Message

From: Erma R. [redacted]
([redacted])
Sent: 9/28/2020 1:25:42 AM
To: Nick Clegg [redacted]@fb.com]
CC: Tim [redacted]@fb.com]; Page [redacted]@fb.com]; Jess [redacted]@fb.com]
Subject: SCHEDULE: Monday, September 28
Attachments: 09.28.2020 - Stella Kyriakides.pdf; 09.28.2020 - Event with Muslim@ + Allies .pdf; FW_ Kids note.msg; Child Safety - State of Play (720).pdf; Weekly Team Clegg Sync Agenda .pdf

- CJ will be on standby for the Kyriakides meeting and the Muslim @ Q&A.
- Please plan to leave the Kyriakides meeting a few minutes before 9:00 am to give enough time to log into the Muslim @ Q&A.

8:12 AM — **Log-into WebEx for Kyriakides Meeting – CJ on standby**
Link: WebEx:
Meeting number (access code): [redacted]
Meeting password: [redacted]

8:15 – 8:58 AM — **MTG: EU Commissioner for Health Stella Kyriakides** (Brief attached)
Link: WebEx:
Meeting number (access code): [redacted]
Meeting password: [redacted] from phones)

NOTE: You will need to leave this call a few minutes before 9:00 am, so you can log into the Muslim@ Q&A a little early.

8:58 AM — **Log-into BlueJeans for Q&A with Muslim@ – CJ on Standby**
BJN: [redacted]
Meeting ID: [redacted]

9:00 – 10:00 AM — **EVENT: Leadership Q&A with Muslim@ and Allies** (Brief attached, Tim to send an update on staffing later tonight)
BJN: [redacted]
Meeting ID: [redacted]

10:00 – 11:00 AM — Break (60 min)

11:00 – 11:15 AM — **CALL: COVID Pre-screening COVID for Flu Shot Appt [Scheduled 9/29]**
A nurse will call you: [redacted]

11:15 – 11:30 AM — Break (15 mins)

11:30 – 12:15 PM	**MTG: Joel 1:1** (Joel to send) Rooms: [redacted]
12:15 – 1:00 PM	Break (45 min)
1:00 – 1:30 PM	**SYNC: Antigone | Tim | Nick** (Pre-read & Tim's Email attached) Rooms: [redacted]
1:30 – 2:00 PM	**MTG: Team Clegg Weekly Sync** (Agenda attached) [redacted]
2:00 – 2:30 PM	**MTG: Nick | Stacey 1:1** Rooms: [redacted]
2:30 – 3:00 PM	**ZOOM MTG: FTC Comms Plan (David [redacted])** Zoom: [redacted] Zoom ID: [redacted] Passcode: [redacted]
3:00 – 3:30 PM	Break (30 mins)
3:30 – 4:00 PM	**SYNC: Lara 1:1** (Lara to send) Rooms: [redacted]
4:00 – 4:15 PM	**SYNC: Tim, Page, Nick [Daily - 15 mins]** Nick to connect via phone
4:30 PM	**Depart for de Young Museum of San Francisco Event @ 5:30 PM** Golden Gate Park, 50 Hagiwara Tea Garden Drive, San Francisco, CA 94118

Message

From: Tim [redacted]
Sent: 9/10/2020 7:23:51 PM
To: Erma [redacted]@fb.com]
Subject: FW: Kids note

Flag: Follow up

Hi Erma – could you find 30 mins for me, Nick and Antigone, ideally the week after next?

T

From: Nick Clegg [redacted]@fb.com>
Date: Tuesday, September 1, 2020 at 11:11 PM
To: Tim [redacted]@fb.com>
Subject: Re: Kids note

Yes, could we meet with Antigone so I can understand this more fully. Parts of this are worrying. N

From: Tim [redacted]@fb.com>
Date: Tuesday, September 1, 2020 at 12:49 PM
To: Nick Clegg <nclegg@fb.com>
Subject: Kids note

Hi Nick

In the madness of the last few weeks I managed to forget to send you the note that Antigone and her team put together on where we are on the various strands of work that relate to kids on our platforms.

