Tennessee v. Meta
Case No. 23-1364-IV
PX-01055A

Child Safety Operations Audit

Company confidential - for intended recipients only

Martin Schneider
Tiffany [redacted]
Jason [redacted]
Eoin [redacted]
Kimi [redacted]
Pratima [redacted]

August 2021

FACEBOOK

Table of Contents

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

Executive Summary

Objective:

Facebook's Child Safety Operations XFN team aims to build a safe community for children by enforcing and reporting on complex and high severity violation types such as Child Exploitative Imagery (CEI), Inappropriate Interactions with Children (IIC), and Minor Sexualization (MS). Appropriate design and operating effectiveness of controls are imperative to mitigating real-world harm and the company's reputational risk, as well as adhering to timely reporting of verified violations to the National Center for Missing and Exploited Children (NCMEC) as mandated by the law. Internal Audit (IA) partnered with Community Product Operations (CPO), Central Integrity (CI), Data Science, Policy, and Legal teams to perform an assurance audit of child safety processes and controls across Facebook, Messenger, and Instagram (IG).

Scope and Audit Summary: *Significant Improvements Needed*

During the audit, a loss of training data and outdated classifier models caused a significant degradation in the precision of IIC Tier 2 classifiers. A SEV was launched, and the Child Safety XFN subsequently completed a 8-week War Room exercise to perform a comprehensive risk assessment and remediation. As the War Room's efforts overlapped with IA's audit scope (i.e., detection, enforcement, banking, and monitoring & reporting), IA's scope was increased to assess the War Room exit criteria and validate the closure of 29 high priority workstreams. **A 'Significant Improvements Needed' rating was assigned based on two High, 12 Medium and 15 Low risk observations.** Six of the 29 observations have been remediated and validated by IA, and the remaining have been assigned management action plans and will be remediated as part of future roadmaps. Notable High and Medium observations are below:

Detection

- No alerting exists for CEI pipeline failures associated with the IG Profiles classifier
- A list of surfaces covered by proactive detection classifiers is not maintained
- A minimum cadence or assessment criteria to re-train CEI classifiers has not been established

Enforcement

- The MAT2 tool used to review potential CEI message threads does not consistently render content during rep review, resulting in a backlog of cases
- Current SLAs for enforcement turnaround time (TAT) are not designed to optimize and measure for high-priority content based on viewport-views, severity types, or classifier scores

Monitoring & Reporting

- The NCMEC backlog includes a significant number of reports (~250K) that are pending submission due to a lack of backlog monitoring
- NCMEC reporting failures are not consistently reviewed or resolved timely
- Facebook's ability to consistently meet reporting obligations to NCMEC may be constrained by team resources

War Room Validation

IA validated the closure of 25 of 29 high priority workstreams but noted instances when the exit criteria were not properly validated or remediation that remain open for completion in H2:

- Alerts to identify anomalies in enforcement volume of recidivist accounts have not been implemented
- Complex entity alerts (e.g., strikes graph, latency, and SEV criteria) for complex entities have not been implemented
- Classifier Estimated Prevalence (CEP) failure detector, which is used to flag errors in IIC T2 victim prevalence calculation, included an incorrect validation code
- Data queries supporting the IIC precision metric in key dashboards incorrectly recognized reported decisions used to create the metrics
- Known bugs affecting complex objects review were not fully resolved

