Tennessee v. Meta
Case No. 23-1364-IV
PX-00021

Child Safety State of Play (9/24)

Note: This document is out-of-date. It reflects where are were with the various policy discussions relating to children on our apps as of 9/24.

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, and South Africa).

Legislative/Regulatory Action

Trends:

- **Child Safety concerns leading to expansion/conflation of illegal and harmful content regulation.** Following UK's Online Harms Paper and Germany NetzDG, Australia has proposed expanding the eSafety Commissioner's remit to enforce on expanded harms (e.g bullying & harassment) not only for children but also adults and is also extending the scope of harms from illegal to harmful content. Countries such as Brazil, the Philippines and Indonesia are all considering misinformation and child safety in the same category in recent legislations. In Pakistan child safety issues are being used to push the Citizen Online Protection Rules to curb free speech.
- **Dedicated child safety regulators being proposed.** Australia and Fiji already have an eSafety Commissioner. Legislators in Ireland, Taiwan, the Philippines, and South Korea are considering similar offices and India has proposed an eSafety Commissioner for each state.
- **Multiple reporting requirements:** India, South Korea, and Portugal are introducing legislations that will require us report CSAM directly to local authorities and not accepting NCMEC reports.
- **Competing legislation exposing safety/privacy tensions**: The ePrivacy Directive v eSafety strategy, and secrecy of communications provisions in Japan v Japan's Amended Telecommunications Broadcasting Act guidelines on SoC provisions and and secrecy of communications provisions Korea v intermediary liability law for NCII and CEI that requires collecting and sharing of more user information with local authorities/civil complainants.

Specific Bills:

- US: EARN-IT, DoJ/White House Section 230 reforms, Platform Accountability and Consumer Transparency Act, Online Freedom and Viewpoint Act, House Appropriations Bill on Labor, HHS, and Education funding, Transparency about Online Child Sexual Abuse Material Act
- UK: Cloud Act negotiations, TCN, Online Harms Bill
- AU: ongoing consultation on Online Safety Act, e-Safety Commissioner Safety by Design Principles
- Five-Eyes: Voluntary CSEA Principles
- India: traceability, amendments to "Protection of Children from Sexual Offenses Rules, 2020" requiring direct reporting w/liability
- Brazil: proposed amendments to "good samaritan provision" in Marco Civil
- EU: EUIF technical requirements and best practices, EU CSAM hub, Audio Visual Media Services Directive, Digital Services Act, CSAM Legislation
- Philippines: Anti-ATIP bill (intermediary liability, duty not to host CEI, mandates turn-around-times, requires platforms to preserve evidence and proactively report CEI directly to local LE. Safer Internet Bill (parental consent for U18s , TATs, 'harmful' content regulations); Anti-ATIP bill (intermediary liability, mandated TAT, evidence preservation requirements and proactively report CEI directly to local LE)
- South Korea: intermediary liability law for NCII and CEI
- Japan: National Police's annual report on Child Sexual Exploitation and Sexual Solicitation of Minors; Amended

Telecommunications Broadcasting Act guidelines on SoC provisions; Provider Limited Liability Act (provision of user information to civil complainants, incl phone numbers)

- Belgium: NCII legislation 6 hr takedown or liability
- Germany: NetzDG
- Ireland: Online Safety and Media Bill
- Kenya: Cybercrime Bill 2018
- Zimbabwe: Data protection and Cybercrimes Bill 2020
- Other: S.Africa, Japan, and Taiwan also have legislative activity

Where we think we are most vulnerable and need to step up our efforts.

Immediate Product Vulnerabilities

- CEI classifiers: low accuracy prevents auto actioning
- Grooming classifiers: lack global efficacy
- Rooms: livestreaming abuse
- Ephemerality: negative impact on ability to report
- Interop: increased discoverability and reachability across apps, particularly for IG where we have few child safety protections in place
- NCMEC reporting: spike in volume, poor metric, new safety mitigations have not yet driven down numbers and resistance to measures that likely will
- WhatsApp: link sharing on social platforms + WA group discovery apps in app store enable CSAM discovery and sharing
- COVID review constraints
- IG: unconnected adults being able to find and message minors in Instagram Direct, coupled with significant implicit sexualization of minors and evidence of potential trafficking and solicitation

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 (some of which are underway)

- **CSAM:**
  - improve CEI classifier, use Google safety API to help prioritize and launch CEI solicitation classifier
  - contextualize NCMEC numbers, stand up better metrics
  - close enforcement gap w/r/t CEI recipients
  - [redacted] safety notices, friending limitations,...
  - demonstrate through upselling of reporting and packaging of metadata and info from public spaces, etc. that we can still aid/enable more investigations than LE currently pursues
- **Minor Sexualization:**
  - launch stricter policies and enforcement w/r/t the sexualization of minors on our services
  - develop specific stricter policies for groups (e.g., group names, comments,...)
  - improve minor sexualization classifier.
- **Trafficking:** build 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, real time calls and other remote presence products:** develop content independent CSAM detection tech, wiretap capability, in-call reporting/reportability with evidence capture to ensure reviewability upon report, user education/self-remediation interventions before, during, or after (e.g., safety notices, upsell reporting/blocking)
- **Interop:** strong discoverability and reachability controls, including further restrictions to PYMK and GYSJ
- **Ephemerality:** adult<>minor carveout for Vanish Mode and/or some way to enable content reporting and review
- **Classifier global resilience and efficiency:** comprehensive audit to assess and ensure child safety classifiers are effective globally. Improve global efficacy of our text classifiers that detect inappropriate interactions with children, which would help enable automation of hard actions (disable/report) against groomers.
- **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
- Private by Default for Teens

- **Reporting:**
  - Launch "Involves a Child" feedback tag where missing e.g., in FB Groups and IG comments.
  - Build access to emergency support into reporting
- **Groups:**
  - stricter policies w/r/t naming, about,... and admin responsibilities
  - 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, as well as global program
  - Well-funded global digital citizenship and wellbeing program similar to Google's Be Internet Awesome
  - Safety education video for minors at sign-up or early in account and potentially to unlock certain features
- **Review Capacity**
  - Legal signed off on WFH review, fast tracking discussions with GenPac
- **Big Bets:**
  - We also should 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.
  - Develop a white paper following on our content regulation white paper that outlines a robust way to successfully and measurably address industry efforts to thwart CSAM

Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged.

Internal Readiness/State of Play

With a strong PM, the pressure of e2e, and a fundamental commitment to child safety, we are making real progress across many of the areas of improvement listed above, BUT we could be doing more faster if we had a dedicated pillar within the company and an executive level sponsor.

	A	B	C	D	E	F
1	**Area**	**Additional notes**				
2	**Ready; work is complete / ongoing and on track**					
3	Sexualisation of minors policy	Ongoing: Expanded minor sexualisation policy in October				
4	CSAM profile level reporting on IG	Completed: profile-level reporting option for CEI				
5	CEI Search Intervention	Shipped deterrance search intervention. Work ongoing to ingest new keywords, adjust to adversarial action and make available on additional products (IG hashtags, IG handle names, IG direct thread names and discovery search)				
6	Rooms	Completed: Safety mitigations include 1) feature gating high-risk users (NCMEC reported and predicted groomers) 2) proactive detection of CEI associated Room names 3) we do not recommend Rooms to unconnected adults if they contain minors (and vice versa) 4) we prevent unconnected adults from inviting minors directly into Rooms 5) we block Room links from comments in a minors' post if the commenter is an unconnected adult				

		6) we block Room creation by Groups with membership of 1 and photos, and 7) we also log and retain identifying metadata on all users, including anonymous joiners for 90 days				
7	Video Chat	Completed: Suspected groomers are blocked from video chatting with minors.				
8	NCMEC numbers contextualization	Ongoing: prioritized by Child Safety DS				
9	Interop minor protections	Ongoing: Just as unconnected FB adults cannot currently initiate messaging with FB minors, IGD adults will not be able to initiate messages with FB minors (ETA ahead of Interop global launch); preventing unconnected FB adults with a high IIC score from being able to reach out to IG minors. Completed: most conservative/private reach and messaging controls for minors and high volume recipients.				
10	Messenger Safety Tips	Ongoing: Safety Tips for Minors when interacting with suspected groomers is planned for cross-app communications. ETA for IG: October/November				
11	Tincan /Secret Conversations and Vanish Mode carve-out	Completed: Suspected groomers are now blocked from Secret Conversations and will be blocked for Vanish Mode. Users will be educated on their ability to report Vanish Mode messages for 1 hour (on Messenger) and 14 days (on IG Direct) after the message is seen. Needed understand work on 1 hour retention period and CO review capacity to upsell reporting in addition to education				
12	IIC classifier use on IG	Completed: IIC Classifier calibrated and operational allowing child safety mitigations such as gating suspected groomers from Vanish Mode.				
13	Improving text classifiers to detect grooming	[redacted]				
14	[redacted]					
15	COVID review challenges	In Progress: Legal has signed off on WFH review of graphic content on a voluntary basis, thereby increasing internal review capacity and allowing us to bring on vendor(s) whose employees are voluntarily willing to on CSAM review. CO has initiated conversations with vendors (in particular Genpact) but it will take weeks to get this up and running.				
16	Google API	In Process: started using Google's Safety API, which includes a powerful child nudity classifier enabling us to more rapidly enqueue and review CEI and significantly improve our action rate.				
17	**Somewhat ready; work is ongoing but more resources are needed**					
18	Interop: discoverability and reachability	Needs more resources: Strong discoverability and reachability controls, including further restrictions to PYMK and GYSJ				
19	Interop minor protections	Planned: Hiding minors' context lines from unconnected adults when they message minors in cross-app communications and blocking				

		message initiation from suspected IG groomers to IG minors.				
20	RTC and Rooms	On-going: In-call reporting with evidence capture work; RTC has commited to bringing themselves into parity with Rooms safety mitigations, which at minimum includes feature gating high-risk users. *In discussion*: user education/self-remediation interventions (i.e. safety notices, upsell reporting/blocking) and gating NCMEC reported users.				
21	FB Live and livestream video	Ongoing: development of a CEI video classifier for FB Live. *Needs more resources*: Child Safety CI team's CEI video classifier isn't great on performance				
22	X-industry video hash sharing	Needs more resources: Initial conversations underway to develop universal video hash sharing in collaboration with key partners like Thorn who have done early work and our opensourcing of PDQ-TMKF in 2019.				
23	[redacted]	[redacted] can demonstrably reduce sharing of CEI (a recent test showed that [redacted] reduced our NCMEC Cybertip report numbers by 7% in test countries) and could help us push back on calls for backdoor legislation. We are currently escalating the [redacted] conversation up through Monika to meet with MLT and could use your support. We are also getting ready to meet with Jay [redacted] to discuss how best to present [redacted] to MLT.				
24	Classifier global resilience	Needed comprehensive audit to assess and ensure child safety classifiers do not operate with bias; Ongoing: Child Safety engineer work to update text classifiers to more easily internationalize and integrate new languages for proactive grooming detection. Needed: Similar efforts for language-agnostic, image-based classifiers.				
25	IG Private by Default for Minors	Two-week test running to explore the viability of Private by Default for all minors				
26	Recidivism	Ongoing: Experimenting with SUMA (Single User Multiple Accounts) to identify links between accounts controlled by the same person who has been identified as as groomer and then applying friending restrictions to new accounts created by the same actor. Results are promising (prevalence delta between test and control is -45%) and show that actor-level restrictions can amplify our ability to prevent recidivism on our platform. Need More Resources: We currently don't device disable accounts flagged for grooming if it is associated with 5 active, non-fake accounts. However, preliminary data analysis has shown that the majority of devices we were not able to disable were grooming violators and were not disabled precisely because of this 5 account rule.				
27	Not ready; work has not been prioritised and more resources are needed					
28	[redacted]					

	[redacted]					
29	Aggregated content on IG surfaces	Build classifiers to detect and take action on problematic aggregated content on surfaces like Explore, hashtag pages, Reels				
30	Detecting minor sexualization	Minor sexualization policies launching in October, however AI tooling is not mature enough to proactively detect this content at scale				
31	Block reach of unconnected adults to IG minors	Unlike MSGR, on OG messages can be initiated by unconnected adults to minors. For Interop, this rule is preserved in MSGR, but not for cross-network messages sent by FB adults to IGD minors or IGD adult to IGD minors. Grooming on IG ~38x as much as FB. Growth however concerned about blocking FB parents and older relatives from being able to reach out to their IGD children/relatives.				
32	Reporting CSAM	Launch "Involves a Child" feedback tag where missing e.g., in FB Groups and IG comments.				
33	Upsell reporting for Vanish Mode	Vanish Mode messages sent in IGD are retained for 14 days and in MSGR are retained for 1 hour. The bulk of user reports for IIC (57%) and CEI (46%) in MSGR and IGD are sent within the first hour. After the first hour, the number of reports drop off dramatically (2-5% made within each subsequent hour). However, product does not have the CO resource capacity to support reporting upsells within Vanish Mode to educate the user that even ephemeral messages can be reported and that they have a limited window of time in which to do so. Once our transparency report is released, governments will seize upon spikes in our NCMEC reporting numbers during COVID-19 to fuel calls for exceptional access legislation. We should make clear to product and CO that during a time when we have increased CEI activity on our platform, we should prioritize reporting upsells for Vanish Mode before global launch.				
34	Sex trafficking classifier	We have one engineer dedicated to this work and we have no data science support to do some basic understand work to lower the strikes threshold for prostitution and sexual solicitation stikes (currently set at 17x, which is too high). This area is severely under-resourced given our SESTA-FOSTA liability.				
35	Proactive CEI video and image classifier	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.				