See attached. It's long, a little rambling, and quite worrying in places. Overall, it tends to confirm the impression I have that this area of work is often sidelined and Antigone doesn't have the resources or consistent leadership support that she needs.

A few things that jumped out at me:

- Her warning that our numbers of NCMEC referrals are likely to show a sharp increase due to COVID
- No PM appointed to lead on youth privacy despite agreeing that this work is important
- Backlogs of actionable information about individuals who have failed age gates
- The prospect that the ePrivacy directive will prevent us from scanning for CEI without the consent of the individual account holders
- The idea of commissioning an audit of special protections for minors – which sounds promising.

NB a few things have evolved in the weeks since the note was written – we are implementing forward limits on Messenger and IG (though not going as far as Antigone would like), and Karina is going to follow up with us shortly imminent decisions around age defaults for certain features on IG.

Would you like to set up time with Antigone to discuss? FYI I'm going to set up a monthly 1:1 with her to try to stay more closely involved with all of this.

Tim

Child Safety - State of Play (7/20)

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

Child Exploitation

Policy Landscape

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

Legislative/Regulatory Action

- US: EARN-IT, Lawful Access to Encrypted Data 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, EC to set up EU CSAM hub (note: much in direct contradiction to ePrivacy Directive)
- 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.
- South Korea: intermediary liability law for NCII and CEI
- Japan: National Police's annual report on Child Sexual Exploitation and Sexual Solicitation of Minors

- Belgium: NCII legislation 6 hr takedown or liability
- Other: Germany, Ireland, S.Africa, Japan, and Taiwan also have legislative activity

Immediate Product Vulnerabilities

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

Areas of Improvement

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

- **NCMEC Reports:**
  - contextualize these numbers
  - meaningfully drive down the number of reports before encryption through preventative work including 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,...)
- **Trafficking:** sex trafficking classifier with a focus on minor sex trafficking
- **TC Project Protect:** fast track some key wins like universal video hash sharing, research on client-side solutions, independent accountability system that tracks to CSEA vol principles
- **Livestream:** develop content independent CEI detection tech, wiretap capability, reportability and reviewability upon report
- **Interop:** strong discoverability and reachability controls, including further restrictions to PYMK and GYSJ

- **Ephemerality:** adult<>minor carveout for disappearing mode and/or some way to enable content reporting and review
- **IG:** Extend existing MSGR Safety tips for scams, impersonation, child safety/IIC in IGD; Expand policies on borderline content re minor sexualization and build classifiers to detect and take action on problematic aggregated content on surfaces like Explore, hashtag pages, Reels, better age-related logic for preventing conversations between unconnected adults and minors
- **Groups:**
    - remove private invite links and group chat functionality for high risk FB groups
    - robust system for removing apps that categorize and make searchable group invite links to WA groups in violation of our terms of service
    - robust system notifying platforms when users share group invite links for alleged WA CSAM groups
- **User Education:** robust ongoing in product user education on safety and privacy tools
- **RTC and Remote Presence products:** in-call reporting with evidence capture, feature gate high-risk users, user education/self-remediation interventions (i.e. safety notices, upsell reporting/blocking)

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

Recommendation

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

Ahead of Interop, work is in progress to block unconnected adults on FB and IG from initiating messages with minors on FB, however work is not planned to prevent FB and IG adults from initiating messages to IG minors. Limited country tests and data analyses have shown that forward limits and Group size limits can demonstrably reduce CEI sharing but product leadership has been unwilling to prioritize these mitigations. We are currently escalating the Forward Limits conversation up through Monika to meet with MLT. We are also getting ready to meet with Jay Sullivan 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. We have progress on blocking predicted groomers from using Disappearing Mode but need CO support to apply the same protection in

IG. Work has been completed on basic safety mitigations for Rooms but work still needs to be done to enable in-call reporting with evidence capture in both Rooms and RTC (aka Video Chat). RTC, which is the predecessor to Rooms, is not yet in parity with Rooms (i.e. feature gating high-risk users, user education/self-remediation interventions) and will require stronger leadership buy-in to accomplish even bare minimum protections this half. We continue to advocate for strong discoverability and reachability controls in anticipation of Interop, including further restrictions to PYMK and GYSJ. Work to develop a CEI video classifier for FB Live is currently ongoing but we need more resources for this in addition to engineering and review resources to develop an operational sex trafficking classifier. Please see the chart below for detailed summary of the specific related work and its status.

