

Universidade do Minho
Escola de Economia e Gestão

Waleed Khalid

Human and Systemic Factors Affecting Operational Efficiency and Decision Consistency in UK-Regulated Fintech Fraud Management Teams: Mixed Data Analysis

April 2026

Human and Systemic Factors Affecting Operational Efficiency and Decision Consistency in UK-Regulated Fintech Fraud Management Teams: Mixed Data Analysis

Waleed Khalid

UMinho | 2026

Universidade do Minho
Escola de Economia e Gestão

Waleed Khalid

Human and Systemic Factors Affecting Operational Efficiency and Decision Consistency in UK-Regulated Fintech Fraud Management Teams: Mixed Data Analysis

Master's Dissertation
Master's in Business and Management

Work conducted under the guidance of
Dr. António Joaquim Araújo Azevedo

April 2026

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.
Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.
Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho



Attribution-NonCommercial-NoDerivatives

CC BY-NC

https://creativecommons.org/licenses/by-nc-nd/4.0/

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

DECLARATION ON THE USE OF ARTIFICIAL INTELLIGENCE (AI) TOOLS

I hereby declare that having acted this academic work with integrity and confirm that I have not resorted to any form of undue use of AI tools, nor to the falsification or omission of information, sources, data, results or processes in any of the stages leading to its elaboration.

I further declare that any use of AI tools was done in a transparent and responsible manner, being duly noted in the work, and that I critically verified the accuracy of the content produced with their support, assuming full responsibility for what I submit.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho, as well as the applicable standards of academic integrity and data protection.

ABSTRACT

Este estudo examina como fatores humanos e sistémicos influenciam a consistência das decisões em equipas de fraude e crime financeiro em empresas fintech reguladas no Reino Unido. Embora os sistemas automatizados de deteção de fraude possam alcançar um forte desempenho técnico, as decisões finais sobre suspeitar, escalar ou atuar sobre alertas continuam a ser tomadas por analistas humanos. Na prática, estas decisões podem variar em função da carga de trabalho, da formação, da clareza dos procedimentos operacionais padrão (SOP), das orientações da gestão, dos precedentes da equipa e do julgamento individual. Esta investigação aborda a lacuna entre forte capacidade de deteção técnica de fraude e a inconsistência na tomada de decisão humana, medindo a consistência com que os analistas avaliam cenários de fraude idênticos e explorando os fatores associados a essa consistência.

O estudo utiliza um desenho de métodos mistos, combinando um inquérito anónimo baseado em cenários com entrevistas semiestruturadas. O inquérito foi respondido por 33 profissionais de fraude, AML e crime financeiro que trabalham em ambientes fintech regulados no Reino Unido ou que prestam serviços nesse pais. Foram recolhidos dados sobre carga de trabalho, formação, clareza dos SOP, confiança em sistemas automatizados, vieses cognitivos e respostas a oito cenários de casos de fraude. O acordo entre analistas foi medido coeficiente kappa de Fleiss, enquanto a consistência individual das decisões foi avaliada através de um índice de consistência. A análise de correlação de Spearman foi utilizada para examinar as relações entre a consistência e fatores como formação, carga de trabalho, clareza dos SOP e influência social. Adicionalmente, foram realizadas quatro entrevistas para fornecer contexto qualitativo aos resultados do inquérito.

Os resultados demonstram que a consistência das decisões é limitada. A pressão da carga de trabalho, a formação limitada, SOP pouco claros e a dependência de instruções da gestão ou de precedentes da equipa parecem contribuir para a variabilidade das decisões. O estudo também evidencia que os analistas recorrem frequentemente à intuição e às normas locais da equipa quando as orientações formais são ambíguas. De forma geral, a investigação demonstra que a inconsistência na tomada de decisão em fraude não é apenas uma questão técnica ou individual, mas também organizacional e processual. O estudo contribui para a literatura sobre conformidade em crime financeiro e acordo interavaliadores, fornecendo evidência empírica a partir de um contexto fintech.

ABSTRACT

This study examines how human and systemic factors influence decision consistency in fraud and financial crime teams within UK-regulated fintech firms. Although system generated fraud detection systems can achieve strong technical performance, the final decisions on whether to suspect, escalate, or act on alerts are still made by human analysts. In practice, these decisions may vary depending on workload, training, SOP clarity, managerial guidance, team precedent, and individual judgment. This research addresses the gap between strong technical fraud detection and inconsistent human decision-making by measuring how consistently analysts evaluate identical fraud scenarios and by exploring the factors associated with that consistency.

