Filed Under Seal

EXHIBIT B

e2ee: an overview of safety mitigations for Child Safety & Terrorism

Friday, 27th March 2020

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053452
METANMAG-002-00285147

Objectives for today:

1. Provide an overview of the safety mitigation strategy (and execution against that strategy) for child safety and terrorism.
    a. Align on the extent to which we can (and cannot) mitigate these safety challenges post e2ee
2. Assess impacts of safety work on pressure from key stakeholders and implications for landing encryption externally
    a. Align on prioritisation of e2ee safety mitigation work vs. other company priorities (Interop), and timeline for execution

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053453
METANMAG-002-00285148

TL;DR:

1. There is a gap between what we think we need to be able to say and where we are in the safety mitigations for E2EE — both currently and in the expected end state. This could jeopardize the launch of E2EE across messaging services in the form of government action or other external pressure that blocks our ability to do so.

    a. Part of the gap will never be closed with the current technical implementation and product privacy decisions on how we will implement e2ee on messaging (e.g., no client-side scanning or backdoor).

    b. Part of the gap will be closed, but, at the current investment rate, we don't foresee adding any additional talking points for external engagements before Q4 2020 due to competing priorities (interop, profile+, etc).

    c. The Policy team has outlined the claims they think we will need to be able to make prior to launching encryption across our messaging platforms to help mitigate the risk (slide 27).

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053454
METANMAG-002-00285149

TL;DR:

2. Based on the CI work done to date, and at the current investment rate, we assess that:

   a. We will lose visibility over 82% (13M/y) of current CEI reports to NCMEC across the family of apps, based on a near-total loss of Messenger-derived reports. We believe we will be able to recharacterize half of the reported content as non malicious after we ship the intent framework. Policy and product agree this will have limited impact on government and safety expert pressure.

   b. We expect to be able to escalate ~30% of reports in DOI Credible threat review flows with the loss of inbox. We will aim to further increase that number as we begin to roll out E2EE optimizations to tooling, rep training and using other indicators, but we are still likely to miss 50% true positives.

   c. We don't foresee adding additional talking points for government and other external engagements before Q4 2020.

   d. Re-prioritization of integrity effort, including mitigations for COVID-19, will most likely see the priority on safety mitigations work decreasing, and consequently the timeline stretching further into 2021.

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053455
METANMAG-002-00285150

Agenda

For each of A) Child Safety and B) Terrorism:

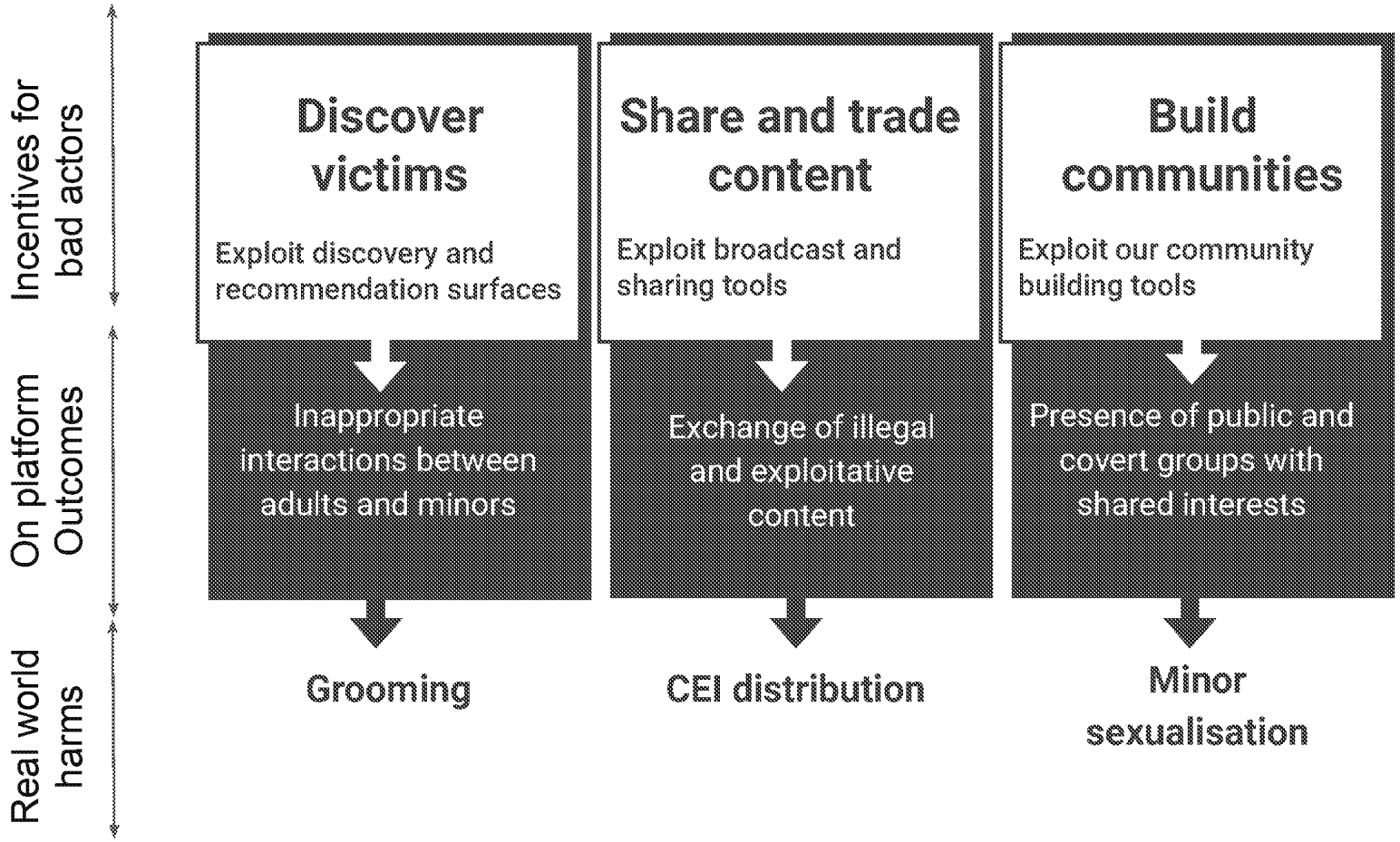
1. What are the **harms** related to e2ee?
2. What is our **product strategy** to address these harms? How is our ability to address **impacted by e2ee** and how much can we mitigate the impact?
3. How far along are we on **execution**?
4. To what extent do these safety impacts address **external pressures** we face from our key stakeholders?
5. Where and on what timeline should we **prioritize mitigation efforts** going forward?

