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

DX0669

IG Youth audit remediations

This document outlines the remediations implemented for the issues surfaced in IG Youth Wellbeing Assurance audit report from Dec 2022.
We completed 17/17 mitigations and submitted them for verification by the internal audit team.

TL;DR

Area	Issue	Status	IA verification
Limiting interactions	***Inconsistent filtering for IIC account recommendations***	Completed	Approved
Limiting interactions	***Inconsistent follow list hiding***	Completed	Approved
Age gating	***Incomplete coverage of recommendation filtering on HIGH Minor-Restricted (RGS) content***	Completed	Approved
Limiting interactions	***Inconsistent effectiveness of IIC content recommendation filtering***	Completed	Approved
Limiting interactions	***Undocumented exemption of accounts***	Completed	Approved
Sensitive Content Control (SCC 2.0)	***Lack of SSI Signal Retention Documentation***	Completed	Approved
Sensitive Content Control (SCC 2.0)	***Inconsistent non-recommendable content flagging***	Completed	Approved
Sensitive Content Control (SCC 2.0)	***Lack of monitoring on content filtering leakage***	Completed	Approved

Sensitive Content Control (SCC 2.0)	***Lack of monitoring on account filtering leakage***	Completed	Approved
Sensitive Content Control (SCC 2.0)	***Missing alert on non-recommendable account prevalence***	Completed	Approved
Sensitive Content Control (SCC 2.0)	***Lack of Logging of Content Demotion***	Completed	Approved
Sensitive Content Control (SCC 2.0)	***Inconsistent content filtering and demotion monitoring***	Completed	Approved
Sensitive Content Control (SCC 2.0)	***Lack of input monitoring of P(Severe) v2 model***	Completed	Approved
Sensitive Content Control (SCC 2.0)	***Lack of performance monitoring of P(Severe) v2 model***	Completed	Approved
Sensitive Content Control (SCC 2.0)	***Incomplete logging of account demotion***	Completed	Approved
Sensitive Content Control (SCC 2.0)	***Lack of account demotion monitoring on Search surface***	Completed	Approved
Age-based Gating, Filtering and Demotion	***Lack of monitoring on age-based recommendation filtering***	Completed	Approved

Limiting interactions

Inconsistent filtering for IIC account recommendations

Implementation POC: [HYPERLINK "mailto:[redacted]@meta.com" \h] (IGWB Teen Safety)
[HYPERLINK "https://docs.google.com/document/d/1syA39kHWmKTMRWvNIiaEhK0ALh_3ApsyLiJw0HLK2AE/edit" \l "heading=h.iar006cvcy8x" \h]

Management Action Plan	**Result**
Perform validation testing upon MaCSA migration and investigate any recommendation leakage of IIC potential violator accounts to teen accounts and vice versa	MaCSA migration was fully rolled out on 2/28 in conjunction with infra migration.
Investigate any potential delays in the propagation of predicted age information from Core Dimensions and mitigate discrepancy, if needed.	Conducted thorough [HYPERLINK "https://docs.google.com/document/d/1G_6JA-g9zGSt6QToVvt921r8Xq4LWEnKablaVLzrpYQ/edit?usp=sharing" \h] on age prediction model landing time. Verified the delay on the predicted age signal is trivial and is within 24hrs. Identified delays on the stated age signal which is [HYPERLINK "https://www.internalfb.com/diff/D43781171" \h]
Implement remediation to further limit exposure between teens and potential violators through AYMF recommendations.	Migrated to an advanced filtering infra (SAKE). With the migration, we are using state-of-art filtering infra and achieved best available performance in integrity enforcements. **Final outcome:** 2.4% potential violators recommended to teens; 0.9% teens recommended to violators. See further details on the remaining gap in [HYPERLINK "https://docs.google.com/document/d/1M_7o_qJi3skgSV1y7GW2qXHdelOO2ZFCsmDNLIH-KUU/edit" \h]

