaboration is ongoing to bring down the prevalence of grooming instances in IGDM's this half.
      - **This work will require robust data science work and product intervention simulations and tests to verify what's most effective.** For example, we believe that preventing suspected groomers (or adults who share no mutual follows with the minor) from messaging, friending, or viewing the profile of minors might be an effective intervention to drive down grooming prevalence on IG Direct. IG, however, does not have its own Child Safety data scientists and progress on work will likely suffer without any additional data science resources.
      - **It is a known gap that IG does not provide the same protections offered to minors on Messenger**, which is to prevent unconnected adults (who is neither Friends nor Friends of Friends with a minor) from initiating messages with a minor. We have spoken publicly about this protection (including in Jay Sullivan's Senate Judiciary Committee testimony and attached QFR's) and should strive to bring IG up to par. In Cross-App Communications, IGD and MSGR adults will be prevented from messaging MSGR minors, but the same will not be true of MSGR adults who message IG minors.
  - **Bring the NCMEC reported signal to Instagram:** This signals to product teams users who have shared CEI but not been disabled. Not all accounts that share CEI get disabled (if viral content, accounts get checkpointed and accrue a strike). We share this signal to product teams for informing decisions around whether users should be allowed to use our features. It's currently running on Rooms and Messenger Kids, and bringing it to Instagram would provide an additional signal for our recommendation surfaces like Explore, and new products like Reels.

- **Improved age detection models:** These will significantly help our actor classifier on Instagram. As we roll out improved age affinity models, this will be valuable classification data for our metric to perform better.
- **Disrupt the ability for predators to discover minors on our surfaces:** Unlike Facebook's People You May Know model, which prevents unconnected adults from discovering or initiating conversations with minors, Instagram has no similar logic for Accounts You May Follow. We're in the process of deep-diving into true positive prevalence labels to determine the most common surfaces where violators find their victims, so we can measure the impact of our efforts to reduce minor discoverability on selected surfaces (e.g. search, suggested accounts).
- **Understand actors who engage in violating behaviour on Instagram:** IGWB Research is developing archetypes of child safety violators for Instagram partnering with members of the Child Safety XFN to incorporate the Intent Taxonomy X Persona framework that has already been developed.
- **Increase reviewer capacity for suicide / self-harming content on Instagram:** When people search for suicidal or self-harming content on Instagram, we block hashtags that are likely to be violating and offer support language over hashtags that may contain sensitive content. However, accounts with these names continue to be surfaced as search results, due to the accounts not having been proactively detected, reviewed, and taken down for violating our policies. We need more reviewer capacity as well as the ability to bring our proactive detection technology to the EU, so that we can proactively scan for and take action on violating accounts / content.
- **Bring our SSI proactive detection technologies to Europe:** We are limited from running our self-harm classifiers that account for both text and media signals and determine if a post has a high likelihood of being actionable for suicide or self-harm.
- **Virality on Instagram:** Minors who go viral on Instagram will likely experience increased interactions with unknown people. As their content and by extension their account becomes surfaced to more unconnected people, they would benefit from additional protections that they can turn on, like limiting who can send them private DMs or follow requests. We should also consider limiting their visibility to adults who score above certain thresholds on our IIC classifiers. More broadly, we need more research and product commitment to specifically addressing how virality on Instagram impacts young people, and their mitigations for this impact.
- **Public accounts:** Instagram as a place for discovering new creators also means a large percentage of minors on the platform live their lives out in the public eye. Encouraging more use of private accounts for minors would allow them to creatively express themselves without immediately opening them up to strangers or unknown interactions. We currently have a two-week test running to explore the viability of Private by Default for all minors; we should strongly advocate from the policy side that the safety, privacy, and legal benefits of defaulting minors to private accounts should outweigh any Growth concerns. Private by Default for Teens would provide a tangible example of IG not only generally considering the interests of minors, but actually implementing distinct product features/experiences catering to their needs. This could go a long way towards helping us address regulatory pressure in Europe and avoid legislation requiring us to implement even more restrictive default settings.
- **Media perception of new Instagram products like Reels:** We are generally perceived as being 'behind' our competitors - especially TikTok - in youth safety. For example, they do not allow anyone under 16 to receive direct messages, and have parental controls that allow a parent to link their TikTok account to their teen's and set controls. As we advance our commitment in Reels and make it the next big bet for Instagram, we need more visible interventions and product commitments to avoid fuelling this external narrative.
- Child Safety landscape - Instagram for more details

* **Age**
  - **The best performing age model (the "Adult Classifier") is not yet globally operational, which leads to a patchwork of integrity measures across the company that leverage age**, including our Child Safety models that provide protections for minors on our platform. We need more review resources to establish ground truth data on age in priority markets.
  - **Age verification pressures:** We will likely continue to face external pressure to verify age. YouTube announced this week that they will ask European users to provide proof that they're above 18 if they seek to access age-restricted videos. The goal is to help YouTube comply with the EU's audiovisual media services directive, which requires platforms to protect minors against harmful material. We could take a similar approach where we use age prediction to put people into additional age verification checkpoints based on whether a piece of content or product feature is age-gated. Or treating all users as u18 for any of these age-gated features or content until our age models generate a prediction that they're over 18. The challenge is getting our age models globally operational and justifying such an approach to product growth teams.
  - **Our biggest enforcement gap in relation to age for child safety is currently our inability to automate decisions accurately to enforce against groomers.** The Adult Classifier — while it is our best available age model to determine whether someone is a teen or non-teen — cannot determine age gaps at a granular level which hampers our ability to action our policies, which rely on making specific age determinations, i.e. if the victim is 14 years old or younger and the offender is 18, disable and escalate but if the victim is 15 and the offender is 18, checkpoint the offender.
  - **Age models do not work as well on IG** because IG only started to collect age data for new sign ups in

December 2019. Thus, we currently only have stated age for ~67% of IG users. Age collection on IG has previously been criticized as being for the purpose of targeting ads, and intrusive to users who do not necessarily need to bring their authenticated selves to the platform. Relying on stated age on similar apps in the industry also have also resulted in age lying and misrepresentations, and fines for collecting children's personal information. We recommend engaging with IG leadership on building better age affinity models to predict age of our users, and prioritise age model maturity on IG.

- **We have de-prioritized enforcing against u13s on our platform.** We are neither building a proactive u13 classifier nor prioritizing enforcing against reported u13s. Recently, Identity Integrity did agree to enable cross-enforcement of u13s (so now, when a user on IG is reported for being u13 and is check-pointed, we will carry this enforcement over to their hard-linked FB account). However, there's also a current u13 backlog (~700K job backlog) that must be addressed via workload allocation and automation (i.e. auto-accept all other age change requests, which is a combo of people going from 13-18 > 18 and people going from >18 to 13-18). This backlog is currently not resourced.

Content Policies

These behaviours are either only enforced against upon escalation or not enforced against at all right now, and would be helpful gaps in our Implementation Standards to close:

- Child sexual exploitation:
  - Empty IG accounts that primarily leave explicitly sexualising comments on pictures of minors - currently, upon escalation where we confirm the account-holder is regularly posting sexualizing comments on images of minors, the account will be disabled.
  - Empty IG accounts that primarily leave implicitly sexualising comments on pictures of minors
  - Empty IG accounts that only engage in behavioural actions e.g. likes, saves of images of minors
- Suicide and self-injury:
  - Better definitions for eating disorders - While eating disorders currently reflect a small % of our reports under suicide and self-injury, communities around eating disorders do not tend to report one another, and we see many accounts, hashtags, and comments that continue to promote eating disorders.
  - Some of the hypotheses that we've been thinking through is that ED content is too narrowly defined by our policies so we under-enforce on what the broader public might consider to be ED content. ED is also rarely a single piece of content, and more a collection of content and behavioural signals. We need more research in this space.

Enforcement

- Add and staff an "involves a child" reporting option across all reportable surfaces e.g. account, comment, DM, hashtags
  - Tests and launches of the "involves a child" reporting option at the post and profile levels have increased the precision of detecting and enforcing against child safety. Adding this reporting option across all surfaces will allow posts to be prioritised for moving out of the broader Nudity & Pornography queue and being reviewed for child safety violations.
  - Current reporting structures hinder our ability to measure the true prevalence of CEI or other child safety violations, as reports are aggregated along with other nudity and pornography violations.
  - Include more granular reporting options to capture different types of child safety violations e.g. distinguishing between a comment soliciting CEI and a comment implicitly sexualising a minor
  - Move child safety reporting options out of the "Nudity & Sexual Activity" category to appear earlier in the reporting flow, with more granular options to capture a wider range of behaviours

Resources / Education

There are a number of opportunities to fill gaps in education and resources related to child safety:

- Create a well-funded digital citizenship and wellbeing program combining Get Digital (Global), We Think Digital (APAC) and other outliers (e.g., My Digital World - SSA) into a robust, global learning platform engaging families, educators and youth to take an active role in online safety.
- Address the general lack of knowledge regarding safety tools currently available on the platform:
  - Elevate positioning of existing online safety resources (e.g., Facebook Safety Center) for increased awareness and usage. Create greater visibility within the Facebook platform - linked from the profile page and elsewhere.
  - Embrace opportunity to proactively promote safety tools and resources pushing out through house ads and other internal mechanisms to increase visibility.
- Create a Child Safety Hub - an online, global resource that explains Facebook's approach to combating child exploitation - targeting Policymakers, CSE experts, industry leaders, and the general public (perhaps housed in the Facebook Safety Center).

    - Build team to take action on initiatives that come of global Child Safety coalition meetings (e.g., WeProtect).
- Enhance product safety messaging and provide educational resources for teens (13+ products) and children (under 13 products) and parents that are globally informed and relevant.
    - Develop consistent, expert informed safety messaging for multi-audiences (e.g., built out safety tips) across multi-platforms/products.
    - Create in-product, point-of-use safety messaging for parents and children to fully understand opt-in/opt-out communication/interactive features and functionality.
    - Relevant localization of safety messages for global product rollouts (e.g., MKids Resource Page - does not exist outside of US).
    - Platform-wide PSA to push community safety/empowerment of users.
    - In conjunction with youth create a coordinated campaign for young people by young people about how to stay safe online.
- Invest in a digital wellness lab that commits to funding third party research and civil society organisations that are tackling the issues of body image, social comparison, and eating disorders. Some initial topics of immediate priority:
    - **Longitudinal effects of AR filters:** We continue to build AR products into Instagram, which continues to be the platform for young people. As scrutiny over body image and social comparison grows, there is increasing attention to how young people use filters on Instagram. When we announced that we were reversing our ban on cosmetic filters on Instagram, we justified it by flagging that this is still an evolving space with limited longitudinal research available on the impact of face altering effects. Proactively funding and speaking about our commitment to understanding this space would put us on the front foot.
    - **Eating disorders (ED) and body image:** Everything that we've done around ED so far has revolved around programmatic outreach like Guides and campaigns with eating disorder associations, but if we want to get ahead of the issue and tackle ED effectively, we need research into how we should be better defining ED, detecting ED, and intervening. Most of our work so far has revolved around programmatic outreach like Wellness Guides and campaigns with eating disorder associations, but if we want to demonstrate our commitment to getting ahead of the issue, a public commitment to research and partnerships in this space would put us on the front foot.
    - **Social comparison and wellbeing**

What we need

- IG Private by default
- A big investment in body image, social comparison, and eating disorder issues
- **A new European approach to regulatory pressure**. Given the breadth of European regulation, privacy and safety, we need a much more coherent plan in our dealings with the European Commission and Member States. The increasing interplay between these regulations is establishing some worrying precedences beyond Europe. This is further complicated by Five Eyes and other international organisations, with their own agendas, timelines and priorities.
- Currently we bring a range of internal stakeholders to these meetings with regulators (privacy, safety, apps, country-level). I think we need, depending on the policy response or aims, to have single representatives around which teams can coalesce.
- **A new European approach to regulatory pressure**. Given the breadth of European regulation, privacy and safety, we need a much more coherent plan in our dealings with the European Commission and Member States. The increasing interplay between these regulations is establishing some worrying precedences beyond Europe. This is further complicated by Five Eyes and other international organisations, with their own agendas, timelines and priorities.
- Currently we bring a range of internal stakeholders to these meetings with regulators (privacy, safety, apps, country-level). I think we need, depending on the policy response or aims, to have single representatives around which teams can coalesce.
- As ePD has demonstrated too, we do not seem to have a coherent approach to European Union Directives and therefore, face trying to respond too late in the cycle to shape the final regulation. I would imagine the same might apply at Member State level.
- **We need to decide on what our plan is for encryption**. We need to decide if we are fully committed to Mark's privacy-focused vision. If so, it strikes me that we should just get on and implement the encryption of Messenger at the earliest opportunity. After more than eighteen months, I am not aware we have any technical solutions that will provide sufficient mitigation in terms of the drop in report to NCMEC.
- If there is a technical solution, like device-side scanning, then we should proceed with this and make it clear it's our intention. I have not heard however, from the expert community, that device-side scanning is a viable technical option. The EUIF may conclude this but now that we have been invited to contribute to the technical paper (albeit in writing), we need to decide what our stance should be. Otherwise they will decide for us and bake it into regulation.
- And a final remark. **We need a PhotDNA moment**. Facebook must have the capability to deliver a technology step forward, which everyone acknowledges is a game changer. It might be device-side scanning, or behavioural classifiers but

something that is beyond proactive CEI scanning. For better or worse, our success around scanning is not being processed that way. Time for a rethink.