The study uses a mixed-methods design combining an anonymous scenario-based survey with semi-structured interviews. The survey was completed by 33 fraud, AML, and financial crime professionals working in UK-regulated or UK-serving fintech environments. It collected data on workload, training, SOP clarity, trust in automated systems, cognitive biases, and responses to eight fraud case scenarios. Inter-analyst agreement was measured using Fleiss' kappa, while individual decision consistency was assessed through a consistency index. Spearman correlation analysis was used to examine the relationships between consistency and factors such as training, workload, SOP clarity, and social influence. In addition, four interviews were used to provide qualitative context for the survey findings.

The findings show that decision consistency is limited. Workload pressure, limited training, unclear SOPs, and dependence on manager instructions or team precedent all appear to contribute to variability in decisions. The study also shows that analysts often rely on intuition and local team norms when formal guidance is ambiguous. Overall, the research demonstrates that inconsistency in fraud decision-making is not only a technical or individual issue, but also an organizational and procedural one. The study contributes to the literature on financial crime compliance and inter-rater agreement by providing empirical evidence from a fintech context.

Keywords:

Fintech, Fraud Decision-Making, Human Bias, Cognitive Bias, Systematic Factor, Fleiss' Kappa

Table of Contents

ABSTRACT ..... **IV**

ABSTRACT ..... **V**

KEYWORDS: v

1 INTRODUCTION **1**

2 LITERATURE REVIEW **3**

2.1 LITERATURE SEARCH AND SELECTION STRATEGY: 3

2.1.1 Search sources and keywords *3*

2.1.2 Filters and time scope *3*

2.1.3 Inclusion and exclusion criteria *4*

2.1.4 Selection and finalization of included studies *4*

2.2 THEMATICALLY ORGANIZED LITERATURE REVIEW: 5

2.3 RESEARCH GAP AND MOTIVATION 10

3 RESEARCH DESIGN, DATA COLLECTION, AND ANALYTICAL MODEL **12**

3.1 RESEARCH PHILOSOPHY AND DESIGN 12

3.2 STUDY POPULATION AND SAMPLING 12

3.3 DATA COLLECTION: SURVEY DESIGN AND IMPLEMENTATION 13

3.4 SURVEY STRUCTURE AND CONTENT 14

3.4.1 Section A (Demographics and role): *14*

3.4.2 Section B (Workload and operational practices): *14*

3.4.3 Section C (Decision protocols, trust, and biases): *14*

3.4.4 Section D (Red-flag patterns and email identifiers): *15*

3.4.5 Section E (Scenario-based judgement tasks): *15*

3.5 TECHNICAL AND ADMINISTRATIVE DETAILS 15

3.6 DATA COLLECTION: QUALITATIVE INTERVIEWS 15

3.7 INTERVIEW QUESTIONS AND THEIR LINK WITH THE STUDY 16

3.8 ANALYTICAL MODEL AND DATA ANALYSIS METHODS 20

3.8.1 Inter-analyst agreement (Fleiss' kappa) *20*

3.8.2 Correlation analysis with human and systemic factors *20*

3.8.3 Software and visualization *21*

3.9 ETHICAL CONSIDERATIONS 21

4 CHAPTER 4: PRESENTATION OF RESULTS **22**

4.1 SUMMARY OF KEY FINDINGS 22

4.2 RESPONDENT PROFILE (CONTEXT FOR FINDINGS) 22

4.3 THREE CORE FINDINGS 24

4.4 KEY DESCRIPTIVE FINDINGS: THE CONTEXT OF INCONSISTENCY 24

4.4.1 Systemic Evidence: SOP Clarity and Interpretation 24
4.4.2 Human Evidence: Reliance on Informal Decision Protocols 25
4.4.3 Human Amplifier: Workload Pressure and Cognitive Bias 26
4.4.4 Spearman Correlations Result 28
4.4.5 Scenario-Specific Agreement Patterns 31
4.5 INTERPRETATION: HUMAN VS. SYSTEMIC FACTORS 32
4.5.1 Primary Driver: Systemic Factors (The Root Cause) 32
4.5.2 Mechanism: Human Factors (The Amplifier) 34
4.5.3 Training: The Missing Systemic Lever 35
4.6 QUALITATIVE RESULTS: INTERVIEW DATA STRUCTURE 36
5 CONCLUSION, LIMITATIONS AND FUTURE RESEARCH 37
5.1 CONCLUSION 37
5.2 LIMITATIONS OF THE STUDY 38
5.3 FUTURE RESEARCH 41
6 REFERENCES 43
ANNEXURE A 45
ANNEXURE B 49

