Tennessee v. Meta
Case No. 23-1364-IV
PX-00970

Document Provided in Native Format

Strategy to address IIC

Child Safety CI & Safety Policy
Chris [redacted], Malia [redacted], Sarah [redacted]
September 10, 2021

FACEBOOK

1. Context setting
2. The state of play today and E2EE readiness plans
3. Risks and gaps

1. Context setting
2. The state of play today and E2EE readiness plans
3. Risks and gaps

Context: we are trying to launch IGY amidst a multi-government campaign using child safety concerns to try and thwart our plans to end-to-end encrypt messaging

No matter what we do to address child exploitation, certain stakeholders will not be satisfied as their goal is to thwart encryption and protect government and law enforcement access to messaging content. BUT a significant portion of safety stakeholders just want to know we have put in place meaningful mitigations and a robust system for reporting when bad things happen.

1 E2EE Supporters	2 Potential to convert with mitigations backed by efficacy data	3 Potential to convert with the same + child safety client side protections similar to muted words	4 Must provide for full client side scanning, homomorphic hashing, access to unencrypted backups, backdoor,..., -- something that keeps detection very close to existing functionality	5 All the same as 4 but not limited to child safety
National Network to End Domestic Violence (US) *NetSafe (NZ)* *SaferNet (BR)*	*UNICEF (Global)* *End Violence Against Children (Global)* *ConnectSafely (US)* *Netherlands government*	*Thorn (US)* *ECPAT (Global)* *Internet Watch Foundation (UK)* *ChildNet (UK)* *INHOPE (Global)*	*NCMEC (US)* *NSPCC (UK)* *US DoJ*	*Law Enforcement* *NSA (UK)* *Indian government* *AU government* *UKHO*

4

H1 Strategy and Execution Recap

Strategy	Execution	Impact
Correct false narrative of 20M victims driving policy regulatory momentum and media coverage by getting NCMEC cybertip contextualization into conversations with policy makers and partner briefings.	NRP on 2/23 New Safety Center Hub on child protection Contextualization briefings for key govt stakeholders (UK, AU) and ~100 safety stakeholders on contextualization	MSFT has reached out to further this research and focus on prevention based approach. EVAC has already put out research and wants to collaborate to ensure their next 2021 funding round (~USD 16M) is informed by our research w/r/t countries/regions to target and issues to focus on.
Demonstrate our commitment to building a robust system for child protection within private messaging.	Briefings on safety mitigations for key govt stakeholders (UK, AU) and ~90 safety stakeholder June CyberSafety Month Messenger and MKids SMTs in the US	UNICEF wants to work with us on their H2/ 2022 planning to focus on mirroring on the ground prevention projects.
Drive safety stakeholder/public policy **shift to prevention based approach** to thwarting online child exploitation through engagements, partnerships and white paper.		Thorn has stepped away from threats of large public campaign fueled by Ashton Kutcher celebrity.
Demonstrate our **commitment to combating online child sexual abuse broadly.**	Project Protect w/ Technology Coalition Child safety viral CSAM PSA in top CSAM sharer countries Supporting global helpline system w/Tech Matters Offender diversion resources global expansion	Positive partnerships with two dozen NGOs, government ministries, government officials, child rights advocates and high profile influencers. Positive press coverage in launch countries.
In key markets, **push back against onerous legislation** and identify and support alternative child safety legislation (the effectiveness of this is dependent on success of 2 and 3).	E.g., In Brazil, a dozen safety orgs, led by Safernet, released a letter to protest against a Presidential decree to amend Marco Civil that would take away platforms' ability to proactively enforce content and remove bad actors without a legal order.	