	A	B	C	D	E	F
1	**Area**	**Additional notes**				
2	**Ready; work is complete / ongoing and on track**					
3	Sexualisation of minors on IG	Ongoing: Expanded minor sexualisation policy in H2; holistic IG profile review in 2021				
4	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				
5	NCMEC numbers contextualization	Ongoing: prioritized by Child Safety DS				
6	Block reach of unconnected adults to FB minors	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)				
7	Messenger Safety Tips	Ongoing: Launch Messenger Safety Tips on IGD (ETA: Q3); Cross-network threads (ETA: November; prior to Interop global launch)				
8	Tincan and Disappearing Mode carve-out	Ongoing: Use the IIC actor level classifier to block our riskiest users from using Tincan and Disappearing Mode (the successor to Tincan), which is a high-risk feature for child safety				
9	**Somewhat ready; work is ongoing but more resources are needed**					
10	Interop: discoverability and reachability	Needs more resources: Strong discoverability and reachability controls, including further restrictions to PYMK and GYSJ				
11	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)				
12	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				
13	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.				
14	Fwd Limits and Group Thread Size Limits	Forward Limit and Group Thread Size Limits can demonstrably reduce sharing of CEI (a recent test showed that forward limits 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 Forward Limits conversation up through Monika to meet with MLT and could use your support. We are also getting ready to meet with Jay Sullivan to discuss how best to present group thread size limits to MLT.				
15	**Not ready; work has not been**					

	prioritised and more resources are needed					
16	Aggregated content on IG surfaces	Build classifiers to detect and take action on problematic aggregated content on surfaces like Explore, hashtag pages, Reels				
17	Block reach of unconnected adults to IG minors	Currently, on MSGR, we do not deliver messages initiated by unconnected adults to minors. In IGD, we do not have this protection in place. For Interop, product has committed to preserving this rule in MSGR, but not for cross-network messages sent by FB adults to IGD minors. Nor will they protect IGD minors from receiving messages from unconnected IGD adults. In light of the fact that grooming on IG occurs most likely ~2x as much as FB, their main reason not to build this protection for IGD minors is less than compelling: we do not want to block FB parents and older relatives from being able to reach out to their IGD children/relatives. They claim this is one of the main value props for Interop and a big growth bet.				
18	IIC classifier use on IG	Product and Policy have aligned that using the IIC classifier is the best available safety mitigation for minors in Disappearing				

		Mode, planned for global launch in November. However, the actor-level groomer classifier is not yet operational on IG, which means that we cannot use it to block groomers from accessing high-risk features like Disappearing Mode. We need about a week or two of CO review capacity in order to calibrate the classifier for IG use. This is low hanging fruit and fully supported by Child Safety CI and IG Policy. We just need help getting CO to prioritize this.				
19	Upsell reporting for Disappearing Mode	Disappearing 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 Disappearing 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 Disappearing Mode before global launch.				
20	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 strikes (currently set at 17x, which is too high). This area is severely under-resourced given our SESTA-FOSTA liability.				

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 U18s)
- 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)
- 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.