	See monitoring in [HYPERLINK "https://fburl.com/unidash/yln15v86" \h]

Inconsistent follow list hiding

Implementation POC: [HYPERLINK "mailto:[redacted]@meta.com" \h] (IGWB Teen Safety)

Management Action Plan	**Result**
Perform validation testing upon MaCSA migration and investigate teen accessibility to discover IIC potential violators via follower and following lists of other accounts and vice versa.	MaCSA migration was fully rolled out on 2/28 in conjunction with infra migration. [HYPERLINK "https://fburl.com/daiquery/uzxqdr2t" \h]
Investigate any potential delays in the propagation of predicted age information from Core Dimensions and mitigate discrepancy, if needed.	Conducted thorough [HYPERLINK "https://docs.google.com/document/d/1G_6JA-g9zGSt6QToVvt921r8Xq4LWEnKablaVLzrpYQ/edit?usp=sharing" \h] on age prediction model landing time. Verified the delay on the predicted age signal is trivial and is within 24hrs. Identified delays on the stated age signal which is [HYPERLINK "https://www.internalfb.com/diff/D43781171" \h]
Implement remediation to further limit teens' exposure to IIC potential violator accounts via follow list and vice versa.	No gaps are found on the server-side. We fixed several client impression logging issues ([HYPERLINK "https://www.internalfb.com/diff/D43696934?entry_point=16" \h], [HYPERLINK "https://www.internalfb.com/diff/D44102324?entry_point=16" \h]4)on Android and updated the audit query. (Thanks to [HYPERLINK "mailto:[redacted]@meta.com" \h] from Instagram Graph Experiences team for implementing Android logging fixes and providing support identifying the correct impression events of the follow list impression

	table!) **Final outcome:** out of ~2.0M IIC potential violators who viewed the follow list surface, 0.03% (6054) were shown at least one unconnected teen account. See monitoring in [HYPERLINK "https://fburl.com/unidash/yln15v86" \h]

Inconsistent effectiveness of IIC content recommendation filtering

Implementation POC: [HYPERLINK "mailto:[redacted]@meta.com" \h] (IGWB Teen Safety)
[HYPERLINK "https://docs.google.com/document/d/1Y8x1AJBNfrAoUpOHQeA3Pf-DSzjdiKeoU9d42WO-hCA/edit" \l "heading=h.iar006cvcy8x" \h]

Management Action Plan	**Result**
Perform validation testing upon Malicious Child Safety Actor (MaCSA) model migration and investigate any leakage of IIC potential violator created content to teens on the recommendation surfaces, and vice versa.	MaCSA migration was fully rolled out on 2/28 in conjunction with infra migration. We have performed the [HYPERLINK "https://fburl.com/daiquery/wfpa5r2r" \h] and saw the number increasing for Explore surface, the explanation is given in the section below.
Implement remediation to further limit teens' exposure to IIC potential violator created content, and vice versa.	Migrated to an advanced filtering infra (SAKE). With the migration, we are using state-of-art filtering infra and achieved best available performance in integrity enforcements. **Final outcome:** 2.6% potential violators content recommended to teens; 5.3% teens content recommended to violators. See further details on the remaining gap in [HYPERLINK "https://docs.google.com/document/d/1M_7o_

	qJi3skgSV1y7GW2qXHdelOO2ZFCsmDNLlH-KUU/edit" \h] See monitoring in [HYPERLINK "https://fburl.com/unidash/yln15v86" \h]

Undocumented exemption of accounts

Implementation POC: [HYPERLINK "mailto:[redacted]@meta.com" \h] (IGWB Teen Safety)

Management Action Plan	**Result**
Reassess the blanket exemption on verified accounts and document the decision and rationale.	Policy approved the removal of the exemption in [HYPERLINK "https://www.internalfb.com/launch_manager/reviews/L1171065INT" \h]
Implement MaCSA model for potential violator identification based on the reassessment decision.	Implementation amended to remove exemption in [HYPERLINK "https://www.internalfb.com/diff/D42983535" \h]

Age based filtering

Incomplete coverage of recommendation filtering on HIGH Minor-Restricted (RGS) content

Implementation POC: [HYPERLINK "mailto:[redacted]@meta.com" \h] (IGWB Teen Safety)
Note: This is already approved by the audit team.