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053456
METANMAG-002-00285151

A) Child Safety

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053457
METANMAG-002-00285152

Child Safety: how bad actors use our platforms?



CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053458
METANMAG-002-00285153

Context: CEI distribution

1. In 2019 we sent 16M reports (84% from IG Direct / FB Messenger in 2019) of CEI to NCMEC - This is the number which is most often quoted externally as indicative of the size of the problem of CEI sharing on Facebook's platforms.
2. We do not believe that this is an accurate representation of the problem and we are currently working on being able to better contextualise this data publicly and give more clarity to NCMEC and Law enforcement about how much actual harm exists on the platform
    a. We have anecdotal evidence of CEI sharing without malicious intent - we intend to publicly ship a quantitative breakdown of intent (i.e. malicious vs. non-malicious sharing).
    b. We have some evidence that we may be over reporting even against our current liberal reporting guidelines.
    c. We are better understanding how CEI is copied/re-uploaded onto our platforms in order to give more signal about where the actual abuse is taking place.

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053459
METANMAG-002-00285154

Child Safety: Impact of Encryption

Incentive	Safety harms	Strategy to mitigate harms	Impact of encryption
Discover victims	Grooming	1. Prevent victim discovery 2. Detecting and reporting inappropriate interactions 3. Encourage reporting 4. Support victims	1. Will lose ability to detect grooming where the adult and minor already knew each off-platform. 2. Will continue to be able to leverage all 4 strategy levers for bad actors discovering minors on our platform. 3. Will only be able to send detailed reports to NCMEC (LE) for reported cases
Share and trade content	CEI distribution	1. Disrupting CEI networks 2. Aggressive content moderation 3. Prevent CEI seeking behaviour	1. We won't be able to auto-delete and report CEI shared within messaging (84% from IG Direct / FB Messenger in 2019) 2. Will be able to detect networks of bad actors trading CEI via metadata + public square signals.
Build communities	Minor sexualisation	1. Prevent discovery via recommendation surfaces 2. Aggressive content moderations: delete violating content and reduce exposure to borderline content	1. Will continue to be able to detect, delete and prevent minor sexualisation across public surfaces (there is limited presence of this abuse on private messaging surfaces)

Post e2ee enforcement coverage <25% ~50% >75%

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053460
METANMAG-002-00285155

Context on Child Safety mitigations

1. In Q3-2019 we set up an E2EE taskforce focussing on a small collection of encryption techniques in order to hit a timeline for encryption of Q2-2020
2. As encryption is now expected until late 2021 at the earliest, we have decided to prioritise foundational work to understand the intent of actors committing child safety abuses on our platforms (i.e. IIC and CEI).
    a. We believe that this will give us the best chance of mitigating the impact of encryption when we come to ship it
3. Direct follow ups coming out of the the taskforce are still being worked on, albeit at a slower pace, and are likely to be completed by the end of 2020.
4. Our focus on Child Safety continues making progress on public surfaces (ie.: Minor Sexualization) while keeping the lights on MSG detection (maintaining our escalation rate to LE for Tier 1 and CEI).

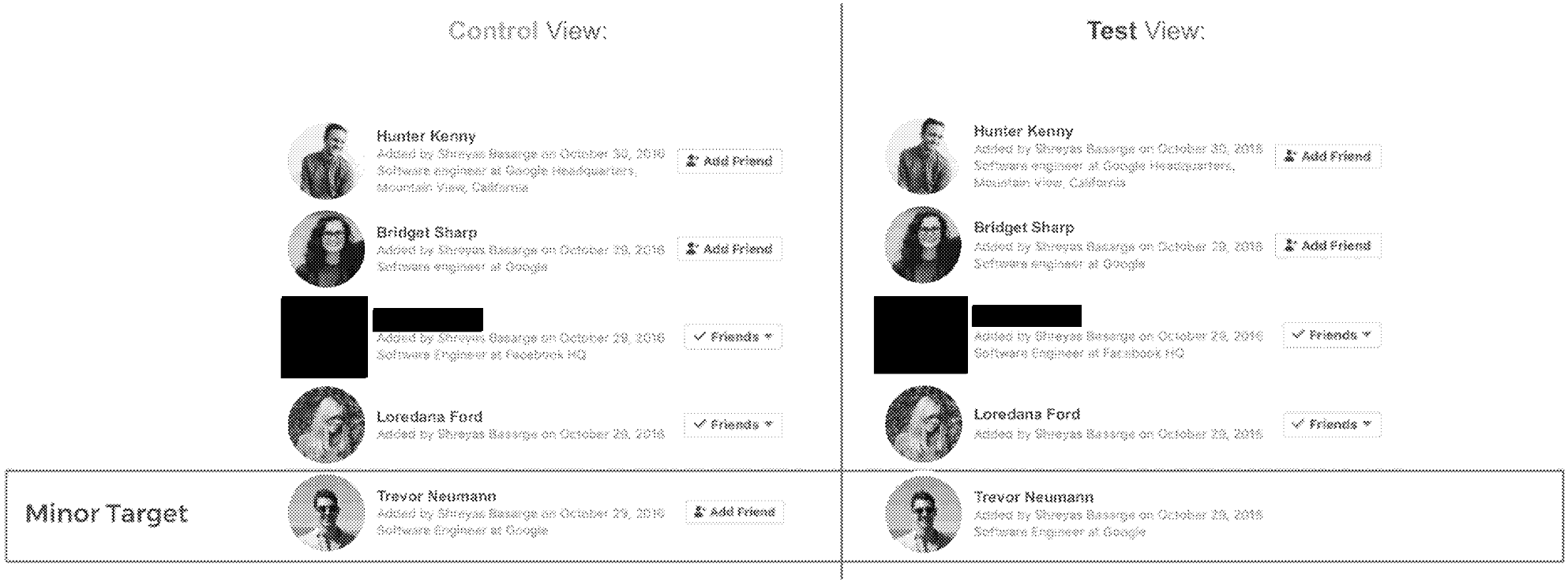
CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053461
METANMAG-002-00285156

Child Safety: Mitigating 'grooming'

Strategy	Tactics	Stage of Execution	Effectiveness
Preventing victim discovery.	Blocking friending between 'suspicious' adults and unconnected minors.	**LAUNCHED**: Experiment launched to 100%. Effect is measurable.	~30% decline in prevalence of IIC victims *(caveat: measurement is not yet e2ee resilient)*.
Proactively detecting & reporting inappropriate interactions.	Behavioural classification of adults with suspicious public square activity/ messaging patterns. NCMEC reports based on limited meta data and public square signals.	**RISK**: Not actively improving behavioural classification or improving metadata based NCMEC reporting *(readiness for interop is currently being prioritised ahead of e2ee)*.	Behavioural classification bounded by ability to measure recall/precision without inbox NCMEC reports may not provide enough 'evidence' to be directly actionable.