Detailed Results and Management Action Plans

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Enforcement	*System bug affecting CEI content review* The Message Access Tool (MAT) 2 tool used to review CEI message threads on FB and IG does not consistently render content for rep review due to a system bug, leading to a backlog of cases. As a result, violating message threads may not be reviewed and enforced on timely.	H	Implement a fix for the content rendering issue in the MAT2 tool and clear existing case backlog.	10/15/21	Dhiren [redacted] Software Engineer, Child Safety
Monitoring & Reporting	*Insufficient report backlog monitoring* The NCMEC reporting backlog is not consistently reviewed or timely resolved. At the time of testing, IA noted ~247K reports enqueued between May 2017 and July 2021 that were in 'pending submission' status. Additionally, IA noted ~4.5K reports that were assigned with a priority ranking of 0, which is an invalid rating within the report prioritization framework. As a result, valid reports may not be prioritized or timely submitted to NCMEC.	H	Conduct a root-cause analysis of high volume of backlogs and invalid report priority ranking and implement a fix, including clearing the existing backlog. Implement a control to periodically review and resolve reporting backlogs.	10/15/21 10/15/21	Mireille [redacted] Software Engineer, Integrity

Detailed Results and Management Action Plans

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Monitoring & Reporting	*Insufficient reporting failure monitoring* NCMEC reporting failures are not consistently reviewed. Upon reviewing NCMEC reporting failures for the audit period (Sept 2020 - Feb 2021), IA noted 77,829 failures aged between three to 10 months that had not been resolved. As a result, eligible reports may not be timely submitted to NCMEC.	M	Investigate and resolve reporting failures. Implement a control to consistently review and resolve NCMEC reporting failures.	Closed 10/31/21	Mireille [redacted] Software Engineer, Integrity
Monitoring & Reporting	*Insufficient resources for reporting controls* Currently, one FTE is dedicated to the NCMEC reporting process, which may not be sufficient for Facebook to consistently meet reporting requirements or maintain an effective control environment to detect, monitor, and resolve reporting discrepancies (e.g., resolving reporting failures, clearing backlogs, monitoring cancelled reports, resolving reporting SEVs). As a result, Facebook may not consistently meet or sustain compliance with NCMEC reporting requirements.	M	Assess resourcing requirements based on NCMEC reporting responsibilities and controls.	12/31/21	Michael [redacted] Software Engineering Manager, Integrity Mireille [redacted] Software Engineer, Integrity

Detailed Results and Management Action Plans

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Detection	*Lack of alerts for classifier pipeline failures* No alerting exists for CEI pipeline refresh failures associated with the IGPR model. Pipeline table failed to refresh twice during IA's testing on 6/1/21 and 6/29/21 and did not generate scores to detect violating IG profiles. As a result, Facebook may not timely detect and enforce on potential violating IG profiles.	M	Child Safety Engineering will implement automated alerting to notify of any pipeline failures. Child Safety Engineering will update the pipeline infrastructure to reduce refresh failures.	9/30/21 9/30/21	Haotian [redacted] Software Engineer, Child Safety
Enforcement	*No volume alerts for recidivist accounts* The CS team does monitor or have alerts to identify anomalies in enforcement volume of recidivist accounts*. As a result, recidivist accounts may not be restricted from discovering and interacting with minors via the People You May Know (PYMK) functionality. **A recidivist account is defined as an account controlled by the same violator whose previous account(s) were disabled for inappropriate interaction with children.*	M	Child Safety Engineering will implement alerts on recidivist account enforcement volume.	9/30/21	Soukaina [redacted] Software Engineer, Child Safety

Detailed Results and Management Action Plans

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Detection	*Incomplete list of surfaces covered by classifiers* The Child Safety team does not have visibility on the surfaces covered by proactive detection classifiers that identify CEI, IIC, and MS related violating content on the platform. As a result, violating content may not be identified and actioned on.	M	Child Safety Engineering will perform a manual audit of surfaces. Child Safety Engineering will formalize process to maintain the list of surfaces.	9/30/21 9/30/21	Dhiren [redacted] Software Engineer, Child Safety
Detection	*Undefined criteria to retrain classifiers* Classifiers are not re-trained on a periodic basis. In addition, Child Safety (CS) Engineering has not documented a minimum cadence or assessment criteria to re-train classifiers for CEI classifiers. As a result, classifiers may degrade over time* impacting the Child Safety team's ability to detect violating content. *For instance due to changes in imagery or new formats	M	Child Safety Engineering will document the criteria to trigger a classifier to be retrained. Child Safety Engineering will periodically re-train classifiers based on the updated criteria.	8/31/2021 10/31/201	Haotian [redacted] Software Engineer, Child Safety