Suggestions on how we might kick start conversations on this work.

Recommendation

Overall, we have XFN alignment on the areas of improvement and priorities and with a strong PM a good bit of this work is in process and on a relatively fast track. The biggest challenges here are structure and resourcing

On the product side child safety sits inside broader integrity efforts, which means priorities are balanced against all harms. As a result, we sometimes have to deprioritize child safety when certain issues take precedence (e.g., US Elections, racial injustice,...). We should consider setting child safety up as a separate vertical similar to privacy, with requisite staffing and funding. Assuming the vertical had an executive level sponsor, doing so also could help with weighting v growth and privacy when making decisions [redacted] age modeling, client side scanning). We might even want to consider a structural change closer to what we have done with health where we pull together the integrity challenges (child safety) and opportunities (youth centered features and apps) work; perhaps this could unlock all kinds of product opportunities (e.g., Messenger Kids, tween apps).

That said, if pushing for this is too big a lift, perhaps policy leadership could initiate a comparative analysis of how we prioritize child safety versus other top areas of concern like privacy (e.g., headcount, research funding,...) and then use that to have a more robust internal conversation regarding our commitment to child safety; and in the meantime, to deal with the most immediate existential threats outlined above policy leadership could:

provide HC for a child safety policy SME who focuses solely on driving a child safety strategy

push for product shifts that would dramatically decrease the number of cybertips ahead of e2e launch (e.g., [redacted]

push for whatever HC and resources are needed by child safety team in CI to help them do more, faster ahead of e2ee; *note: this isn't something I have run by Guy*

push to improve our age models and make them globally operational since all child safety mitigations rely on underlying age models; *note: some of this was discussed as part of age verification discussion, and there was commitment to do some of this but*

help unlock exploration of the most privacy sensitive client-side scanning in case we need it as an off ramp

consider a large scale investment @$20-40M in the development of robust external ecosystem to thwart CSAM and support victims ("Big Safety Investment")

Age Assurance/Appropriateness

Where are we with the various policy discussions relating to children on our apps.

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, as well as significant negative stakeholder feedback and PR. With a few key product changes, regulatory collaboration and/or investment in research we can prevent ill-considered legislation (e.g., requiring id age verification at sign up, 18+ requirement for social media,...), secure the support and confidence of child safety stakeholders, including policymakers parents/caregivers, and reduce bad press. 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.

Legislative/Regulatory Action

Trends:

- **Growing momentum for mandatory child safety by design:** Prompted by UK's ICO Age Appropriate Design Code and Australia's eSafety Commissioner's Safety By Design principles', more and more policymakers around the global are looking at child safety by design requirements
- **Increasing number of countries adopting >13 age of digital consent, parental consent and age verification laws:** Policymakers in India have proposed legislation requiring parental consent for those under 18; policymakers in various African countries exploring the same; and in Korea where such legislation already exists they are exploring mechanisms for verifying parental consent.

Specific Bills:

- US: FTC to update COPPA rules (expected in 2020)
- UK: Age Appropriate Design Code (likely implemented in H1 2021 subject to Parliamentary approval), Online Harms Bill

- Ireland: Child Safety and Media Regulation
- Australia: eSafety Commissioner's Safety by Design principles, amendments to the Online Safety Act
- Brazil: Internet Steering Committee ("CGI") focused on ads for minors and use of minors data (e.g. Bill 1746/2015)
- Germany: Federal and Interstate Youth Law (likely to come into effect in H2 2020; proposes browser-level age verification for German citizens)
- South Africa: Protection of Personal Information Act 4 of 2013 (July 1, 2021)
- Philippines: Safer Internet Bill (parental consent for U18s)
- India: Personal Data Protection Bill, 2019 (requires age verification and parental consent)
- SSA: Zimbabwe's Cyber crimes bill (parental consent and robust age verification), South Africa Films and classification amendment act 2020 (registration, self classification, age verification for pornographic sites), Kenya Data Protection Act 2019 (age verification for under 18's on social media platforms)
- Indonesia and Philippines: Draft bills w/r/t age-verification to ensure minors not exposed to "inappropriate/immoral content"
- South Korea and Japan: Proposals requiring we verify age through mobile phone numbers
- Colombia/Peru/Paraguay/CENAM: routine bills on mandating blocking content for kids (video games, social media, educational, etc)

Where we think we are most vulnerable and need to step up our efforts.

Immediate Product Vulnerabilities

- We lack a U13 classifier, insufficient means to detect and remove u13s at scale
- We also have deprioritized u13 enforcement because COPPA compliance is at risk when we sit on knowledge of u13, incentivizes not knowing
- Age prediction models are not that accurate, reducing the efficacy of classifiers like the grooming classifier and requiring more reliance on human review
- Age prediction models are not globally operational; as a result, integrity measures that leverage age are inconsistent and unreliable, including child safety protections
- Age prediction models are not globally resilient, they tend to perform better in Western markets and worse in emerging markets
- Age models do not work as well on IG; as a result, we have no ability to more granularly gate minors out from age-related experiences on Instagram.
- As a result of all of the above, are age-gating doesn't work as well as it should yet we offer age-gated products like Watch Together, FB Dating,... and we have minors who appear as young as 12 years old.
- Lack of u18 checkpoint: Risk having u18 users on 18+ products like FB Dating, Marketplace, Mentorship, or as under-age parents on Messenger Kids
- Age verification: We remain reactive on age verification, pushing for a device side solution, when we're getting increasing policymaker pressure to be proactive while our competitors improve their approaches

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 should fast track that effort. We also should improve our age identifying classifiers and expand "age appropriate" protections and experiences. Below are specific suggestions:

- **U13 enforcement:**
  - Prioritize cross-enforcement of u13s on IG who have hard-linked FB accounts.
  - Clear current u13 backlog (~700K job backlog) which is currently not resourced.
- **U18 enforcement:**
  - Create an u18 checkpoint to effectively manage gating u18s from 18+ products
- **Age verification:**
  - Use age prediction in combination with age verification checkpoints to gate based on age-gated product or content.
- **Age Model improvements:**
  - Expand best available age models to more markets
  - Improve age model accuracy in emerging markets
  - Improve age models on IG, we currently only have stated age for ~67% of IG users and age models don't work as well, need IG leadership to consider collecting age for the remaining users for whom we have no stated age.
- **Age Audit:**

- Review our default settings and gating for minors, pressure test them identify improvements (e.g. GYSJ and PYMK restrictions), and implement them

- **Safety by Design:**
  - Develop a "safety by design" process or product safety review similar to privacy review
- **Education:**
  - In product U18 specific user education on safety and privacy tools (e.g., unique user education at sign up)
- **Age appropriate experiences:**
  - Develop 13-16 yo, 16-18 yo and 18+ experiences

Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged.

Internal Readiness/State of Play

Overall we are quite vulnerable when it comes to age management on our apps. To some degree that is because leadership has decided to live with certain risks at least in the near term. For example in March, we decided we would not build a classifier to detect minors under 13 years old. Identity Integrity identified a PM who will own age verification but we will only prioritize this work when we face clear legal obligations to do so. Instead, we will push industry partners for device/OS level age verification and management solutions as they are best placed to do this well. At the same time, leadership did not rule out improving our age classifiers and we are building age specific experiences, but that work is not advancing with the same speed as external pressures on this subject.

	A	B	C	D	E	F
1	**Area**	**Additional notes**				
2	**Ready; work is complete / ongoing and on track**					
3	Preventing unconnected adults from initiating messages with FB minors	Using a combination of stated age, age prediction models, and hard-matched or high-confidence soft matching, product work is ongoing and on track to prevent unconnected IG adults from initiating messaging with FB minors in cross-app communications.				
4	Parental Controls	Work is ongoing to bring parental controls to Oculus -- the only gaming platform that doesn't have parental controls. Unification across the company and account linking puts Oculus on notice of users' ages. But as a platform that will largely pull from FB data to predict age, backend protections for minors on Oculus will also rely upon improving our age models.				
5	**Somewhat ready; work is ongoing but more resources are needed**					
6	IG Age Affinity Models	Core Dimensions is building out age affinity models to drive the appropriate usage of age on IG through better onboarding, usability and enforcement. However, 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 age for ~65% of IG global Monthly Active Users using stated age (using either an IG stated or FB hard match). 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. While IG leadership has been periodically engaged with Core Dimensions, we'd recommend accelerating progress on these improvements and have strong leadership backing to prioritise age model maturity on IG.				
7	U13 enforcement	Ongoing: Recently, Identity Integrity did agree to enable cross-enforcement of u13s on IG who have hard-linked FB accounts since basic COPPA compliance is at risk when product does not prioritize checkpointing and disabling u13s. There is a chance, however, that this fails to get resourced because of engineering constraints on that team. However, there's also a current u13 backlog (~450K 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 and has been escalated to Ops Resources. In the current state, we would not be able to reduce the backlog to zero for another year. The cross-enforcement of u13s may also be at risk if the 2020 elections need more investment.				
8	Global use of age models	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 provide protections for minors on our platform. We need more (a) review resources to establish ground truth data on age in priority markets, which takes headcount and capital, as well as (b) more engineering resources to invest in model improvements where accuracy is poor (e.g., India). We could start by engaging with Core Dimensions (the product team who owns the age models) on what markets we've identified and stack ranked as a Policy org to prioritize with their available H2 resources.				
9	Accounts You May Follow (Instagram)	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, largely in part because we historically did not have age-related data for IG accounts. 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).				
10	Not ready; work has not been prioritised and more resources are needed					
11	Age data on IG	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 remaning ~33% and prioritizing age model maturity on IG.				
12	Addressing 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. However, with a lack of age data, we cannot effectively develop solutions at scale that address virality of minors on Instagram without also generating a large number of false positives / false negatives.				
13	Inability to automate groomer disables	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. However, single-year granularity is really difficult to achieve with machine learning models. Thus, we'd also need additional human reviewer investment to get better at this.				
14	U18 checkpoint	We do not yet have product support or resources for an u18 checkpoint where we can efficiently manage u18 users who are reported or detected on 18+ products like FB Dating, Marketplace, Mentorship, or as under-age parents on Messenger Kids. Identity Integrity, which owns underage checkpoints, has no plans to invest in age management in H2 and will instead prioritize impersonation work. Central Privacy has decided to				

		stand up a small team on youth privacy but they have not yet hired a PM and have offered to prioritize a few low business cost/high policy impact projects for H2. If Identity Integrity cannot prioritize age because of U.S. GE 2020, our best opportunity might be with Central Privacy's willingness to help and we could push for immediate hire of a PM to make some progress on this work in H2.				
15	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. We could also treat 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.				

How we might kick start conversations on this work.

Recommendation

Our overarching recommendation is the same as mentioned above, create an independent child safety vertical, and include this issue as a priority within that vertical to ensure it the proper focus and resources . Again, if that is too big a lift, to deal with the most immediate vulnerabilities w/r/t to age assurance, management and appropriate design policy leadership could:

provide HC for a child safety policy SME who focuses solely on driving a child safety strategy

sign off to conduct an expert audit of our default settings and protections for minors, identify improvements (e.g. GYSJ and PYMK restrictions), and implement them.

ensure product and CO have the resources need 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, etc..

establish as basic company principle that if we're using predicted age for business purposes, we should do the same for enforcing on age/safety

support improving our age classifiers and operationalizing them based on a stack ranking of our most sensitive policy markets

prioritize rapid improvement of IG age models given younger demographic

urge Central Privacy to fast track hiring of a youth privacy PM and to prioritize a few low business cost/high policy impact projects for H2

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. At the same time, some policymakers, most notably the IDPC, 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.

