

FILED UNDER SEAL

EXHIBIT 9

Messenger, Instagram Direct, and WhatsApp Enforcement Summary

Messaging and Safety Policy,
MWB

April 2021



HIGHLY CONFIDENTIAL (COMPETITOR)
HIGHLY CONFIDENTIAL (ATTORNEY'S EYES ONLY)

META3047MDL-224-00001065
METANMAG-059-00001065

Goals:

- 1) Provide an overview of Messenger and Instagram Direct policy enforcement, with a focus on *reactive* enforcement (aka, in response to reports)
- 2) Surface potential enforcements options under consideration for e2ee world
- 3) Identify others areas of concern / priority issues to address with Product teams
- 4) Get initial guidance on approach to cross-app enforcement

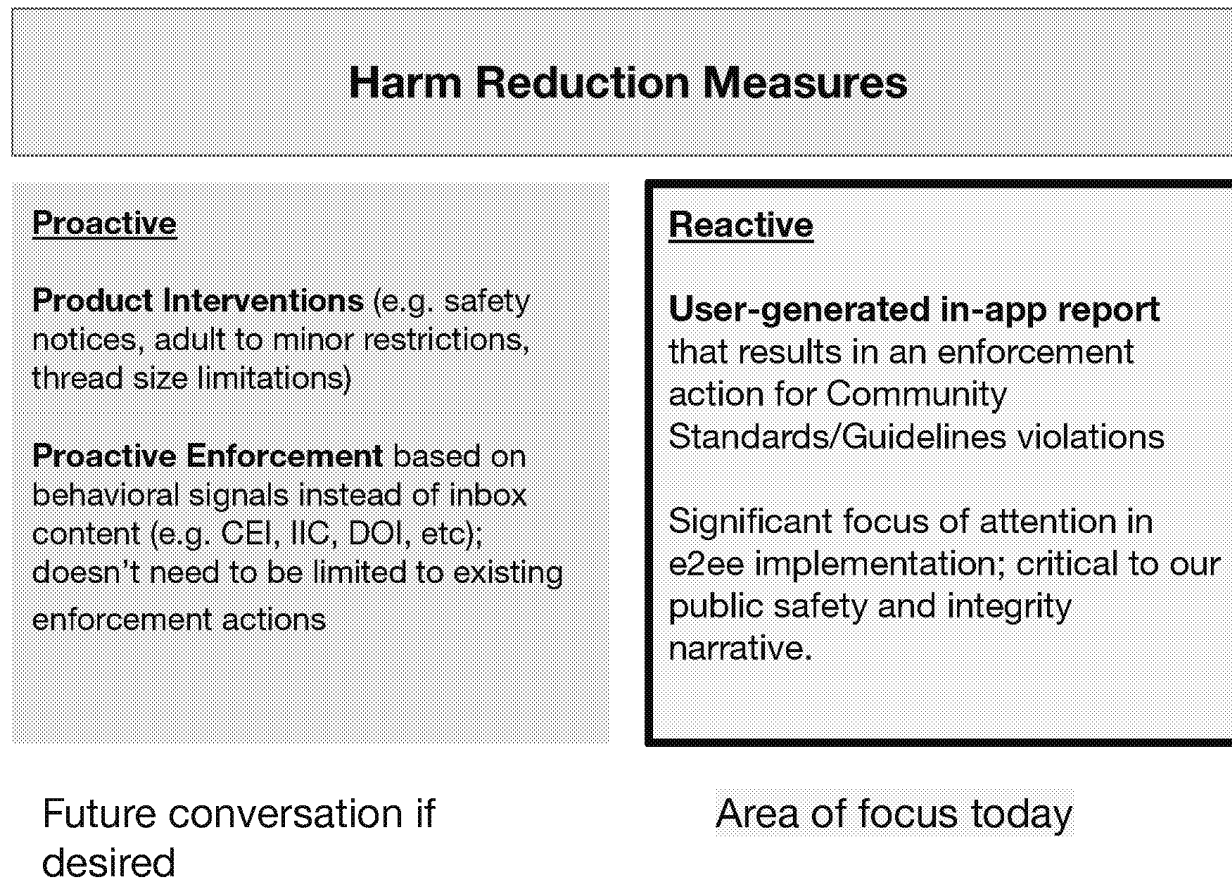
Non-goals:

- 1) Provide a detailed understanding of user controls, *proactive* enforcement and/or *proactive* product interventions in messaging ← *will address in future meetings if desired*
- 2) Revisit decisions about client-side solutions for enforcement
- 3) Revisit basic approach to e2ee

Summary

1. Defining “reactive” enforcement
2. In-app reporting and review data
3. Reporting entry points (MSGR/IGD)
4. Existing review constraints and planned improvements (MSGR/IGD)
5. MSGR/IGD enforcement actions explained
6. Community Standards enforcement actions mapping (MSGR/IGD)
7. Improving MSGR/IGD enforcement
8. WA Policies, Reporting, and Enforcement
9. Key challenges and questions in the policy enforcement space
10. Appendix