- **Need for more programme funding.** In Europe, compared to many of our technology competitors, our safety programmes are underfunded. Google and Microsoft have significant, often multi-year, multi-country safety programmes with very specific policy goals and KPIs. TikTok too, is beginning to do the same, although their motives and lack of coherence are questionable.
- **Specific funding for safety mitigation strategy.** Other than Report Remove pilot, within Europe, we have no significant investments in our e2ee safety encryption safety mitigation. Already, Report Remove has been a valuable way to demonstrate our investment around reporting. We need the same around prevention and victim support. There were also no funded projects from the 2019 UK Hackathon, which law enforcement and other stakeholders remind us of.
- It may be that the Tech Coalition Project Protect, is the appropriate place for some initiatives but cross-industry bodies often suffer from inertia. I also think we need clearly definable Facebook investments/initiatives that we can point to as a demonstration of our leadership. Yes, we may be the catalyst for a broader cross-industry initiative but we should take the credit.
- Increased investment in improving DIY tooling
- More PR support - we have to find a way to counter the pressure and narrative to push back and really showcase our resources and products that can help with child safety.
- AMET: Need for more funding to support online safety awareness programmes in the region. We currently do not have any safety programmes budget support for this. Our competitors have made huge investments and are currently leading the efforts in developing localised programmes to address the digital literacy issues in the region.

July 2020

- Contact Support

Conversation History

Liz [redacted] opened your conversation with Jennifer [redacted] Amber [redacted], Sarah [redacted], Ravi Sinha, Karuna [redacted], Antigone Davis, Holly [redacted], Deepali [redacted] Sylvia [redacted] Vaishnavi J, MC [redacted], Sebastian [redacted] and David [redacted] *Sep 24, 2020 at 8:50 pm*

Sarah [redacted] edited *Sep 24, 2020 at 7:49 pm*

- Educating users about their ability to report is likely to be a low return on investment without extending the 1 hour retention period of VM… of the first reported message sent. For CEI, it's even worse — only 46% of CEI reactive reports were sent within the hour. However, the data also shows that after the first hour, the number of reports drop off dramatically (2-5% made within each subsequent hour). We should further explore upselling reporting so and provide sufficient CO review capacity to work the queues in a timely manner.

Amber [redacted] *Sep 24, 2020 at 6:33 am*

Ha sorry it was so late... meant media classification board

Karuna [redacted] *Sep 24, 2020 at 6:29 am*

Amber Hawkes what does this mean?

Karuna Nain edited *Sep 24, 2020 at 9:07 am*

Child Safety State of Play *Interim update - September 2020*

Interim update - September 2020
Where are we with the various policy discussions relating to children on our platforms, and where we think we are most vulnerable and need to step up our efforts.

Child Exploitation

Policy Landscape

Our encryption announcement has created tremendous anxiety among CSAM safety stakeholders, and generated significant legislative activity across the globe centered around exceptional access and other technical requirements and/or intermediary content liability, as well as calls to break us up. Multi-national and multi-sector bodies, including the Five Eyes, the EUIF, and the WePROTECT Global Alliance, are driving a coordinated campaign against our message encryption plans, which has caught the attention of our advertising clients as well as large foundations such as the Oak Foundation and Skoll, and driven shareholder inquiries. A spike in CSAM reporting and increased dependence on online technologies during the pandemic has made this issue COVID 19 resilient. There is some risk we become isolated from the rest of industry, which could result in FB specific solutions (e.g., Apple has tried to differentiate FB e2e messaging based on social graph). Priority Countries: US, UK, Australia, India, EU and Brazil (other potential hot spots due to high incidence rates and/or active child exploitation legislation Philippines, S. Korea, Taiwan, Japan, Argentina, Chile and South Africa).



Immediate Product Vulnerabilities

- Rooms: livestreaming abuse
- Ephemerality: negative impact on ability to report
- Interop: discoverability and reachability
- NCMEC reporting: missing context
- WhatsApp groups: link sharing on social platforms + apps on iOS and Android store enable CSAM discovery and sharing
- COVID review constraints
- Unconnected adults being able to find and message minors in Instagram Direct
- Implicit sexualisation of minors across multiple posts and comments on Instagram
- Sex trafficking and sexual solicitation networks on Instagram

Areas of Improvement

While some governments will not be satisfied with anything short of a backdoor, offering limited client-side scanning will satisfy leading safety stakeholders (NCMEC, IWF, NSPCC, Thorn,...) and many governments. It will, however, alienate some privacy stakeholders. Other areas where we can step up our efforts to help manage our vulnerabilities and push back against efforts to require encryption backdoors are listed below.

- **NCMEC Reports:**
  - contextualize these numbers
  - meaningfully drive down the number of reports before encryption through preventative work including [redacted] safety notices, friending limitations,...
  - demonstrate through the upselling of reporting and packaging of metadata and info from public spaces, etc. that we can still aid/enable more investigations than LE currently does
- **Minor Sexualization:** much stricter policies and enforcement w/r/t the sexualization of minors on our services (e.g., group names, comments,...)
- **Trafficking:** sex trafficking classifier with a focus on minor sex trafficking
- **TC Project Protect:** fast track some key wins like universal video hash sharing, research on client-side solutions, independent accountability system that tracks to CSEA vol principles
- **Livestream:** develop content independent CEI detection tech, wiretap capability, reportability and reviewability upon report
- **Interop:** strong discoverability and reachability controls, including further restrictions to PYMK and GYSJ
- **Ephemerality:** adult<>minor carveout for disappearing mode and/or some way to enable content reporting and review
- **IG:** Extend existing MSGR Safety tips for scams, impersonation, child safety/IIC in IGD; Expand policies on borderline content re minor sexualization and build classifiers to detect and take action on problematic aggregated content on surfaces like Explore, hashtag pages, Reels, better age-related logic for preventing conversations between unconnected adults and minors
- **Groups:**
  - remove private invite links and group chat functionality for high risk FB groups
  - robust system for removing apps that categorize and make searchable group invite links to WA groups in violation of our terms of service
  - robust system notifying platforms when users share group invite links for alleged WA CSAM groups
- **User Education:** robust ongoing in product user education on safety and privacy tools
- **RTC and Remote Presence products:** in-call reporting with evidence capture, feature gate high-risk users, user education/self-remediation interventions (i.e. safety notices, upsell reporting/blocking)

We also should seriously consider establishing ourselves as the industry's world leader on child safety through a large scale investment @$20-40M) in the development of robust external ecosystem to thwart CSAM and support victims ("Big Safety Investment"), would include initiatives like global child help textline, global case management tool that aggregates all case data for investigations and helps with prioritization, and scaleable automated system for x-industry threat information sharing

Recommendation

Internal readiness to rise to the challenge. Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged, suggestions on how we might kick start conversations on this work.

Ahead of Interop, work is in progress to block unconnected adults on FB and IG from initiating messages with minors on FB, however work is not planned to prevent FB and IG adults from initiating messages to IG minors. Limited country tests and data analyses have shown that [redacted] can demonstrably reduce CEI sharing but product leadership has been unwilling to prioritize these mitigations. We are currently escalating the [redacted] conversation up through Monika to meet with MLT. We are also getting ready to meet with Jay [redacted] to discuss how best to present [redacted] to MLT. We need leadership support to have product prioritize this work because it will allow us to make a much stronger case with our external critics and stakeholders that we are doing good work on this front. We have progress on blocking predicted groomers from using Disappearing Mode but need CO support to apply the same protection in IG. Work has been completed on basic safety mitigations for Rooms but work still needs to be done to enable in-call reporting with evidence capture in both Rooms and RTC (aka Video Chat). RTC, which is the predecessor to Rooms, is not yet in parity with Rooms (i.e. feature gating high-risk users, user education/self-remediation interventions) and will require stronger leadership buy-in to accomplish even bare minimum protections this half. We continue to advocate for strong discoverability and reachability controls in anticipation of Interop, including further restrictions to PYMK and GYSJ. Work to develop a CEI video classifier for FB Live is currently ongoing but we need more resources for this in addition to engineering and review resources to develop an operational sex trafficking classifier. Please see the chart below for detailed summary of the specific related work and its status.

Age Assurance

Policy Landscape

With increasing frequency safety stakeholders are demanding that we do our utmost to keep U13 users off our platform and that we appropriately account for U18 users in the design of our services. Gaps in this area help fuel online harms legislation and regulation of various kinds, particularly outside the U.S. With a few key product changes and/or regulatory collaboration we can prevent ill-considered legislation (e.g., requiring id age verification at sign up, 18+ requirement for social media,...) and secure the support and confidence of child safety stakeholders, including policymakers parents/caregivers. Priority Countries: countries of jurisdiction (US and Ireland), and countries that are highly vocal and influential in region AU, UK, Brazil, South Africa and India.



Immediate Product Vulnerabilities

- Lack of U13 proactive age identification and enforcement mechanism
- Need for improved age identification models that allow us to better predict age of users and offer appropriate age-related filtering and protections (esp. w/r/t IG where still lack age data for most users, which prevents us from implementing a number of age-related protections e.g. restricting unconnected minors and adults from interacting, full rollout of Safety Warnings (currently only planned for those users whose age we know)

Areas of Improvement

The development of a device-side system for age verification to block U13s from our services will go a long way to addressing the concerns of child safety stakeholders, but we need to fast track that effort. We also need to improve our age identifying classifiers and expand "age appropriate" protections and experiences (e.g., coming soon IG account will be private by default for minors).

- Looking back: work with a trusted expert/group of experts to review our default settings, protections,... for minors, identify improvements (e.g. GYSJ and PYMK restrictions), and implement them
- Looking forward:
    - develop a "safety by design" process or product safety review similar to privacy review
    - in product U18 specific user education on safety and privacy tools (e.g., unique user education at sign up)

Recommendation

Internal readiness to rise to the challenge

Internal readiness to rise to the challenge. Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged, suggestions on how we might kick start conversations on this work.

In March it was decided that we would not build a classifier to detect minors under 13 years old. Instead, we would push industry partners for device/OS level age verification and management solutions. Since then, Central Privacy has stood up a small team on youth privacy but a PM is yet to be hired. They have offered to prioritize a few low business cost/high policy impact projects for H2. Identity Integrity has also identified a PM who would own age verification but they will not prioritize this work until we have clear legal obligations to do so. We need leadership support to push product to act on knowledge we already have, i.e. clear our backlogs, disable

verified u13s, build an u18 checkpoint to review minors who have been reported on 18+ products like FB Dating. Against this backdrop, our enforcement is not consistent across all apps for a single known person. In addition, age models perform inconsistently across different markets. We recommend that Policy leadership stack rank and prioritize our most sensitive policy driven markets and make clear to product that we should operationalize those markets first. We also need leadership to support the basic principle that if we're using a signal to predict age for business purposes, it should be used to enforce on age. For example, we need to use age prediction to gate minors from alcohol ads and other high-risk features where we currently rely solely on stated age. Age models also do not work as well on IG because IG only started to collect age data in December 2019 and we currently only have stated age for ~67% of IG users. We need leadership support to convince IG leadership to collect age for the remaining users for whom we have no stated age. We also would like leadership sign off to conduct an audit of special protections for minors and then bring in a trusted expert/group of experts to review our default settings and protections for minors, identify improvements (e.g. GYSJ and PYMK restrictions), and implement them.

SSI

Policy Landscape

While SSI experts support our existing policies and our commitment to ongoing review of our policies, the need to explain the ever increasing rate of death by suicide and concerns about the impact of social media on youth well-being more broadly keep us under heightened scrutiny in this area — most notably in the UK. Specifically, this translates into concern among some policymakers and stakeholders that almost any exposure to this content for minors "normalizes" it and puts them at risk, and these concerns which help fuel online harms legislation to ensure we behave responsibly. Additionally, some policymakers and health policy experts are raising privacy related questions w/r/t our work in this area, creating a tension between those who push for increased scanning and those who want us to take a more privacy sensitive approach. This tension begs for a forward leaning approach involving both sets of stakeholder to come to agreed upon norms.



