Tennessee v. Meta
Case No. 23-1364-IV
PX-01384A

Instagram Youth Well-being Assurance Audit - Phase 1 Audit Report

Company confidential - for intended recipients only

Martin Schneider
Anna Tchernina
Kalyan Upadhyaya
Yanan Li
Pratima Paudel

December, 2022

Assurance Audit provides an objective and independent evaluation of the design and operating effectiveness of internal controls.

Meta

TABLE OF CONTENTS

1. Executive Summary
2. Detailed Results
3. Appendix
   - Communication Channels
   - Audit Report Rating Guidelines
   - Observation Remediation Guidelines

EXECUTIVE SUMMARY

Objective:

Instagram (IG) leadership has publicly committed to raising the industry standards for helping protect youth well-being and safety online. As part of this work, in 2021 and 2022, IG announced a stricter approach to cater recommendations and implement additional processes for youth. The implementations focused on further reducing the teens' exposure to potential violating or sensitive content or accounts in recommendation surfaces (e.g., Reels, Explore), which extends the basic protection provided by the Cross-Meta Integrity (XI) enforcement that takes down violating content and accounts.

Internal Audit (IA) assessed the design and operating effectiveness of the key controls for the IG processes (based on internal guidelines) implemented to further reduce youths' exposure to potential violating or sensitive content and accounts.

Scope:

IA independently assessed the key controls of IG Youth Well-being in the following 3 areas:

- **Age-based Gating, Filtering & Demotion**: Reduce the visibility of potential violating content on connected and unconnected surfaces, e.g., graphic violence, and restricted goods & services.
- **Limiting Interactions**: Protect teens from negative interactions with potential violators, including the disabled 'follow' action.
- **Sensitive Content Control**: Limit exposure of users to sensitive content and accounts in recommendation surfaces, e.g., posts that may be sexually suggestive.

Results Summary: Significant Improvements Needed*

While the potential violating or sensitive content and accounts recommended were notably reduced after the above processes were implemented and the age-based content gating framework was established, based on the testing completed as of 11/30/2022, IA identified 3 High risk and 14 Medium risk observations during the audit. Based on the overall effectiveness of controls for in-scope areas, an "Orange" rating is assigned to this report. The results and observation ratings in this report are based on the design and operating effectiveness of internal controls only. Key observation themes are summarized below:

- Inconsistent filtering on potential Inappropriate Interaction with Children (IIC) accounts and content.
- Incomplete coverage of recommendation filtering on Minor-Restricted Regulated Goods and Services (RGS) content (e.g., diet products) for teens.
- Lack of documented rationale for exemption of verified accounts and Suicide & Self Injury (SSI) signal retention period.
- Missing comprehensive monitoring and alert of filtering volume, leakage, and machine learning performance.

* This rating is assigned in accordance with the IA's report rating guidelines in Slide 18

DETAILED RESULTS AND MANAGEMENT ACTION PLANS

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Limiting Interactions	***Inconsistent filtering of IIC account recommendation*** *The IG Well-being Teen Safety team does not always limit teen accounts from being recommended to the IIC potential violators* and vice versa. As a result, teens' accounts may be recommended to potential violators and vice versa via account recommendation surfaces.* *For example, on a sampled day tested, on the Accounts You May Follow (AYMF) surface, IA identified:* *1) 7.2% (~1.4M out of ~19M) of potential violator accounts in AYMF were recommended to at least one teen.* *2) 4.0% (~3.4M out of ~85M) of AYMF recommendations to potential violators were recommending teen accounts.* ** This control applies to IIC Medium potential violators, defined by the product team as users above 18 without verified tag, having 1 or more Child Sexual Exploitation (CSE) violation in past 90 days or a Community Integrity (CI) classifier for IIC score score greater than or equal to 0.3.*	HIGH	*1. Perform validation testing upon MaCSA migration and investigate any recommendation leakage of IIC potential violator accounts to teen accounts and vice versa.* *2. Investigate any potential delays in the propagation of predicted age information from Core Dimensions and mitigate discrepancy, if needed.* *3. Implement remediation to further limit exposure between teens and potential violators through AYMF recommendations.*	*1&2. 2/28/2023* *3. 3/31/2023†*	*Brian Thomas (Software Engineer, IG Well-being Teen Safety)* *Soukaina Nait Hmid (Software Engineer, IG Well-being Teen Safety)*

† Due to complexity of identifying the root cause of the leakage and the migration from the original IIC classifier to the new MaCSA encryption resilient detection method, the IGWB Teen Safety team has requested a 30-day extension to IA's standard 60-day remediation guideline for High risk observations.