Legislative/Regulatory Action

- UK: [redacted]lue (pre-inquest review hearing September 25); Online Harms Act;-UK Samaritans guidance for industry released 9/10 recommending a single PoC, clearer policies to prohibit "normalizing" SSI and no encryption* to ensure can scan for this content in messaging
- Ireland/IDPC/EU: IDPC has continuing privacy concerns around processing of mental health data when running our SSI classifiers, as a result we are taking a far more limited approach in the EU which means we are susceptible to more press fires
- EU: under the e-privacy directive scheduled to take effect in December 2020, OTT communication services have to get explicit individual opt ins (~GDPR) for the classifiers we run, including SSI classifiers.
- Denmark: Documentary on Danish group suicides leading to proposals from Danish regulators for regulating social media
- Korea: bullying of celebrities in Korea (K-Pop) driving regulatory initiatives, incl. liability for platforms failing to remove content, or compelling data disclosure to victims in civil claims.
- Japan: Proposed 'defamation' and cyber-bullying' regulation following celebrity cyber bullying case connected to a suicide.

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; facing particular scrutiny on this w/r/t eating disorders
- lack different resources and specific actions for users who are using our platform to promote versus admit suicidal ideation
- 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, without e2e resilient detection mechanisms
- Borderline content: borderline SSI policy in place but still waiting on a classifier; labelling exercise required to inform the classifier is on hold due to limited COVID-19 reviewer capacity
- Product infrastructure limitations: inhibit our ability to make necessary guarantees around protection of mental health related data; resourcing and prioritizing Purpose Policy Framework for protection of health information

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.

Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged.

Internal Readiness/State of Play

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...).

	A	B	C	D	E	F
1	Area	**Additional notes**				
2	**Ready; work is complete / ongoing and on track**					
3	SSI industry guidelines	Launched, with a focus on providing tech companies with a				

		menu of options rather than prescriptive guidance				
4	Detection and actioning of Borderline SSI content based on new policy	Ongoing (ETA: H1)				
5	Splitting out our violation types between more narrow categories within Admission and Promotion and action appropriately	Ongoing (ETA: end of Q3)				
6	Prioritization to weigh various safety risks associated with the types of SSI videos.	Ongoing (ETA: Q4) All SSI live videos currently carry same weight for reviewers, need to assess how to streamline more severe ones and how defined.				
7	Addressing harmful content by understanding if a frequency based prevalence metric can measure harm	Ongoing (ETA: H2)				
8	Updating search interstitials and content takedown messaging incorporating more helpful, educational, and empathetic language	Ongoing (ETA: Q4)				
9	**Somewhat ready; work is ongoing but more resources are needed**					
10	Establish processes to look for SSI misses, improve escalation connections and protocols for LIVE, improve our escalation recall, and increase the number of escalations to Law Enforcement	Better understand and respond to content indicating credible intent of suicide; potential constraints due to classifier limitation in the EU				
11	Addressing gaps in resources and coverage for specific types of self-harm, breaking apart eating disorders, cutting, suicide, etc.	Address current resource provision for all forms of SSI centering around suicide; particularly relevant for eating disorders on IG				
12	Address gaps in LIVE SSI escalation protocols	Exploring feature blocking in specific circumstances, removing VOD, better defining movement from passive to active attempt)				
13	Adoption of self-applied warning screens for people demonstrating an intention to share self-harm.	Testing adoption, need clearer policy education for violations without appearing to punish those using platform for admission and recovery purposes				
14	Detection and actioning of Borderline SSI content based on new policy	Ongoing (ETA: Q3)				
15	**Not ready; work has not been prioritised and more resources are needed**					
16	Spaces like IGD used to encourage or promote suicide in the EU	Build classifiers to detect and take action on problematic aggregated content on surfaces like Explore, hashtag pages, Reels				
17	Bullying of celebrities	Revisiting definitions of public				

	driving regulatory initiatives	figures with OCP and making the case for more narrower scoping, but we may not make movement on policy in time and past efforts to narrow what can be said towards public figures have been challenging				

How we might kick start conversations on this work.

Recommendation

To the extent we want to make faster headway on the "areas of improvement" w/r/t SSI and well-being, safety policy needs more HC to support our health policy lead as she currently manages SSI, health and well-being policy. The product teams have grown in these areas and need more policy support.

Well-Being

Policy Landscape

While we have made product adjustments to address concerns in this area (e.g., changes to Newsfeed, hiding likes on IG,...) thought leaders and academics continue to ask questions about the impact of our apps on the social and emotional development and well-being of young people (e.g., recent release of the "Social Dilemma"). At the same, the physical distancing requirements of the COVID-19 pandemic have generated greater 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, managing loneliness, etc.). 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. Of course, 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.

Legislative/Regulatory Action

- US: FTC to update COPPA rules (expected in 2020)
- UK: Women and Equalities Select Committee Changing the Perfect Picture: an inquiry into body image to study impact of advertising and social media on body image
- Denmark: concerns around wellbeing and loneliness have led to calls for regulating content on social media.
- Zimbabwe: Cyber Security and Data Protection Bill
- Mexico: Persistent bullying issues in schools spilling over online, President's son bullied publicly raising questions of censorship on social media.

Immediate Product Vulnerabilities

- Some core features of products connect to larger societal challenges around body image and objectification of women (e.g., AR face altering filters on Instagram, likes, etc.)
- Some core features of products connect to larger societal issues around social comparison and picture perfect lives (e.g, likes and comments)
- Prominence of eating disorder-related content in searches and recommendation surfaces like Explore because of both (a) limited policy definitions around what constitutes ED content and (b) no ED-related classifiers that we can run across our surfaces.
- Some gaps in tools to protect against bullying:
    - IG DM: minimal control re: who can send you a DM request
    - IG DM: no keyword filter tool
    - FB: no mass comment moderation tool

Areas of Improvement

There is a need for stronger research around how our apps and social media broadly impacts the social and emotional development and well-being of young people, particularly w/r/t to issues of social comparison. Until we have that we will be chasing the narrative and those who want to capitalize on the issue. To lean into the issue, not only can we support research efforts, which will take time, in the short term we also can 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, funding CSOM and the connected generations initiative.

- **Research**: support internal and external research around social comparison and loneliness w/r/t social media, and any

potential connections to SSI, bullying, and harassment

- **Product:**
  - Proactive bullying AI comment filters on FB to get parity w/ IG and mass comment moderation tools
  - Need to ensure global mental health efforts are culturally sensitive and align with existing approaches towards mental health;
- **Policies:** Refining public figure definitions and developing more protections from harassment for public figures, including protecting them from attacks on their physical appearance, gender, race, or other protected characteristics; including focusing on U18 public figures as this is an area of vulnerability for IG
- **Resources:** Expanding mental health resource pages that are easily accessible to users and regionally and age applicable

Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged.

Internal Readiness/State of Play

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.

	A	B	C	D	E	F
1	Area	**Additional notes**				
2	**Ready: work is complete / ongoing and on track**					
3	Mental health resources aimed at addressing specific challenges experienced during pandemic	Fuller set of tips and resources covering a range of mental health topics and needs appearing on COVID-19 information center				
4	New mental health resources landing page under development	Ongoing: Pilot in subset of countries to ensure cultural risks are mitigated				
5	Wellness Guides	In-app, invite-only product for mental health organisations to share wellness and mental health tips. Partnerships with IG creators to amplify Guides				
6	**Somewhat ready; work is ongoing but more resources are needed**					
7	Social comparison	Social comparison product work was deprioritised in H2 in favour of election preparation; we need to reinvest in it in H1				
8	Expanding crisis service over messenger	Expansion with one US partner under way, product resources secured; should explore global expansion tied to concurrent Connected Care health work				
9	Creating loneliness focused product experiences	More important than ever to help connect people during this time; research continuing, goal for product ideas Q4				
10	Eating disorders	Build classifiers to detect and take action on eating disorder content specifically on both FB and IG; Commission research into eating disorders and				

		how people engage with this content on Instagram and Facebook (pending final agreement and funding)				
11	Hiding Likes	Users commonly compare validation metrics such as like & follower counts to trigger negative social comparison. While we initially experimented with Project Daisy (hiding Likes), broader rollout on Instagram has been put on hold while Facebook considers bringing Daisy to the Blue app. We need to prioritise rollout on Instagram as soon as possible given the younger demographics.				

How we might kick start conversations on this work.

Recommendation

Given the lack of solid research in this area, making it hard for us to move quickly and thoughtfully on the product side amid continuing pressure from thought leaders and influencers in countries of jurisdiction, we'd recommend a high profile demonstration of our commitment to the issue that also helps us advance strong actionable research such as supporting the launch of the Digital Wellness Lab with the Center for Media and Health at Boston's Children Hospital and Harvard Medical School. A commitment of this kind, alongside efforts like hiding Likes, building ED classifiers for content removal, expanding support resources, should help us lean in to this issue and be seen as acting responsibly.

Racial Justice

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

Additional Quips/Docs

- Child Safety State of Play Interim update - September 2020
- Child Safety - State of Play (7/20)
- H12020 Global Safety, Health and Well-Being Policy Plan
- H12020 Plans (Review)
- Safety Policy Product Asks
- H2 Plans
- Safety Legislative / Regulatory Tracker
- Managing CBI Reviews During COVID-19

Contact Support

Conversation History

Amber [redacted] opened your conversation with Anna [redacted], Vaishnavi J, Liz [redacted], Sarah [redacted], Jennifer [redacted], Sylvia [redacted], Claire [redacted], Antigone Davis, Deepali [redacted], Sebastian [redacted], Chris [redacted] (CI), Holly [redacted], David [redacted], Daniele [redacted], Karuna [redacted], Abe [redacted] and MC [redacted] *Mar 1, 2021 at 3:53 pm*

Karuna [redacted] edited *Feb 26, 2021 at 4:28 pm*
Where Note: This document is out-of-date. It reflects where are we were with the various policy discussions relating to children on our apps as of 9/24.

Sebastian [redacted] opened your conversation with Anna [redacted], Vaishnavi J, Liz [redacted], Sarah [redacted], Jennifer [redacted], Sylvia [redacted], Claire [redacted], Antigone Davis, Deepali [redacted], Chris [redacted] (CI), Holly [redacted], David [redacted], Daniele [redacted], Karuna [redacted], Abe [redacted], MC [redacted] and Amber [redacted] *Dec 9, 2020 at 1:30 pm*

Daniele [redacted] opened your conversation with Anna [redacted], Vaishnavi J, Liz [redacted], Sarah [redacted], Jennifer [redacted], Sylvia [redacted], Claire [redacted], Antigone Davis, Deepali [redacted], Sebastian [redacted], Chris [redacted] (CI), Holly [redacted], David [redacted], Karuna [redacted], Abe [redacted], MC [redacted] and Amber [redacted] *Nov 30, 2020 at 8:23 pm*

Abe [redacted] opened your conversation with Anna [redacted], Vaishnavi J, Liz [redacted], Sarah [redacted], Jennifer [redacted], Sylvia [redacted], Claire [redacted], Antigone Davis, Deepali [redacted], Sebastian [redacted], Chris [redacted] (CI), Holly [redacted], David [redacted], Daniele [redacted], Karuna [redacted], MC [redacted] and Amber [redacted] *Nov 5, 2020 at 5:31 pm*

Antigone Davis added Abe [redacted] to the thread *Nov 5, 2020 at 1:17 pm*

Antigone Davis added you to a document *Nov 5, 2020 at 1:12 pm*

Jennifer [redacted] opened your conversation with Anna [redacted], Vaishnavi J, Liz [redacted], Sarah [redacted], Sylvia [redacted], Claire [redacted], Antigone Davis, Deepali [redacted], Sebastian [redacted], Chris [redacted] (CI), Holly [redacted], David [redacted], Daniele [redacted], Karuna [redacted], Abe [redacted], MC [redacted] and Amber [redacted] *Nov 3, 2020 at 4:51 pm*

Sarah [redacted] opened your conversation with Anna [redacted], Vaishnavi J, Liz [redacted], Jennifer [redacted], Sylvia [redacted], Claire [redacted], Antigone Davis, Deepali [redacted], Sebastian [redacted], Chris [redacted] (CI), Holly [redacted], David [redacted], Daniele [redacted], Karuna [redacted], Abe [redacted], MC [redacted] and Amber [redacted] *Oct 30, 2020 at 5:50 pm*