Immediate Product Vulnerabilities

- only one classifier running in the EU → escalations, heightened risk on IG specifically because of how #tags work
- lack self-harm support resources (e.g. eating disorders, cutting); currently, we send suicide prevention resources and tips to people engaging in any kind of self-harm
- lack different resources and specific actions for users who are using our platform to promote versus admit suicide
- lack population/risk specific resources on the SSI checkpoint
- IG:
    - lack classifiers to detect and take action on problematic aggregated content on surfaces like Explore, hashtag pages, Reels
    - spaces like IGD being used to encourage or promote suicide e.g., group suicides in Norway that occurred last year and we don't scan there

Areas of Improvement

Resolving the conflict with the IDPC so we can use our existing SSI classifiers in the EU will go a long way to addressing the policy heat we face in the UK in particular. That said, we should continue to stress test our SSI policies and implementation standards including on borderline content and SSI promotion vs. admission, and build a more context sensitive response system with differentiated actions and resources. We also have an opportunity to lead here by 1) working with privacy and safety stakeholders together to develop best

practices with regard to AI, responses, data retention,... and 2) helping build a global crisis text line system

Recommendation

Internal readiness to rise to the challenge

Internal readiness to rise to the challenge. Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged, suggestions on how we might kick start conversations on this work.

On SSI, we are making progress on policy development re boderline content and strengthening our engagement x-industry and with experts. Our key limitation here remains the inability to use our existing SSI classifiers in the EU. We also don't have work underway to build classifiers on IG to detect and take action on problematic aggregated content on surfaces like Explore, hashtag pages, Reels. In addition we need to continue to refine our ability to provide people with more local/ culturally relevant resources depending on the crisis they are experiencing (self harm vs eating disorders vs suicidal ideation...).

Well-Being

Policy Landscape

The physical distancing requirements of the COVID-19 pandemic have generated greater awareness of and concern for issues related to mental health and well-being (e.g., anxiety and loneliness), as well as an appreciation among policy makers and the press for the role technology can play in supporting mental health and well-being (e.g., keeping people connected). This shift of focus from how tech may create, increase or exacerbate mental health and well-being issues to how tech can assist in combatting some of them is an opportunity for us, but a failure to execute well could surface an even worse "tech backlash" around mental health and well-being issues negatively associated with technology (e.g., passive time spent, pressure to be perfect, FOMO, etc.) and fuel some ongoing legislative/regulatory efforts.

[redacted]

Immediate Product Vulnerabilities

- FB+IG: Lack of self harm support resources (e.g. eating disorders, cutting); currently, we send suicide prevention resources and tips to people engaging in any kind of self-harm,
- Product infrastructure limitations: inhibit our ability to make necessary guarantees around protection of mental health related data
- FB - Bullying and harassment: Lack of mass comment moderation tools

Areas of Improvement

There is a need for stronger research around social comparison and loneliness on our platforms. We have an opportunity to launch x-industry Digital Wellness Lab with Center for Media and Child Health at Harvard to seed independent research in this space. We also have an opportunity here to demonstrate our commitment to building a supportive community through efforts such as building group admin support learning units, expanding WA business API for crisis support orgs, CSOM and the connected generations initiative.

- **IG**
    - **Policies:** we need to stress test our policies w/r/t bullying of U18 public figures, given IG public figures are younger and harassment targeting them is gaining more scrutiny
    - **Research**: conduct new research around social comparison and loneliness on Instagram to tackle some of the root causes of SSI, bullying, harassment; more foundational mental health research being done in non US/UK markets
- **Facebook:**
    - **Product:** Proactive bullying AI comment filters on FB to get parity w/ IG
    - **Policies:** Refining public figure definitions and developing more protections from harassment for public figures
- **Product:** Need to ensure global mental health efforts are culturally sensitive and align with existing approaches towards mental health; resourcing and prioritizing Purpose Policy Framework for protection of health information
- **Resources:** Expanding mental health resource pages that are easily accessible to users and regionally and age applicable

Recommendation

Internal readiness to rise to the challenge

Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged, suggestions on how we might kick start conversations on this work

We're making progress connecting people with resources on the COVID Information Center and with a new mental health resources landing page. We need more product work on issues such as social comparison, bullying of minor public figures. We are expanding CSOM but need more resources on product given the need to demonstrate value of encrypted surfaces for connecting people with safe, private and secure support.

Racial Justice

Also suggest internal audit to ensure that all safety classifiers are globally resilient and operate without bias.

Some detailed product asks:

- Lock in safety mitigations for Interop, including:
    - Strong discoverability and reachability controls, i.e. default minors to the most private and safe settings;
    - Extend existing MSGR messaging and Friending restrictions for unconnected adult<>minor in IGD and ensure cross-network messaging will not degrade current protections in place for minors;
    - Adult<>minor carveout for Disappearing Mode and prevent use of feature by predicted groomers;
    - User reporting education and message download for Unsent and Disappearing Mode messages with tombstones:
        - Ensure all content is reportable and reviewable by CO;
            - Increase retention period for unsent and ephemeral messages;
            - Add thread-level reporting in IG;
            - Call out in MSGR that unsent messages are reviewable;
        - Educate users about Impermanence at key product experience points
            - Tell sender+receiver that unsent messages can be reported for X first uses
        - Make Reports available to Reporters and give them options to save a copy or access reported content, including tombstones within context
    - Robust user controls for x-network reachability/discoverability.
    - Ensure existing safety controls such as Restrict and Ignore are not degraded by cross network messaging and maintain user expectations.
- Strong integrity measures to reduce harmful content sharing and interactions:
    - Remove private invite links and group chat tab in harmful content associated FB groups
    - [redacted]
    - Add "Who can…
- Sex Trafficking classifier with a focus on minor sex trafficking
- Self-remediation tools and backend measures that promote safety while respecting users' privacy for RTC and Remote Presence products, such as:
    - In-call reporting with evidence capture;
    - Screen-sharing static notification either in roster or as part of UI control panel;
    - Feature gate high-risk users;
    - User education/self-remediation interventions (i.e. safety notices, upsell reporting/blocking);
    - Understand abuse prevalence / useful proxies in IG RTC + MSGR RTC.
- Support our Global Classifier Resiliency and Fairness initiative for safety classifiers that are critical to our issue areas.
- Age Assurance / Verification
    - Proactive age verification and enforcement / u13 classifier; u18 checkpoint
    - Expanding age appropriate experiences, i.e. identify and build special protections (default settings, content restrictions, blocked products) that should apply to u13 and u18 users.
- Suicide and Self Harm
    - Self harm support resources (e.g. eating disorders, cutting; currently, we send suicide prevention resources and tips to people engaging in any kind of self-harm)
    - Providing varying resources and actions to people using our platform to promote versus admit suicidality; policy development and classifiers needed to action differently and appropriately.
    - IDPC classifier approval for Europe

- SSI checkpoints on FB and IG need to be more relevant and population/risk specific
- Crisis Service Over Messenger engineering support

X-Industry Sharing:

- Video DNA x-industry banks

User Education and Marketing:

- In-product user education campaign to drive awareness of privacy and safety tools, settings,..
- Design and Eng resources for the Safety Center
- Better product walkthrough for new users, teens (13-15) to flag how to report content, connect to the safety centre and help centre

Additional Quips/Docs

- 1H2020 Global Safety, Health and Well-Being Policy Plan
- H12020 Plans (Review)
- Safety Policy Product Asks
- H2 Plans
- Safety Legislative / Regulatory Tracker

Rough

- Our European data protection regulator (the IDPC) has opened two new inquiries concerning Instagram's processing of children's data
- **French bill to regulate the commercial exploitation of the image of children under 16 on online platforms:** The bill has a few safety-specific provisions. Namely, it requires that platforms inform and raise awareness of minors about their rights and the consequences of their online exposure, encourage users to report content involving minors that would undermine their dignity or their moral or physical integrity, and work with child protection associations.

[redacted]

- **Germany youth protection legislation:** Three new pieces of youth protection legislation are on the horizon, with different areas of focus - some are focussed on regulating content that is harmful to minors, while others focus on structural youth protection e.g. settings around youth defaults.
- **Response to Unilever and 5Rights around how we handle age-appropriate design:** We prepared a written response to a number of concerns raised by the 5Rights Foundation. The IG-specific concerns revolved around: (a) public profiles on Instagram increasing risk to minors, (b) private messaging on Instagram being open to anyone, even unconnected people, and (c) a perceived lack of age verification on FB and IG.

Recent media coverage
Note: Updates have been ported over to Child Safety State of Play (9/24). Please highlight any additions and tag Karuna Nain in comments.

July document Child Safety - State of Play (7/20)

Antigone Davis opened your conversation with Jennifer [redacted], Amber [redacted], Sarah [redacted], Ravi Sinha, Karuna [redacted], Holly [redacted], Deepali [redacted], Liz [redacted], Sylvia [redacted], Vaishnavi J, MC [redacted], Sebastian [redacted] and David [redacted] *Sep 24, 2020 at 3:33 am*

Jennifer [redacted] edited *Sep 24, 2020 at 2:02 am*

In the United States, we continue to see legislation focused on child sexual exploitation. The Department of Justice, in close coordination with the White House, is advocating for Section 230 reforms and also uses public opportunities to raise concerns around child safety and encryption. There are very negative sentiments about Facebook ahead of the election from both… around anti-conservative "censorship" through content moderation and have been addressing this through legislation, letters, public statements, and coordinated efforts across the government. These proposals would impact our ability to moderate safety-related content and misinformation. Child safety remains an area of bi-partisan agreement and we have seen continued interest in legislating on child sexual abuse material and Section 230 reforms throughout the summer. We also hear regularly through letters and individual Congressional office inquiries about child safety.

Sarah [redacted] edited *Sep 24, 2020 at 1:22 am*

- **Inability to auto-action using our proactive CEI classifiers:** Our CEI video and photo proactive classifiers do not perform well enough to auto-action content because we don't meet required hard action (delete) or soft action (downrank/quarantine/friction) thresholds. This has implications for company priority products and the future of their child safety measures:
    - Rooms: Product continues to push for growth and plans to allow Rooms creation directly from WA and IG without FB log-in…
        - message is seen; (3) educating users about their ability to report vanish mode messages; (4) VM is only available in existing threads.
            - Educating users about their ability to report is likely to be a low return on investment without extending the 1 hour retention period of VM messages. Recent data analysis shows that only 57% of grooming reports occurred within the hour of the first reported message sent. For CEI, it's even worse — only 46% of CEI reactive reports were sent within the hour.

- **Close enforcement gaps revealed by NCMEC reports data analysis:**
    - Child Safety data science work is ongoing to contextualize NCMEC report numbers. One… and/or under-enforcing against recipients. Interestingly, Asian and MENA dominate both uploaders and receivers and only 3% are from the U.S.
    - Based on this ongoing work, **we should refine the way we currently gate or feature block NCMEC reported users from products and product experiences.** Currently, only Messenger Kids and Rooms are using a filter to block NCMEC reported users who have shared viral or severe meme CEI content but have not yet been disabled from our platform. This is a stop-gap measure to protect our most vulnerable users and risky products but in the near future, we need to a) revise our enforcement policy against CEI recipients and include malicious CEI recipients in these product gating filters as well as b) refine our ability to target malicious NCMEC reported uploaders to make consistent product gating decisions across FB and IG.
- **COVID review constraints:**
- **Nudity & Pornography (N&P) review reductions:** Due to the lack of graphic content review capacity across Community Operations, there is a plan to reduce N&P's review capacity by ~80%. Child Safety teams rely heavily on N&P queues to catch novel CEI content. Short term, this will reduce volumes of escalations, reduce novel CEI banking and increase banking latency. Longer term, the lifting of graphic content WFH restrictions will help to address this.
- **Need more investment in proactive detection of Minor Sex Trafficking:** Need a proactive classifier that allows us target minor sex trafficking and distinguish it from adult prostitution activity. Also need…
- bias. There is ongoing Child Safety engineer work to update our text classifiers to more easily internationalize and integrate new languages for proactive grooming detection but we should do the same global work even for our language-agnostic, image-based classifiers.
- **Improve our text classifiers that detect inappropriate interactions with children:** Child Safety CI is working to develop a text model across both FB and IG that will capture conversations with minors where

someone solicits images from them - proactively . Proactively detecting this type of conversations within posts and comments would be incredibly helpful. In addition, there is ongoing work to improve our existing text classifiers that detect grooming to achieve two goals: 1) help us land automation of hard actions (disable/report) against groomers who are making plans to meet minors in real life for sex and 2) eliminate the need for new engineering when adding new markets because these new text models will perform well even in languages for which we have no training data.

- **Improve ability for users to report child safety issues anywhere on our platform.** Child Safety CI has continued to charge forward on their goal to ensure that anywhere someone could see or experience child sexual abuse on our platform, they could tell us about it. As part of that effort, we recently shipped a user report feedback tag, "Involves a Child," on IG Profiles. This was met by significant praise from stakeholders who had petitioned for this improvement. However, we need to continue closing existing gaps in reporting coverage and launch the "Involves a Child" feedback tag where it is missing across the platform, for example, in FB Groups and IG comments.