Proactive versus reactive enforcement



Messenger and IG Direct Reporting

Summary of reporting entry points

	1:1 Message reporting <i>(Ability to select or long-press a message to be reported; +/- 15 messages viewable to CO)</i>	1:1 Thread reporting <i>(Ability to report the last 30 messages in a conversation)</i>	Message reporting in a group conversation <i>(Ability to select or long-press a message to be reported; +/- 15 messages viewable to CO)</i>	Thread reporting in a group conversation <i>(Ability to report the last 30 messages in a conversation; must select a person to report)</i>
Messenger iOS	✗	✓	✗	✓
Messenger Android	✗	✓	✗	✓
IG Direct iOS	✓	✓	✓	✗
IG Direct Android	✓	✓	✓	✗

**Reporting experience mocks available in Appendix*

Messenger/IGD Report Review

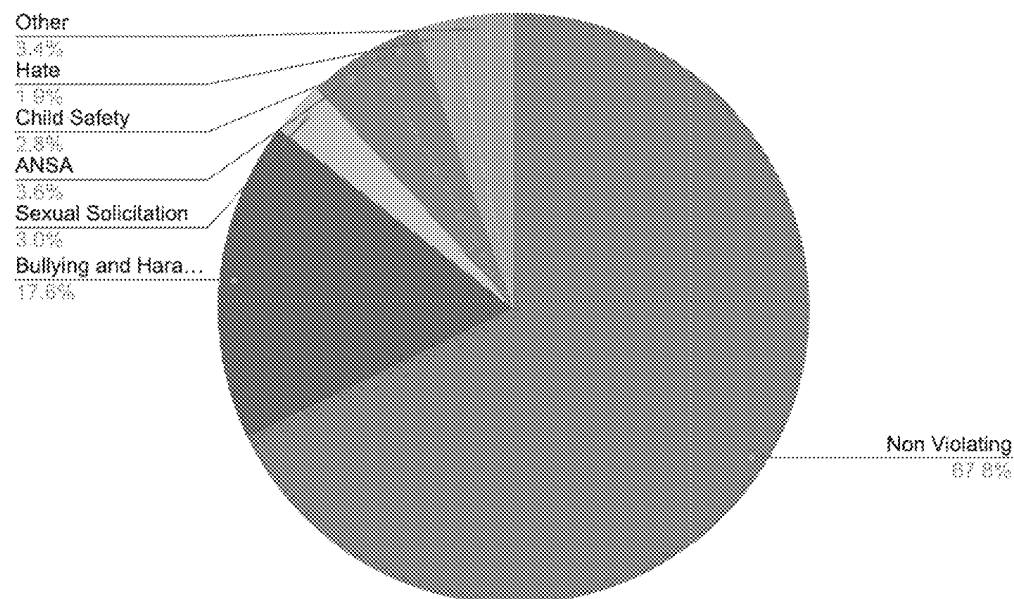
Tracked here...

1.3M Reports per week

735K Reviewed (56% of submitted reports)*

256K Enforced (35% of reviewed reports)

28K Account Disables
228K Strikes



*current report review backlog: 3.7m
80% older than 90 days and about to be bulk closed

HIGHLY CONFIDENTIAL (COMPETITOR)
HIGHLY CONFIDENTIAL (ATTORNEY'S EYES ONLY)

META3047MDL-224-00001071
METANMAG-059-00001071

Current Messenger/IGD report review constraints and efforts to address them

Reactive enforcement depends on review of reports

Current resource constraints allow for review of just over half of submitted reports.

Increased reporting friction designed to limit inflow

We have deliberately added steps and difficulty to reporting flow to reduce the overall number, increase % of actionable reports, and prioritize reporting of higher severity violations. Frictionless reporting would require 50-100x current review capacity.

Increasing role of automation, particularly to identify high severity reports for review

Currently we will not action without human review, but we are investing in automation; automated enforcement for B&H could start Q2. Automation investments will help us focus limited human review capacity on the reports most likely to lead to high severity abuses/enforcement. Low severity reports (deprioritised by ML) are closed without review.

Upselling reporting at critical moments

Large investment in ML required. In an e2ee environment, increased actionable reporting is a valuable signal.

Investing in user controls

MWB is focusing on prevention and creating tools to allow users to control their experience.