Facebook company

Company confidential - for intended recipients only

CONFIDENTIAL

META3047MDL-034-00375746

Detailed Results and Management Action Plans

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Detection	*Content not enqueued for human review* Updates to upstream data to comply with the ePrivacy Directive caused some IIC T2* instances of EU actors and threads not to be enqueued. IA noted the enqueuement volume was ~50% of the daily target. As a result, violating accounts may not be reviewed and enforced on timely. **IIC T2 - Inappropriate Interaction with Children Tier 2 Classifier*	M	Perform a root cause analysis for the IIC T2 enqueuement process and resolve identified issues.	8/31/21	Haotian [redacted], Software Engineer, Child Safety Soukaina [redacted], Software Engineer, Child Safety
Enforcement	*Ineffective operational SLAs* Current SLAs for enforcement turnaround time (TAT) are not designed to optimize and measure for high-priority content. SLAs are currently established at 24 hour TAT for complex objects (e.g., Group, Page, Profiles) and 48 hours for simple objects (e.g., Photo, Video, Post, Comments), which does not account for severity types, VPVs, or classifier scores. As a result, operational performance and timeliness metrics may not be effective for monitoring.	M	Define and update TAT for higher priority content (e.g., higher classifier score, VPVs, Live content).	12/31/21	Swithin [redacted], Project Manager, Community Product Ops

Facebook company

Company confidential - for intended recipients only

CONFIDENTIAL

META3047MDL-034-00375747

Detailed Results and Management Action Plans

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
War Room	*Incomplete War Room remediation* A War Room remediation item was not completed. Complex entity alerts (e.g., strikes graph, latency, and SEV criteria) have not been implemented. As a result, complex entity key metrics may not be appropriately monitored for failures or spikes.	M	Implement alerts leveraging strikes graph. Establish an implement SEV criteria for complex entities. Implement alerts for latency.	Closed 10/31/21 12/31/21	Shilpa [redacted], Software Engineer, Child Safety
War Room	*Inadequate validation of War Room exit* FB Classifier Estimated Prevalence (CEP) failure detector, which is used to flag errors in IIC T2 victim prevalence calculation, included an incorrect validation code. The passing condition was erroneously hard coded as PASS = 0 in the upstream Integrity Measurement & Insights (IMI) table, causing incorrect returned results. As a result, CEP validation failures may not be identified and escalated in a timely manner.	M	Fix the CEP query to leverage XDB tables that will accurately produce alerts.	Closed	Soukaina [redacted], Software Engineer, Child Safety

Facebook company

Company confidential - for intended recipients only

CONFIDENTIAL

META3047MDL-034-00375748

Detailed Results and Management Action Plans

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
War Room	*Inadequate validation of War Room exit* In data queries supporting the IIC precision metric, "not a groomer" decisions (negative label) were erroneously counted as "groomer" (positive label), which caused the dashboard to incorrectly recognize reported decisions used to create the metrics. As a result, the precision metric may be inflated*, and the IIC team may track and make decisions on incorrect performance actions. *The dashboard showed precision at ~90% however IA noted precision calculated by the IIC T2 IG specific dash showed ~70%	M	Update the text extraction method in the source query to remediate the counting issue.	Closed	Dhiren [redacted] Software Engineer, Child Safety