- **IG**
  - **Reduce prevalence of grooming on IG:**
    - The daily number of minor victims having sex talk conversations in IGDM with adults are approximately 38x (though we should note that there are 7x…
      - MSGR adults will be prevented from messaging MSGR minors, but the same will not be true of MSGR adults who message IG minors.
  - **The best performing age model (the "Adult Classifier") is not yet globally operational, which leads to a patchwork of integrity measures across the company…** that provide protections for minors on our platform. We need more review resources to establish ground truth data on age in priority markets.
  - **Age verification pressures:** We will likely continue to face external pressure to verify age. YouTube announced this week that they will ask European users to provide proof…
  -
  - **Age models do not work as well on IG** because IG only started to collect age data for new sign ups in December 2019. In addition Thus, age we currently only have stated age for ~67% of IG users. Age collection on IG has previously been criticised criticized as being for the purpose of targeting ads, and intrusive to users who do not necessarily need to bring their authenticated selves to the… on stated age on similar apps in the industry also have also resulted in age lying and misrepresentations, and fines for collecting children's personal information. Thus, we currently only have stated age for ~67% of IG users. We recommend engaging with IG leadership on building better age affinity models to predict age of our users, and prioritise age model maturity…
  - of people going from 13-18 > 18 and people going from >18 to 13-18). This backlog is currently not resourced.

Karuna [redacted] *Sep 23, 2020 at 11:04 pm*

Do you mean the Download Your Information (DYI) feature?

Karuna [redacted] *Sep 23, 2020 at 11:04 pm*

David Miles do you have specific projects that you think we should propose for funding? Can we list them instead please and put a $ amount next to each

---

Karuna [redacted] edited *Sep 23, 2020 at 11:03 pm*

- CDA 230
  - EARN IT
  - UK Online Harms
- QAnon
- COVID-19

European Context Europe

- FB EMEA leadership sees a concerted effort to go after tech in Europe and Facebook's business model specifically.
    - Google and Microsoft all considered to have too much power in the European and Global marketplace.
    - Surveillance narrative (as expounded by Zuboff) broadly accepted in Europe and US tech has not been able to mount an effective counter narrative.
- Safety, due to the consensus around the importance of protecting children in the digital space, being used successfully in Europe, by the EU and Member States, through regulation, in part, to reassert control of Europe's digital destiny.
- Our decision to encrypt Messenger is seen as symptomatic of a pivot to privacy, at the cost of children's safety. That US notion of data privacy, is not the same in Europe and therefore exacerbates the situation.
    - The European Union Internet Forum for example, consciously excluded Facebook from its draft technical paper on encryption.
- A number of separate stakeholder groups (predominantly privacy, safety and security but including some of our competitors) are mounting well-coordinated attacks, leveraging the regulatory momentum both within countries (member states) and across the European Union.
    - Our competitors (Microsoft in particular and Google to a lesser degree) are accentuating that approach and along with so-called experts, seek a EU legislative response during Q2 2021, that could be detrimental to our plans and in Microsoft's case, have little impact.
- Regulations:
    - Member States
        - Belgium - NCII Law (July 2020)
        - Germany - Federal (December) and Interstate Youth Protection Law (September)/NetzDG (November - 2.0/3.0)
        - Ireland - Online Safety and Media Bill (January 2021)
        - UK - Online Harms Bill (October 2020)/Age Appropriate Design Code (September 2020)
        - Danish government announced plans to regulate content on social media but no concrete plans yet.
        - In the UK, the Women and Equalities Select Committee Changing the Perfect Picture asked us to prepare a submission for their inquiry into body image (https://urldefense.proofpoint.com/v2...=) to study impact of advertising and social media on body image

We are not alone in this. Facebook, Google and Microsoft are all considered to have too much power in the European and Global marketplace. The surveillance narrative (as expounded by Zuboff) is broadly accepted in Europe and US tech has not been able to mount an effective counter narrative. We have come to symbolise all that is wrong in terms of the European digital marketplace.

Safety, because of the consensus around the importance of protecting children in the digital space, is now being used successfully in Europe, by the European Union and Member States, through regulation, in part, to reassert control of Europe's digital destiny.

Our decision to encrypt Messenger is seen as symptomatic of a pivot to privacy, at the cost of children's safety. That US notion of data privacy, is not the same in Europe and therefore exacerbates the situation.

The European Union Internet Forum for example, consciously excluded Facebook from its draft technical paper on encryption. It suggests to me that Facebook is perceived as the problem. Our competitors (Microsoft in particular and Google to a lesser degree) are accentuating that approach and along with so-called experts, seek a EU legislative response during Q2 2021, that could be detrimental to our plans and in Microsoft's case, have little impact.

[redacted]

What we need

- Representation of the region in the Safety Advisory Board and Youth Advisory Board. We do not have any

representatives from the AMET region in our safety advisory board which has led to underrepresentation of the region safety issues at the SAB level.

- AMET: Need for more funding to support online safety awareness programmes in the region. We currently do not have any safety programmes budget support for this. Our competitors have made huge investments and are currently leading the efforts in developing localised programmes to address the digital literacy issues in the region.

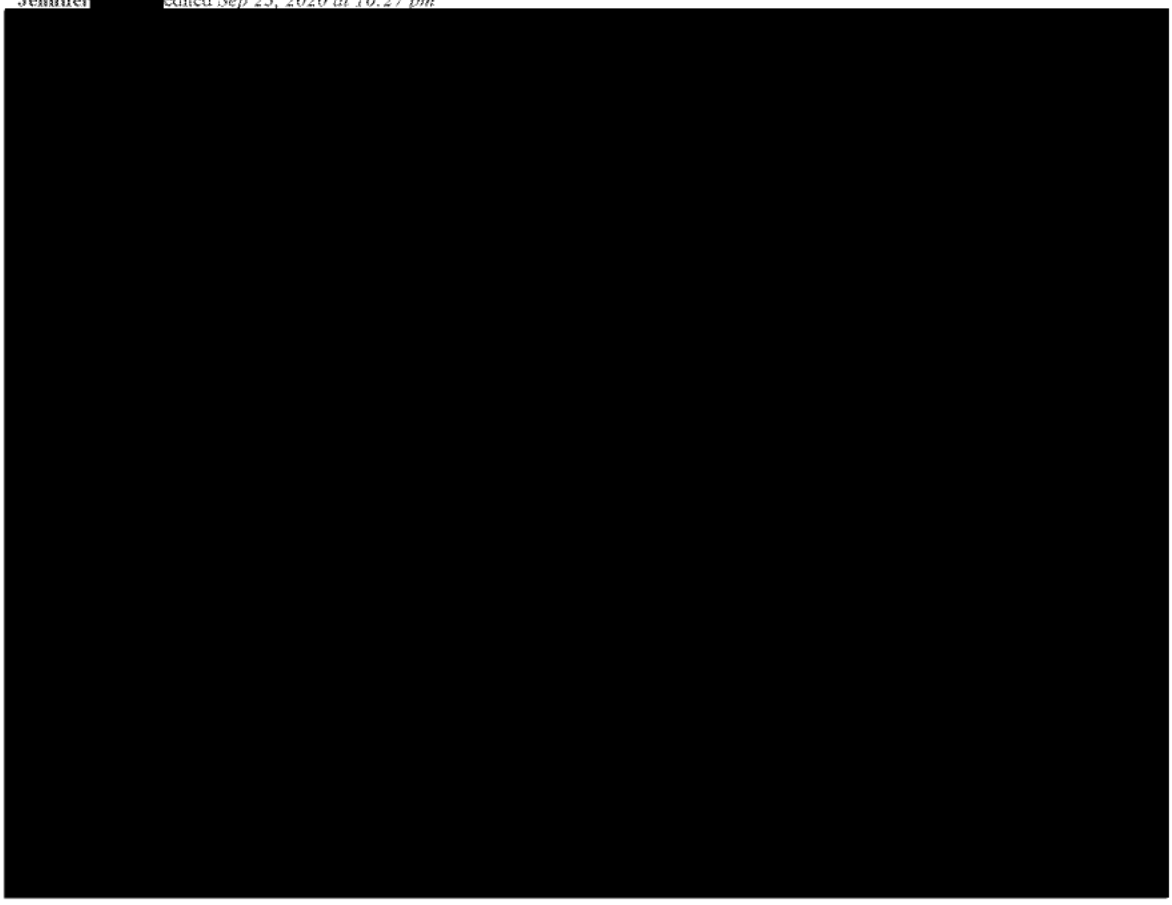
Jennifer *Sep 23, 2020 at 10:18 pm*

Agree on this and really feel that we need to be able to deliver on some of the ideas from concept hacks or something big to show that we are still minimizing CSAM when we move to encryption. We've spent a lot of time saying we will have safety mitigations but need to demonstrate more.

Jennifer *Sep 23, 2020 at 10:14 pm*

I only included bills that are new since the last update or have had some movement.

Jennifer edited *Sep 23, 2020 at 10:27 pm*



Jennifer opened your conversation with Amber , Sarah , Ravi Sinha, Karuna , Antigone Davis, Holly , Deepali , Liz , Sylvia , Vaishnavi J, MC , Sebastian and David *Sep 23, 2020 at 9:28 pm*

Jennifer *Sep 23, 2020 at 9:28 pm*

Yes and this would help with visibility and awareness issue.

Vaishnavi J *Sep 23, 2020 at 7:31 pm*

David Karuna Might be helpful to include Denmark and UK IG-related moves:

The Danish government has announced it will regulate content on social media but has not presented anything concrete yet. There is a risk it will copy parts of the German NetzDG model and we are therefore following the process closely and have engaged with the responsible ministries to influence the proposal. The recent SSI case in Denmark on IG made several MPs ask for self harm content to be included in the coming legislation

In the UK, the Women and Equalities Select Committee Changing the Perfect Picture asked us to prepare a submission for their inquiry into body image (https://urldefense.proofpoint.com/v2...) to study impact of advertising and social media on body image

---

Sarah *Sep 23, 2020 at 6:08 pm*

Karuna Defer to you on whether better to mention up here or down below with the rest of Age where it's flagged as improvement needed on IG

Sarah edited *Sep 23, 2020 at 6:08 pm*

Sarah *Sep 23, 2020 at 6:07 pm*

Karuna Vaishnavi J Already flagged above as work ongoing in first IG bullet.

Vaishnavi J *Sep 23, 2020 at 5:55 pm*

Sarah reframed this to provide more context for IG reasoning - could we also get the breakdown for what % of IG users we have stated age from, and what % we can infer through hard matched FB accounts?

Vaishnavi J edited *Sep 23, 2020 at 7:29 pm*

- **Improve IIC prevalence measurement:** There is ongoing work by Child Safety CI, Messenger Wellbeing, and IG Wellbeing to improve IIC prevalence measurement on IGD.
- **Increase reviewer capacity for suicide / self-harming content on Instagram:** When people search for suicidal or self-harming content on Instagram, we block hashtags that are likely to be violating and offer support language over hashtags that may contain sensitive content. However, accounts with these names continue to be surfaced as search results, due to the accounts not having been proactively detected, reviewed, and taken down for violating our policies. We need more reviewer capacity as well as the ability to bring our proactive detection technology to the EU, so that we can proactively scan for and take action on violating accounts / content.
- **Bring our SSI proactive detection technologies to Europe:** We are limited from running our self-harm classifiers that account for both text and media signals and determine if a post has a high likelihood of being actionable for suicide or self-harm.
- **Virality on Instagram:** Minors who go viral on Instagram will likely experience increased interactions with unknown people. As their content and by extension their account becomes surfaced to more unconnected people, they would benefit from additional protections that they can turn on, like limiting who can send them

private DMs or follow requests. We should also consider limiting their visibility to adults who score above certain thresholds on our IIC classifiers. More broadly, we need more research and product commitment to specifically addressing how virality on…

- for discovering new creators also means a large percentage of minors on the platform live their lives out in the public eye. Encouraging more use of private accounts for minors would allow them to creatively express themselves without immediately opening them up to strangers or unknown interactions. We currently have a two-week test running to explore the viability of Private by Default for all minors; we should strongly advocate from the policy side that the safety, privacy, and legal benefits of defaulting minors to private accounts should outweigh any Growth concerns. Private by Default for Teens would provide a tangible example of IG not only generally considering the interests of minors, but actually implementing distinct product features/experiences catering to their needs. This could go a long way towards helping us address regulatory pressure in Europe and avoid legislation requiring us to implement even more restrictive default settings.
- **Media perception of new Instagram products like Reels:** We are generally perceived as being 'behind' our competitors - especially TikTok - in youth safety. For example, they do not allow anyone under 16 to receive direct messages, and have parental controls that allow a parent to link their TikTok account to their teen's and set controls. As we advance our commitment in Reels and make it the next big bet for Instagram, we need more visible interventions and product commitments to avoid fuelling this external narrative.