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053462
METANMAG-002-00285157

Experiment Example: Limit connect-ability



CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053463
METANMAG-002-00285158

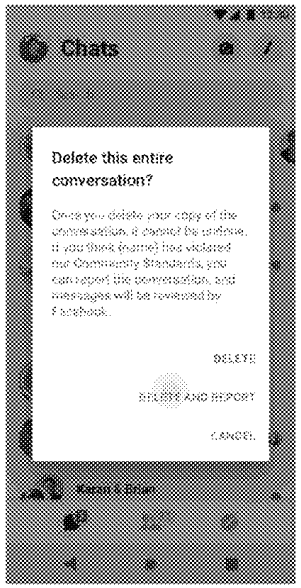
Child Safety: Mitigating 'grooming'

Strategy	Tactics	Stage of Execution	Effectiveness
Encouraging user reporting and other forms of self-remediation.	Increasing reporting and self remediation entry points (more visible, contextual upsells).	**EXPERIMENTING**: Ran several successful experiments, but blocked from launching due to capacity constraints. Launched 'proactive warnings' which up-sells blocking to IIC victims.	~50x Increase in IIC victims who report from 0.1% to 4.8%. Increase in perpetrator coverage to ~18%.
Supporting victims.	Ship support material to confirmed victims of Tier 1 IIC over Facebook and Messenger.	**EXPERIMENTING**: On the Child Safety roadmap for H1 work planned for Q2.	Targeting 20% engagement from IIC victims with support material.

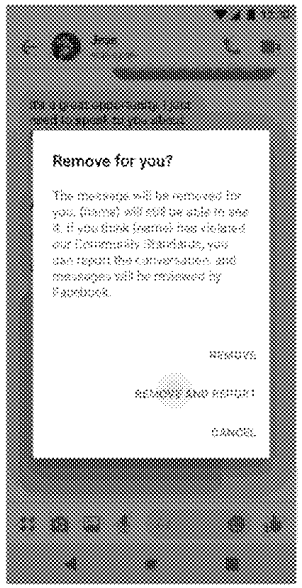
CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053464
METANMAG-002-00285159

Experiment Example: Increasing high-quality reports

Report after thread delete



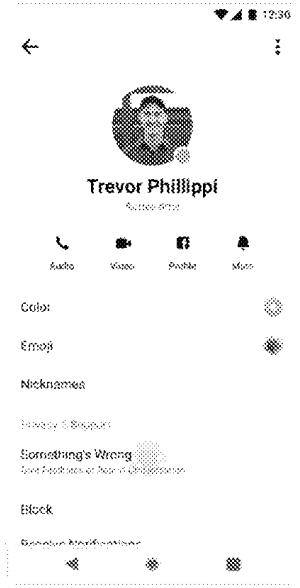
Report after message remove



Long press to report

Report from inbox

Reorder thread details



Report after block

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053465
METANMAG-002-00285160

Child Safety: Mitigating 'CEI distribution'

Strategy	Tactics	Stage of Execution	Effectiveness
Disrupting CEI sharing networks	Developing heuristics to distinguish malicious from non-malicious CEI sharing.	**UNDERSTAND:** We are in the analysis stage of understanding actor intent.	We have promising early results (up to 50% of sharing is non-malicious) but it's too soon to tell how effective we will be.
	Developing actor and network level classification to detect malicious CEI sharing.	**UNDERSTAND:** We have additional work to do to understand actors and use more e2ee resilient behavioral signals.	We believe that can get to a similar precision to WhatsApp (c. 80%) and there we may be able to go beyond that.

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053466
METANMAG-002-00285161

Child Safety: Mitigating ‘minor sexualisation’

Strategy	Tactics	Stage of Execution	Effectiveness
Aggressive enforcement of Child Safety on unencrypted surfaces.	Progressing the maturity of our unencrypted surfaces for Minor Sexualisation.	We are have improved proactive detection for and shipped reporting for Minor Sexualisation in comments in Q1 - with groups and profiles likely to land in Q2.	We have better detection but in comparison to the rest of Child Safety action volumes for this abuse remain comparatively low.

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053467
METANMAG-002-00285162

External pressure from stakeholders - child safety

Constituencies	What they are concerned about?	Outcomes they would like to see?	Satisfaction with what we can provide post e2ee	Implications
Law enforcement [& intelligence]	* Content for investigations (retrospective) and surveillance (prospective) * Useful referrals for investigation and rescue	* No e2ee * Backdoor (device-side, server-side) * Other tech "solutions" (e.g., ghost protocol) * Client/device-scanning with reporting		* Unlikely to be assuaged by anything short of access to content/encryption backdoor and/or client side with reporting
Safety NGOs	* Prevalence/sharing of CSAM * Useful referrals for LE investigation and rescue * Overall child safety ecosystem	* Limit e2ee (e.g., for kids) * Scanning (server, client, or device side) * Demonstrated commitment to address child safety ecosystem		* Would likely be satisfied with client/device-side scanning * Some (e.g., Thorn) may be amenable to outcomes short of client/device-side, i.e. robust and demonstrable claims about e2ee resilient capabilities (reduce prevalence, prevent connections, on-platform actions, actionable reports, victim support)
Others (e.g., policymakers, tech, media)	* "Win" over tech * Overall child safety ecosystem * NCMEC numbers	* Transparency/accountability about the impacts of encryption * Demonstrable improvement on e2ee resilient capabilities		* Some encryption skeptics may be satisfied with client/device-side scanning to address child safety concerns * Others be amenable to robust and demonstrable claims e2ee resilient capabilities (reduce prevalence, prevent connections, on-platform actions, actionable reports, victim support) * We should be candid about safety tradeoffs alongside continued articulation of affirmative benefits of encryption, other eco-system efforts

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053468
METANMAG-002-00285163

What we need to be able to say prior to launch - Child Safety

Based on the estimated impacts on our ability to make reports to NCMEC, **these are the claims we think we need to be able make about our post E2EE safety capabilities**:

- We have **significantly reduced the prevalence** of CSAM on our platform: "Using tools that do not rely on inbox content, including efforts to stop bad actors from connecting with other bad actors, we have significantly reduced CSAM on our platform."
    - *Ex. measurement: a significant drop in CyberTips to NCMEC ahead of launching end-to-end encryption, attributable to new integrity efforts*

- We have enhanced our ability to **prevent potentially harmful connections with minors:** "Using tools that do not rely on inbox content, we are able to identify accounts that may be seeking to engage in IIC and prevent interaction with/harm to minors"
    - *Ex. measurement: we prevent [X#] of adults from friending minors on our platform*
      **Note this could count from not offering friend button and from upselling blocking, and any other integrity efforts*

- We are still able to **identify and remove bad actors** from our platform: "Using tools that do not rely on inbox content we remove…
    - *Ex. measurement: [X#] accounts per month for suspected child exploitation across all of our messaging apps.*
    - *Ex. measurement: [X#] accounts per month for suspected child exploitation on Messenger.*
    - *Ex. measurement: 250,000] accounts per month for suspected child exploitation on WhatsApp.*
    - *Ex. measurement: [X#] accounts per month for suspected child exploitation on IG DM*

- We will still provide ample ***actionable* reporting to NCMEC and LE:** "Using tools that do not rely on inbox content, we will provide more CyberTips that include CEI content within end-to-end encryption than what LE investigates today."
    - *Ex. measurement: X times current # of CyberTips with CEI content currently accessed by LE*

- We have introduced new world-class **victim response** measures: "In addition to everything we are doing to prevent, detect and report child sexual exploitation, we have built a world class system for user education, victim response and support (~to suicide/self injury)."

Assessment of safety mitigations to support child safety claims

5. Prioritization

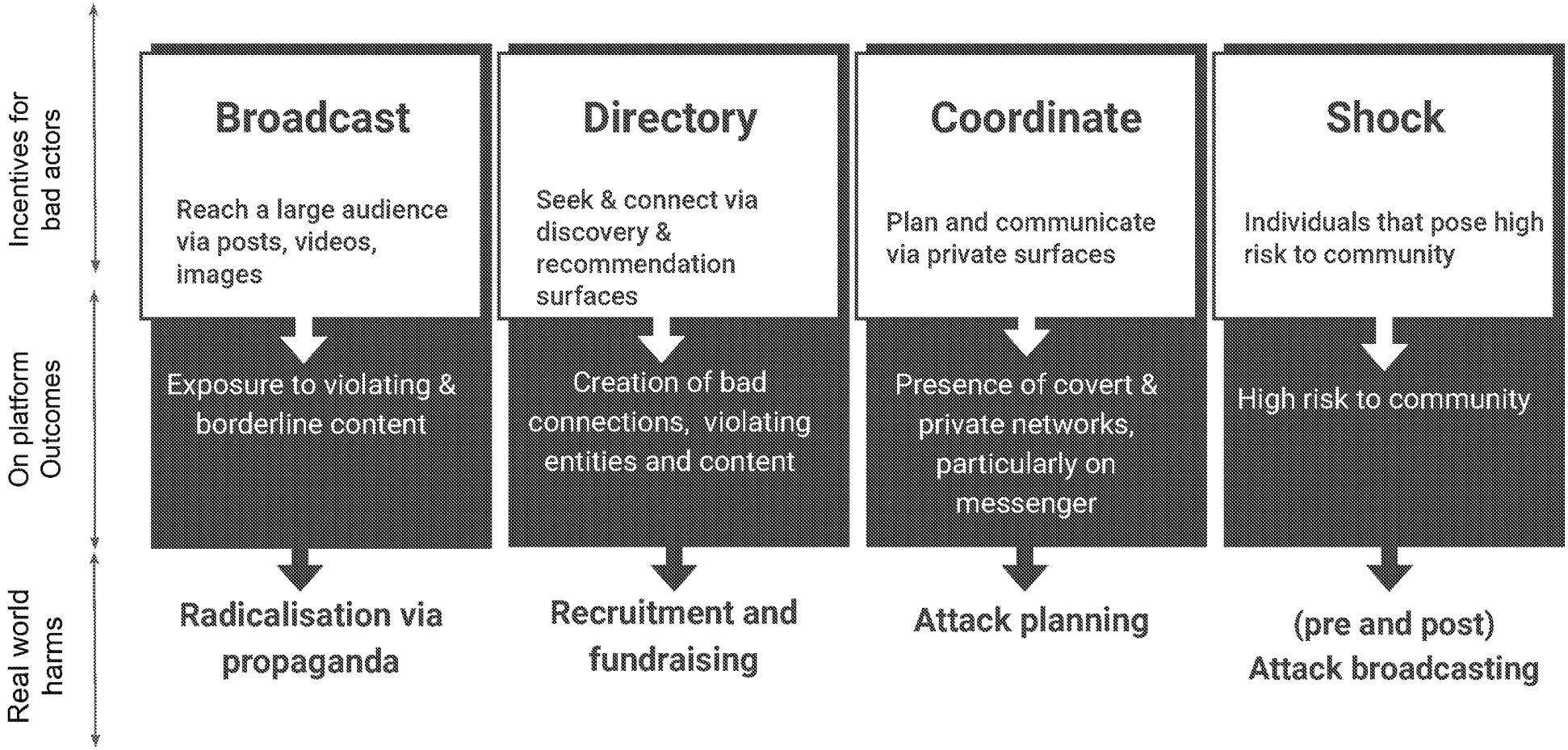
Mitigation	Claim	Ability to substantiate based on e2ee mitigation work		Comments
		Work to date	Est. End state	
Reduce prevalence	We have significantly reduced (CSAM)...			Likely single-digit effect; will be recharacterizing existing prevalence metrics
Prevent harmful connections	We increasingly identify and take action against suspicious accounts before messages are exchanged...			More limited for actors that already know each other; could get to green but with growth implications (friending)
Remove bad actors	We remove accounts per month on par/above WhatsApp...			
Actionable reporting	We will be able to provide actionable reports on par with existing LE actioned NCMEC reports for CSAM...			# of reports likely exceed those acted on by LE, but contain less actionable information
Victim response	We have built a world-class system...			

Ability to substantiate claim	**Not at all / little**	**Somewhat**	**Mostly**	**Entirely**

B) Terrorism

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053471
METANMAG-002-00285166

Terrorism: how bad actors use our platforms?



CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053472
METANMAG-002-00285167

Terrorism: strategy to mitigate these harms

Incentive	Safety harms	Strategy to mitigate harms	Impact of encryption
Broadcast	Radicalisation	**Super-charge content moderation:** 1. Expand automated actions (limit reach) 2. Utilize broad soft-actions	• Majority harm manifests on public surface. • Limited impact from encryption.
Directory	Recruitment and fund-raising	**Prevent abusive community building** 1. Disrupt connections between seekers & entities which are enabled by FB tools a. Expand detection of overt entity level signals b. Prevent 'motivated seekers' from utilising our search & discovery surfaces.	• Majority harm manifests on public surface. • Limited impact from encryption.
Communicate	Attack planing	**Disrupt existing networks:** 1. Ensuring high reportability of threats to LE 2. Takedown networks using non-inbox signals	• Loss of inbox leads to increase reliance on user metadata (group membership, non-inbox signals).
High Risk Individuals	Attack broadcasting	**Preemptively identify high risk individuals:** 1. Utilise content & behavioral signals to identify individuals	• Increased reliance on public square behavioural and content signals.

Post e2ee enforcement coverage <25% ~50% >75%

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053473
METANMAG-002-00285168

Terrorism: mitigating ‘attack planning’

Strategy	Tactics	Stage of Execution	Effectiveness
Ensuring high reportability (of threats)	Move to a no-inbox threat review flow Continuously improve recall for user reports	**EXPERIMENTING**: Completed tooling for E2EE optimized review flow, ETA: end H1. Working to setup Don’t Miss a thing archetype approved metrics on recall.	Improve user report recall
Takedown existing networks	Eliminate the perception of safety in private surfaces by using strategic takedowns of networks (like CIB)	**MEASURED**: We launched the V1 draft of the Network Disruption Playbook, improved existing recidivism detection and scoped out a baseline success measurement.	Taken down 5 networks (Proud Boys, FARC..) with 100s of FB, IG users, groups and pages. Additional 5 networks by end of H1. Tracker

Terrorism: mitigating ‘attack broadcasting’

Strategy	Tactics	Stage of Execution	Effectiveness
Utilise content and behavioral signals to identify and disrupt high risk individuals	Build policies that allow identification of probable high risk actors Take disruptive actions to limit risk to community	**UNDERSTAND:** We are currently in understand phase (mapping actor signals indicating risk) and have mapped possible interventions. Policy development is an H1 goal.	Recall/Precision of high risk individual detection

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053475
METANMAG-002-00285170

WIP - External pressure from stakeholders - CT/DOI

Constituencies	What they are concerned about?	Outcomes they would like to see?	Satisfaction with what we can provide post e2ee	Implications
Law enforcement [& intelligence]	* Content for investigations (retrospective) and surveillance (prospective) * Useful referrals for investigation and rescue	* No e2ee * Backdoor (device-side, server-side) * Other tech "solutions" (e.g., ghost protocol) * Client/device-scanning with reporting		Unlikely to be assuaged by anything short of access to content/encryption backdoor
NGOs?	* Prevalence/sharing of CSAM * Useful referrals for LE investigation and rescue * Overall child safety ecosystem	* Limit e2ee (e.g., for kids) * Scanning (server, client, or device side) * Demonstrated commitment to address child safety ecosystem		Likely satisfied with client/device-side scanning Some (e.g., Thorn) may be amenable to claims about e2ee resilient capabilities (reduce prevalence, prevent connections, on-platform actions, actionable reports,
Others (e.g., policymakers, tech, media)	* "Win" over tech * Overall child safety ecosystem * NCMEC numbers	* Transparency/accountability about the impacts of encryption * Demonstrable improvement on e2ee resilient capabilities		May be amenable to claims about e2ee resilient capabilities (reduce prevalence, prevent connections, on-platform actions, actionable reports) Alongside continued articulation of affirmative benefits of encryption, other eco-system efforts

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053476
METANMAG-002-00285171

What we need to be able to say prior to launch - CT/DOI

Based on the estimated impacts on our ability to make reports to LE and remove DOI/CT content on the platform, **these are the claims we think we need to be able make about our post E2EE safety capabilities**:

- We are **reducing the prevalence** of violating content**:** "As a result of a new framework for soft actions on likely violating content and behaviour, users will be less exposed violating content":
    - *Ex. measurement: Our soft actioning framework reduces the prevalence of likely violating content by [X%]*
- We continue to invest in **detecting and removing violating content**: "By increasing the reach of automated detection and constantly improving our automation (e.g. precision, recall), we will continue to enforce aggressively against terrorism and hate content":
    - *Ex. measurement: X pieces of terrorism content removed , out of which X% were removed as a result of proactive detection.*
    - *Ex, measurement: estimated reduction in X% of false negatives detected for terrorism, , and likewise reduction by X% of false positives in content removed for terrorism*
- We are doing more to **disrupt potential radicalisation and recruitment efforts** for terrorist activity: "Through strategies to disrupt radicalisation and recruitment pathways, we increasingly limit potential bad actors' ability to connect with others or broadcast their agenda":
    - *Ex. measurement: We impose soft actions on [X#] of profiles per month*
- We will improve cross-platform enforcement to **identify and remove bad actors**: "Using tools that do not rely on inbox content, we remove a significant number of accounts for terrorism violations; interoperability improves our cross-platform enforcement":
    - *Ex. measurement: Ahead of end-to-end encryption, we are removing [X%] of accounts for terrorism violations without inbox signals; [X%] of these users are simultaneously disabled on WhatsApp.*
    - *Ex. measurement: At full interoperability we expect these types of cross-platform terrorism disables to increase by [X%]*
- We will still provide ***actionable* referrals to law enforcement:** "By leveraging behavioral signals, we will be better able to detect [X] types of profiles and provide law enforcement with information that furthers their investigations":
    - *Ex. measurement: We proactively refer [X%] of the reports for credible threats that we did before e2ee*

Assessment of safety mitigations to support CT/DOI claims

Mitigation	Claim	Ability to substantiate based on e2ee mitigation work		Comments
		Work to date	Est. End state	
Reduce prevalence of potentially violating content	We have seen a decline in users engaging with potentially violating content.			
Detect and remove violating content	We will continue to enforce aggressively on violating content			
Disrupt radicalisation and recruitment	Potentially bad actors' will be limited in their ability to connect with others.			
Remove bad actors across platform	At full interoperability, we will be better at enforcing on violating actors across platform.			
Actionable referrals	We are better able to detect [X] types of profiles; refer X% of reports for credible threats...			