DETAILED RESULTS AND MANAGEMENT ACTION PLANS

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Limiting Interactions	***Inconsistent follow list hiding*** *The IG Well-being Teen Safety team does not always hide teen accounts from IIC potential violators* and vice versa on the "followers and following" surface. As a result, teen accounts may be displayed to potential violators and vice versa via the follow lists.* *For example, on 2022-11-25, IA observed that out of ~2.5M IIC potential violators who viewed the follow list surface, 37% (924K) were shown at least one unconnected teen account.* ** This control applies to IIC Medium potential violators*	HIGH	*1. 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.* *2. Investigate any potential delays in the propagation of predicted age information from Core Dimensions and mitigate discrepancy, if needed.* *3. Implement remediation to further limit teens' exposure to IIC potential violator accounts via follow list and vice versa.*	*1&2. 2/28/2023* *3. 3/31/2023†*	*Brian Thomas (Software Engineer, IG Well-being Teen Safety)* *Soukaina Nait Hmid (Software Engineer, IG Well-being Teen Safety)*

† Due to complexity of identifying the root cause of the leakage and the migration from the original IIC classifier to the new MaCSA encryption resilient detection method, the IGWB Teen Safety team has requested a 30-day extension to IA's standard 60-day remediation guideline for High risk observations.

DETAILED RESULTS AND MANAGEMENT ACTION PLANS

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Age-based Gating, Filtering and Demotion	***Incomplete coverage of recommendation filtering on Minor-Restricted (RGS) content**** *The IG Well-being Teen Safety team has not completely incorporated Minor-Restricted RGS signals into age-based content recommendation filtering enforcement. As a result, age inappropriate content (e.g., alcohol) may get recommended to teens.* ** Minor-restricted RGS content includes diet products, alcohol and tobacco.*	HIGH	*Complete QE testing and re-launch classifiers for age-based filtering on Minor-Restricted RGS content on Reels, Explore and Feed Recs.*	*12/16/2022* *Pending Validation*	*Alexander Shan (Software Engineer, IG Well-being Teen Safety)*

DETAILED RESULTS AND MANAGEMENT ACTION PLANS

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Limiting Interactions	***Inconsistent effectiveness of IIC content recommendation filtering*** *The IG Well-being Teen Safety team does not always limit recommendations of teen created content to the IIC potential violators* and vice versa. As a result, teens' content may be recommended to IIC potential violators via content recommendation surfaces and vice versa.* *For a sampled week tested:* *1) 2.2% (~27K out of ~1.2M) of IIC potential violator created content on the Explore surface was recommended to at least one teen, accumulating ~466K teen impressions.* *2) 1.6% (~83K out of ~5.2M) of contents recommended to IIC potential violators on the Explore surface were created by teens, accumulating ~200K potential violator impressions.* ** This control applies to IIC Medium potential violators*	MEDIUM	*1. 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.* *2. Implement remediation to further limit teens' exposure to IIC potential violator created content, and vice versa.*	*1. 2/28/2023* *2. 3/31/2023*	*Brian Thomas (Software Engineer, IG Well-being Teen Safety)* *Soukaina Nait Hmid (Software Engineer, IG Well-being Teen Safety)*

DETAILED RESULTS AND MANAGEMENT ACTION PLANS

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Limiting Interactions	***Undocumented exemption of accounts*** *The IG Well-being Teen Safety team exempted verified accounts from being classified as IIC potential violators without documented assessment of the risk or rationale. As a result, potential violating verified accounts may not be classified as IIC potential violators.*	MEDIUM	*1. Reassess the blanket exemption on verified accounts and document the decision and rationale.* *2. Implement MaCSA model for potential violator identification based on the reassessment decision.*	*3/31/2023*	*Brian Thomas (Software Engineer, IG Well-being Teen Safety)* *Soukaina Nait Hmid (Software Engineer, IG Well-being Teen Safety)* *Swithin Tauro (Product Policy Manager, IG Product Policy)*