Offline message: "Hi Alex, hope your week is going great! Wanted to let you know that we went through the MAP validation for Minor Restricted RGS- Due to computation limitations on the queries, we prioritized and completed our verification for alcohol, tobacco and diet products classifiers on the Clips surface. We observed a leakage of <1% on a daily basis as well as overall during the first 15 days of February, which we think is immaterial given the volume of overall enforcement. Therefore, we will close this observation, no action is needed from your end □"

Lack of monitoring on age-based recommendation filtering

Implementation POC: [HYPERLINK "mailto:[redacted]@meta.com" \h] (IGWB Teen Safety)

Management Action Plan	Result
Set up monitoring and alert to track filtering volumes of ANSA & RGS content and get notifications on significant changes.	Dashboard: [HYPERLINK "https://internalfb.com/intern/unidash/dashboard/igwb_age_gating_infra/enforcement_metrics" \h] Alerts: - Age based filtering:[HYPERLINK "https://fburl.com/onedetection/b1npbfwt" \h][HYPERLINK "https://fburl.com/onedetection/b1npbfwt" \h] - Age gating:[HYPERLINK "https://fburl.com/onedetection/qfe5g4ig" \h][HYPERLINK "https://fburl.com/onedetection/qfe5g4ig" \h]
Create a dataswarm pipeline to measure input signal coverage and set up monitoring alert on the coverage.	Pipeline: Alert:[HYPERLINK "https://fburl.com/onedetection/wpwwmcjj" \h][HYPERLINK "https://fburl.com/onedetection/wpwwmcjj" \h] Dashboard:[HYPERLINK "https://fburl.com/widget/1070330443924709" \h][HYPERLINK "https://fburl.com/widget/1070330443924709" \h]

Sensitive Content Control (SCC 2.0)

Lack of SSI Signal Retention Documentation

Legal POC [HYPERLINK "mailto:[redacted]@meta.com" \h] (Data Protection Legal)
Note: This is already approved by the audit team.

Management Action Plan	Result
The Legal team will document the data retention requirement specifically for SSI signal.	Legal provided impact assessment and retention will remain 30 days.

The XI team will implement the data retention period on SSI signals based on updated Legal guidance.	Implementation unchanged based on impact assessment from legal.

Inconsistent non-recommendable content flagging

Implementation POC: [HYPERLINK "mailto[redacted]@meta.com" \h] (Recommendation integrity)

Management Action Plan	**Result**
Investigate and remediate the leakage of content with nudity exceeding the predetermined non- recommendable threshold.	Leakage fixed following SAKE migration[HYPERLINK "https://docs.google.com/spreadsheets/d/1zorSflbfmHcehceAaP7nKtj9K4FEy51GQB56Lle-A_8/edit" \l "gid=329113400" \h]
Set up monitoring and alert on the volume of content exceeding the non-recommendable thresholds.	Dashboard: [HYPERLINK "https://fburl.com/unidash/7417q1ps" \h] Alert: [HYPERLINK "https://www.internalfb.com/onedetection/detector?observer_id=1148124925862659" \h]

Lack of monitoring on content filtering leakage

Implementation POC: [HYPERLINK "mailto[redacted]@meta.com" \h] (IG Relevance Integrity)

Management Action Plan	**Result**
Set up monitoring and alert on the non-recommendable content prevalence across all SCC settings.	Dashboard: [HYPERLINK "https://fburl.com/unidash/gnaiuxn9" \h] Alerts: [HYPERLINK "https://www.internalfb.com/onedetection/detector?observer_id=239193275150808" \h]