Facebook company

Company confidential - for intended recipients only

CONFIDENTIAL

META3047MDL-034-00375749

PX-01055A, Page 10 of 22

Detailed Results and Management Action Plans

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
War Room	*Incomplete War Room remediation* War Room remediation for eliminating bugs affecting complex objects review was not completed. The remaining bugs include: 1. One of the alerts "number of child safety complex objects jobs enqueued drops/raises" does not have appropriate filters applied within OneDetection (alerting platform). 2. Time periods monitored for source_name and creation_source differ. Source_name is monitored for 2 week time periods, whereas creation_source is monitored for 1 week time periods. 3. The dashboard's query for the "job creation" metric includes deprecated queues that should be excluded. As a result, complex object alerts may not be effective, and dashboards may not be accurate or reflect the correct time period.	M	Resolve remaining bugs affecting complex objects.	Closed	Shilpa [redacted] Software Engineer

Facebook company

Company confidential - for intended recipients only

CONFIDENTIAL

META3047MDL-034-00375750

Detailed Results and Management Action Plans

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
MMS Banking	*Insufficient documentation for banking alert changes* Changes to banking alerts are not supported by documented rationales. A quarterly review of banking alerts is performed; however, changes made to banking alerts in OneDetection are not supported by evidence or historical context. As a result, banking alerts may be changed inappropriately and/or historical changes may not be traceable.	L	Document and retain evidence/rationale for changes made to banking alerts during the quarterly refresh process.	10/31/21	Shilpa [redacted], Software Engineer
War Room	*Incomplete War Room remediation* Banking SEV criteria per the War Room Banking Project Plan have not been implemented. As a result, SEVs may not be created and investigated when banking metrics fall out of expected bounds.	L	Complete the implementation of banking SEV criteria as defined in the Banking War Room Project Plan.	9/30/21	Shilpa [redacted], Software Engineer

Facebook company

Company confidential - for intended recipients only

CONFIDENTIAL

META3047MDL-034-00375751

PX-01055A, Page 12 of 22

Detailed Results and Management Action Plans

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
War Room	*Incomplete War Room remediation* SEV criteria for production intervention (e.g., Search and PYMK) have not been formally established. As a result, SEVs may not be created and investigated when product intervention metrics fall significantly out of expected bounds.	L	Establish and implement product intervention SEV criteria.	9/30/21	Soukaina [redacted], Software Engineer, Child Safety
Monitoring & Reporting	*Insufficient reporting access review* Access to cancel enqueued NCMEC reports is not periodically reviewed. Access is not revoked when a user(s) no longer requires access. IA noted nine users from the Sales Analytics Team with access to cancel NCMEC reports, which were temporarily granted during the COVID-19 business continuity operations. As a result, inappropriate cancellations of NCMEC reports may occur.	L	Remove access for the identified nine Sales Analytics individuals. Implement a periodic user access review to ensure reporting access is controlled.	9/30/21 9/30/21	Mireille [redacted] Software Engineer, Integrity

Detailed Results and Management Action Plans

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Monitoring & Reporting	*Insufficient report cancellation review* Cancelled NCMEC reports are not reviewed periodically to ascertain the validity of cancellations. As a result, invalid report cancellations may be undetected and not reported to NCMEC.	L	Perform periodic reviews of NCMEC report cancellations to identify and resolve invalid report cancellations.	10/31/21	Mireille [redacted], Software Engineer, Integrity
Monitoring & Reporting	*Lack of human review tag for NCMEC reports* Reports submitted to NCMEC through the Single Review Tool (SRT) did not include a tag to indicate a human review as required by NCMEC. IA noted 32 out of 583 sampled violations that were not marked as human reviewed. As a result, incomplete or unvalidated reports may be submitted to NCMEC.	L	Conduct a root-cause analysis for violations that were missing human review tag, and implement a control to ensure human review tag is applied prior to report submission.	12/31/21	Mireille [redacted], Software Engineer, Integrity