DETAILED RESULTS AND MANAGEMENT ACTION PLANS

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Sensitive Content Control (SCC 2.0)	***Lack of SSI Signal Retention Documentation*** *The X-Meta Integrity team has not collaborated with the Legal team* PRIV **PRIV**	MEDIUM	*1. The Legal team will* **PRIV** *2. The XI team will* PRIV **PRIV** *Legal guidance.*	*1. 3/15/2023* *2. 4/30/2023*	*1. Saverio Puddu (Associate General Counsel, Data Protection)* *Victoria Bradshaw (Lead Product Counsel, Product (Central Platform) - Legal)* *Sonny Carpenter (Associate General Counsel, Privacy Counseling, Integrity)* *2. Boris Belyaev (Software Engineer, X-Meta Integrity)* *Shiou Tian Hsu (Research Scientist, X-Meta Integrity)*

DETAILED RESULTS AND MANAGEMENT ACTION PLANS

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Sensitive Content Control (SCC 2.0)	***Inconsistent non-recommendable content flagging*** *The IG Relevance Integrity team does not always flag non-recommendable media when the classifier scores exceed predetermined thresholds. As a result, users may be exposed to sensitive content via content recommendation surfaces.* *For a sampled day tested, IA identified 0.2% of content with nudity classifier scores exceeding the integrity threshold that was not flagged as non-recommendable therefore, not enforced.*	MEDIUM	*1. Investigate and remediate the leakage of content with nudity exceeding the predetermined non-recommendable threshold.* *2. Set up monitoring and alert on the volume of content exceeding the non-recommendable thresholds.*	*3/31/2023*	*Lucas Saloumi (Software Engineer, IG Relevance Integrity)* *Bryan Sagredo (Software Engineer, XI)*
Sensitive Content Control (SCC 2.0)	***Lack of monitoring on content filtering leakage*** *The IG Relevance Integrity team does not monitor the leakage level of non-recommendable content getting recommended to users. As a result, the team may not promptly identify issues if an excessive number of sensitive content is displayed on recommendation surfaces.*	MEDIUM	*Set up monitoring and alert on the non-recommendable content prevalence across all SCC settings.*	*3/31/2023*	*Lucas Saloumi (Software Engineer, IG Relevance Integrity)*

DETAILED RESULTS AND MANAGEMENT ACTION PLANS

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Sensitive Content Control (SCC 2.0)	***Lack of monitoring on account filtering leakage*** *The IG Well-being Connections Safety team does not monitor the leakage level of non-recommendable accounts getting recommended on the AYMF surface. As a result, the team may not promptly identify issues if an excessive number of sensitive accounts are displayed on AYMF.*	MEDIUM	*Set up automated alerts to monitor filtering leakage and maintain an on-call runbook.*	*3/31/2023*	*Boyi He (Software Engineer, IG Well-being Connections Safety)*
Sensitive Content Control (SCC 2.0)	***Missing alert on non-recommendable account prevalence*** *The IG Well-being Connections Safety team has not set up automated alerts to monitor non-recommendable account prevalence on the AYMF surface. As a result, if the prevalence of potential violating accounts on the surface increases, the issue may not be identified and resolved promptly.*	MEDIUM	*1. Investigate discrepancies in impressions count for users with the 'Allow' setting on the 'IG AYMF non-rec prevalence' dashboard.* *2. Set up automated alerts to monitor non-recommendable account prevalence on AYMF.* *3. Develop and maintain an on-call runbook to address and remediate alerts.*	*3/31/2023*	*Boyi He (Software Engineer, IG Well-being Connections Safety)*