Karuna [redacted] edited *Oct 30, 2020 at 5:26 pm*

Claire [redacted] opened your conversation with Anna [redacted], Vaishnavi J, Liz [redacted], Sarah [redacted], Jennifer [redacted], Sylvia [redacted], Antigone Davis, Deepali [redacted], Sebastian [redacted], Chris [redacted] (CI), Holly [redacted], David [redacted], Daniele [redacted], Karuna [redacted], Abe [redacted], MC [redacted] and Amber [redacted] *Oct 30, 2020 at 5:18 pm*

Karuna [redacted] added Claire [redacted] to the thread *Oct 30, 2020 at 5:22 pm*

Karuna [redacted] added you to a document *Oct 30, 2020 at 5:17 pm*

Liz [redacted] opened your conversation with Anna [redacted], Vaishnavi J, Sarah [redacted], Jennifer [redacted], Sylvia [redacted], Claire [redacted], Antigone Davis, Deepali [redacted], Sebastian [redacted], Chris [redacted] (CI), Holly [redacted], David [redacted], Daniele [redacted], Karuna [redacted], Abe [redacted], MC [redacted] and Amber [redacted] *Oct 21, 2020 at 2:16 pm*

Vaishnavi J opened your conversation with Anna [redacted], Liz [redacted], Sarah [redacted], Jennifer [redacted], Sylvia [redacted], Claire [redacted], Antigone Davis, Deepali [redacted], Sebastian [redacted], Chris [redacted] (CI), Holly [redacted], David [redacted], Daniele [redacted], Karuna [redacted], Abe [redacted], MC [redacted] and Amber [redacted] *Oct 20, 2020 at 11:11 pm*

Sylvia [redacted] opened your conversation with Anna [redacted], Vaishnavi J, Liz [redacted], Sarah [redacted], Jennifer [redacted], Claire [redacted], Antigone Davis, Deepali [redacted], Sebastian [redacted], Chris [redacted] (CI), Holly [redacted], David [redacted], Daniele [redacted], Karuna [redacted], Abe [redacted], MC [redacted] and Amber [redacted] *Oct 15, 2020 at 3:55 pm*

Deepali [redacted] opened your conversation with Anna [redacted], Vaishnavi J, Liz [redacted], Sarah [redacted], Jennifer [redacted], Sylvia [redacted], Claire [redacted], Antigone Davis, Sebastian [redacted], Chris [redacted] (CI), Holly [redacted], David [redacted], Daniele [redacted], Karuna [redacted], Abe [redacted], MC [redacted] and Amber [redacted] *Oct 14, 2020 at 9:09 am*

Holly [redacted] opened your conversation with Anna [redacted], Vaishnavi J, Liz [redacted], Sarah [redacted], Jennifer [redacted], Sylvia [redacted], Claire [redacted], Antigone Davis, Deepali [redacted], Sebastian [redacted], Chris [redacted] (CI), David [redacted], Daniele [redacted], Karuna [redacted], Abe [redacted], MC [redacted] and Amber [redacted] *Oct 13, 2020 at 2:23 pm*

David [redacted] opened your conversation with Anna [redacted], Vaishnavi J, Liz [redacted], Sarah [redacted], Jennifer [redacted], Sylvia [redacted], Claire [redacted], Antigone Davis, Deepali [redacted], Sebastian [redacted], Chris [redacted] (CI), Holly [redacted], Daniele [redacted], Karuna [redacted], Abe [redacted], MC [redacted] and Amber [redacted] *Oct 13, 2020 at 6:45 am*

Karuna [redacted] added you to a document *Oct 12, 2020 at 5:16 pm*

Karuna [redacted] added you to a document *Oct 12, 2020 at 5:13 pm*

Karuna [redacted] added you to a document *Oct 12, 2020 at 5:13 pm*

Karuna [redacted] added you to a document *Oct 12, 2020 at 5:13 pm*

Karuna [redacted] added you to a document *Oct 12, 2020 at 5:13 pm*

Karuna [redacted] added you to a document *Oct 12, 2020 at 5:13 pm*

Karuna [redacted] added you to a document *Oct 12, 2020 at 5:13 pm*

Karuna [redacted] added you to a document *Oct 12, 2020 at 5:13 pm*

Karuna [redacted] added you to a document *Oct 12, 2020 at 5:13 pm*

Karuna [redacted] added you to a document *Oct 12, 2020 at 5:13 pm*

Karuna [redacted] added Amber [redacted], Daniele [redacted], David [redacted], Deepali [redacted], Holly [redacted], Jennifer [redacted], Liz [redacted] Sarah [redacted] Sebastian [redacted], Sylvia [redacted] and Vaishnavi J to the thread *Oct 12, 2020 at 5:18 pm*

Karuna [redacted] added you to a document *Oct 12, 2020 at 5:13 pm*

Karuna [redacted] enabled link sharing · View *Oct 12, 2020 at 5:17 pm*

MC [redacted] opened your conversation with Anna [redacted] Vaishnavi J, Liz [redacted], Sarah [redacted] Jennifer [redacted], Sylvia [redacted] Claire [redacted], Antigone Davis, Deepali [redacted], Sebastian [redacted] Chris [redacted] (CI), Holly [redacted], David [redacted], Daniele [redacted] Karuna [redacted], Abe [redacted] and Amber [redacted] *Oct 12, 2020 at 4:10 pm*

Chris [redacted] **(CI)** opened your conversation with Anna [redacted] Vaishnavi J, Liz [redacted] Sarah [redacted] Jennifer [redacted] Sylvia [redacted] Claire [redacted] Antigone Davis, Deepali [redacted] Sebastian [redacted] Holly [redacted], David [redacted], Daniele [redacted] [redacted], Karuna [redacted], Abe [redacted], MC [redacted] and Amber [redacted] *Oct 1, 2020 at 4:57 pm*

Antigone Davis added Chris [redacted] (CI) to the thread *Oct 1, 2020 at 5:00 pm*

Antigone Davis added you to a document *Oct 1, 2020 at 4:55 pm*

Karuna [redacted] edited *Sep 28, 2020 at 7:36 pm*

Antigone Davis edited *Sep 28, 2020 at 3:21 pm*

- **Big**
- **Bet**
- **Bets:**
- We also should 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.
-
    - We also should 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.
    - Develop a white paper following on our content regulation white paper that outlines a robust way to successfully and measurably address industry efforts to thwart CSAM
- Product infrastructure limitations: inhibit our ability to make necessary guarantees around protection of mental health related data; resourcing and prioritizing Purpose Policy Framework for protection of health information

There is a need for stronger research around how our apps impact and social media broadly impacts the social and emotional development and well-being of young people, particularly w/r/t to issues of social comparison. Until we have done that we will be chasing the narrative and loneliness those who want to capitalize on our platforms the issue. We have an opportunity to launch x-industry Digital Wellness Lab with Center for Media and Child Health at Harvard to seed independent To lean into the issue, not only can we support research efforts, which will take time, in this space. We the short term we also have an opportunity here can 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, funding 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
- support internal and external research around social comparison and loneliness
    - on Instagram
- w/r/t social media, and any pontential connections to
    - tackle some of the root causes of
- SSI, bullying, and 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 and mass comment moderation tools
-
    - Proactive bullying AI comment filters on FB to get parity w/ IG and mass comment moderation tools
    - Need to ensure global mental health efforts are culturally sensitive and align with existing approaches towards mental health;
- **Policies:** Refining public figure definitions and developing more protections from harassment for public figures, including protecting them from attacks on their physical appearance, gender, race, or other protected characteristics
    - .
- ; including focusing on U18 public figures as this is an area of vulnerability for IG
- **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

Given the lack of solid research in this area, making it hard for us to move quickly and thoughtfully on the product side amid continuing pressure from thought leaders and influencers in countries of jurisdiction, we'd recommend a high profile demonstration of our commitment to the issue that also helps us advance strong actionable research such as supporting the launch of the Digital Wellness Lab with the Center for Media and Health at Boston's Children Hospital and Harvard Medical School. A commitment of this kind, alongside efforts like hiding Likes, building ED classifiers for content removal, expanding support resources, should help us lean in to this issue and be seen as acting responsibly.

Antigone Davis edited *Sep 28, 2020 at 5:13 am*

provide HC for a child safety policy SME who focuses solely on driving this work on the policy side a child safety strategy

push for product shifts that would dramatically decrease the number of cybertips ahead of e2e launch (e.g., forward limits, …

off ramp

consider a large scale investment @$20-40M in the development of robust external ecosystem to thwart CSAM and support victims ("Big Safety Investment")

- **Growing momentum for mandatory child safety by design:** Prompted by UK's ICO Age Appropriate Design Code and Australia's eSafety Commissioner's Safety By Design principles' increasing concern globally among , more and more policymakers that with lack of age verification / detection mechanisms, platforms cannot deliver on around the global are looking at child safety and age appropriate experiences for minors. by design requirements
- **GDPR fueling Increasing number of countries adopting different ages >13 age of digital consent, parental consent and age verification laws:** In Africa, policy makers are increasingly critical of the 13 y/o age limit as children are defined as "under 18" Policymakers in most countries in the region. Countries like India are proposing have proposed legislation that would require requiring parental consent for U18s while those under 18; policymakers in South various African countries exploring the same; and in Korea, which where such legislation already requires exists they are exploring

mechanisms for verifying parental consent when processing children's data, the next step is verifying the consent.

- Germany: Federal and Interstate Youth Law (likely to come into effect in H2 2020; proposes browser-level age verification for German citizens); ~~UK~~ citizens)
- South Korea and Japan: Proposals requiring we verify age through requiring mobile phone numbers
    - Improve age models on IG, we currently only have stated age for ~67% of IG users and age models don't work as well, need IG leadership to consider collecting age for the remaining users for whom we have no stated age.

Overall we remain are quite vulnerable when it comes to age management on our apps. To some degree that is because leadership has decided to live with certain… in the near term. For example in March, we decided 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 as they are best placed to do this well. Identity Integrity identified a PM who will own age verification but we will only prioritize this work when we face clear legal obligations to do so. Still Instead, we will push industry partners for device/OS level age verification and management solutions as they are best placed to do this decision left room for us to continue to work on well. At the same time, leadership did not rule out improving our age classifiers and we are building age appropriate specific experiences, but that work is not advancing with the same speed as external pressures on this subject.

Our overarching recommendation is the same as mentioned above, create an independent child safety vertical, and include this issue as a priority within that vertical to ensure it the proper focus and resources . Again, if that is too big a lift, to deal with the most immediate vulnerabilities w/r/t to age assurance, management and appropriate design policy leadership could:

provide HC for a child safety policy SME who focuses solely on driving a child safety strategy

sign off to conduct an expert audit of our default settings and protections for minors, identify improvements (e.g. GYSJ and PYMK restrictions), and implement them.

ensure product and CO have the resources need 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, etc..

establish as basic company principle that if we're using predicted age for business purposes, we should do the same for enforcing on age/safety

support improving our age classifiers and operationalizing them based on a stack ranking of our most sensitive policy markets

prioritize rapid improvement of IG age models given younger demographic

urge Central Privacy to fast track hiring of a youth privacy PM and to prioritize a few low business cost/high policy impact projects for H2

At that time, we have decided to address certain risks. For example, Central Privacy stood up a small team on youth privacy and hey plan to prioritize a few low business cost/high policy impact projects for H2, which could help on the safety front as well, BUT a PM is yet to be hired.

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.

(e.g., coming soon IG account will be private by default for minors).
We should consider setting child safety up as a separate vertical similar to privacy, with requisite staffing and funding. Doing so could help with weighting v growth (e.g., in-product user education for minors) and weighting v privacy (e.g., age modeling). We might even want to consider a structural change closer to what we have done with health where pull we together the integrity challenges (child safety) and opportunities (youth centered features and apps) work; perhaps this could unlock all kinds of product opportunities (e.g., Messenger Kids, tween apps).

At that time, we have decided to address certain risks. For example, Central Privacy stood up a small team on youth privacy and hey plan to prioritize a few low business cost/high policy impact projects for H2, which could help on the safety front as well, BUT a PM is yet to be hired.
That said, if pushing for this is too big a lift, to deal with the most immediate existential threats outlined above perhaps policy leadership could:
While SSI experts support our existing policies and our commitment to ongoing review of our policies, the need to explain the ever increasing rate… 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 At the same time, some policymakers, most notably the IDPC, 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… a more privacy sensitive approach. This tension begs for a forward leaning approach involving both sets of stakeholder to come to agreed upon norms.