Ability to substantiate claim	**Not at all / little**	**Somewhat**	**Mostly**	**Entirely**

APPENDIX

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053479
METANMAG-002-00285174

We have made a number of public commitments that we need to show progress on:

Mark Zuckerberg (March 2019)

"There are real safety concerns to address before we can implement end-to-end encryption across all of our messaging services."

"...we [will]take the time to build the appropriate safety systems that stop bad actors as much as we possibly can within the limits of an encrypted service.

but

"...we will never find all of the potential harm we do today when our security systems can see the messages themselves."

[redacted] ***Guy Rosen (2H 2019)***

- **Prevent:**
  - *We are working to better understand how bad actors connect with their victims in the first place, and we're developing ways to prevent that initial contact.*
  - *We are building tools to look for signals and patterns of suspicious activity so that we can stop abusers from reaching potential victims.*
  - *Each person should have control over who can find and initiate communication with them, who can see if they are online and what those people can find out about them.*
- **Detect:**
  - *We are getting better at using non-content-based signals to identify bad behavior.*
  - *We are particularly focused on working across the entire Facebook family of apps.*
- **Respond:**
  - *We recently began testing new ways to help more minors report adults who send unwanted messages, and so far, our results show a significant increase in reporting.*
  - *Law enforcement will still receive valuable information in response to lawful requests.*

We continue to receive significant external pressure:

In addition to longstanding law enforcement calls for encryption backdoors:

- Governments and safety NGOs demand **no reduction in user safety**
- They want **proof that our AI works** as intended

NSPCC

"I'm urging you not to introduce end-to-end encryption unless you can guarantee that children's safety will not be compromised."

Home Office

How will AI working on meta data only detect similar levels of suspicious activity to current methods and what are the expected levels of detection when content is not accessible?

Governments – particularly in UK, Australia, US – are pressuring us to show concrete progress on safety as part of "technical consultations"

- Australia – 2-day session postponed from March 8, to be rescheduled
- UK – abbreviated video conference held March 19, follow up [TBD]
- US – meeting with FBI on Jan. 22

Although we are unlikely to placate government concerns in these consultations, showing progress on our commitments helps limit the likelihood of government action to block our encryption plans

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053481
METANMAG-002-00285176

Deconstructing LE & safety NGO encryption concerns

When law enforcement and safety NGOs express concerns about "encryption," they are talking about impacts to three things:

	Who cares?	What they want to be able to do?	What tech solutions they would like to see?	What we can/will provide post e2ee
Investigation (retrospective)	* Law enforcement	* Obtain information relevant to existing investigations	(0) Server access (backdoor) (1) Device access (2) Cloud access (backups)	* Metadata and content from unencrypted surfaces (retrospective* and/or prospective) * Investment in law enforcement capabilities/resources * Better coordination among companies
Surveillance (prospective)	* Law enforcement *[Intelligence]	* Conduct surveillance on suspected bad actors * [To collect intelligence on targets]	(0) Server access (backdoor) (3) "Ghost protocol" (4) Lawful hacking	
Tips & scans (proactive)	* Law enforcement * Child safety	* Receive useful referrals for rescue and investigations *Prevent sharing of CSAM and other harmful content	(0) Server access (scanning) (5) Client/device-side scanning a. OS w/reporting b. OS w/o reporting c. App w/ reporting d. App w/o reporting	* Continued scanning of unencrypted services and surfaces * Better upstream prevention of harmful activity * Better signals-based detection * Enhanced reporting tools for users * Better support for victims * Better coordination among companies * Big investment across child safety ecosystem

**Not currently stored or provided on WhatsApp*

Tech solutions sought by various stakeholders

"Lawful" access (with a search warrant or court order)

0. server access: LE access to data on/passing thru servers (I.e., backdoor/no e2ee)
1. device access: LE access to e2e encrypted data on device
2. cloud access: no e2e encrypted backups
3. "ghost protocol": LE added to e2e encrypted conversations
4. lawful hacking: LE exploits vulnerabilities to access data

Client and device-side scanning

5a. scanning w/in operating system without reporting (~malware)

5b. scanning w/in operating system with reporting

5c. scanning w/in app without reporting

5d. scanning w/in app with reporting

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053483
METANMAG-002-00285178

UNUSED / GRAVEYARD

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053484
METANMAG-002-00285179

Interop Safety Mitigations and Encryption Concerns Part II

1. Complete review of safety mitigations

2. Explain Law Enforcement/NGO safety concerns and desired tech solutions by audience (e.g. client-side scanning)

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053485
METANMAG-002-00285180

Safety mitigations – overview of existing work

Prevent

1. **Restricting adult -> minor friending**
2. **Analysis of "seeker" and "producer" behavior** to collapse bad actor ecosystem
3. **Fake minor account network analysis** to prevent inappropriate interactions with children

Detect

1. **Inbox blind test** of 65 Dangerous Orgs cases reported to LE
2. **Detecting CEI group threads** without inbox content

Respond

1. Testing new **reporting entry points and prompts**
2. **No-inbox proactive reports** for NCM EC
3. **Victim support** – testing with sextortion victims

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053486
METANMAG-002-00285181

Non-inbox Reports to NCMEC

- **Goal:** Deliver actionable information to NCMEC in an e2ee-protective way

- **Test**: We are testing a new NCMEC reporting format that a) removes content, b) adds new non-inbox information, and c) makes reports easier to read. We have sent four of these new reports to LE for feedback. **Changes include:**
  - Table that identifies how all the accounts listed in the report are linked to the main suspect e.g. via IP or email
  - Group Message thread titles for all suspect accounts
  - Comments and Profile posts that are suspect
  - Details on Page and Group memberships of suspect accounts
  - Details on types of searches the suspect accounts are running

- **Findings:** Feedback is that the new changes are "super helpful," but if stacked against other reports with inbox content, likely won't be prioritized. Example feedback:
  - Attribution chart: Using this LE has been able to tie individual reports to other NCMEC reports (either from us or other companies) and identify whether accounts listed are linked to a person of interest in separate investigations (via geolocation)
  - Group chat name: LE can look these up in device searches and link them together
  - Comments and Profile Posts: Provides helpful context to better understand the offending user
  - Pages and Group membership: LE can use this information for other avenues of investigation
  - Search history: LE can use this during their investigation, and potentially learn new trends or identify minor targets they haven't connected with yet

- **Next steps:** Continue to seek feedback and iterate on reports

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053487
METANMAG-002-00285182