Lack of monitoring on account filtering leakage

Implementation POC: [HYPERLINK "mailto:[redacted]@meta.com" \h] (IG Connections Safety)

Management Action Plan	Result
Set up automated alerts to monitor filtering leakage and maintain an on-call runbook.	Dashboard:[HYPERLINK "https://fburl.com/unidash/intub1xa" \h][HYPERLINK "https://fburl.com/unidash/intub1xa" \h] Alerts: [HYPERLINK "https://fburl.com/onedetection/a09b5vja" \h]

Missing alert on non-recommendable account prevalence

Implementation POC: [HYPERLINK "mailto:[redacted]@meta.com" \h] (IG Connections Safety)

Management Action Plan	Result
Investigate discrepancies in impressions count for users with the 'Allow' setting on the 'IG AYMF non-rec prevalence' dashboard.	Dashboard: [HYPERLINK "https://fburl.com/unidash/p3fd26fa" \h] Investigation document; [HYPERLINK "https://docs.google.com/document/d/1e6wgYI_bz_GL46RmTF7C4iuSKZealeQ8NnQrhoK9IVw/edit" \h]
Set up automated alerts to monitor non-recommendable account prevalence on AYMF	Alert [HYPERLINK "https://fburl.com/onedetection/4agcljzq" \h]
Develop and maintain an on-call runbook to address and remediate alerts.	[HYPERLINK "https://www.internalfb.com/intern/wiki/Instagram_Well-being/Connections_Safety/oncall_guide/AYMF_Oncall_Runbook/filter_leak/" \h]

Lack of Logging of Content Demotion

Implementation POC: [HYPERLINK "mailto:[redacted]@meta.com" \h] (IG Relevance Integrity)

Management Action Plan	Result
Migrate late stage demotion onto Soft Action Kaleidoscopic Enforcement (SAKE) platform.	SAKE was not needed for implementing demotion logging.

Implement logging of the real-time demotion score and dynamic threshold.	Dashboard for real-time score and threshold [HYPERLINK "https://fburl.com/unidash/t0azir1h" \h] Table [HYPERLINK "https://fburl.com/scuba/ig_barkeep_lua_stats_aggregation/ujxg2aae" \h] **Note:** When any piece of content gets ranked, we log its integrity score to LUA stats. LUA stats aggregates all these scores to compute the global quantiles on a given classifier and stores them in ig_barkeep_lua_stats_aggregation. It gives us the mapping (integrity_score <-> quantile) mapping. This table (ig_barkeep_lua_stats_aggregation) provides in real-time the scores global distribution of all content. When the demotion event happens, we read from this table to apply the demotion.

Inconsistent content filtering and demotion monitoring

Implementation POC: [HYPERLINK "mailto:[redacted]@meta.com" \h] (IGWB Enforcement & Classification Platform)

Management Action Plan	**Result**
Implement and maintain an on-call shift summary tracker to: *- log alerts with assigned oncall information, and* *- document relevant details on the issue/ resolution to provide context for future alerts*	Runbook added [HYPERLINK "https://www.internalfb.com/intern/wiki/Instagram_Well-being/Enforcements_&_Classifications_Platform/Developer_Resources/Runbook/" \h] Oncall shift summary tracked: [HYPERLINK "https://www.internalfb.com/omh/view/ig_ecp/shift_summary" \h]

Establish SLA to close alerts promptly.	SLA established in runbook [HYPERLINK "https://www.internalfb.com/intern/wiki/Instagram_Well-being/Enforcements_&_Classifications_Platform/Developer_Resources/Runbook/" \l "ecp-team-oncall-slas" \h]

Lack of input monitoring of P(Severe) v2 model

Implementation POC: [HYPERLINK "mailto:[redacted]@meta.com" \h] (IGWB Teen Safety)