In order to continue to make progress with safety stakeholders we will need

Must haves:

- Continued progress on safety mitigations, in particular upstream visible prevention measures and visible robust victim supports.
- Efficacy data to demo effectiveness

Likely will need:

- Some sort of client side CEI safety tools (along the lines of muted words)

May need:

- More robust client side for CEI

1. Context setting
2. The state of play today and E2EE readiness plans
3. Risks and gaps

Mission

~~Build a safe community for children~~

Prevent harm to children in our community

Strategy

Prevent: Pivot resources upstream to prevent harm. Stop possible violators (identified with e2ee-resilient behavioral means) from interacting with children, with one another, and with child sexual abuse material (CSAM).

Control: Empower minors with options to control their experience, so they can better recognize harm, and remove friction for minors to report harm they experience to us.

Care: Increase recall and volume of reports, and provide support to minors who are victimized on our platform.

{More detail here}

Inappropriate Interactions with Children (IIC): **Definitions**

1. **IIC Tier 1 (Real-world harm)** : Imminent or ongoing harm to a minor, such as intent to meet a child for sex or sextortion. This is illegal and we report to NCMEC as per our legal obligations.

1. **IIC Tier 2 (Online harm)** : Sexual conversation with a minor (colloquially known as 'grooming').

E2EE requires a new playbook for addressing IIC

- **IIC Tier 1**: We will upsell reports to children, aiming for a 10x volume increase, and focus on correctly actioning and escalating these reports following the *don't miss a thing* (DMaT) playbook.

- **IIC Tier 2**: We will shift from using victim prevalence metrics (labelled using message content) to using UNiS proxy measurement and will develop a playbook for making principled enforcement decisions based on 'risk'.

IIC detection, measurement & enforcement today vs post E2EE

Problem	Playbook	Status Quo (What happens today)	Consequences of E2EE (What happens when we lose message content)	Mitigations expected to be in place by M6 (Colour predicts extent of safety win provided by each mitigation, compared to today's status quo)	Confidence that we will deliver these mitigations by M6 (High/medium/low)
IIC T1	Detection	Text classifiers proactively scanning 'risky' conversations in several languages (not in the EU). FB only.	These classifiers will no longer function outside of user reports	The focus here will be on dramatically increasing user report voume (10x).	We already validated this is possible in the E2EE task force, we just need to execute.
	Measurement	DMaT (escalation recall & latency).	DMaT will apply only to user reports, no longer to proactively detected conversations	We will be measuring escalation recall, escalation latency, prevalence within reports, human review market coverage.	We already set up the frameworks to measure these in the war room, these metrics will be delivered within H2.
	Enforcement	Disable & manually report to NCMEC.	Currently the large majority of T1 reports are proactive, we will no longer be able to find these.	We will expand human review capacity to cover 100% of eligible reports (up from ~1% today) so that we are maximising reactive T1.	This relies on Ops Readiness, which isn't currently accepting requests for net new restricted demand. Moving these metrics will be tricky when we have 10x the volume of reports, and many different languages to cover.
IIC T2	Detection	- Text classifiers proactively scanning 'risky' conversations. FB only, - Behavioural classifiers (trigged without message data, but trained using message content)	Text classifiers will no longer work, and behavioural classifiers will drift over time	1) Using a broader range of principle-based rules in place of classifiers. 2) Using the pairwise classifiers developed by UnIS to increase recall for similar levels of precision by acting on pairs, not just actors 3) Upselling reports (10x)	This relies on the principles-based rules being validated and report upselling working, but both are relatively straightforward. UnIS detection has not yet been validated in production, so results may vary.
	Measurement	'Victim prevalence', measured by sampling message threads between minors and adults meeting certain criteria.	Victim prevalence will cease to be a functional metric.	Pivot to use UnIS proxy measurement for prevalence on FB & IG: - Defining a 'risky' interaction based on E2EE safe profile level characteristics of bad actors and potential targets. - Counting volume of potentially 'risky' messages between bad actors <> targets	Preliminary work in H1 has confirmed that UNiS measurement tracks with victim prevalence, but additional work is required to fully validate this. (A susequent Armadillo deep-dive will focus on UNiS specifically).
	Enforcement	- Auto-disable at 90% precision. - Gate access to 'risky' products at 60% precision, and remove 'follow' button on minor profiles (IG). - Remove children from PYMK & search (FB) and all major surfaces on IG at 30% precision. - Remove 'friend' button on minor profiles for adults who have received NFX from children in the previous 14 days.	We will need to get comfortable making principled enforcement decisions based on 'risk' (as determined by message metadata and non-message interactions such as search history, content consumption etc).	Expand discoverability protections: - IG: Prevent 'risky' adults from finding & connecting with children on 8 surfaces (eg. Explore, Reels, Follow lists). - FB: Prevent risky adults from finding & connecting with children within friend list and via profile information (bio, photo).	There is limited precedent for making enforcement & ship/no-ship decisions using principles-based rules, although we did this for preventing unconnected adults from messaging minors on IG Direct. IGTS have already rolled out removing minors from all major surfaces for 'risky' adults (where we have age model coverage).