Victim Support

- **Goal:** Increase engagement and effectiveness of support

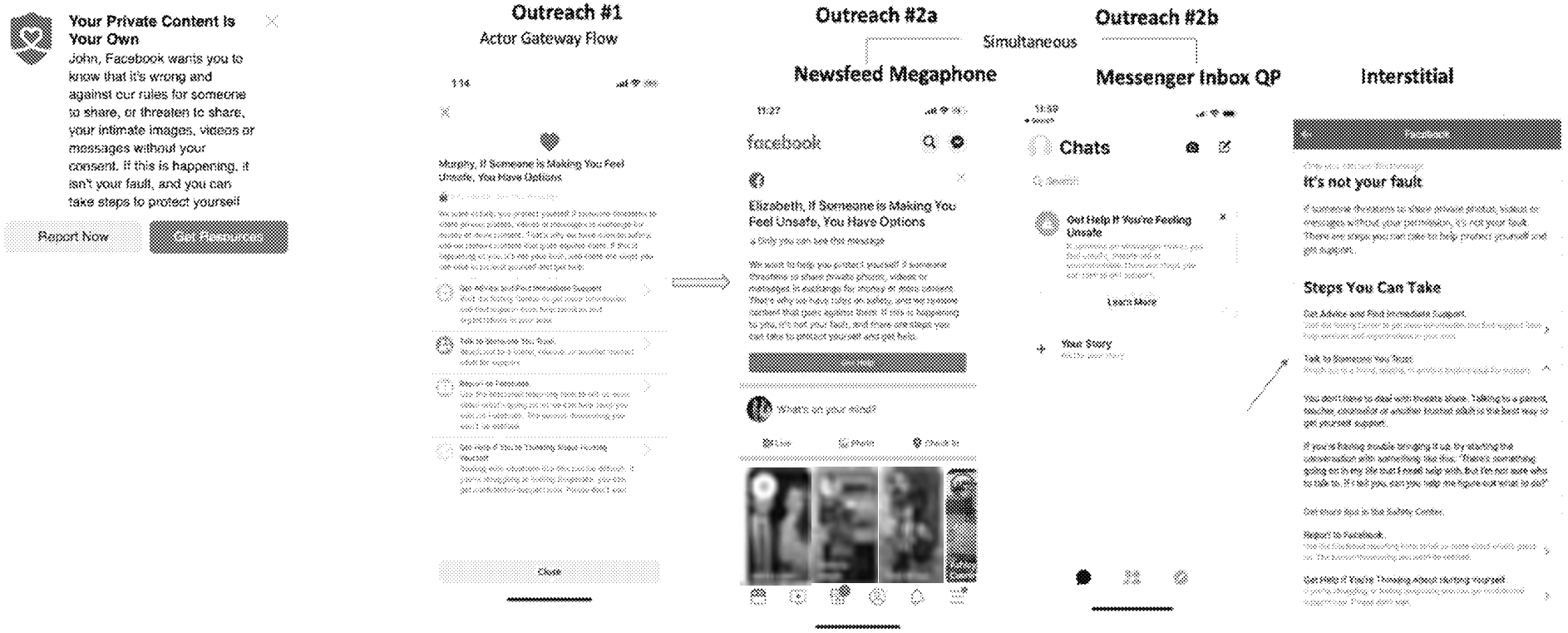
- **Test:** Launched an updated version of the supportive experience that involves an updated entry point with a more informative UI. Currently only child victims of sextortion, where a manual NCMEC report has been filed, receive support.

- **Findings:**
  - **4.27K users have seen the expanded (V2) support with an overall CTR (click through rate) on resources is ~18%**
  - **Of the users who clicked on any options:**
    - 39% of users clicked "Get Advice and Find Immediate Support"
    - 35% of users clicked "Report to Facebook" (note: only 3% of this group actually reported)
    - 25% clicked on "Talk to Someone You Trust"
    - 25% clicked on "Get Help if You're Thinking About Hurting Yourself"

- **Next steps:** (1) evaluate which populations to expand support to, (2) evaluate whether moving the support closer to the event will lead to more positive outcomes, (3) explore why reporting conversion is so low

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053488
METANMAG-002-00285183

Sextortion Victim Support- Suite of Options



v1

v2

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053489
METANMAG-002-00285184

	Server Access	Device Access	Cloud Access	Ghost Protocol	Lawful Hacking	OS Scan + Report	OS Scan, No Report	App Scan + Report	App Scan, No Report
LE	Favored	Favored	Support	Support	Currently doing	Support	Not against, but not of value to them	Support	Not against, but not of value to them
Security Community	Concerns (cyber security)	Openness (Carnegie report)		Unknown	Mixed views				
Thorn	Do not support	Do not support	No stated position	Do not support	Do not support	Favored	Sellable	Favored, prefer on OS	Sellable
NCMEC	No stated position	No stated position	No stated position	No stated position	No stated position	Favored	Sellable	Favored, prefer on OS	Sellable
Privacy NGOs	Strongly opposed	Opposed	Opposed	Strongly opposed	Opposed				

	Server Access	Device Access	Cloud Access	Ghost Protocol	Lawful Hacking	OS Scan + Report	OS Scan, No Report	App Scan + Report	App Scan, No Report
Definitional Concerns	Not e2ee - company has the key (I.e., backdoor)	Preserves e2ee for data in motion, but is a back-door for at rest	Not e2ee	Third party has accessed the data	Third party has accessed the data	Not end user to end user	Not end user to end user, but no access (~malware)	Not end user to end user	Not end user to end user, but no access
Security Concerns	Introduces vulnerability by giving another party the key	Introduces vulnerability by giving another party the key	Data not e2e encrypted as is, no new vulnerability	Introduces vulnerability	Exploits a vulnerability, which we would not be able to correct until later			Concerns about reverse engineering algorithm	Concerns about reverse engineering algorithm
US Constitutional Concerns	Limited as long as search warrant	Limited as long as search warrant	Limited as long as search warrant	Limited as long as court order	Limited as long as search warrant	Potential 4th amendment if reporting mandated	Not if limited to CEI	Potential 4th amendment if reporting mandated	Not if limited to CEI
User Trust	Undermined as company would have to keep key	Undermined as company would have to keep key	Little impact, data not e2e encrypted as is	Undermined to extent don't trust due process in relevant country; and increases vulnerability of all users engaging with suspect party	All non suspect users would want us to fix the vulnerability upon learning of it.	Some "creepiness" factor	Some "creepiness" factor	Some "creepiness" factor	Some "creepiness" factor

Terrorism: mitigating 'radicalisation'

STRATEGY	TACTICS	STAGE OF EXECUTION	EFFECTIVENESS
Expand auto-actions to detect and enforce against propaganda	Launch InfoFirst reviews to expand automated actions. Classifier updates (*Mtml, CNN, PyText, and WPIE.*)	**EXPERIMENTING**: We have updated guidelines for CO. New GTRs was launched as experiment and requires follow ups. COVID capacity loss means we can only roll out to FTCs in limited queues.	Automation rates: ~92% for terrorism 75% for Hate Orgs
Utilize Soft-actions to detect and enforce against propaganda	Feed demotions for probable violating content	**OPERATIONAL**: Launched approved feed demotions on FB for Hate orgs.	Reduced bVPVs by over 1.1M daily for Hate orgs.