- UK: [redacted]due (pre-inquest review hearing September 25); Online Harms Act; ~~upcoming~~ UK Samaritans guidance for industry released 9/10 recommending a single PoC, clearer policies to prohibit "normalizing" SSI and no encryption* to ensure can scan for this content in messaging ~~(*experts will split on encryption, some will not want us to use it so we can scan private messages, others will want us only to scan in public spaces and to encrypt private messaging so people can safely and securely discuss mental health issues)~~
- Ireland/IDPC/EU: IDPC has continuing privacy concerns around processing of mental health data when running our SSI classifiers, as a result we are…
- e-privacy directive scheduled to take effect in December 2020, OTT communication services have to get explicit individual opt ins (~GDPR) for the classifiers we run. For example, each user will have to agree to allow us to run photoDNA across their photos. If a user doesn't give that consent and we use PhotoDNA on the images they share, that will be considered unlawful. *Strategically, very interesting as this suggests encrypting messaging aligns with expectations including SSI classifiers.
- Denmark: Documentary on Danish group suicides leading to proposals from Danish regulators for regulating social media
- Korea: bullying of celebrities… remove content, or compelling data disclosure to victims in civil claims.
- Japan: Proposed 'defamation' and cyber-bullying' regulation following celebrity cyber bullying case that connected to a suicide.

- 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; facing particular scrutiny on this w/r/t eating disorders
- lack different resources and specific actions for users who are using our platform to promote versus admit suicide suicidal ideation
- lack population/risk specific resources on the SSI checkpoint
- IG:
    - lack classifiers to detect and take action on problematic aggregated content on…
    - IGD being used to encourage or promote suicide e.g., group suicides in Norway that occurred last year and we don't scan there, without e2e resilient detection mechanisms
- Borderline content: borderline SSI policy in place. Labelling but still waiting on a classifier; labelling exercise required to inform the classifier (on is on hold due to limited COVID-19 reviewer capacity) capacity
- Product infrastructure limitations: inhibit our ability to make necessary guarantees around protection of mental health related data

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… 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.

Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged.

Internal Readiness/State of Play

How we might kick start conversations on this work.

Recommendation

To the extent we want to make faster headway on the "areas of improvement" w/r/t SSI and well-being, safety policy needs more HC to support our health policy lead as she currently manages SSI, health and well-being policy. The product teams have grown in these areas and need more policy support.
The While we have made product adjustments to address concerns in this area (e.g., changes to Newsfeed, hiding likes on IG,...) thought leaders and academics continue to ask questions about the impact of our apps on the social and emotional development and well-being of young people (e.g., recent release of the "Social Dilemma"). At the same, 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) connected, managing loneliness, etc.). 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 . Of course, 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.

- Denmark: concerns around wellbeing and loneliness have led to calls for regulating content on social media.
- 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,
- Negative
- Some core features of products connect to larger societal challenges around body image and objectification of women (e.g., AR face altering filters on Instagram, likes, etc.)
- Some core features of products connect to larger societal issues around social comparison and the perception of our products as being picture-perfect views of people's picture perfect lives - amplified by the existence of validating tools like Likes (e.g. likes and Comments comments)
- Prominence of eating disorder-related content in searches and recommendation surfaces like Explore because of both (a) limited policy definitions around what constitutes ED content and (b) no ED-related classifiers that we can run across our surfaces.
- Inability
- Some gaps in tools to
- control who can send you a DM request
- protect against bullying:
- Lack of keyword filters in DMs, similar to the
    - IG DM: minimal control re: who can send you a DM request
    - IG DM: no keyword filter tool
- available for Comments
    -
- FB - Bullying and harassment: Lack of mass comment moderation tools
    - FB: no mass comment moderation tool

There is a need for stronger research around how our apps impact the social and emotional development and well-being of young people, particularly w/r/t to issues of social comparison. Until we have done that we 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… efforts such as building group admin support learning units, expanding WA business API for crisis support orgs, CSOM and the connected generations initiative.

    - ~~**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…~~
    - health research being done in non US/UK markets
- **Facebook:**
    - **Product:** Proactive bullying AI comment filters on FB to get parity w/ IG and mass comment moderation tools

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
Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged.

Internal Readiness/State of Play

How we might kick start conversations on this work.

Also suggest internal audit to ensure that all safety classifiers are globally resilient, globally operational 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 Vanish Mode and prevent use of feature by predicted groomers;
    - User reporting education and message download for Unsent and Vanish 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
    - Messaging Group size limits, forward limits
    - 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

Antigone Davis edited *Sep 28, 2020 at 12:30 am*

- **Safety Mechanisms ineffective without strong mechanisms Growing momentum for detecting age liars mandatory child safety by design:** Prompted by UK's ICO Age Appropriate Design Code and Australia's eSafety Commissioner's Safety By Design principles' increasing concern globally among policymakers that with lack of age verification / detection mechanisms, platforms cannot deliver on safety and age appropriate experiences for minors.

Karuna [redacted] *Sep 27, 2020 at 7:45 pm*

Done

Karuna [redacted] edited *Sep 27, 2020 at 7:46 pm*

- **Safety Mechanisms ineffective without strong mechanisms for detecting age liars:** Prompted by UK's ICO Age Appropriate Design Code and Australia's eSafety Commissioner's Safety By Design principles' increasing concern globally among policymakers that with lack of age verification / detection mechanisms, platforms cannot deliver on safety and age appropriate experiences for minors.
- **GDPR fueling countries adopting different ages of digital consent:** In Africa, policy makers are increasingly critical of the 13 y/o age limit as children are defined as "under 18" in most countries in the region. Countries like India are proposing legislation that would require parental consent for U18s while in South Korea, which already requires parental consent when processing children's data, the next step is verifying the consent.
- Germany: Federal and Interstate Youth Law (likely to come into effect in H2 2020) 2020; proposes browser-level age verification for German citizens); ~~UK~~
- South Korea and Japan: Proposals requiring we verify age through requiring mobile phone numbers

---

Antigone Davis edited *Sep 27, 2020 at 6:52 pm*

- **Legislations Competing legislation exposing safety/privacy tensions**: The ePrivacy Directive v eSafety strategy, and secrecy of communications provisions in Japan and Korea v Japan's Amended Telecommunications Broadcasting Act guidelines on SoC provisions and South and secrecy of communications provisions Korea: v intermediary liability law for NCII and CEI that requires collecting and sharing of more user information with local authorities/civil complainants.
- Rooms: livestreaming abuse
- Ephemerality: negative impact on ability to report
- Interop: increased discoverability and reachability across apps, particularly for IG where we have few child safety protections in place
- NCMEC reporting: losing context spike in volume, poor metric, new safety mitigations have not yet driven down numbers and resistance to measures that likely will

While some governments will not be satisfied with anything short of a backdoor, offering limited client-side scanning will satisfy leading safety stakeholders (NCMEC,… where we can step up our efforts to help manage our vulnerabilities and push back against efforts to require encryption backdoors are listed below (some of these which are underway)

  - improve CEI classifier, use Google safety API to help prioritize and launch CEI soliciation classifier
  - 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 pursues
  - ~~remove private invite links and group chat functionality for high risk CD groups~~

- **Review Capacity**
    - Legal signed off on WFH review, fast tracking discussions with GenPac

Ahead With a strong PM, the pressure 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 and we recently shipped [redacted] for Messenger in [redacted] ~~but product leadership has shown some resistance~~ a fundamental commitment to ~~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 child safety, we are doing good work on this front. ~~We have~~ making real progress ~~on blocking predicted groomers from using Disappearing Mode but need CO support to apply the same protection in IG.~~ Work has completed to block suspected groomers from Secret Conversations and Vanish Mode. On Video chat, product and policy are aligned on blocking suspected groomers (ETA H2) and work is ongoing for feature gating NCMEC reported users, user education and self-remediation interventions. 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, and have agreed to launch more private messaging controls for minors and high volume recipients. ~~including~~ We will continue to explore further restrictions to PYMK and GYSJ. [redacted] 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 across many of the specific related work areas of improvement listed above, BUT we could be doing more faster if we had a dedicated pillar within the company and its status an executive level sponsor.

Overall, we have XFN alignment on the areas of improvement and priorities and with a strong PM a good bit of this work is in process and on a relatively fast track. The biggest challenges here are structural, structure and resourcing and funding.

On the product side child safety sits inside broader integrity efforts, which means priorities are balanced against all harms. As a result, …..). We should consider setting child safety up as a separate vertical similar to privacy, with requisite staffing and funding. Doing Assuming the vertical had an executive level sponsor, doing so also could help with weighting v growth and privacy when making decisions (e.g., [redacted] and weighting v privacy (e.g., age modeling, client side scanning). We might even want to consider a structural change closer to what we have done with health where we pull we together the integrity challenges (child safety) and opportunities (youth centered features and apps) work; perhaps this could unlock all kinds of product opportunities (e.g., Messenger Kids, tween apps).

That said, if pushing for this is too big a lift, perhaps policy leadership could initiate a comparative analysis of how we prioritize child safety versus other top areas of concern like privacy (e.g., headcount, research funding,...) and then use that to have a more robust internal conversation regarding our commitment to child safety; and in the meantime, to deal with the most immediate existential threats outlined above perhaps policy leadership could:

> provide HC for a child safety policy SME who focuses solely on driving this work on the policy side
>
> push for product shifts that would dramatically decrease the number of cybertips ahead of e2e launch (e.g., [redacted]
>
> push for whatever HC and resources are needed by child safety team in CI to help them do more, faster ahead of e2ee; *note: this isn't something I have run by Guy*
>
> push to improve our age models and make them globally operational since all child safety mitigations rely on underlying age models; *note: some of this was discussed as part of age verification discussion, and there was commitment to do some of this but*
>
> help unlock exploration of the most privacy sensitive client-side scanning in case we need it as an off ramp
>
> consider a large scale investment @$20-40M in the development of robust external ecosystem to thwart CSAM and support victims ("Big Safety Investment")

2) push for whatever investment is needed to a) rapidly improve CEI classifier to enable more automation and b) ensure global efficacy of grooming classifier c) improving age models and making them globally operational since all child safety mitigations rely on underlying age models

3) unlock exploration of the most privacy sensitive client-side scanning in case we need it as an off ramp

4) consider a large scale investment @$20-40M in the development of robust external ecosystem to thwart CSAM and support victims ("Big Safety Investment")

- ~~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)~~
- Lack of
- We lack a U13 classifier

- : Insufficient
- , insufficient means to detect and remove u13s at scale
- and on products like Dating
-
- De-prioritized We also have deprioritized u13 enforcement: Basic because COPPA compliance is at risk when we sit on knowledge of u13, incentivizes not knowing
- Age prediction models are not that accurate, reducing the efficacy of classifiers like the grooming classifier and requiring more reliance on human review
- Age prediction models are not globally operational: Creates inconsistent ; as a result, unreliable integrity measures across the company that leverage age are inconsistent and unreliable, including our Child Safety models that provide child safety protections for minors on our platform.
- Global resiliency of age Age prediction models: Tend are not globally resilient, they tend to perform better in Western markets and worse in emerging markets
- IG:Age models do not work as well on IG. As ; as a result, we have no ability to more granularly gate minors out from age-related experiences on Instagram, since our best models still only distinguish between teens and non-teens .
- Grooming detection: Lack of better age models and human review investment leads to inaccurate groomer detections.
- Insufficient Age-gating: Examples include As a result of all of the above, are age-gating doesn't work as well as it should yet we offer age-gated products like Watch Together, which features 14+, 18+ or 21+ content and can't effectively age-gate and FB Dating where .... and we have found minors who appear as young as 12 years old.
- Lack of u18 checkpoint: Risk having u18 users on 18+ products like FB Dating, Marketplace, Mentorship, or as under-age parents on Messenger Kids.
- Age verification: We remain reactive on age verification, pushing for a device side solution, when we're getting increasing policymaker pressure to be proactive while our competitors improve their approaches

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 should fast track that effort. We also need to should 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). Below are specific suggestions:

- **U13 enforcement:**
  - Prioritize cross-enforcement of u13s on IG who have hard-linked FB accounts.
  - Clear current u13 backlog (~700K job backlog) which is currently not resourced.
- **U18 enforcement:**
  - Create an u18 checkpoint to effectively manage gating u18s from 18+ products
- **IG:**
  - Improve age models on IG
- **Age verification:**
  - Use age prediction in combination with age verification checkpoints to gate based on age-gated product or content.
- **Age Model improvements:**
  - Expand best available age models to more markets
  - Improve age model accuracy in emerging markets
  - This would also improve our groomer classifiers
  - Improve age models on IG
- **Age Audit:**
  - Review our default settings and gating for minors, pressure test them identify improvements (e.g. GYSJ and PYMK…
  - .g., unique user education at sign up)
- **Age appropriate experiences:**
  - Develop 13-16 yo, 16-18 yo and 18+ experiences

In Overall we remain quite vulnerable when it comes to age management on our apps. To some degree that is because leadership has decided to live with certain risks at least in the near term. For example in March it was , we 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 as they are best placed to be hired. They have offered to prioritize a few low business cost/high policy impact projects for H2 do this well. Identity Integrity has also identified a PM who would will own age verification but they we will not only prioritize this work until when we have face clear legal obligations to do so. We need leadership support Still, this decision left room for us to push product continue to act work on knowledge we already have improving our age classifiers and age appropriate experiences, 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 but

that work is not consistent across all apps for a single known person advancing with the same speed as external pressures on this subject. 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.