- Age models do not work as well on IG because IG only started to collect age data for new sign ups in December 2019. In addition, Adam Mosseri will not support upselling age collection on IG has previously been criticised as being for existing the purpose of targeting ads, and intrusive to users because he sees it as intrusive data collection who do not necessarily need to bring their authenticated selves to the platform. Relying on stated age on similar apps in the industry also have also resulted in age lying and misrepresentations, and fines for collecting children's personal information. Thus, we currently only have stated age for ~67% of IG users. We recommend engaging with IG leadership on collecting building better age for the remaining ~33% affinity models to predict age of our users, and prioritizing prioritise age model maturity on IG.

- Empty IG accounts that primarily leave explicitly sexualising comments on pictures of minors - currently, upon escalation where we confirm the account-holder is regularly posting sexualizing comments on images of minors, the account will be disabled.
- Child sexual exploitation:
  - Empty IG accounts that primarily leave explicitly sexualising comments on pictures of minors - currently, upon escalation where we confirm the account-holder is regularly posting sexualizing comments on images of minors, the account will be disabled.
  - Empty IG accounts that primarily leave implicitly sexualising comments on pictures of minors
  - Empty IG accounts that only engage in behavioural actions e.g. likes, saves of images of minors
- Suicide and self-injury:
  - Better definitions for eating disorders - While eating disorders currently reflect a small % of our reports under suicide and self-injury, communities around eating disorders do not tend to report one another, and we see many accounts, hashtags, and comments that continue to promote eating disorders.
  - Some of the hypotheses that we've been thinking through is that ED content is too narrowly defined by our policies so we under-enforce on what the broader public might consider to be ED content. ED is also rarely a single piece of content, and more a collection of content and behavioural signals. We need more research in this space.

- Add and staff an "involves a child" reporting option across all reportable surfaces e.g. account, comment, DM, hashtags
  - Tests and launches of the "involves a child" reporting option at the post and profile levels have increased the precision of detecting and enforcing against child safety. Adding this reporting option across all surfaces will allow posts to be prioritised for moving out of the broader Nudity & Pornography queue and being reviewed for child safety violations.
  - Current reporting structures hinder our ability to measure the true prevalence of CEI or other child safety violations, as reports are aggregated along with other nudity and pornography violations.
  - Include more granular reporting options to capture different types of child safety violations e.g. distinguishing between a comment soliciting CEI and a comment implicitly sexualising a minor
  - Move child safety reporting options out of the "Nudity & Sexual Activity" category to appear earlier in the reporting flow, with more granular options to capture a wider range of behaviours

- Invest in a digital wellness lab that commits to funding third party research and civil society organisations that are tackling the issues of body image, social comparison, and eating disorders. Some initial topics of immediate priority:
  - **Longitudinal effects of AR filters:** We continue to build AR products into Instagram, which continues to be the platform for young people. As scrutiny over body image and social comparison grows, there is

increasing attention to how young people use filters on Instagram. When we announced that we were reversing our ban on cosmetic filters on Instagram, we justified it by flagging that this is still an evolving space with limited longitudinal research available on the impact of face altering effects. Proactively funding and speaking about our commitment to understanding this space would put us on the front foot.

- **Eating disorders (ED) and body image:** Everything that we've done around ED so far has revolved around programmatic outreach like Guides and campaigns with eating disorder associations, but if we want to get ahead of the issue and tackle ED effectively, we need research into how we should be better defining ED, detecting ED, and intervening. Most of our work so far has revolved around programmatic outreach like Wellness Guides and campaigns with eating disorder associations, but if we want to demonstrate our commitment to getting ahead of the issue, a public commitment to research and partnerships in this space would put us on the front foot.
- **Social comparison and wellbeing**

Vaishnavi J *Sep 23, 2020 at 5:44 pm*

That sounds great!

Sarah [redacted] *Sep 23, 2020 at 5:36 pm*

Vaishnavi J The current NCMEC reported signal is more of a stop-gap measure and we think there is a better way to do this, especially by including CEI recipients who have shown sufficient signal for bad intent. Instead of porting over the current Rooms/MK NCMEC filter to IG, I'll revise this to say we should refine our NCMEC reported user filter/signal and a) revise our enforcement policy against CEI recipients and b) use that revised enforcement policy to make product gating decisions across FB and IG. cc Karuna Nain

---

Sarah [redacted] edited *Sep 23, 2020 at 6:07 pm*

- Our CEI video and photo proactive classifiers do not perform well enough to auto-action content because we don't meet required hard action (delete) or soft action (downrank/quarantine/friction) thresholds:
- Before we move to e2ee, we should prioritize reducing CEI shares VPV's (views) and preventing bad actor connections: Data analysis has shown that large message threads may harbor a higher concentration of bad users and harmful content… move to Cross-App Communications, we need to invest in data science and product intervention experiments to understand what solutions best achieve these goals.
- The daily number of minor victims having sex talk conversations in IGDM with adults are approximately 38x (though we should note that there are 7x as many minors on IG as on FB, and at least 7x as many victims) more than on FB Messenger.
- We will likely continue to face external pressure to verify age. YouTube announced this week that they will ask European users to provide proof that they're above 18 if they seek to access age-restricted videos. The goal is to help YouTube comply with the EU's audiovisual media services directive, which requires platforms to protect minors against harmful material. We could take a similar approach where we use age prediction to put people into additional age verification checkpoints based on whether a piece of content or product feature is age-gated. Or treating all users as u18 for any of these age-gated features or content until our age models generate a prediction that they're over 18. The challenge is getting our age models globally operational and justifying such an approach to product growth teams.

Amber [redacted] *Sep 23, 2020 at 5:03 pm*

+1 for APAC

Vaishnavi J edited *Sep 23, 2020 at 5:23 pm*

- **Improve our text classifiers that detect inappropriate interactions with children**: Child Safety CI is working to develop a text model across both FB and IG that will capture conversations with minors where someone solicits images from them - proactively detecting this type of conversations within posts and comments would be incredibly helpful.

- **Bring the NCMEC reported signal to Instagram:** This signals to product teams users who have shared CEI but not been disabled. Not all accounts that share CEI get disabled (if viral content, accounts get checkpointed and accrue a strike). We share this signal to product teams for informing decisions around whether users should be allowed to use our features. It's currently running on Rooms and Messenger Kids, and bringing it to Instagram would provide an additional signal for our recommendation surfaces like Explore, and new products like Reels.
- **Improved age detection models:** These will significantly help our actor classifier on Instagram. As we roll out improved age affinity models, this will be valuable classification data for our metric to perform better.
- **Disrupt the ability for predators to discover minors on our surfaces:** Unlike Facebook's People You May Know model, which prevents unconnected adults from discovering or initiating conversations with minors, Instagram has no similar logic for Accounts You May Follow. We're in the process of deep-diving into true positive prevalence labels to determine the most common surfaces where violators find their victims, so we can measure the impact of our efforts to reduce minor discoverability on selected surfaces (e.g. search, suggested accounts).
- **Understand actors who engage in violating behaviour on Instagram:** IGWB Research is developing archetypes of child safety violators for Instagram partnering with members of the Child Safety XFN to incorporate the Intent Taxonomy X Persona framework that has already been developed.
- **Improve IIC prevalence measurement:** There is ongoing work by Child Safety CI, Messenger Wellbeing, and IG Wellbeing to improve IIC prevalence measurement on IGD.
- Child Safety landscape - Instagram for more details

These behaviours are either only enforced against upon escalation or not enforced against at all right now, and would be helpful gaps in our Implementation Standards to close:

- Empty IG accounts that primarily leave explicitly sexualising comments on pictures of minors - currently, upon escalation where we confirm the account-holder is regularly posting sexualizing comments on images of minors, the account will be disabled.
- Empty IG accounts that primarily leave implicitly sexualising comments on pictures of minors
- Empty IG accounts that only engage in behavioural actions e.g. likes, saves of images of minors

---

Amber [redacted] edited *Sep 23, 2020 at 5:03 pm*

- This sits alongside a proposed **Expansion/conflation of illegal and harmful content regulation.** Proposed expansion of eSafety's powers (AUS) to enforce on expanded harms (e.g bullying & harassment) for not only children but adults, (mirroring the UK and Germany to… strictly illegal content is also apparent in proposed regulation in the Philippines which now has in its sights content that is developmentally harmful to children. Other countries considering harmful content regulation (ex. Indonesia) considering misinformation and child safety in the same category.

[redacted]

- eSafety's**Safety by Design**. eSafety's **SbD** approach continues to gain traction with an increasingly global audience (collaboration with VCs and Alex Stamos in the US, multi-lateral/INGOs such as UNICEF and… APAC/South Asian countries). Phase 2 moves beyond principles to best practice recommendations on internal processes, integrity measures and settings to protect minors. While these are still voluntary, the proposed AU Online Safety Act already points to SBD as the basis for minimum requirements/expectations of intermediaries, so could form the basis of legal obligations in future and will continue to have visibility.

- Age-verification is also going to become a dominant feature across APAC and this has been accelerated by CoVID where we are seeing an
- **Safety/Privacy tension increasingly apparent**: secrecy of communications provisions in Japan and Korea are calling into question the legality of our scanning for harmful content while other regulatory proposals in both countries require either collecting and sharing of more user information with local authorities/civil complainants or increasing efforts in

proactive detection of violating illegal content. (mirrors tension in EU with ePrivacy directive?)

- **Push for age-verification increasing**
- number of CEI/grooming escalations and where in countries lik
- . In South Korea and Japan
- the push is generally to
- current proposals requiring we verify age through mobile phone number verification.
- In
- Philippines Dep of Ed issuing advisories for u13s to use Messenger for home-based learning leading to clear exposure to us having millions of u13s on the
- Philippines recommendations Pressure is coming from grooming
- platform,
- CEI and cyberbullying escalations
- which we aren't currently detecting/enforcing against.
- **CEI/NCII prevalence risks fueling throttling or unhelpful data regulation:** Pakistan (X platforms) already banned due to "immoral content" and FB child safety issues used to fuel Citizen Online Protection Rules/control of speech.
- **Livestreaming abuse and coordination with 5 eyes.** Local momentum on public calls by senior gov (Dutton etc) is

[redacted]

Immediate Product Vulnerabilities

Sylvia [redacted] edited *Sep 23, 2020 at 4:00 pm*

AMET Policy and Regulatory Context

There are neutral to negative sentiments generally around Facebook in the region with concerns over our platforms impact on misinformation, hate speech, privacy and its effect on young users in the region. Our market domination in the continent is also seen as largely problematic by policy makers who are churning out social media, tax and competition legislation to deal with this.

Child safety concerns are growing in the region especially after the increased use of digital platforms by children because of school closures due to covid-19. Access issues and digital literacy form the majority of concerns as the level of online safety awareness is very low and younger users are exposed to inappropriate content. Facebook's efforts to address this have come under scrutiny from CSO partners discontent with our digital literacy investments in the region. Our competitors, Google are investing heavily on child safety in the region: Be Internet awesome and they recently announced a 1 million USD fund to support safety NGO's in Africa

[redacted]

What we need

- Need for more funding to support online safety awareness programmes in the region. We currently do not have any safety programmes budget support for this. Our competitors have made huge investments and are currently leading the efforts in developing localised programmes to address the digital literacy issues in the region.
- Representation of the region in the Safety Advisory Board and Youth Advisory Board. We do not have any representatives from the AMET region in our safety advisory board which has led to underrepresentation of the region safety issues at the SAB level.

Ravi Sinha edited *Sep 23, 2020 at 3:34 pm*

- Minor Sex Trafficking: Need a proactive classifier that allows us target minor sex trafficking and distinguish it from adult prostitution activity. Also need further definition of what constitutes trafficking and where our policies draw the line between it and prostitution.

Amber [redacted] *Sep 23, 2020 at 2:40 pm*

Following up on more detail on this... heard from LE outreach this would be a win in ANZ as LE use our tooling once they have access to device etc and it helps with investigations. Sentiment is that focusing on improving quality of NCMEC reports won't gain as much traction in countries like Australia but this could be a win.