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053492
METANMAG-002-00285187

Terrorism: mitigating 'recruitment and fundraising'

STRATEGY	TACTICS	STAGE OF EXECUTION	EFFECTIVENESS
Expand detection of overt entity level signals	Remove entities with clear support and representation Build actor level classification	**LAUNCHED**: Launched detection & review flows to enforce on profiles. In H1, we are developing entity level scores for profiles, groups & pages.	Disabled 41K profiles that represent terrorist entities on their public profiles in Q2 2019
Prevent 'motivated seekers' from utilising our search & discovery surfaces.	Search filtering, demotions, tiering Filtering on Recommendation surfaces Downranking in Feed	**LAUNCHED**: We have reduced searchability of Terrorist content and now have measurement for entities. Expect to impact by end of H1. Expect to be at measured for non-search discoverability by end of H1.	Reduced searchability by over 80% on FB content in H2 2019, plan to reduce profiles & pages in H1. Also developing seeker success metric & experiments.

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053493
METANMAG-002-00285188

Context: external pressure from stakeholders

When law enforcement and safety NGOs express concerns about "encryption," they are talking about impacts to three things:

	Who cares?	What they want to be able to do?	What tech solutions they would like to see?*
Investigation (retrospective)	Law enforcement	Obtain information relevant to existing investigations	1. Server access (backdoor) 2. Device access 3. Cloud access (backups)
Surveillance (prospective)	Law enforcement [Intelligence]	Conduct surveillance on suspected bad actors [To collect intelligence on targets]	1. Server access (backdoor) 2. "Ghost protocol" 3. Lawful hacking
Tips & scans (proactive)	Law enforcement	Receive useful referrals for rescue and investigations	1. Server access (scanning) 2. Client/device-side scanning a. OS w/ reporting b. OS w/o reporting c. App w/ reporting d. App w/o reporting
	Child safety non-profits	Prevent sharing of CSAM and other harmful content	

*Tech solutions described in more detail on slide 25 in the appendix

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053494
METANMAG-002-00285189

What we need to be able to say prior to launch

Purpose: Following are the claims we need to make prior to launching encryption to avoid government action that blocks our ability to do so. We are delineating these claims in order to work with the product team to develop a plan for externally demonstrating measurable incremental progress toward these claims over the time it takes to fully move to encrypted interoperable messaging.

Whether or not we communicate it in this way externally, ***we need to align internally on the baseline***:

"Without scanning the content of people's messages, we will be able to make (a) [1%] of the Messenger and IG Direct reports of CSAM to NCMEC ([15%] of total Facebook family of apps) and (b) [~30%] of the reports of CT and DOI to law enforcement that we currently make.

- *Ex. measurement: Number of reports to NCMEC compared to prior years*
- *Ex. measurement: Number of CT/DOI reports to law enforcement compared to prior years*

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053495
METANMAG-002-00285190

What we need to be able to say prior to launch

Purpose: These are the claims we need to make prior to launching encryption to avoid government action that blocks our ability to do so. We are delineating these claims in order to work with the product team to develop a plan for externally demonstrating measurable incremental progress toward these claims over the time it takes to fully move to encrypted interoperable messaging.

We need to pivot to strong, metrics-driven proof of the impact of new mitigations:

- *We have **reduced prevalence** of the bad activity*: "Without scanning the content of people's messages, we have (a) reduced the sharing of CSAM and (b) seen no increase in terrorist/DOI content on our platform."
    - *Ex. measurement: a significant drop in CyberTips to NCMEC ahead of launching end-to-end encryption, attributable to prevention measures*
    - *Ex. measurement: steady or decline in CSER numbers for CT/DOI attributable to prevention measures*
- *We have enhanced our ability to **prevent harmful connections:*** "Using tools that do not rely on inbox content, we can identify and take action against accounts that may be (a) seeking to engage in IIC or CEI trading ~~sharing~~ (b) seeking to radicalize/recruit for terrorist activity."
    - *Ex. measurement: [TBD]*
- *We are still able to **identify and remove bad actors** from our platform*: "Using tools that do not rely on inbox content we remove (a) [##] accounts per month for suspected child sexual exploitation and (b) [##] for suspected terrorist-related activity."
    - Ex. measurement: similar to current WA takedowns
- *We will still provide ample *actionable* reporting to NCMEC and LE:* "Using content from the public square and user reports, we will provide (a) more CyberTips that include CEI content within end-to-end encryption than LE currently accesses and (b) continued actionable reports to LE on terrorist/dangerous org threats identified on the platform.
    - *Ex. measurement: X times current # of CyberTips with CEI content currently accessed by LE*
    - *Ex. measurement: X times current # of actionable CT/DOI reports*
- *We introduced new **victim response** measures*: "In addition to everything we are doing to prevent, detect and report child sexual exploitation, we have built a world class system for user education, victim response and support (~to SSI)."

CONFIDENTIAL
CONFIDENTIAL
META3047MDL_REPROD_VOL021
METANMAG_REPROD_VOL016
META3047MDL-014-00053496
METANMAG-002-00285191

What we need to be able to say prior to launch

Purpose: The following pages include the claims related to child safety and CT/DOI that we think we will need to be able make prior to launching encryption to avoid government action blocking our ability to launch e2e.

Backdrop: At present, product anticipates that post e2ee we will be able to...

- For child safety: make [1%] of the Messenger and IG Direct reports of CSAM to NCMEC (or, put another way, [15%] of total reports by the Facebook family of apps)
- For CT/DOI: make [~30%] of the reports of CT/DOI to law enforcement that we currently make, and remove [X%] of the CT/DOI-related content that we do today across the platform

We are delineating these claims in order to work with the product team to assess feasibility and develop a plan for demonstrating measurable incremental progress externally toward these claims over the time it takes to fully move to encrypted interoperable messaging.