DETAILED RESULTS AND MANAGEMENT ACTION PLANS

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Sensitive Content Control (SCC 2.0)	***Lack of Logging of Content Demotion*** *The IG Well-being Enforcements & Classifications Platform team does not log the real-time demotion score* and dynamic threshold for non-recommendable content. As a result, the demotions applied cannot be measured or compared against the original expectation.* ** demotion score indicates the expected demotion strength*	MEDIUM	*1. Migrate late stage demotion onto Soft Action Kaleidoscopic Enforcement (SAKE) platform.* *2. Implement logging of the real-time demotion score and dynamic threshold.*	*1. 4/30/2023* *2.5/31/2023†*	*Adam Parker (Software Engineering Manager, IG Well-being Enforcements & Classifications Platform)*
Sensitive Content Control (SCC 2.0)	***Inconsistent content filtering and demotion monitoring*** *The IG Well-being Enforcements & Classifications Platform team does not always investigate and address filtering and demotion performance alerts promptly. As a result, content filtering/demotion issues may not be identified or resolved promptly.*	MEDIUM	*1. 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* *2. Establish SLA to close alerts promptly.*	*3/31/2023*	*Adam Parker (Software Engineering Manager, IG Well-being Enforcements & Classifications Platform)*

† Due to complexity of migration to SAKE, collaborative development platform for integrity enforcements across Instagram, the IGWB Enforcements & Classification Platform team has requested a 30-day extension to IA's standard 120-day remediation guideline for Medium risk observations.

DETAILED RESULTS AND MANAGEMENT ACTION PLANS

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Sensitive Content Control (SCC 2.0)	***Lack of input monitoring of P(Severe) v2 model*** *The IG Well-being Teen Safety team does not monitor the input signal coverage of the P(Severe) v2 model* to assess its reasonableness. As a result, low coverage of specific signals (e.g., profile picture signal) may go undetected, leading to deteriorated model output.* ** P(Severe) v2 model is the machine learning model that uses profile element approach to identify violating and non-recommendable profiles.*	MEDIUM	*Set up monitoring and alert on ML input signals for the P(Severe) v2 model in production.*	*3/31/2023*	*Pranav Makhijani (Software Engineer, Machine Learning, IG Well-being Teen Safety)*
Sensitive Content Control (SCC 2.0)	***Lack of performance monitoring of P(Severe) v2 model*** *The IG Well-being Teen Safety team does not monitor the performance (e.g., precision, recall) of the P(Severe) v2 model. As a result, deteriorating model performance over time may go undetected.*	MEDIUM	*1. Set up monitoring and alert on performance (Precision/ Recall) of the P(Severe) v2 model.* *2. Establish model refresh cadence based on performance monitoring of P(Severe) v2 model.*	*3/31/2023*	*Pranav Makhijani (Software Engineer, Machine Learning, IG Well-being Teen Safety)*