Amber [redacted] opened your conversation with Jennifer Hanley, Sarah Chang, Ravi Sinha, Karuna Nain, Antigone Davis, Holly Hawkins, Deepali Liberhan, Liz Lagone, Sylvia Musalagani, Vaishnavi J, MC Capelo, Sebastian Sposito and David Miles *Sep 23, 2020 at 2:33 pm*

Amber [redacted] edited *Sep 23, 2020 at 3:59 pm*

APAC Trends

- eSafety's Safety by Design approach continues to gain traction with an increasingly global audience (collaboration with VCs and Alex Stamos in the US, multi-lateral/INGOs such as UNICEF and ICTU and seeding approach with APAC/South Asian countries). Phase 2 moves beyond principles to best practice recommendations on internal processes, integrity measures and settings to protect minors. While still voluntary, the proposed AU Online Safety Act already points to SBD as the basis for minimum requirements/expectations of intermediaries, so could form the basis of legal obligations in future and will continue to have visibility.
- This sits alongside a proposed expansion of eSafety's powers to enforce on expanded harms (e.g bullying & harassment) for not only children but adults, (mirroring the UK and Germany to extending the scope of harms from illegal to harmful content). The expansion beyond strictly illegal content is also apparent in proposed regulation in the Philippines which now has in its sights content that is developmentally harmful to children.
- Several countries are considering establishing a dedicated child safety regulator. Australia and Fiji already have one, the model (which Australia continues to spruik) is gaining some traction among CSOs and legislators in Taiwan and to some extent, the Philippines, at present. Various government entities in South Korea are also attempting to stake claim over online child safety (and NCII). While updated regulation and government coordination in these countries could be a welcome improvement, we are seeing multiple proposals, largely in conflict with US law, to expand our reporting obligations and data sharing to local LE and gov and - in the case of South Korea - require us to proactively detect/prevent uploads of content the gov deems violating without the ability to verify what the content is (i.e ingest CSAM/NCII hashes from gov not just other industry/NGOs).
- Age-verification is also going to become a dominant feature across APAC and this has been accelerated by CoVID where we are seeing an increasing number of CEI/grooming escalations and where in countries lik In South Korea and Japan the push is generally to verify age through mobile phone number verification. In the Philippines recommendations Pressure is coming from grooming, CEI and cyberbullying escalations.
- It may be that the Tech Coalition Project Protect, is the appropriate place for some initiatives but cross-industry bodies often suffer from inertia.… a demonstration of our leadership. Yes, we may be the catalyst for a broader cross-industry

initiative but we should take the credit.

- Increased investment in improving DIY tooling

Sylvia [redacted] opened your conversation with Jennifer [redacted], Amber [redacted], Sarah [redacted] Ravi Sinha, Karuna [redacted] Antigone Davis, Holly [redacted], Deepali [redacted] Liz [redacted], Vaishnavi J, MC [redacted], Sebastian [redacted] and David [redacted] *Sep 23, 2020 at 2:24 pm*

Sylvia Musalagani edited *Sep 23, 2020 at 2:24 pm*

- Create a Child Safety Hub - an online, global resource that explains Facebook's approach to combating child exploitation - targetingPolicymakers targeting Policymakers, CSE experts, industry leaders, and the general public (perhaps housed in the Facebook Safety Center).

David [redacted] opened your conversation with Jennifer [redacted], Amber [redacted], Sarah [redacted], Ravi Sinha, Karuna [redacted], Antigone Davis, Holly [redacted], Deepali [redacted], Liz [redacted], Sylvia [redacted] Vaishnavi J, MC [redacted] and Sebastian [redacted] *Sep 23, 2020 at 1:22 pm*

David [redacted] edited *Sep 23, 2020 at 1:29 pm*

-

European Context

Facebook's EMEA leadership view is that Facebook is facing a much more concerted effort to go after tech in Europe and Facebook's business model specifically.

We are not alone in this. Facebook, Google and Microsoft are all considered to have too much power in the European and Global marketplace. The surveillance narrative (as expounded by Zuboff) is broadly accepted in Europe and US tech has not been able to mount an effective counter narrative. We have come to symbolise all that is wrong in terms of the European digital marketplace.

Safety, because of the consensus around the importance of protecting children in the digital space, is now being used successfully in Europe, by the European Union and Member States, through regulation, in part, to reassert control of Europe's digital destiny.

Our decision to encrypt Messenger is seen as symptomatic of a pivot to privacy, at the cost of children's safety. That US notion of data privacy, is not the same in Europe and therefore exacerbates the situation.

The European Union Internet Forum for example, consciously excluded Facebook from its draft technical paper on encryption. It suggests to me that Facebook is perceived as the problem. Our competitors (Microsoft in particular and Google to a lesser degree) are accentuating that approach and along with so-called experts, seek a EU legislative response during Q2 2021, that could be detrimental to our plans and in Microsoft's case, have little impact.

[redacted]

*Estimated dates of being enacted or being laid in parliament

European Union

ePrivacy Directive (21st December 2020)
Audio Visual Media Services Directive (AVMSD) (September 2020)
Digital Services Act (July 2020)
CSAM Legislation (March 2021)
*Estimated dates of being enacted of being laid in parliament)



Ravi Sinha *Sep 22, 2020 at 11:06 pm*

Sarah [redacted]

Ravi Sinha edited *Sep 22, 2020 at 11:05 pm*

- Minor Sex Trafficking: Need a proactive classifier that allows us target minor sex trafficking and distinguish it from prostitution activity. Also need further definition of what constitutes trafficking and where our policies draw the line between it and prostitution.

Holly [redacted] edited *Sep 22, 2020 at 9:35 pm*

There are a number of opportunities to fill gaps in education and resources related to child safety:

- Create a well-funded digital citizenship and wellbeing program combining Get Digital (Global), We Think Digital (APAC) and other outliers (e.g., My Digital World - SSA) into a robust, global learning platform engaging families, educators and youth to take an active role in online safety.
- Address the general lack of knowledge regarding safety tools currently available on the platform:
  - Elevate positioning of existing online safety resources (e.g., Facebook Safety Center) for increased awareness and usage. Create greater visibility within the Facebook platform - linked from the profile page and elsewhere.
  - Embrace opportunity to proactively promote safety tools and resources pushing out through house ads and other internal mechanisms to increase visibility.
- Create a Child Safety Hub - an online, global resource that explains Facebook's approach to combating child exploitation - targetingPolicymakers, CSE experts, industry leaders, and the general public (perhaps housed in the Facebook Safety Center).
  - Build team to take action on initiatives that come of global Child Safety coalition meetings (e.g., WeProtect).
- Enhance product safety messaging and provide educational resources for teens (13+ products) and children (under 13 products) and parents that are globally informed and relevant.
  - Develop consistent, expert informed safety messaging for multi-audiences (e.g., built out safety tips) across multi-platforms/products.
  - Create in-product, point-of-use safety messaging for parents and children to fully understand opt-in/opt-out communication/interactive features and functionality.
  - Relevant localization of safety messages for global product rollouts (e.g., MKids Resource Page - does not exist outside of US).
  - Platform-wide PSA to push community safety/empowerment of users.
  - In conjunction with youth create a coordinated campaign for young people by young people about how to stay safe online.

Holly [redacted] edited *Sep 22, 2020 at 6:08 pm*

Sarah [redacted] edited *Sep 22, 2020 at 8:14 pm*

- Immature minor sexualization classifier: Implicit Minor sexualization policies may launch this half but our AI tooling is not mature enough to proactively detect complex objects, where (where a page/group/profile shows a collection of borderline to benign image is images of minors accompanied by minor sexualizing comments/captions comments/captions) because our page/group/profile-level detection has not been set up yet. In addition, we've been detecting and can currently only look for auto-actioning explicit minor sexualizing comments after launching a policy to enforce against explicit minor sexualization last half. However, our detection only runs on comment text in comments , which can lead to false positives if the media being commented on is benign/borderline, or false negatives if we do not detect explicit minor sexualizing captions attached to a photo/video featuring a minor.
  -
- Classifier global resilience: Need a comprehensive audit to assess whether our child safety classifiers are globally resilient and do not operate with bias. There is ongoing Child Safety engineer work to update our text classifiers to more easily internationalize and integrate new languages for proactive grooming detection.
  - This work will require robust data science work and product intervention simulations and tests

to verify what's most effective. For example, we believe… minor) from messaging, friending, or viewing the profile of minors might be an effective intervention to drive down grooming prevalence on IG Direct. IG, however, does not have its own Child Safety data scientists and progress on work will likely suffer without any additional data science resources.

- The best performing age model (the "Adult Classifier") is not yet globally operational, which leads to a patchwork of integrity measures across the company that leverage age, including our Child Safety models that detect grooming and determine true age differences between victim and offender provide protections for minors on our platform. We need more review resources to establish ground truth data on age in priority markets.
- Our biggest enforcement gap in relation to age for child safety is currently our inability to automate decisions accurately to enforce against groomers. The Adult Classifier — while it is our best available age model to determine whether someone is a teen or non-teen — cannot determine age gaps at a granular level which hampers our ability to action our policies, which rely on making specific age determinations, i.e. if the victim is 14 years old or younger and the offender is 18, disable and escalate but if the victim is 15 and the offender is 18, checkpoint the offender.

Sarah [redacted] edited *Sep 22, 2020 at 4:34 am*

- Implicit sexualisation of minors across multiple posts and comments on Instagram
- Sex trafficking and sexual solicitation networks on Instagram
- We have de-prioritized enforcing against u13s on our platform. We are neither building a proactive u13 classifier nor prioritizing enforcing against reported u13s. Recently, Identity Integrity did agree to enable cross-enforcement of u13s (so now, when a user on IG is reported for being u13 and is check-pointed, we will carry this enforcement over to their hard-linked FB account). However, there's also a current u13 backlog (~700K job backlog) that must be addressed via workload allocation and automation (i.e. auto-accept all other age change requests, which is a combo of people going from 13-18 > 18 and people going from >18 to 13-18). This backlog is currently not resourced.

Ravi Sinha opened your conversation with Jennifer [redacted] Amber [redacted], Sarah [redacted] Karuna [redacted], Antigone Davis, Holly [redacted] Deepali [redacted], Liz [redacted], Sylvia [redacted] Vaishnavi J, MC [redacted], Sebastian [redacted] and David [redacted] *Sep 22, 2020 at 4:26 am*

Ravi Sinha edited *Sep 22, 2020 at 4:56 am*

- AR/VR devices:
- Minor Sex Trafficking:
  -

Sarah [redacted] *Sep 22, 2020 at 4:11 am*

Ravi Sinha

Karun [redacted] *Sep 22, 2020 at 4:05 am*

Vaishnavi J can we be more specific in what we need here - research, engineering resources

---

Karuna [redacted] edited *Sep 22, 2020 at 4:07 am*

PublicPolicy Public Policy landscape

•

Content Policies

- IG Private by default
- A big investment in body image, social comparison, and eating disorder issues

Rough

- Our European data protection regulator (the IDPC) has opened two new inquiries concerning Instagram's processing of children's data



- **Germany youth protection legislation:** Three new pieces of youth protection legislation are on the horizon, with different areas of focus - some are focussed on regulating content that is harmful to minors, while others focus on structural youth protection e.g. settings around youth defaults.
- **Response to Unilever and 5Rights around how we handle age-appropriate design:** We prepared a written response to a number of concerns raised by the 5Rights Foundation. The IG-specific concerns revolved around: (a) public profiles on Instagram increasing risk to minors, (b) private messaging on Instagram being open to anyone, even unconnected people, and (c) a perceived lack of age verification on FB and IG.

- CNN International reached out to IWF enquiring about whether they've seen an increase in CSAM and included this line in their email: "We are investigating how the reduction in the number of moderators on social media affected the removal of child sexual abuse material online. In particular, we are interested to see whether Facebook failed to prioritise moderating content on Instagram, which has documented the highest number of child sexual abuse material."
- Media cycle about SSI on Instagram: We fielded an enquiry from a tech reporter at Daily Telegraph about self harm and eating disorder (ED) content on Instagram and being recommended accounts that promote it.
- BBC story about drugs on IG: The BBC released an investigation into the sale of drugs on social media, particularly Instagram and Snapchat, and focused on the ease with which young people can find and buy drugs online. They spoke to a group of mothers who have all lost teenage children to drugs; we spoke ahead of time with one of the mothers who said that it was clear Instagram is already putting 'an awful lot of effort into this and have a lot of people working on this specifically' relative to other companies. However, the parents don't believe social media companies like Instagram are doing enough, given the money, resources and technical expertise we have. They believe we should be able to a) stop the content from being posted in the first place; and b) catch it more quickly.
- Danish broadcast about SSI on IG: The Danish Broadcasting Corporation (DR) aired a feature on SSI on Instagram after they found several examples of graphic SSI content. This was positioned as a follow-up to the documentary that aired in February which uncovered an SSI network on Instagram. We used this as an opportunity to highlight the need for our more sophisticated classifiers in the EU and highlighted that COVID-19 meant reduced reviewer capacity. The main criticisms were that we don't care enough to look for and remove this content, we don't do enough to encourage people to report this content, and we're not living up to our responsibility to keep people safe.

- **Escalation - Jonathan Galindo SSI viral challenge (T68917996):** An article in a local newspaper quoted Brazilian police alerting parents about profiles with the name Jonathan Galindo and a 'dog face' profile photo for allegedly inciting suicidal behavior in young people. The case gained traction among the Brazilian press, with main national outlets picking up the news and the policy team received inquiries from policymakers.