At that time, we have decided to address certain risks. For example, Central Privacy stood up a small team on youth privacy and hey plan to prioritize a few low business cost/high policy impact projects for H2, which could help on the safety front as well, BUT a PM is yet to be hired.

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.

(e.g., coming soon IG account will be private by default for minors).

At that time, we have decided to address certain risks. For example, Central Privacy stood up a small team on youth privacy and hey plan to prioritize a few low business cost/high policy impact projects for H2, which could help on the safety front as well, BUT a PM is yet to be hired.

Karuna [redacted] *Sep 25, 2020 at 4:07 am*

Done!

Karuna [redacted] *Sep 25, 2020 at 3:55 am*

Note: I haven't had a chance to edit down the text in this table yet

---

Karuna [redacted] *Sep 25, 2020 at 3:48 am*

Also, she has ED here and under well-being. Suggest keeping there instead

---

Karuna [redacted] *Sep 25, 2020 at 3:44 am*

From VJ: **In the UK, the Women and Equalities Select Committee has asked us to make submissions on body image to study impact of advertising and social media on body image. In Denmark, concerns around wellbeing and loneliness have led to calls for regulating content on social media. And as we continue to build popular AR face altering filters into Instagram, which continues to be the platform for young people, scrutiny over body image and social comparison grows.**

Karuna [redacted] *Sep 25, 2020 at 3:39 am*

Note: VJ had the following that I haven't included: **On issues like eating disorders, given the lack of global hotlines, we could make a big commitment to the space by offering funding for the establishment of global hotlines to tackle EDs.**

Karuna [redacted] *Sep 25, 2020 at 3:36 am*

Note: VJ had the following however it is my understanding this is our product decision and not a gap: ** On private messaging surfaces - IG Direct, Messenger, we do not have ways to proactively detect suicidal or self-harming conversations and we need e2ee-resilient product solutions to address

*

Karuna [redacted] *Sep 25, 2020 at 3:33 am*

Note: VJ had the following in case needed: **We have limited resources around eating disorders to share and limited tailored product support for people looking for this content**

Karuna [redacted] *Sep 25, 2020 at 3:32 am*

From Sarah: correct but to the extent youth privacy overlaps with our issues, i.e. more private controls, TTC labs/safety by design approaches to product design, etc., we should collaborate with them since they're product and they have actual PM's and eng resources

---

Karuna [redacted] *Sep 25, 2020 at 3:22 am*

From Sarah: Group links have been deprecated but we should make clear that these should stay deprecated, and add a trigger to notify MSGR wellbeing if it does get added again

---

Karuna [redacted] *Sep 25, 2020 at 2:50 am*

Additional info from VJ and SC not included here: We have spoken publicly about this existing MSGR protection (including in Jay [redacted]'s Senate Judiciary Committee testimony and QFR's) and should strive to bring IG up to par.

Karuna [redacted] *Sep 25, 2020 at 2:36 am*

Additional infor from Sarah if needed: For example, User A shares a link with User B (depth 1), User B then shares with User C (depth 2), etc. and we would limit subsequent shares after depth X and/or consider removing the share button altogether after X number of depth shares. Data analysis has shown that aThe current proposal for a Forward Limit of 5 is the first step in bringing Messenger closer to WhatsApp parity for forwarding. The second step would be to introduce a depth limit that could significantly reduce the sharing of CSAM and other violating content. Data analysis has also 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. We are chairing a working group to gather data analysis and product experiments to shape a product proposal to discuss with Jay [redacted] and 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. More data science resources are needed to dedicate time to analyze the impact of these interventions on harmful content shares, views, and the formation of CEI dedicated communities and networked behavior.

Karuna [redacted] *Sep 25, 2020 at 2:18 am*

From Sarah: need to demonstrate ability to reduce NCMEC reports before e2ee

---

Karuna [redacted] *Sep 25, 2020 at 2:17 am*

Suggestion from Sarah "+ lack of child protection mitigations for IG"

Karuna [redacted] *Sep 25, 2020 at 1:24 am*

Additional comments from Sarah (not included since covered above): **In general, we should prioritize improving our CEI image and video classifiers whose performance is not accurate enough to auto-action content. In addition, some of the most visible child safety harm on our platform occurs via the sexualization of minors and the solicitation of CEI via benign images of children or in minor sexualizing groups, where bad actors congregate and connect to then transition over to private messaging and share CEI. While our minor sexualization policies will be completed this half, our AI tooling is not mature enough to proactively detect this content at scale which often occurs in a highly contextualized manner. There are also significant enforcement gaps revealed by ongoing data science work to contextualize our NCMEC report numbers. For example, we currently neither detect nor enforce against CEI recipients. However, recent analysis has shown that there are ways to detect malicious actors amongst those who have received CEI as a result of spamming/harassing behavior. Work is planned to refine the way we currently disable, gate or feature block NCMEC reported users from products using this data analysis.**

Karuna [redacted] *Sep 25, 2020 at 1:22 am*

Sarah also had the following but since this info is captured here I haven't included it: further understand abuse prevalence and how to best protect minors from potential harm exposure, and build all the tooling necessary to allow our investigators to send NCMEC any necessary manual reports

---

Karuna [redacted] *Sep 24, 2020 at 10:53 pm*

Quick flag that Sarah has this detailed in her version but I am not porting as it seems too much: **On group size limits and other product interventions for large group message threads, we are chairing a working group to gather data analysis and product experiments to shape a product proposal to discuss with Jay** [redacted] **to discuss how best to present group thread size limits 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.**

Karuna [redacted] *Sep 24, 2020 at 8:29 pm*

Antigone Sarah recommended striking this and added in the **Product has not committed to preventing unconnected FB adults from initiating messages with IG minors this half -- mainly because they do not want to risk creating a high number of false positives during a critical launch moment -- but there is data analysis and impact estimations ongoing to determine whether this is even an impactful intervention to reduce grooming on IG. In parallel, a taskforce is currently investigating how to reduce prevalence of grooming on IG overall. One of the mitigations agreed upon by the taskforce to ship this half is to prevent message requests from suspected IG groomers to IG minors.**

Antigone Davis *Sep 24, 2020 at 8:24 pm*

Karuna, this section feels really weak compared to the others. I wonder if there is a lot we can pull from Vaishnavi's IG safety policy assessment note

Antigone Davis edited *Sep 24, 2020 at 8:23 pm*

Karuna [redacted] *Sep 24, 2020 at 8:21 pm*

Added this Antigone - in red for your sign off

Karuna [redacted] edited *Sep 25, 2020 at 4:07 am*

- **Legislations exposing safety/privacy tensions**: The ePrivacy Directive v eSafety strategy, and secrecy of communications provisions in Japan and Korea deem scanning v Japan's Amended Telecommunications Broadcasting Act guidelines on SoC provisions and South Korea: intermediary liability law for harmful content an illegal intercept at same time.. NCII and CEI that requires collecting and sharing of more user information with local authorities/civil complainants.
- **Ephemerality:** adult<>minor carveout for disappearing mode Vanish Mode and/or some way to enable content reporting and review

Ahead of interop, work is in progress to block unconnected adults on ~~FB and~~ IG from initiating messages with minors on FB, however… initiating messages to IG minors. Limited country tests and data analyses have shown that [redacted] can demonstrably reduce CEI sharing and we recently shipped [redacted] ~~but product leadership has shown some resistance to these mitigations. We are currently escalating the forward limits conversation up through Monika to meet with…. We have progress on blocking predicted groomers from using Disappearing Mode but need CO support to apply the same protection in IG.~~ Work has completed to block suspected groomers from Secret Conversations and Vanish Mode. On Video chat, product and policy are aligned on blocking suspected groomers (ETA H2) and work is ongoing for feature gating NCMEC reported users, user education and self-remediation interventions. 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 RFC (aka Video Chat) Chat). ~~RFC, which is the predecessor to Rooms, is not yet in parity with Rooms (i.e. feature gating high-risk users… 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, and have agreed to launch more private messaging controls for minors and high volume recipients. ~~including~~ We will continue to explore further restrictions to PYMK and GYSJ. Work to develop a CEI video classifier for FB Live is currently ongoing but we need more resources… resources to develop an operational sex trafficking classifier. Please see the chart below for detailed summary of the specific related work and its status.

2) push for ~~whatever~~ investment ~~is needed~~ to a) rapidly improve CEI classifier to enable more automation and b) ensure global efficacy of grooming classifier c) improving age models and making them globally operational since all child safety mitigations rely on underlying age models

- ~~Lack of U13 proactive age identification and enforcement mechanism~~
- ~~Need for improved age identification models that allow us to better predict age of users…g. restricting unconnected minors and adults from interacting, full rollout of Safety Warnings (currently only planned for those users whose age we know)~~
- Lack of U13 classifier: Insufficient means to detect u13s at scale and on products like Dating
- De-prioritized u13 enforcement: Basic COPPA compliance is at risk when we sit on knowledge of u13
- Lack of u18 checkpoint: Risk having u18 users on 18+ products like FB Dating, Marketplace, Mentorship, or as under-age parents on Messenger Kids.
- Age prediction models not globally operational: Creates inconsistent, unreliable integrity measures across the company that leverage age, including our Child Safety models that provide protections for minors on our platform
- Global resiliency of age prediction models: Tend to perform better in Western markets and worse in emerging markets
- Grooming detection: Lack of better age models and human review investment leads to inaccurate groomer detections.
- IG: Age models do not work as well on IG. As a result, no ability to more granularly gate minors out from age-related experiences on Instagram, since our best models still only distinguish between teens and non-teens
- Insufficient Age-gating: Examples include Watch Together, which features 14+, 18+ or 21+ content and can't effectively age-gate and FB Dating where we have found minors who appear as young as 12 years old
- Age verification: We remain reactive on age verification when we're getting increasing policymaker pressure to be proactive while our competitors improve their approaches
- Review our default settings and gating for minors, pressure test them identify improvements (e.g. GYSJ and PYMK

restrictions), and implement them

- Develop a "safety
- **U13 enforcement:**
    - Prioritize cross-enforcement of u13s on IG who have hard-linked FB accounts.
    - Clear current u13 backlog (~700K job backlog) which is currently not resourced.
- **U18 enforcement:**
    - Create an u18 checkpoint to effectively manage gating u18s from 18+ products
- **IG:**
    - Improve age models on IG
- **Age verification:**
    - Use age prediction in combination with age verification checkpoints to gate based on age-gated product or content.
- **Age Model improvements:**
    - Expand best available age models to more markets
    - Improve age model accuracy in emerging markets
    - This would also improve our groomer classifiers
- **Age Audit:**
    - Review our default settings and gating for minors, pressure test them identify improvements (e.g. GYSJ and PYMK restrictions), and implement them
- **Safety by**
- design" process or product safety review similar to privacy review
- **Design:**
- In product U18 specific user education on safety and privacy tools (e.g., unique user education at sign up)
    - Develop a "safety by design" process or product safety review similar to privacy review
- **Education:**
- Develop 13-16 yo, 16-18 yo and 18+
    - In product U18 specific user education on safety and privacy tools (e.g., unique user education at sign up)
- **Age appropriate experiences:**
    - Develop 13-16 yo, 16-18 yo and 18+ experiences

- UK: [redacted] due H2 2020 (may be delayed due to COVID-19) (pre-inquest review hearing September 25); Online Harms Act; ~~upcoming~~ UK Samaritans guidance for industry released 9/10 ~~recommending a single PoC, clearer policies to prohibit "normalizing" SSI and no encryption" to ensure can scan for this content in messaging ("experts will…~~
- on the images they share, that will be considered unlawful. *Strategically, very interesting as this suggests encrypting messaging aligns with expectations.
- Denmark: Documentary on Danish group suicides leading to proposals from Danish regulators for regulating social media

- Borderline content: borderline SSI policy in place. Labelling exercise required to inform the classifier (on hold due to limited COVID-19 reviewer capacity)

- Prominence of eating disorder-related content in searches and recommendation surfaces like Explore because of both (a) limited policy definitions around what constitutes ED content and (b) no ED-related classifiers that we can run across our surfaces.
- Negative social comparison and the perception of our products as being picture-perfect views of people's lives - amplified by the existence of validating tools like Likes and Comments
- Inability to control who can send you a DM request
- Lack of keyword filters in DMs, similar to the keyword filter tool available for Comments

    - **Policies:** Refining public figure definitions and developing more protections from harassment for public figures, including protecting them from attacks on their physical appearance, gender, race, or other protected characteristics.