Messenger and IG Direct Enforcement of Community Standards / Guidelines

Reactive Enforcement Actions: Description, Applicability, Appeals

Number of messages reviewed upon report	Enforcement Action	Description of enforcement action	When is the enforcement action used?	Appeal available?
Primarily human review; 30 messages are viewable to CO	Account Disable	Entire FB or IG Account is disabled.	Severe Community Standards violations (captured in next slides) On escalation only: 1) Disable IG accounts if 2 or more bullying/harassment, hate speech, or violence & incitement violations in IG Direct. For other abuse areas, disable if 5 or more violations in IG Direct. 2) Disable FB Profile under SOP consideration.	Yes No appeal for escalation-only disables
	MSGR or IGD “Read-Only” Feature Limit	Temporary restriction that prevents the violating user from sending messages. Violations outside of messaging do not impact message usage or vice versa. After the first policy violation, the feature limit duration extends. 2 violations = 1 day limit 3 violations = 3 day limit 4 or more violations = 7 day limit; no disable threshold for repeat abuse	Community Standards/Guidelines violations where the abuse doesn’t result in an immediate account disable	No 10

Messaging violations that result in an immediate account disable (FB & IG) through at-scale review

Community Standards violation	Appeal available?
Sexual Exploitation of Children (CEI, IIC)	Yes
Sexual Exploitation of Adults (NCII, Prostitution, Sextortion)	Yes
Human Exploitation (Sex trafficking)	No*
Dangerous Orgs & Individuals	Yes
Fraud & Deception	Yes

*Has been prioritized for the past year but unable to ship due to bandwidth and competing priorities (T84070606)

At-scale review: Reports primarily reviewed at outsourcing locations; no additional context available apart from what's viewable in SRT (review tool).

Through at-scale review, all other Community Standards violations result in the read-only messaging feature limit.

On-escalation: Reports reviewed internally; additional context available through tooling and through individuals who are escalating a case (safety orgs, Comms, Partnerships, etc...)

Community Standards Violations Mapped to Enforcement Actions

Policy Section	Violation	Messenger	IG Direct
Violence and Criminal Behavior	Violence and Incitement	Messaging feature limit	Messaging feature limit during at-scale review; Disable account on escalation-only (2+ strikes)
	Dangerous Individuals & Orgs	Account disable	
	Coordinating Harm & Publicizing Crime	Messaging feature limit	Messaging feature limit during at-scale review; Disable account on escalation-only (5+ strikes)
	Regulated Goods	Messaging feature limit	
	Fraud and Deception	Account disable	

Community Standards Violations Mapped to Enforcement Actions

Policy Section	Violation	Messenger	IG Direct
Safety	Suicide & Self Injury	SSI specific checkpoint with resources	SSI specific checkpoint with resources; Disable account on escalation-only (5+ strikes)
	Sexual Exploitation of Children	Account disable	
	Sexual Exploitation of Adults		
	Bullying and Harassment	Messaging feature limit	Messaging feature limit during at-scale review; Disable account on escalation-only (2+ strikes)
	Human Exploitation	Account disable	
	Privacy Violations & Image Privacy	Messaging feature limit	Messaging feature limit during at-scale review; Disable account on escalation-only (5+ strikes)

Community Standards Violations Mapped to Enforcement Actions

Policy Section	Violation	Messenger	IG Direct
Objectionable Content	Hate Speech	Messaging feature limit	Messaging feature limit during at-scale review; Disable account on escalation-only (2+ strikes)
	Violence and Graphic Content		Messaging feature limit during at-scale review; Disable account on escalation-only (5+ strikes)
	Adult Nudity & Sexual Activity		
	Adult Sexual Solicitation & Sexually Explicit Language		
	Cruel and Insensitive		

Community Standards Violations Mapped to Enforcement Actions

Policy Section	Violation	Messenger	IG Direct
Integrity & Authenticity	Account Integrity & Authenticity	N/A - handled outside of messaging	
	Spam	Messaging feature limit	
	Cybersecurity	N/A - handled outside of messaging	
	Inauthentic Behavior		
	False News	Rely on forwarding limits as upstream product lever	
	Manipulated Media		
	Memorialization	N/A - handled outside of messaging	

Improving enforcement on Messenger/IGD: current areas for exploration

- 1) **Disable FB/IG accounts with repeat messaging violations at scale**
Establish an appropriate threshold for repeat violations and launch an appeals system to enable disabling accounts via scaled enforcement. Explore if strikes outside of messaging and messaging strikes should be counted together towards a single disable threshold.
- 2) **Enforcement on group threads**
It is possible to report group members on Messenger/IGD, but it is not possible to report and enforce on a group itself. When the purpose of a group thread is bad (bullying, CEI, terrorism) we should have the ability to enforce on the group itself. More investigation is required to find the appropriate enforcement actions that are E2EE compliant and consistent with private messaging narrative.
- 3) **Launch Support Inbox**
At present, we do not communicate enforcement outcomes to individuals who have reported something in Messenger or Instagram Direct. Adding transparency to the review process and communicating report outcomes could help build legitimacy around our reporting experience.