Sarah [redacted] *Sep 22, 2020 at 2:14 am*

Ravi Sinha

Sarah [redacted] edited *Sep 22, 2020 at 4:15 am*

    - Our CEI video and photo proactive classifiers do not perform well enough to auto-action content
        - Rooms: Product continues to push for growth and plans to allow Rooms creation directly from WA and IG without FB log-in. They are also currently testing relaxing initial integrity measures, such as requiring an account age of 30 days to see if this leads to growth. If we want to avoid having Zoom's child safety problems, we need to prioritize filling the following product and enforcement gaps:
            - We hope to land alignment amongst policy, legal, and product leadership for an in-call evidence capture solution in H1 2021 but in order to review and enforce against our policies quickly and at scale, we'll need improved CEI proactive video classifiers that can detect child nudity;
            - We are also discussing with product the possibility of moving some of our riskiest users and communities onto public spaces — whether via public Rooms (subject to proactive harmful content scanning much as we do for FB Live) or public Messenger threads (forked off from e2ee MSGR infra) that contain more than 1K users. Any such solutions will require improved CEI proactive image and video detection.
    - Before we move to e2ee, we should prioritize reducing CEI shares and preventing bad actor connections: Data analysis has shown that large message threads may harbor a higher concentration of bad users and harmful content shares. If…
        - our current enforcement regime is capturing an incomplete picture and perhaps incorrectly over-enforcing against uploaders and/or under-enforcing against recipients. Interestingly, Asian and MENA dominate both uploaders and receivers and only 3% are from the U.S.
    - COVID review constraints
        - CO review capacity for CEI is at ~30%; CO Safety suspects this will continue to be true for many more months to come for large pools of queues (125k hr/wk across all teams) with <20% projected supply for the near future.
        - With the exception of MMS and CEI content classifiers, we're not currently running classifiers nor reviews on threads with >2 users in them during COVID b/c of capacity constraints and b/c these have very low action rates with very time consuming review processes
        - Groomer detection remains on for English on both IGD and MSGR and Spanish on FB during COVID. There is no Tier 1 IIC (which detects sextortion or plans to meet up in real life for sex) detection running for French, German, Indian, or Portuguese markets. We are slowly getting back French, German, Portuguese but Indian offices are still offline as they keep having outbreaks. We're also trying to expand to more Tier 1 markets but there's so little review capacity that in order to do so, we'll have to make a tradeoff elsewhere.
    - AR/VR devices:
    - Minor Sex Trafficking:
- **IG**
    - The daily number of minor victims having sex talk conversations in IGDM with adults are approximately 38x more than on FB Messenger.
        - Child Safety CI, Messenger Wellbeing, and IG Wellbeing collaboration is ongoing to bring down the prevalence of grooming instances in IGDM's this half.
        - This work will require robust data science work and product intervention simulations and tests to verify what's most effective. For example, we believe that preventing suspected groomers (or adults who share no mutual follows with the minor) from messaging, friending, or viewing the profile of minors might be an effective intervention to drive down grooming prevalence on IG Direct.

        - It is a known gap that IG does not provide the same protections offered to minors on Messenger, which is to prevent unconnected adults (who is neither Friends nor Friends of Friends with a minor) from initiating messages with a minor. We have spoken publicly about this protection (including in Jay [redacted]'s Senate Judiciary Committee testimony and attached QFR's) and should strive to bring IG up to par. In Cross-App Communications, IGD and MSGR adults will be prevented from messaging MSGR minors, but the same will not be true of MSGR adults who message IG minors.
    - Implicit sexualisation of minors across multiple posts and comments on Instagram
    - Sex trafficking and sexual solicitation networks on Instagram
- **Age**
    - The best performing age model is not yet globally operational, which leads to a patchwork of integrity measures across the company that leverage age, including our Child Safety models that detect grooming and determine true age differences between victim and offender. We need more review resources to establish ground truth data on age in priority markets.
    - Age models do not work as well on IG because IG only started to collect age data in December 2019. In addition, Adam Mosseri will not support upselling age collection for existing users because he sees it as intrusive data collection. Thus, we currently only have stated age for ~67% of IG users. We recommend engaging with IG leadership on collecting age for the remaining ~33% and prioritizing age model maturity on IG.

Vaishnavi J opened your conversation with Jennifer [redacted] Amber [redacted] Sarah [redacted] Ravi Sinha, Karuna [redacted] Antigone Davis, Holly [redacted], Deepali [redacted], Liz [redacted] Sylvia [redacted] MC [redacted] Sebastian [redacted] and David [redacted] *Sep 21, 2020 at 11:20 pm*

Vaishnavi J edited *Sep 21, 2020 at 11:54 pm*

- Our European data protection regulator (the IDPC) has opened two new inquiries concerning Instagram's processing of children's data

[redacted]

- **Germany youth protection legislation:** Three new pieces of youth protection legislation are on the horizon, with different areas of focus - some are focussed on regulating content that is harmful to minors, while others focus on structural youth protection e.g. settings around youth defaults.
- **Response to Unilever and 5Rights around how we handle age-appropriate design:** We prepared a written response to a number of concerns raised by the 5Rights Foundation. The IG-specific concerns revolved around: (a) public profiles on Instagram increasing risk to minors, (b) private messaging on Instagram being open to anyone, even unconnected people, and (c) a perceived lack of age verification on FB and IG.

Recent media coverage

- CNN International reached out to IWF enquiring about whether they've seen an increase in CSAM and included this line in their email: "We are investigating how the reduction in the number of moderators on social media affected the removal of child sexual abuse material online. In particular, we are interested to see whether Facebook failed to prioritise moderating content on Instagram, which has documented the highest number of child sexual abuse material."
- Media cycle about SSI on Instagram: We fielded an enquiry from a tech reporter at Daily Telegraph about self harm and eating disorder (ED) content on Instagram and being recommended accounts that promote it.
- BBC story about drugs on IG: The BBC released an investigation into the sale of drugs on social media, particularly Instagram and Snapchat, and focused on the ease with which young people can find and buy drugs online. They spoke to a group of mothers who have all lost teenage children to drugs; we spoke ahead of time with one of the mothers who said that it was clear Instagram is already putting 'an awful lot of effort into this and have a lot of people working on this

specifically' relative to other companies. However, the parents don't believe social media companies like Instagram are doing enough, given the money, resources and technical expertise we have. They believe we should be able to a) stop the content from being posted in the first place; and b) catch it more quickly.

- Danish broadcast about SSI on IG: The Danish Broadcasting Corporation (DR) aired a feature on SSI on Instagram after they found several examples of graphic SSI content. This was positioned as a follow-up to the documentary that aired in February which uncovered an SSI network on Instagram. We used this as an opportunity to highlight the need for our more sophisticated classifiers in the EU and highlighted that COVID-19 meant reduced reviewer capacity. The main criticisms were that we don't care enough to look for and remove this content, we don't do enough to encourage people to report this content, and we're not living up to our responsibility to keep people safe.
- **Escalation - Jonathan Galindo SSI viral challenge (T68917996):** An article in a local newspaper quoted Brazilian police alerting parents about profiles with the name Jonathan Galindo and a 'dog face' profile photo for allegedly inciting suicidal behavior in young people. The case gained traction among the Brazilian press, with main national outlets picking up the news and the policy team received inquiries from policymakers.

- Private by default
- A big investment in body image, social comparison, and eating disorder issues

---

Sarah Chang opened your conversation with Jennifer Hanley, Amber Hawkes, Ravi Sinha, Karuna Nain, Antigone Davis, Holly Hawkins, Deepali Liberhan, Liz Lagone, Sylvia Musalagani, Vaishnavi J, MC Capelo, Sebastian Sposito and David Miles *Sep 21, 2020 at 6:33 pm*

Holly [redacted] edited *Sep 21, 2020 at 6:14 pm*

Holly [redacted] opened your conversation with Jennifer [redacted] Amber [redacted] Sarah [redacted] Ravi Sinha, Karuna [redacted] Antigone Davis, Deepali [redacted] Liz [redacted] Sylvia [redacted] Vaishnavi J, MC [redacted], Sebastian [redacted] and David [redacted] *Sep 21, 2020 at 5:35 pm*

Karuna [redacted] added you to a document *Sep 21, 2020 at 5:17 pm*

Karuna [redacted] added you to a document *Sep 21, 2020 at 5:17 pm*

Karuna [redacted] added you to a document *Sep 21, 2020 at 5:17 pm*

Karuna [redacted] added you to a document *Sep 21, 2020 at 5:17 pm*

Karuna [redacted] added you to a document *Sep 21, 2020 at 5:17 pm*

Karuna [redacted] added Antigone Davis, Deepali [redacted], Holly [redacted], Liz [redacted] Sarah [redacted] and Vaishnavi J to the thread *Sep 21, 2020 at 5:23 pm*

Karuna [redacted] added you to a document *Sep 21, 2020 at 5:17 pm*

Karuna [redacted] enabled link sharing · View/Edit *Sep 21, 2020 at 5:20 pm*

Karuna Nain edited *Sep 21, 2020 at 5:22 pm*

Child Safety State of Play

Interim update - September 2020

PublicPolicy landscape

- CDA 230
    - EARN IT
    - UK Online Harms
- QAnon
- COVID-19

Product

- **Facebook**
- **IG**

Policies

Enforcement

Resources / Education

What we need

July 2020

Where are we with the various policy discussions relating to children on our platforms, and where we think we are most vulnerable and need to step up our efforts.

Child Exploitation

Policy Landscape

Our encryption announcement has created tremendous anxiety among CSAM safety stakeholders, and generated significant legislative activity across the globe centered around exceptional access and other technical requirements and/or intermediary content liability, as well as calls to break us up. Multi-national and multi-sector bodies, including the Five Eyes, the EUIF, and the WePROTECT Global Alliance, are driving a coordinated campaign against our message encryption plans, which has caught the attention of our advertising clients as well as large foundations such as the Oak Foundation and Skoll, and driven shareholder inquiries. A spike in CSAM reporting and increased dependence on online technologies during the pandemic has made this issue COVID 19 resilient. There is some risk we become isolated from the rest of industry, which could result in FB specific solutions (e.g., Apple has tried to differentiate FB e2e messaging based on social graph). Priority Countries: US, UK, Australia, India, EU and Brazil (other potential hot spots due to high incidence rates and/or active child exploitation legislation Philippines, S. Korea, Taiwan, Japan, Argentina, Chile and South Africa).

[redacted]

Immediate Product Vulnerabilities

- Rooms: livestreaming abuse
- Ephemerality: negative impact on ability to report
- Interop: discoverability and reachability
- NCMEC reporting: missing context
- WhatsApp groups: link sharing on social platforms + apps on iOS and Android store enable CSAM discovery and sharing
- COVID review constraints
- Unconnected adults being able to find and message minors in Instagram Direct
- Implicit sexualisation of minors across multiple posts and comments on Instagram
- Sex trafficking and sexual solicitation networks on Instagram

Areas of Improvement

While some governments will not be satisfied with anything short of a backdoor, offering limited client-side scanning will satisfy leading safety stakeholders (NCMEC, IWF, NSPCC, Thorn,...) and many governments. It will, however, alienate some privacy stakeholders. Other areas where we can step up our efforts to help manage our vulnerabilities and push back against efforts to require encryption backdoors are listed below.

- **NCMEC Reports:**
  - contextualize these numbers
  - meaningfully drive down the number of reports before encryption through preventative work including [redacted] safety notices, friending limitations,...
  - demonstrate through the upselling of reporting and packaging of metadata and info from public spaces, etc. that we can still aid/enable more investigations than LE currently does
- **Minor Sexualization:** much stricter policies and enforcement w/r/t the sexualization of minors on our services (e.g., group names, comments,...)
- **Trafficking:** sex trafficking classifier with a focus on minor sex trafficking
- **TC Project Protect:** fast track some key wins like universal video hash sharing, research on client-side solutions, independent accountability system that tracks to CSEA vol principles
- **Livestream:** develop content independent CEI detection tech, wiretap capability, reportability and reviewability upon report
- **Interop:** strong discoverability and reachability controls, including further restrictions to PYMK and GYSJ
- **Ephemerality:** adult<>minor carveout for disappearing mode and/or some way to enable content reporting and review
- **IG:** Extend existing MSGR Safety tips for scams, impersonation, child safety/IIC in IGD; Expand policies on borderline content re minor sexualization and build classifiers to detect and take action on problematic aggregated content on surfaces like Explore, hashtag pages, Reels, better age-related logic for preventing conversations between unconnected adults and minors
- **Groups:**

- remove private invite links and group chat functionality for high risk FB groups
- robust system for removing apps that categorize and make searchable group invite links to WA groups in violation of our terms of service
- robust system notifying platforms when users share group invite links for alleged WA CSAM groups

- **User Education:** robust ongoing in product user education on safety and privacy tools
- **RTC and Remote Presence products:** in-call reporting with evidence capture, feature gate high-risk users, user education/self-remediation interventions (i.e. safety notices, upsell reporting/blocking)

We also should seriously consider establishing ourselves as the industry's world leader on child safety through a large scale investment (@$20-40M) in the development of robust external ecosystem to thwart CSAM and support victims ("Big Safety Investment"), would include initiatives like global child help textline, global case management tool that aggregates all case data for investigations and helps with prioritization, and scaleable automated system for x-industry threat information sharing

Recommendation

Internal readiness to rise to the challenge. Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged, suggestions on how we might kick start conversations on this work.