Support experiences

IIC Safety Notices [MWB]	Shipped on open threads. In progress for MSGR Armadillo Dec 2021, IGD Armadillo Jan 2022.
	Scoping for Real-Time calling.
Improve reporting for minors while increasing reporting volume: Upsell reporting, reduce friction in reporting flow & increase discoverability of reporting [MWB]	On track for IGD Late Sept, MSGR October; more concepts being explored with Thorn and plans to start experimenting in Q4.
Provide users easy ways to sync blocks across accounts to prevent recontact from an account they've already blocked from another account [MWB]	On track for XAC in Sept and IG in-network Sept/Oct
Provide block sync tools using SUMA to block any new accounts created by offenders [MWB]	On track, Oct/Nov
Victim Support and age-appropriate, contextual safety education/resources [MWB + Online Safety]	In progress, likely Q4 project
Understand and build product explorations for on-device, client-initiated nudity/sexually explicit content filter in messaging that respects e2ee principles, similar to Muted Words [MWB]	Scoping
Build digital safety and literacy education in-app and off-platform [MK + Youth XFN]	MK Pledge Activities, On track
	Youth XFN Transparency & Education workstream, Funded and Scoping

1. Context setting
2. The state of play today and E2EE readiness plans
3. Risks and gaps

Risks and gaps

Risk	*Example*	*What's needed to unblock*
1. Maturity gaps on IG & MK: Child Safety detection and enforcement is poor or missing on IG/MK compared to FB.	Currently, there is no proactive text detection of IIC on MK/IG and we do not know how well FB's text detection will perform on IGY.	Eng & labelling capacity
2. Limited age prediction coverage: labelling limitations make it difficult to protect all minors, and find some of the most serious offenders, e.g., adult age-liars.	We don't have stated age for ~30% of IG users and currently have no way to detect age-lying adults who impersonate minors for grooming purposes.	Eng & labelling capacity (Core Dims P1 goal this half)
3. Lack of human review capacity, quality and routing: Reactive user reporting from private surfaces are not always reviewed appropriately.	For example, on IG, reviewers have no tool available to report IIC to NCMEC from reactive report queues, and coverage is sparse outside of English.	Tooling and eng work + review capacity + training (CIP 1 goal this half)
4. Limited Surface coverage: We have limited understanding of the scope of Child Safety issues on several important surfaces.	We do not know the extent of the Child Safety problem on Reels, Stories, Live, Ads, etc.	Investigator, Eng and DS/IS capacity
5. Heuristics for safety: We cannot continue to rely on models validated against message content (e2ee is coming).	We propose making broader use of principles-based rules, e.g. NFX received from minors, to limit interactions between risky adults and minors, but there is limited precedent for doing this, and we need to establish a playbook to validate these rules alongside Growth.	DS work to validate heuristics + Growth/Policy alignment

Appendix

CIPCO & Partner Team collaboration model