Detailed Results and Management Action Plans

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
War Room	*Incomplete War Room remediation* The system for receiving user-initiated appeals is currently not functional due to COVID-related resourcing constraints and will remain offline for H2 2021. As a result, appeals system functionality could not be verified during War Room.	L	Contingent on the reinstatement of user appeals, complete system validation to ensure appeals functionality.	1/31/22	Omer [redacted] Software Engineering Manager, Child Safety
Enforcement	*Inability to specify no bank decisions* The 'No Bank' decision tree in SRT currently does not include an option for reps to specify a reason for not banking content (e.g., due to deleted content vs. non-violating content). Deleted content occurs when users take down content, and hence no content is available for reps to bank. As a result, banking performance metrics may be skewed or trigger inaccurate alerts.	L	Implement an option to specify reason for 'No Bank' decision (e.g., due to deleted content vs. non-violating content) to more accurately measure banking performance metrics.	12/31/21	Matt [redacted] Project Manager, Community Product Operations

Detailed Results and Management Action Plans

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
War Room	*Incomplete War Room remediation* The IIC T1 latency metric within the Child Safety dashboard does not only focus on flows that end up with NCMEC outbound nodes. As a result, latency monitoring of IIC T1 key metrics may not be effective.	L	Validate the latency metric for NCMEC nodes once implemented by the I2 team.	11/30/21	Malia [redacted] Data Scientist, Product
Detection	*Cross Family of Apps (FoA) signals not consistently used for detection* The Child Safety team utilizes limited cross-FoA signals in proactive detection for FB & IG CEI complex entities (e.g., Groups and Pages). As a result, recidivist accounts* may not be appropriately identified and restricted from identifying and interacting with minors. **A recidivist account is defined as an account controlled by the same violator whose previous account(s) were disabled for inappropriate interaction with children (applies to CEI, IIC, and MS violating content).*	L	Child Safety Engineering will perform an impact analysis of not using the cross FoA signals, and if required, create a roadmap to address the issue in H1 2022.	12/31/21	Rory [redacted] Software Engineering Manager, Child Safety

Detailed Results and Management Action Plans

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Detection	*Potential Minor Sexualization (MS) violations not enqueued for human review* Alerts for data pipeline failures are not in place. Several Groups flagged for potential MS violations were not enqueued for review due to bugs in the upstream data pipeline between 6/1/21 and 6/10/21. As a result, Facebook may not timely enforce on MS violations.	L	Child Safety Engineering will implement a fix for the MS Group pipeline to enqueue content for review. Child Safety Engineering will implement an automated alerting over MS Group pipeline failure.	Closed 12/31/21	Soukaina [redacted], Software Engineer, Child Safety
Detection	*Content not enqueued for human review* Enqueuing performance by classifiers is not monitored for accuracy. Content identified by the CEI IG Photo classifier that met the human review threshold was not enqueued in SRT. IA's testing of all jobs created by the classifier on 6/14/21 indicated an IG post that met the threshold of 0.6 that was not enqueued for review. As a result, CEI violations may not be reviewed for enforcement.	L	Child Safety Engineering re-ran the classifier and validated that the error did not reproduce, and the content was appropriately enqueued. Child Safety Engineering will establish a monitoring control to identify instances of content meeting thresholds but not enqueued.	Closed 12/31/21	Haotian [redacted], Software Engineer, Child Safety

Detailed Results and Management Action Plans

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
War Room	*Inadequate validation of War Room exit* Exit criteria for the PYMK product intervention remediation (i.e., Actors with > 30% IIC score and has been previously disabled for IIC violations will not be able to access PYMK) was calculated using an inaccurate number of restricted accounts caused by malfunctioning PYMK functionality. As a result, effectiveness of PYMK product intervention for recidivist accounts may be inaccurate.	L	Child Safety Engineering revalidated the remediation closure with correct data.	Closed	Soukaina [redacted] Software Engineer, Child Safety
War Room	*Incomplete War Room remediation* Exit criteria to remediate issue where the IIC escalation flow was not producing reactive reports to NCMEC was not met. As a result, reactive reports may not be timely escalated to NCMEC.	L	Fix escalation issues to validate the production of reactive reports to NCMEC.	Closed	Brian [redacted] Software Engineer, Child Safety