Ahead of Interop, work is in progress to block unconnected adults on FB and IG from initiating messages with minors on FB, however work is not planned to prevent FB and IG adults from initiating messages to IG minors. Limited country tests and data analyses have shown that [redacted] can demonstrably reduce CEI sharing but product leadership has been unwilling to prioritize these mitigations. We are currently escalating the [redacted] conversation up through Monika to meet with MLT. We are also getting ready to meet with Jay [redacted] to discuss how best to present [redacted] to MLT. We need leadership support to have product prioritize this work because it will allow us to make a much stronger case with our external critics and stakeholders that we are doing good work on this front. We have progress on blocking predicted groomers from using Disappearing Mode but need CO support to apply the same protection in IG. Work has been completed on basic safety mitigations for Rooms but work still needs to be done to enable in-call reporting with evidence capture in both Rooms and RTC (aka Video Chat). RTC, which is the predecessor to Rooms, is not yet in parity with Rooms (i.e. feature gating high-risk users, user education/self-remediation interventions) and will require stronger leadership buy-in to accomplish even bare minimum protections this half. We continue to advocate for strong discoverability and reachability controls in anticipation of Interop, including further restrictions to PYMK and GYSJ. Work to develop a CEI video classifier for FB Live is currently ongoing but we need more resources for this in addition to engineering and review resources to develop an operational sex trafficking classifier. Please see the chart below for detailed summary of the specific related work and its status.

Age Assurance

Policy Landscape

With increasing frequency safety stakeholders are demanding that we do our utmost to keep U13 users off our platform and that we appropriately account for U18 users in the design of our services. Gaps in this area help fuel online harms legislation and regulation of various kinds, particularly outside the U.S. With a few key product changes and/or regulatory collaboration we can prevent ill-considered legislation (e.g., requiring id age verification at sign up, 18+ requirement for social media,...) and secure the support and confidence of child safety stakeholders, including policymakers parents/caregivers. Priority Countries: countries of jurisdiction (US and Ireland), and countries that are highly vocal and influential in region AU, UK, Brazil, South Africa and India.

[redacted]

content"

- Colombia/Peru/Paraguay/CENAM: routine bills on mandating blocking content for kids (videogames, social media, educational, etc)

Immediate Product Vulnerabilities

- Lack of U13 proactive age identification and enforcement mechanism
- Need for improved age identification models that allow us to better predict age of users and offer appropriate age-related filtering and protections (esp. w/r/t IG where still lack age data for most users, which prevents us from implementing a number of age-related protections e.g. restricting unconnected minors and adults from interacting, full rollout of Safety Warnings (currently only planned for those users whose age we know)

Areas of Improvement

The development of a device-side system for age verification to block U13s from our services will go a long way to addressing the concerns of child safety stakeholders, but we need to fast track that effort. We also need to improve our age identifying classifiers and expand "age appropriate" protections and experiences (e.g., coming soon IG account will be private by default for minors).

- Looking back: work with a trusted expert/group of experts to review our default settings, protections.... for minors, identify improvements (e.g. GYSJ and PYMK restrictions), and implement them
- Looking forward:
    - develop a "safety by design" process or product safety review similar to privacy review
    - in product U18 specific user education on safety and privacy tools (e.g., unique user education at sign up)

Recommendation

Internal readiness to rise to the challenge

Internal readiness to rise to the challenge. Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged, suggestions on how we might kick start conversations on this work.

In March it was decided that we would not build a classifier to detect minors under 13 years old. Instead, we would push industry partners for device/OS level age verification and management solutions. Since then, Central Privacy has stood up a small team on youth privacy but a PM is yet to be hired. They have offered to prioritize a few low business cost/high policy impact projects for H2. Identity Integrity has also identified a PM who would own age verification but they will not prioritize this work until we have clear legal obligations to do so. We need leadership support to push product to act on knowledge we already have, i.e. clear our backlogs, disable verified u13s, build an u18 checkpoint to review minors who have been reported on 18+ products like FB Dating. Against this backdrop, our enforcement is not consistent across all apps for a single known person. In addition, age models perform inconsistently across different markets. We recommend that Policy leadership stack rank and prioritize our most sensitive policy driven markets and make clear to product that we should operationalize those markets first. We also need leadership to support the basic principle that if we're using a signal to predict age for business purposes, it should be used to enforce on age. For example, we need to use age prediction to gate minors from alcohol ads and other high-risk features where we currently rely solely on stated age. Age models also do not work as well on IG because IG only started to collect age data in December 2019 and we currently only have stated age for ~67% of IG users. We need leadership support to convince IG leadership to collect age for the remaining users for whom we have no stated age. We also would like leadership sign off to conduct an audit of special protections for minors and then bring in a trusted expert/group of experts to review our default settings and protections for minors, identify improvements (e.g. GYSJ and PYMK restrictions), and implement them.

SSI

Policy Landscape

While SSI experts support our existing policies and our commitment to ongoing review of our policies, the need to explain the ever increasing rate of death by suicide and concerns about the impact of social media on youth well-being more broadly keep us under heightened scrutiny in this area — most notably in the UK. Specifically, this translates into concern among some policymakers and stakeholders that almost any exposure to this content for minors "normalizes" it and puts them at risk, and these concerns which help fuel online harms legislation to ensure we behave responsibly. Additionally, some policymakers and health policy experts are raising privacy related questions w/r/t our work in this area, creating a tension between those who push for increased scanning and those who want us to take a more privacy sensitive approach. This tension begs for a forward leaning approach involving both sets of stakeholder to come to agreed upon norms.

[redacted]

[redacted]

Immediate Product Vulnerabilities

- only one classifier running in the EU → escalations, heightened risk on IG specifically because of how #tags work
- lack self-harm support resources (e.g. eating disorders, cutting); currently, we send suicide prevention resources and tips to people engaging in any kind of self-harm
- lack different resources and specific actions for users who are using our platform to promote versus admit suicide
- lack population/risk specific resources on the SSI checkpoint
- **IG:**
    - lack classifiers to detect and take action on problematic aggregated content on surfaces like Explore, hashtag pages, Reels
    - spaces like IGD being used to encourage or promote suicide e.g., group suicides in Norway that occurred last year and we don't scan there

Areas of Improvement

Resolving the conflict with the IDPC so we can use our existing SSI classifiers in the EU will go a long way to addressing the policy heat we face in the UK in particular. That said, we should continue to stress test our SSI policies and implementation standards including on borderline content and SSI promotion vs. admission, and build a more context sensitive response system with differentiated actions and resources. We also have an opportunity to lead here by 1) working with privacy and safety stakeholders together to develop best practices with regard to AI, responses, data retention,... and 2) helping build a global crisis text line system

Recommendation

Internal readiness to rise to the challenge

Internal readiness to rise to the challenge. Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged, suggestions on how we might kick start conversations on this work.

On SSI, we are making progress on policy development re boderline content and strengthening our engagement x-industry and with experts. Our key limitation here remains the inability to use our existing SSI classifiers in the EU. We also don't have work underway to build classifiers on IG to detect and take action on problematic aggregated content on surfaces like Explore, hashtag pages, Reels. In addition we need to continue to refine our ability to provide people with more local/ culturally relevant resources depending on the crisis they are experiencing (self harm vs eating disorders vs suicidal ideation...).

Well-Being

Policy Landscape

The physical distancing requirements of the COVID-19 pandemic have generated greater awareness of and concern for issues related to mental health and well-being (e.g., anxiety and loneliness), as well as an appreciation among policy makers and the press for the role technology can play in supporting mental health and well-being (e.g., keeping people connected). This shift of focus from how tech may create, increase or exacerbate mental health and well-being issues to how tech can assist in combatting some of them is an opportunity for us, but a failure to execute well could surface an even worse "tech backlash" around mental health and well-being issues negatively associated with technology (e.g., passive time spent, pressure to be perfect, FOMO, etc.) and fuel some ongoing legislative/regulatory efforts.

[redacted]

[redacted]

Immediate Product Vulnerabilities

- FB+IG: Lack of self harm support resources (e.g. eating disorders, cutting); currently, we send suicide prevention resources and tips to people engaging in any kind of self-harm,
- Product infrastructure limitations: inhibit our ability to make necessary guarantees around protection of mental health related data
- FB - Bullying and harassment: Lack of mass comment moderation tools

Areas of Improvement

There is a need for stronger research around social comparison and loneliness on our platforms. We have an opportunity to launch x-industry Digital Wellness Lab with Center for Media and Child Health at Harvard to seed independent research in this space. We also have an opportunity here to demonstrate our commitment to building a supportive community through efforts such as building group admin support learning units, expanding WA business API for crisis support orgs, CSOM and the connected generations initiative.

- **IG**
  - **Policies:** we need to stress test our policies w/r/t bullying of U18 public figures, given IG public figures are younger and harassment targeting them is gaining more scrutiny
  - **Research**: conduct new research around social comparison and loneliness on Instagram to tackle some of the root causes of SSI, bullying, harassment; more foundational mental health research being done in non US/UK markets
- **Facebook:**
  - **Product:** Proactive bullying AI comment filters on FB to get parity w/ IG
  - **Policies:** Refining public figure definitions and developing more protections from harassment for public figures
- **Product:** Need to ensure global mental health efforts are culturally sensitive and align with existing approaches towards mental health; resourcing and prioritizing Purpose Policy Framework for protection of health information
- **Resources:** Expanding mental health resource pages that are easily accessible to users and regionally and age applicable

Recommendation

Internal readiness to rise to the challenge

Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged, suggestions on how we might kick start conversations on this work

We're making progress connecting people with resources on the COVID Information Center and with a new mental health resources landing page. We need more product work on issues such as social comparison, bullying of minor public figures. We are expanding CSOM but need more resources on product given the need to demonstrate value of encrypted surfaces for connecting people with safe, private and secure support.

Racial Justice

Also suggest internal audit to ensure that all safety classifiers are globally resilient and operate without bias.

Some detailed product asks:

- Lock in safety mitigations for Interop, including:
  - Strong discoverability and reachability controls, i.e. default minors to the most private and safe settings;
  - Extend existing MSGR messaging and Friending restrictions for unconnected adult<>minor in IGD and ensure cross-network messaging will not degrade current protections in place for minors;
  - Adult<>minor carveout for Disappearing Mode and prevent use of feature by predicted groomers;
  - User reporting education and message download for Unsent and Disappearing Mode messages with tombstones:
    - Ensure all content is reportable and reviewable by CO;
      - Increase retention period for unsent and ephemeral messages;
      - Add thread-level reporting in IG;

        - Call out in MSGR that unsent messages are reviewable;
        - Educate users about Impermanence at key product experience points
            - Tell sender+receiver that unsent messages can be reported for X first uses
        - Make Reports available to Reporters and give them options to save a copy or access reported content, including tombstones within context
    - Robust user controls for x-network reachability/discoverability.
    - Ensure existing safety controls such as Restrict and Ignore are not degraded by cross network messaging and maintain user expectations.
- Strong integrity measures to reduce harmful content sharing and interactions:
    - Remove private invite links and group chat tab in harmful content associated FB groups
    - [redacted]
    - Add "Who can…
- Sex Trafficking classifier with a focus on minor sex trafficking
- Self-remediation tools and backend measures that promote safety while respecting users' privacy for RTC and Remote Presence products, such as:
    - In-call reporting with evidence capture;
    - Screen-sharing static notification either in roster or as part of UI control panel;
    - Feature gate high-risk users;
    - User education/self-remediation interventions (i.e. safety notices, upsell reporting/blocking);
    - Understand abuse prevalence / useful proxies in IG RTC + MSGR RTC.
- Support our Global Classifier Resiliency and Fairness initiative for safety classifiers that are critical to our issue areas.
- Age Assurance / Verification
    - Proactive age verification and enforcement / u13 classifier; u18 checkpoint
    - Expanding age appropriate experiences, i.e. identify and build special protections (default settings, content restrictions, blocked products) that should apply to u13 and u18 users.
- Suicide and Self Harm
    - Self harm support resources (e.g. eating disorders, cutting; currently, we send suicide prevention resources and tips to people engaging in any kind of self-harm)
    - Providing varying resources and actions to people using our platform to promote versus admit suicidality; policy development and classifiers needed to action differently and appropriately.
    - IDPC classifier approval for Europe
    - SSI checkpoints on FB and IG need to be more relevant and population/risk specific
    - Crisis Service Over Messenger engineering support

X-Industry Sharing:

- Video DNA x-industry banks

User Education and Marketing:

- In-product user education campaign to drive awareness of privacy and safety tools, settings,..
- Design and Eng resources for the Safety Center
- Better product walkthrough for new users, teens (13-15) to flag how to report content, connect to the safety centre and help centre

Additional Quips/Docs

- 1H2020 Global Safety, Health and Well-Being Policy Plan
- H12020 Plans (Review)
- Safety Policy Product Asks
- H2 Plans
- Safety Legislative / Regulatory Tracker

Karuna Nain created the document *Sep 21, 2020 at 5:20 pm*