	A	B	C	D	E	F
1	**Area**	**Additional notes**				
2	**Somewhat ready; work is ongoing but more resources are needed**					
3	Age Management	An 18-month long escalation that ended				

		with a Cross-Family Review on 3/3/20 with Nick's input resulted in a decision ***not*** to build a classifier to detect minors under 13 years old. Instead, leadership decided to invest in policy work to push industry partners for device/OS level age verification and management solutions. In the meantime, our basic COPPA compliance is at risk when product does not prioritize checkpointing and disabling u13s. We also do not yet cross-enforce against u13s discovered on IG who have hard-linked FB accounts. In addition, 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.				
4	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. We need more review resources to establish ground truth data on age in priority markets. We should 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.				
5	**Not ready;**					

	work has not been prioritised and more resources are needed					
6	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.				
7						
8						
9						
10						

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: Coroner's report from inquest into the death by suicide of British teenager Molly Russell 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 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	Ongoing: Samaritans have incorporated most of our recommendations (removed encryption warnings, and moved towards a broader set of principles and accompanying factsheets)				
4	Detection and actioning of	Ongoing (ETA: Q3)				

	Borderline SSI content based on new policy					
5	Splitting out our violation types between more narrow categories within Admission and Promtion and action appropriately	Ongoing (ETA: end of Q3)				
6	Addressing harmful content by understanding if a frequency based prevalence metric can measure harm	Ongoing (ETA: H2)				
7	**Somewhat ready; work is ongoing but more resources are needed**					
8	Establish processes to look for SSI misses, 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				
9	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				
10	Prioritization to weigh various safety risks	All SSI live videos currently carry same weight for reviewers,				

	associated with the types of SSI videos.	need to assess how to streamline more severe ones and how defined.				
11	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				
12	**Not ready; work has not been prioritised and more resources are needed**					
13	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				
14	Bullying of celebrities driving regulatory initiatives	Revisiting definitions of public 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				

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.

	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	**Somewhat ready; work is ongoing but more resources are needed**					

6	Social comparison	Product teams brainstorming but more tangible interventions needed				
7	Bullying / harassment	Minor public figures policy in place but need to stress test				
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						
11						
12						
13						

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 add me to Group Threads" control in MSGR (parity with WA)
    - Extend existing MSGR Safety tips for scams, impersonation, child safety/IIC, in IGD
    - Develop strikes and disable thresholds for repeat violations in IGD and MSGR
    - PYMK, GYSJ
        - Prevent suspected IIC adults from friending minors and make PYMK, messaging/friending rules bi-lateral for integrity measures taken against suspected IIC users, i.e. We currently don't surface minors to adults at IIC precision rate 30% so we shouldn't surface these IIC users to minors either.
- Livestream CEI detection tech
        - Reviewability & Reportability, including CEI detection in RTC (which includes co-watch, screen-sharing, VCL, etc.).
- Sex Trafficking classifier with a focus on minor sex trafficking

- Self-remediation tools and backend measures that promote safety while respecting users' privacy for RTC and Remote Presence products, such as:
    - In-call reporting with evidence capture;
    - Screen-sharing static notification either in roster or as part of UI control panel;
    - Feature gate high-risk users;
    - User education/self-remediation interventions (i.e. safety notices, upsell reporting/blocking);
    - Understand abuse prevalence / useful proxies in IG RTC + MSGR RTC.
- Support our Global Classifier Resiliency and Fairness initiative for safety classifiers that are critical to our issue areas.
- Age Assurance / Verification
    - Proactive age verification and enforcement / u13 classifier; u18 checkpoint
    - Expanding age appropriate experiences, i.e. identify and build special protections (default settings, content restrictions, blocked products) that should apply to u13 and u18 users.
- Suicide and Self Harm
    - Self harm support resources (e.g. eating disorders, cutting; currently, we send suicide prevention resources and tips to people engaging in any kind of self-harm)
    - Providing varying resources and actions to people using our platform to promote versus admit suicidality; policy development and classifiers needed to action differently and appropriately.
    - IDPC classifier approval for Europe
    - SSI checkpoints on FB and IG need to be more relevant and population/risk specific
    - Crisis Service Over Messenger engineering support

X-Industry Sharing:

- Video DNA x-industry banks

User Education and Marketing:

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

Additional Quips/Docs

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