    - Adult<>minor carveout for Disappearing Vanish Mode and prevent use of feature by predicted groomers;
    - User reporting education and message download for Unsent and Disappearing Vanish Mode messages with tombstones:

Antigone Davis *Sep 24, 2020 at 8:01 pm*

Karuna, can you check in with Sarah on this. Won't this team focus on data protection and privacy of minors and not on underage accounts or age appropriate experiences?

Antigone Davis *Sep 24, 2020 at 7:48 pm*

Karuna, can you ask Sarah to identify specific product vulnerabilities resulting from our lack of age identification (e.g., dating app issues, % U13 accounts on IG, efficacy of efforts to avoid recommending minors to unconnected adults via PYMK,...). This section should look more like the one above with specific product vulnerabilities.

Antigone Davis *Sep 24, 2020 at 7:31 pm*

Karuna, can you add a trends analysis like you did above? For example, seeing legislation around safety by design, age appropriate content restrictions with liability, liability protection carves outs for minors put in harms way (SSI, solicitation/trafficking,...). OR whatever is accurate in terms of trends. Can even say regulatory activity not at level or pitch of exceptional access discussions but slow and steady succesful efforts...

Anna [redacted] opened your conversation with Vaishnavi J, Liz [redacted], Sarah [redacted], Jennifer [redacted], Sylvia [redacted], Claire [redacted] Antigone Davis, Deepali [redacted], Sebastian [redacted], Chris [redacted] (CI), Holly [redacted], David [redacted], Daniele [redacted] Karuna [redacted], Abe [redacted], MC [redacted] and Amber [redacted] *Sep 24, 2020 at 6:43 pm*

Anna [redacted] edited *Sep 24, 2020 at 6:43 pm*

Antigone Davis *Sep 24, 2020 at 5:48 pm*

Karuna, this chart does not tightly track to suggested areas of improvement; not a huge deal but there are some big recommendations, like improving our CEI classifier, that aren't mentioned; can you ask Sarah to clean that up a bit. For example, the chart doesn't include anything on the Google API to deal with CEI classifier weakness, nor does it mention legal unlocking review WFH review to deal with COVID review challenges.

Antigone Davis *Sep 24, 2020 at 5:44 pm*

Thanks

Antigone Davis *Sep 24, 2020 at 5:40 pm*

Combining :)

Antigone Davis edited *Sep 24, 2020 at 8:00 pm*

- **Livestream, real time calls and other remote presence products:** develop content independent CEI CSAM detection tech, wiretap capability, reportability and in-call reporting/reportability with evidence capture to ensure reviewability upon report, user education/self-remediation interventions before, during, or after (e.g., safety notices, upsell reporting/blocking)
- **RTC and Remote Presence products:** in-call reporting with evidence capture, ~~feature gate high-risk users~~, user education/self-remediation interventions before, during, or after calls (i.e. safety notices, upsell reporting/blocking)

- **Big Bet:** 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… case management tool that aggregates all case data for investigations and helps with prioritization, and scaleable automated system for x-industry threat information sharing.

Ahead of Interop interop, work is in progress to block unconnected adults on FB and IG from initiating messages with minors on FB, however work is not… resources to develop an operational sex trafficking classifier. Please see the chart below for detailed summary of the specific related work and its status.
Overall, we have XFN alignment on the areas of improvement and priorities and a good bit of this work is in process. The biggest challenges here are org structure structural, resourcing and funding.
1) Child Safety On the product side child safety sits inside broader integrity efforts, which means priorities are balanced against all harms. As a result, we sometimes have to deprioritize child safety when certain issues take the foreground precedence (e.g., US Elections, racial injustice,...). We should consider setting it child safety up as a separate vertical similar to privacy, with requisite staffing and funding. This Doing so also could help with weighting v growth (e.g., forward limits, group size limits and weighting v privacy (e.g., age modeling, client side scanning). We might also help even want to consider a structural change closer to what we have done with health where pull we together the integrity challenges (child safety) and opportunities (youth centered features and apps) work; perhaps this could unlock all kinds of product opportunities (e.g., Messenger Kids, tween apps).

That said, if pushing for this is too big a lift, to deal with the most immediate existential threats outlined above perhaps policy leadership could:

1) help push for product shifts that would dramatically decrease the number of cybertips ahead of e2e launch (e.g., forward limits, group size limits)
2) push for whatever investment is needed to a) rapidly improve CEI classifier to enable more automation and b) ensure global efficacy of grooming classifier
3) unlock exploration of the most privacy sensitive client-side scanning in case we need it as an off ramp
4) consider a large scale investment @$20-40M in the development of robust external ecosystem to thwart CSAM and support victims ("Big Safety Investment")

Where are we with the various policy discussions relating to children on our apps.
Trends:

Specific Bills:
Where we think we are most vulnerable and need to step up our efforts.

- Looking back: work with a trusted expert/group of experts to review Review our default settings, protections,... and gating for minors, pressure test them 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
    - 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)
- In product U18 specific user education on safety and privacy tools (e.g., unique user education at sign up)
- Develop 13-16 yo, 16-18 yo and 18+ experiences

Extent to which there is work ongoing to address the 'areas for improvement' and opportunities flagged.

Internal Readiness/State of Play

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.

Internal readiness to rise to the challenge ***How we might kick start conversations on this work.***
We should consider setting child safety up as a separate vertical similar to privacy, with requisite staffing and funding. Doing so could help with weighting v growth (e.g., in-product user education for minors) and weighting v privacy (e.g., age modeling). We might even want to consider a structural change closer to what we have done with health where pull we together the integrity challenges (child safety) and opportunities (youth centered features and apps) work; perhaps this could unlock all kinds of product opportunities (e.g., Messenger Kids, tween apps).

That said, if pushing for this is too big a lift, to deal with the most immediate existential threats outlined above perhaps policy leadership could:

Karuna [redacted] edited *Sep 24, 2020 at 5:00 pm*

- **RTC and Remote Presence products:** in-call reporting with evidence capture, ~~feature gate high-risk users~~, user education/self-remediation interventions before, during, or after calls (i.e. safety notices, upsell reporting/blocking)

Karuna [redacted] *Sep 24, 2020 at 4:59 pm*

Sarah's take is they should be separate - she did have an update here which i am making in red

---

Antigone Davis *Sep 24, 2020 at 4:14 pm*

Karuna, this chart does not tightly track to suggested areas of improvement; not a huge deal but there are some big recommendations, like improving our CEI classifier, that aren't mentioned; can you ask Sarah to clean that up a bit.

Karuna [redacted] *Sep 24, 2020 at 4:12 pm*

Vaishnavi says this is in progress - currently working on preventing unconnected adults with high likelihood of being an IIC violator from finding and messaging minors

---

Antigone Davis *Sep 24, 2020 at 2:52 pm*

Karuna, what is the contradictory legislation in these countries

---

Antigone Davis *Sep 24, 2020 at 2:41 pm*

Karuna, can you ask Sarah if this is duplicative of the bullet above for livestreaming; can these be combined

---

Antigone Davis *Sep 24, 2020 at 2:24 pm*

Karuna, can you check with Vaishnavi whether this has been fixed?

Antigone Davis *Sep 24, 2020 at 1:57 pm*

Karuna, and then would add trends for age gating, verification, ... if any

---

Antigone Davis *Sep 24, 2020 at 1:55 pm*

Karuna, based on inputs could we add a paragraph here on this sets of issues?

Antigone Davis edited *Sep 24, 2020 at 5:00 pm*

Previous edition: Child Safety - State of Play (7/20)

Where are we with the various policy discussions relating to children on our apps.
Our encryption announcement has created tremendous anxiety among CSAM safety stakeholders, and generated significant legislative activity across the globe centered around exceptional access and… 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).

- **Child Safety concerns leading to expansion/conflation of illegal and harmful content regulation.** Following UK's Online Harms Paper and Germany NetzDG, AUS Australia has proposed expanding the eSafety Commissioner's remit to enforce on expanded harms (e.g bullying & harassment) not only for children but also… to push the Citizen Online Protection Rules to curb free speech.
- **Dedicated child safety regulators being proposed.** Australia and Fiji already have an esafety commissioner eSafety Commissioner. Legislators in Ireland, Taiwan, the Philippines, and South Korea are considering similar offices and India has proposed an eSafety commissioner Commissioner for each state.
- **Dual Multiple reporting requirements:** India, South Korea, and Portugal are introducing legislations that will require us report CSAM directly to local authorities and not accepting NCMEC reports.
- **Legislations exposing safety/privacy tensions**: The ePrivacy Directive v eSafety strategy, and secrecy of communications provisions in Japan and Korea deem scanning for harmful content an illegal intercept at same time...

Specific Bills:

- US: EARN-IT, ~~Lawful Access to Encrypted Data Act~~, DoJ/White House Section 230 reforms, Platform Accountability and Consumer Transparency Act, Online Freedom and Viewpoint Act, House Appropriations Bill on Labor, HHS, and Education funding, Transparency about Online Child Sexual Abuse Material Act
- UK: Cloud Act negotiations, TCN, Online Harms Bill
- AU: ongoing consultation on Online Safety Act, e-Safety Commissioner Safety by Design Principles
- Five-Eyes: Voluntary CSEA Principles
- India: traceability, amendments to "Protection of Children from Sexual Offences Offenses Rules, 2020" requiring direct reporting w/liability
- Brazil: proposed amendments to "good samaritan provision" in Marco Civil
- EU: EUIF ~~to explore recommended best practices including~~ technical requirements and best practices, ~~EC to set up~~ EU CSAM hub ~~(note: much in direct contradiction~~ to ePrivacy Directive), Audio Visual Media Services Directive, Digital Services Act, CSAM Legislation
- Philippines: Anti-ATIP bill; (intermediary (intermediary liability, duty not to host CEI, mandates turn-around-times, requires platforms to preserve evidence and proactively report CEI directly to local LE. Safer Internet Bill ~~(parental~~ (parental consent ~~from~~ for U18s age of consent , TATs, 'harmful' content regulations) regulations); Anti-ATIP bill (intermediary liability, mandated TAT, evidence preservation requirements and proactively report CEI directly to local LE)
- South Korea: intermediary…
- ; Amended Telecommunications Broadcasting Act guidelines on SoC provisions; Provider Limited Liability Act (provision of user information to civil complainants, incl phone numbers)
- Germany: NetzDG
- Ireland: Online Safety and Media Bill
- Kenya: Cybercrime Bill 2018
- Zimbabwe: Data protection and Cybercrimes Bill 2020
- Other: ~~Germany, Ireland,~~ S.Africa, Japan, and Taiwan also have legislative activity

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.