Facebook company

Company confidential - for intended recipients only

CONFIDENTIAL

META3047MDL-034-00375757

PX-01055A, Page 18 of 22

Detailed Results and Management Action Plans

Business Area	Observation/Risk Description	Rating	Management Action Plan	Due Date	Owner
Detection	*Insufficient alerting for proactive detection classifier precision* Currently, no alerts exist to identify anomalies in classifier precision for CEI. As a result, CEI violations may not be accurately detected and enforced.	L	Child Safety Engineering will implement alerting of proxy metrics (i.e., action rate) as a signal for classifier precision performance.	Closed	Haotian [redacted] Software Engineer, Child Safety

Facebook company

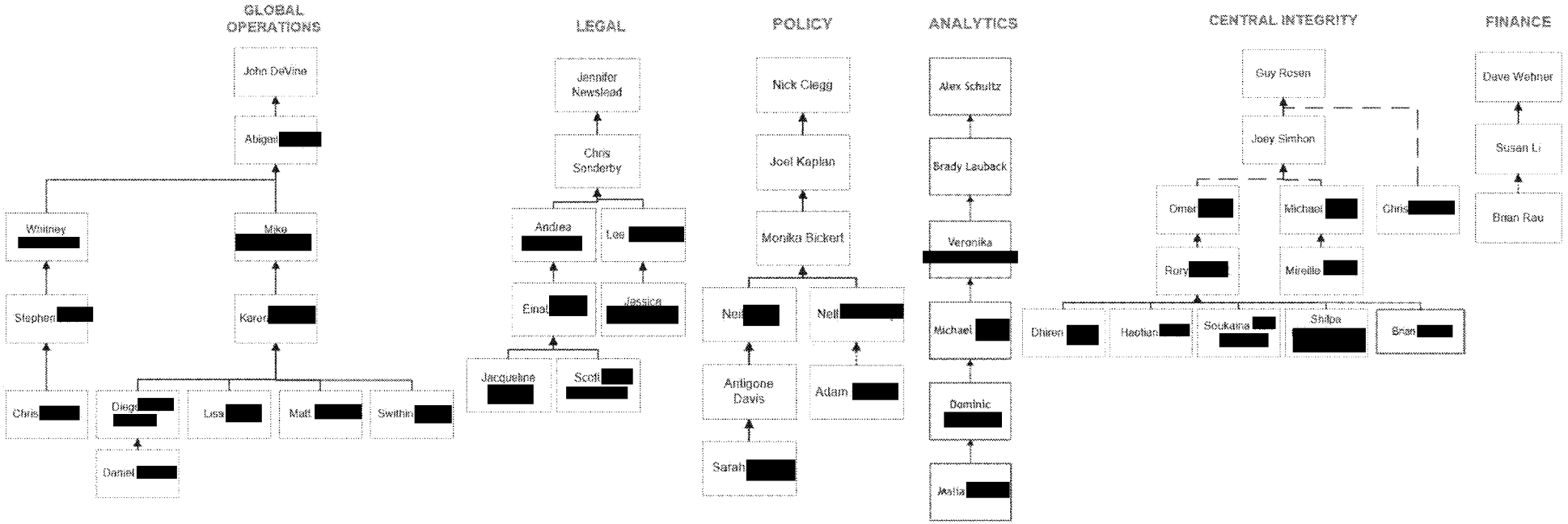
Company confidential - for intended recipients only

CONFIDENTIAL

META3047MDL-034-00375758

PX-01055A, Page 19 of 22

Appendix I - Communication Channel



Facebook company

Company confidential - for intended recipients only

CONFIDENTIAL

META3047MDL-034-00375759

PX-01055A, Page 20 of 22

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.
Unacceptable	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.

Facebook company

Company confidential - for intended recipients only

CONFIDENTIAL

META3047MDL-034-00375760

PX-01055A, Page 21 of 22

Appendix III - 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 Facebook'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 Facebook'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