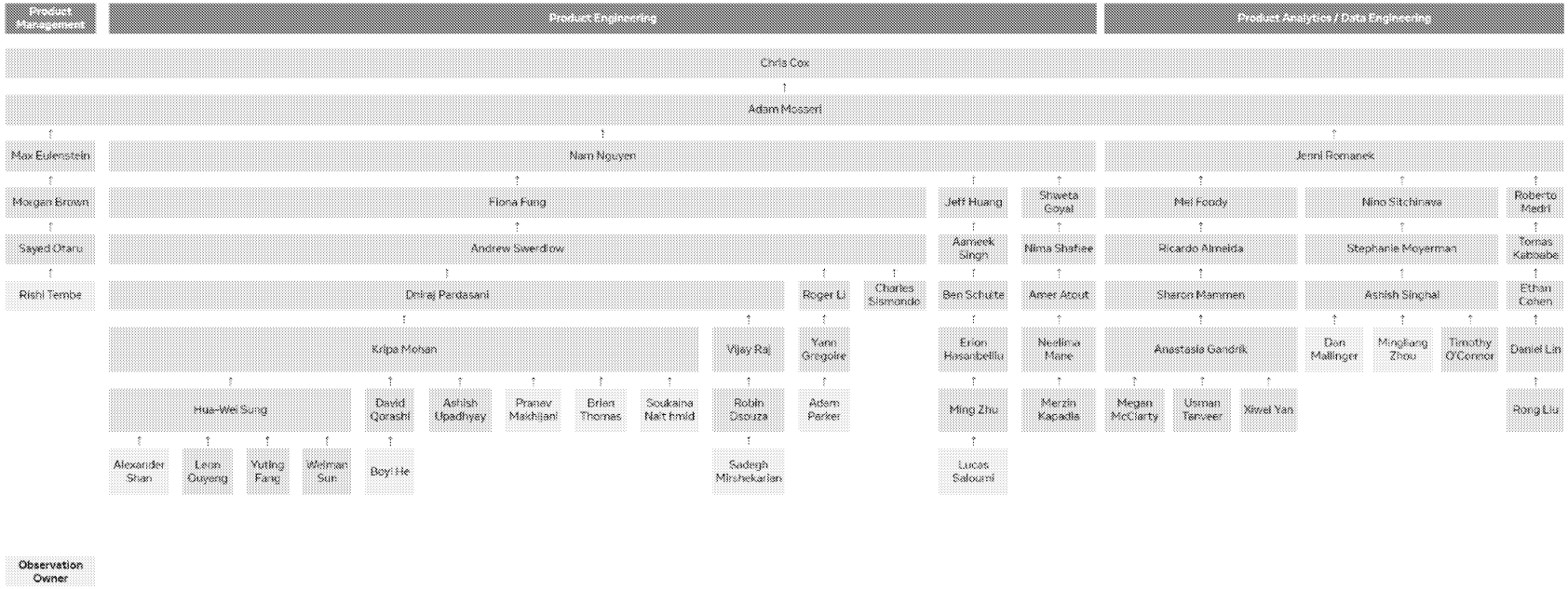
DETAILED RESULTS AND MANAGEMENT ACTION PLANS

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Sensitive Content Control (SCC 2.0)	***Incomplete logging of account demotion*** *The IG Well-being Discovery Safety team does not completely log account demotion on the search surface. As a result, the effectiveness of account demotion cannot be measured or compared against the original expectation.*	MEDIUM	*Set up logging of pagination subrequests in the account demotion logging table.*	*3/31/2023*	*Sadegh Mirshekarian (Research Scientist, IG Well-being Discovery Safety)*
Sensitive Content Control (SCC 2.0)	***Lack of account demotion monitoring on Search surface*** *The IG Well-being Discovery Safety team does not monitor demotion volume on the Search surface. As a result, account demotion leakage may not be identified and resolved promptly.*	MEDIUM	*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.*	*3/31/2023*	*Sadegh Mirshekarian (Research Scientist, IG Well-being Discovery Safety)*