- Inability to auto-action using our proactiveCEI classifiers: low accuracy prevents auto actioning
- Grooming classifiers: lack global efficacy
- Rooms: livestreaming abuse
- Ephemerality: negative impact on ability to report
- Interop: increased discoverability and reachability across apps
- NCMEC reporting: missing losing context, poor metric

- WhatsApp groups: link sharing on social platforms + WA group discovery apps on iOS and Android in app store enable CSAM discovery and sharing
- COVID review constraints
- Unconnected IG: unconnected adults being able to find and message minors in Instagram Direct, coupled with significant implicit sexualization of minors and evidence of potential trafficking and solicitation
- Implicit sexualisation of minors across multiple posts and comments on Instagram
- Sex trafficking and sexual solicitation networks on Instagram

While some governments will not be satisfied with anything short of a backdoor, offering limited client-side scanning will satisfy leading safety stakeholders (NCMEC,… 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. (some of these are underway)

- **NCMEC Reports CSAM:**
    - contextualize these NCMEC numbers, stand up better metrics
    - close enforcement gap w/r/t CEI recipients
    - meaningfully drive down the number of reports before encryption through preventative work including forward limitations, group…
    - 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 pursues
- **Minor Sexualization:**
- much stricter policies and enforcement w/r/t the sexualization of minors on our services (e.g., group names, comments,...) and immature minor sexualization classifier. *Update: Minor Sexualization Policy launching in October*
-
    - launch stricter policies and enforcement w/r/t the sexualization of minors on our services
    - develop specific stricter policies for groups (e.g., group names, comments,...)
    - improve minor sexualization classifier.
- **Trafficking:** build sex trafficking classifier with a focus on minor sex trafficking
- **TC Project Protect:** fast track some key wins like universal video hash sharing…
- mode and/or some way to enable content reporting and review
- **Classifier global resilience and efficiency:** comprehensive audit to assess and ensure child safety classifiers do not operate with bias are effective globally. Improve global efficacy of our text classifiers that detect inappropriate interactions with children by developing text model to , which would help enable automation of hard actions (disable/report) against groomers.
- **IG:**
    - Extend existing MSGR Safety tips for scams, impersonation, child safety/IIC in IGD;
    - Expand policies on borderline content… hashtag pages, Reels,
    - better age-related logic for preventing conversations between unconnected adults and minors
    - Private by Default for Teens
- **Reporting:**
- Launch "Involves a Child" feedback tag where missing e.g., in FB Groups and IG comments.
-
    - Launch "Involves a Child" feedback tag where missing e.g., in FB Groups and IG comments.
    - Build access to emergency support into reporting
- **Groups:**
    - remove private invite links stricter policies w/r/t naming, about,... and group chat functionality for high risk FB groups admin responsibilities
    - 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 in-product user education on safety and privacy tools, as well as global program
    - Well-funded global digital citizenship and wellbeing program similar to Google's Be Internet Awesome
    - Safety education video for minors at sign-up or early in account and potentially to unlock certain features
- **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)

- **Big Bet:** 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.

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.

Recommendation Internal Readiness/State of Play

Ahead of Interop, work is in progress to block unconnected adults on FB and IG from initiating messages with minors on FB, however… 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 shown some resistance to 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… resources to develop an operational sex trafficking classifier. Please see the chart below for detailed summary of the specific related work and its status.

Suggestions on how we might kick start conversations on this work.

Recommendation

Overall, we have XFN alignment on the areas of improvement and priorities and a good bit of this work is in process. The biggest challenges here are org structure, resourcing and funding.

1) Child Safety sits inside broader integrity efforts, which means priorities are balanced against all harms. As a result, we sometimes have to deprioritize child safety when certain issues take the foreground (e.g., US Elections, racial injustice,...). We should consider setting it up as a separate vertical similar to privacy. This might also help

Age Assurance Assurance/Appropriateness

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 as well as significant negative stakeholder feedback and PR. With a few key product changes and/or , regulatory collaboration and/or investment in research 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, and reduce bad press. Priority Countries: countries of jurisdiction (US , US and Ireland) Ireland, and countries that are highly vocal and influential in region AU, UK, Brazil, South Africa and India.

- US: FTC to update COPPA rules (expected in 2020)
- UK: Age Appropriate Design Code (likely implemented in H1 2021 subject to Parliamentary…
- ~~UK~~
- South Africa: Protection of Personal Information Act 4 of 2013 (July 1, 2021)
- Philippines: Safer Internet Bill (parental consent ~~from~~ for U18s age of consent))
- India: Personal Data Protection Bill, 2019 (requires age verification and parental consent)
- SSA: ~~Data Protection Act and Cyber Crimes act in Kenya, South Africa's POPI act and~~ Zimbabwe's Cyber crimes bill (parental consent and robust age verification), South Africa Films and classification amendment act 2020 (registration, self classification, age…
- and Philippines: Draft bills w/r/t age-verification to ensure minors not exposed to "inappropriate/immoral content"
- Colombia/Peru/Paraguay/CENAM: routine bills on mandating blocking content for kids (videogames (video games, social media, educational, etc)

- Child Safety State of Play Interim update - September 2020
- Child Safety - State of Play (7/20)
- H12020 Global Safety, Health and Well-Being Policy Plan
- H12020 Plans (Review)
- Safety Policy Product Asks

- H2 Plans
- Safety Legislative / Regulatory Tracker
- Managing CEI Reviews During COVID-19

Antigone Davis opened your conversation with Anna [redacted], Vaishnavi J, Liz [redacted], Sarah [redacted], Jennifer [redacted], Sylvia [redacted], Claire [redacted], Deepali [redacted], Sebastian [redacted], Chris [redacted] (CI), Holly [redacted], David [redacted], Daniele [redacted], Karuna [redacted], Abe [redacted], MC [redacted] and Amber [redacted] *Sep 24, 2020 at 10:54 am*

Karuna [redacted] enabled link sharing · View/Edit *Sep 24, 2020 at 5:46 am*

Karuna [redacted] added Antigone Davis to the thread *Sep 24, 2020 at 5:46 am*

Karuna [redacted] added you to a document *Sep 24, 2020 at 5:40 am*

Karuna [redacted] edited *Sep 24, 2020 at 9:10 am*

Child Safety State of Play (9/24)

Rough Inputs: Child Safety State of Play 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.

Previous edition: Child Safety - State of Play (7/20)

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).

Trends:

- **Child Safety concerns leading to expansion/conflation of illegal and harmful content regulation.** Following UK's Online Harms Paper and Germany NetzDG, AUS has proposed expanding the eSafety Commissioner's remit to enforce on expanded harms (e.g bullying & harassment) not only for children but also adults and is also extending the scope of harms from illegal to harmful content. Countries such as Brazil, the Philippines and Indonesia are all considering misinformation and child safety in the same category in recent legislations. In Pakistan child safety issues are being used to push the Citizen Online Protection Rules to curb free speech.
- **Dedicated child safety regulators being proposed.** Australia and Fiji already have an esafety commissioner. Legislators in Ireland, Taiwan, the Philippines, South Korea are considering similar offices and India has proposed an eSafety commissioner for each state.
- **Dual reporting requirements:** India, South Korea, Portugal are introducing legislations that will require us report CSAM directly to local authorities and not accepting NCMEC reports.
- **Legislations exposing safety/privacy tensions**: The ePrivacy Directive and secrecy of communications provisions in Japan and Korea deem scanning for harmful content an illegal intercept.

Legislative/Regulatory Action

- US: EARN-IT, ~~Lawful Access to Encrypted Data Act~~; DoJ/White House Section 230 reforms, Platform Accountability and Consumer Transparency Act, Online Freedom and Viewpoint Act, House Appropriations Bill on Labor, HHS, and Education funding, Transparency about Online Child Sexual Abuse Material Act
- UK: Cloud Act negotiations, TCN, Online Harms Bill
- AU: ongoing consultation on Online Safety Act, e-Safety Commissioner Safety by Design Principles
- Five-Eyes: Voluntary CSEA Principles
- India: traceability, amendments to "Protection of Children from Sexual Offences Rules, 2020" requiring direct reporting w/liability
- Brazil: proposed amendments to "good samaritan provision" in Marco Civil
- EU: EUIF ~~to explore recommended best practices including~~ technical requirements best practices, ~~EC to set up~~ EU CSAM hub ~~(note: much in direct contradiction to~~ ePrivacy Directive~~)~~, Audio Visual Media Services Directive, Digital Services Act, CSAM Legislation
- Philippines: Anti-ATIP bill; (intermediary liability, duty not to host CEI, mandates turn-around-times, requires platforms to preserve evidence and proactively report CEI directly to local LE. Safer Internet Bill (~~parental consent from U18s~~ age of consent, TATs, 'harmful' content regulations); Anti-ATIP bill (intermediary liability, mandated TAT, evidence preservation requirements and proactively report CEI directly to local LE)
- South Korea: intermediary liability law for NCII and CEI
- Japan: National Police's annual report on Child Sexual Exploitation and Sexual Solicitation of Minors; Amended Telecommunications Broadcasting Act guidelines on SoC provisions; Provider Limited Liability Act (provision of user information to civil complainants, incl phone numbers)
- Belgium: NCII legislation 6 hr takedown or liability
- Germany: NetzDG
- Ireland: Online Safety and Media Bill
- Kenya: Cybercrime Bill 2018
- Zimbabwe: Data protection and Cybercrimes Bill 2020
- Other: ~~Germany; Ireland~~; S.Africa, Japan, and Taiwan also have legislative activity

Immediate Product Vulnerabilities

- Inability to auto-action using our proactive CEI classifiers
- 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
  - close enforcement gap w/r/t CEI recipients
  - meaningfully drive down the number of reports before encryption through preventative work including forward limitations, group size limitations, 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,...) and immature minor sexualization classifier. *Update: Minor Sexualization Policy launching in October*
- **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
- **Classifier global resilience and efficiency:** comprehensive audit to assess and ensure child safety classifiers do not operate with bias. Improve our text classifiers that detect inappropriate interactions with children by developing text model to enable automation of hard actions (disable/report) against groomers
- **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
    - Private by Default for Teens
- **Reporting:** Launch "Involves a Child" feedback tag where missing e.g., in FB Groups and IG comments.
- **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
    - Well-funded global digital citizenship and wellbeing program similar to Google's Be Internet Awesome
    - Safety education video for minors at sign-up
- **RTC and Remote Presence products: i**n-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. [redacted] 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.

Legislative/Regulatory Action

- US: FTC to update COPPA rules (expected in 2020)
- UK: Age Appropriate Design Code (likely implemented in H1 2021 subject to Parliamentary approval), Online Harms Bill
- Ireland: Child Safety and Media Regulation
- Australia: eSafety Commissioner's Safety by Design principles, amendments to the Online Safety Act
- Brazil: Internet Steering Committee ("CGI") focused on ads for minors and use of minors data (e.g. Bill 1746/2015)
- Germany: Federal and Interstate Youth Law (likely to come into effect in H2 2020)~~, UK~~
- South Africa: Protection of Personal Information Act 4 of 2013 (July 1, 2021)
- Philippines: Safer Internet Bill (~~parental consent from U18~~s age of consent)
- India: Personal Data Protection Bill, 2019 (requires age verification and parental consent)
- SSA: ~~Data Protection Act and Cyber Crimes act in Kenya, South Africa's POPI act and~~ Zimbabwe's Cyber crimes bill (parental consent and robust age verification), South Africa Films and classification amendment act 2020 (registration, self classification, age verification for pornographic sites), Kenya Data Protection Act 2019 (age verification for under 18's on social media platforms)
- Indonesia and Philippines: Draft bills w/r/t age-verification to ensure minors not exposed to "inappropriate/immoral 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.

Legislative/Regulatory Action

- UK [redacted] due H2 2020 (may be delayed due to COVID-19); Online Harms Act; upcoming UK Samaritans guidance recommending a single PoC, clearer policies to prohibit "normalizing" SSI and no encryption* to ensure can scan for this content in messaging (*experts will split on encryption, some will not want us to use it so we can scan private messages, others will want us only to scan in public spaces and to encrypt private messaging so people can safely and securely discuss mental health issues)
- Ireland/IDPC/EU: IDPC has continuing privacy concerns around processing of mental health data when running our SSI classifiers, as a result we are taking a far more limited approach in the EU which means we are susceptible to more press fires
- EU: under the e-privacy directive scheduled to take effect in December 2020, OTT communication services have to get explicit individual opt ins (~GDPR) for the classifiers we run. For example, each user will have to agree to allow us to run photoDNA across their photos. If a user doesn't give that consent and we use PhotoDNA on the images they share, that will be considered unlawful. *Strategically, very interesting as this suggests encrypting messaging aligns with expectations.
- Korea: bullying of celebrities in Korea (K-Pop) driving regulatory initiatives, incl. liability for platforms failing to remove content, or compelling data disclosure to victims in civil claims.
- Japan: Proposed 'defamation' and cyber-bullying' regulation following celebrity cyber bullying case that connected to suicide.

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 borderline 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.

Legislative/Regulatory Action

- US: FTC to update COPPA rules (expected in 2020)
- UK: Women and Equalities Select Committee Changing the Perfect Picture: an inquiry into body image to study impact of advertising and social media on body image
- Zimbabwe: Cyber Security and Data Protection Bill
- Mexico: Persistent bullying issues in schools spilling over online, President's son bullied publicly raising questions of censorship on social media.

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
    - Messaging Group size limits, forward limits
    - 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

- Child Safety - State of Play (7/20)
- 1H2020 Global Safety, Health and Well-Being Policy Plan
- H12020 Plans (Review)
- Safety Policy Product Asks
- H2 Plans
- Safety Legislative / Regulatory Tracker
- Managing CBI Reviews During COVID-19

Karuna [redacted] created the document *Sep 24, 2020 at 5:44 am*