Management Action Plan	**Result**

Set up monitoring and alert on ML input signals for the P(Severe) v2 model in production.	Mitigations documented in [HYPERLINK "https://docs.google.com/document/d/1Ujm-ynvi937jFhuT-G1JqHSNtoDMJ9wkDehHn1sBi9w/edit?usp=sharing" \h]
	[HYPERLINK "https://www.internalfb.com/intern/unidash/dashboard/eqp_classifier_monitoring/ig_eqp_adhoc_dashboards_psevere_classifier_monitoring/" \h]
	Alerts for input signals:
	- Profile picture:
	  - [HYPERLINK "https://fburl.com/onedetection/n1zg1kmh" \h]
	  - [HYPERLINK "https://fburl.com/onedetection/29iimffb" \h]
	- User attributes:
	  - [HYPERLINK "https://fburl.com/onedetection/vnmmrj9j" \h]
	  - [HYPERLINK "https://fburl.com/onedetection/kyovvtq4" \h]

	- [HYPERLINK "https://fburl.com/onedetection/m4r97wta" \h] - [HYPERLINK "https://fburl.com/onedetection/98b5brdl" \h] - [HYPERLINK "https://fburl.com/onedetection/6mbqtvwr" \h] - [HYPERLINK "https://fburl.com/onedetection/zmwehktw" \h]

Lack of performance monitoring of P(Severe) v2 model

Implementation POC: [HYPERLINK "mailto:[redacted]@meta.com" \h] (IGWB Teen Safety)

Management Action Plan	**Result**
Set up monitoring and alert on performance (Precision/ Recall) of the P(Severe) v2 model.	Implemented [HYPERLINK "https://fburl.com/unidash/jlrnkbqr" \h] and [HYPERLINK "https://fburl.com/onedetection/tsr2ahsu" \h] of p(severe) precision and recall[HYPERLINK "https://docs.google.com/document/d/1Ujm-ynvi937jFhuT-G1JqHSNtoDMJ9wkDehHn1sBi9w/edit?usp=sharing" \h]
Establish model refresh cadence based on performance monitoring of P(Severe) v2 model.	Documented [HYPERLINK "https://docs.google.com/document/d/1Ujm-ynvi937jFhuT-G1JqHSNtoDMJ9wkDehHn1sBi9w/edit" \l "bookmark=id.g1n2p7gv8zn5" \h] **Note:** The model refresh cadence based on performance monitoring dashboard will kick-in with the release of P(Severe) v3. As of 3/30, we're working on releasing v3 and deprecating v2 because the performance of v2 has already dropped below P75

Incomplete logging of account demotion

Implementation POC: [HYPERLINK "mailto:[redacted]@meta.com" \h] (IG Discovery Safety)

Management Action Plan	**Result**
Set up logging of pagination subrequests in the account demotion logging table.	Adding pagination_token field to table i[HYPERLINK "https://fburl.com/scuba/ig_search_integrity_logger/tfnp450g" \h] in [HYPERLINK "https://www.internalfb.com/diff/D43362239" \h]

Lack of account demotion monitoring on Search surface

Implementation POC: [HYPERLINK "mailto:[redacted]@meta.com" \h] (IG Discovery Safety)

Management Action Plan	**Result**
Set up monitoring and alert for the Search surface on: *- Demotion volume* *- Demotion rate, and* *- Prevalence of non- recommendable accounts on Search across SCC settings.*	Dashboard: [HYPERLINK "https://fburl.com/unidash/enjlc1pw" \h] Alerts: - Demotion volume with SCC breakdown: [HYPERLINK "https://fburl.com/onedetection/qm3c6rua" \h] - Prevalence of non-rec, overall and with SCC breakdown: [HYPERLINK "https://fburl.com/onedetection/aib6gde7" \h] [HYPERLINK "https://fburl.com/onedetection/i45ynh89" \h] **Note:** Team believes demotion volume and rate are the same hence providing only 1 tracking for volume/rating.