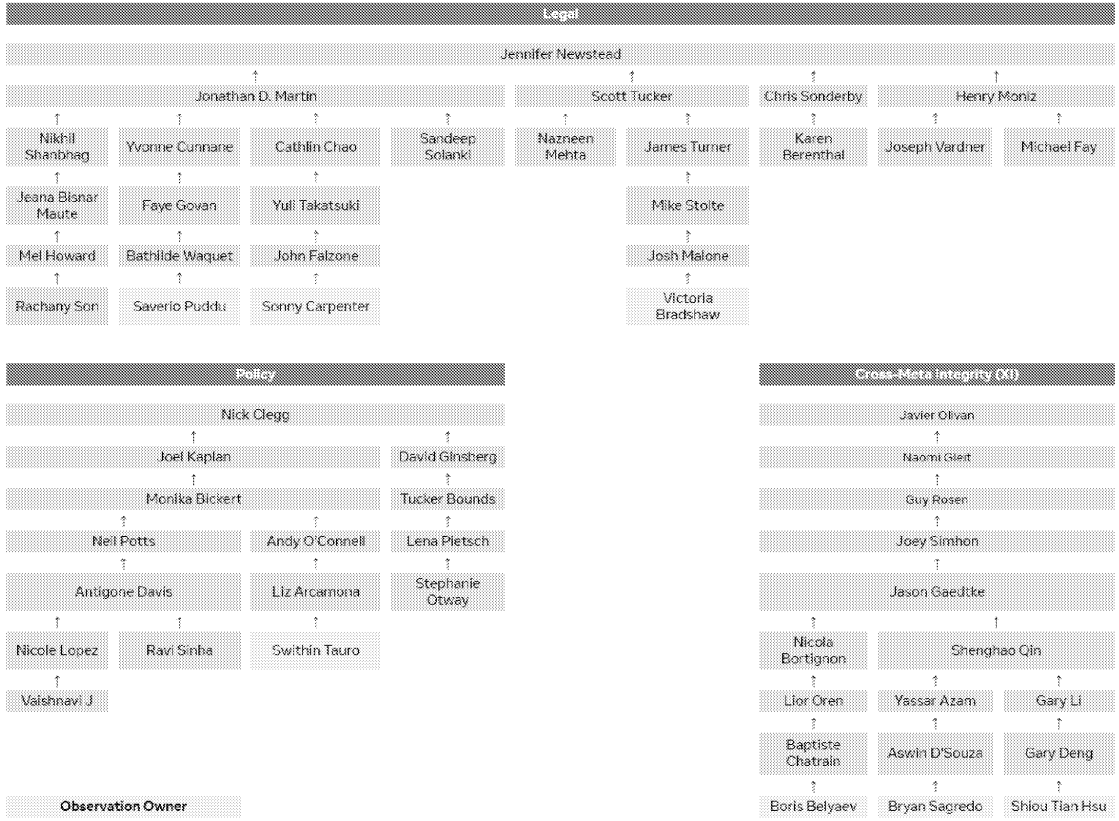
DETAILED RESULTS AND MANAGEMENT ACTION PLANS

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Age-based Gating, Filtering and Demotion	***Lack of monitoring on age-based recommendation filtering*** *The IG Well-being Teen Safety team does not monitor the input signal coverage and volume for age-based filtering of the potential violating Adult Nudity and Sexual Activity (ANSA) & RGS content. As a result, anomalies of filtering may not be detected.*	MEDIUM	1. *Set up monitoring and alert to track filtering volumes of ANSA & RGS content and get notifications on significant changes.* 2. *Create a dataswarm pipeline to measure input signal coverage and set up monitoring alert on the coverage.*	*1. CLOSED* *2.2/28/2023*	*Alexander Shan (Software Engineer, IG Well-being Teen Safety)*

APPENDIX I - COMMUNICATION CHANNEL (1)



APPENDIX I - COMMUNICATION CHANNEL (2)



APPENDIX II- AUDIT REPORT RATING GUIDELINES

Rating	Rating Criteria
Effective	Limited “Low risk” observations: Processes and controls provide reasonable assurance that risks are managed and objectives will be met.
Some Improvements Needed	Limited “Medium risk” observations or a number of “Low risk” observations: Processes and controls require improved effectiveness and efficiency to provide reasonable assurance that risks are managed and objectives will be met.
Significant Improvements Needed	Limited “High risk” observations or a number of “Medium risk” observations: Processes and controls include gaps that require significant improvements to provide reasonable assurance that risks are managed and objectives will be met.
	Several “High risk” observations: Processes and controls include critical gaps that require immediate improvement to provide reasonable assurance that risks are managed and objectives will be met.

OBSERVATION REMEDIATION GUIDELINES

Rating	Rating Definition	Remediation Expectations	AROC Communication
High Risk	Process and control gaps that may result in (i) significant deficiency or material weakness in internal controls, (ii) severe decrease in operational effectiveness/efficiency, (iii) violations in policy, laws or regulations, or (iv) damage to Meta's reputation leading to significant business impact.	Resolve issue within 60 days	Yes
Medium Risk	Process and control gaps that may result in (i) deficiency in internal controls, (ii) significant decrease in operational effectiveness/ efficiency, (iii) non-compliance with policy, laws or regulations, or (iv) damage to Meta's reputation leading to moderate business impact.	Target remediation within 120 days	Yes
Low Risk	Process and control gaps that may result in (i) minor non-compliance with policy, laws or regulations, (ii) deficiency in internal controls, or (iii) decrease in operational effectiveness/efficiency.	Target remediation within 180 days	No