1 INTRODUCTION

Digital financial services have transformed the way consumers and businesses move money, but they have also created new opportunities for financial crime. Fraud detection has become a core operational function for fintech firms, regulated banks, and payment institutions, and many organizations now rely on sophisticated artificial intelligence (AI) and machine-learning and rules-based systems to flag suspicious activity. These systems can achieve high levels of technical performance, with some hybrid models reporting precision and recall values around 0.94 and 0.93, respectively (Bertrand et al., 2024) However, the performance of detection technology does not fully guarantee sound operational outcomes, because human analysts remain the final decision-makers on whether to escalate alerts, block accounts, or file suspicious activity reports (SARs).

Recent UK regulatory and industry reporting shows that fraud, financial crime control and suspicious activity reporting are central operation concerns across the financial services sector, reinforncing the practical relevance of this study. (UK Finance Annual Fraud Report, 2025)

Recent regulatory enforcement actions and industry benchmarks reveal an ongoing problem: human inconsistency in fraud decisions, even when red flags are similar or identical. *(UKFIU, SARs Annual Report, 2024)*

The UKFIU's 2023-2024 annual report recorded 872,048 SARs received, 57,081 DAMLs received, and £240.1m in funds denied from DAML requests, showing the scale of intelligence-led intervention in the UK financial crime system. *(UKFIU, SARs Annual Report, 2024)*

For example, regulators have documented cases at firms where structurally comparable alerts led to different outcomes, suggesting heterogeneous application of escalation thresholds and risk appetite.

The FCA's 2024/25 enforcement data also shows continued operational intensity, including 1,456 firm cancellations, 37 Final Notices, 5 criminal convictions, and £186.4m in financial penalties, which highlights the importance of timely detection and escalation processes. *(FCA Enforcement Data, 2025)*

At the same time, a growing body of work on human factors and cognitive biases highlights how professional decisions under uncertainty are shaped by confirmation bias, automation bias, fatigue, and reliance on heuristics. *((UK Finance Annual Fraud Report, 2024)*

Recent regulatory enforcement actions and industry benchmarks reveal an ongoing problem: human inconsistency in fraud decisions, even when red flags are similar or identical. For example, regulators have documented cases at firms where structurally comparable alerts led to different outcomes, suggesting heterogeneous application of escalation thresholds and risk appetite. (Axelsen et al., 2025)

In parallel, research on inter-rater agreement in clinical decision-making, pathology, and compliance auditing has long relied on metrics such as Fleiss' kappa, and intraclass correlation (ICC) to quantify the degree of agreement between professionals. These studies show that, even in tightly defined clinical protocols or quality-assurance frameworks, substantial disagreement can persist if criteria are ambiguous or experience levels differ. By contrast, in the specific domain of fintech fraud and AML analysts, no primary empirical study has systematically measured inter-analyst agreement on standardized fraud case scenarios. This gap is critical, because without quantitative evidence on how consistently analysts apply rules, it is difficult to attribute poor outcomes to system design, training, workload, or regulatory guidance. At the same time, a growing body of work on human factors and cognitive biases highlights how professional decisions under uncertainty are shaped by confirmation bias, automation bias, fatigue, and reliance on heuristics. These effects are especially pronounced in high-pressure environments where decision-makers must process large volumes of alerts in short time windows. In such settings, two analysts may see the same pattern of velocity, high-risk destinations, sanctions-related entities, or round-trip payment flows, yet interpret them differently depending on prior experience, workload, or team norms. Moreover, organizational factors-such as the clarity of standard operating procedures (SOPs), the presence of ambiguous "interpretation room," and the degree to which analysts defer to team precedent or manager instructions are known to influence decision quality in other operational domains. (Hagen et al., 2025)

This research therefore presents the first primary empirical study that measures:

1. Inter-analyst agreement using Fleiss' kappa across a set of standardized fraud scenarios in a sample of UK-regulated fintech analysts; and
2. the associations between training, workload, SOP clarity, and social influence (team precedent and manager instructions) and individual decision-consistency metrics derived from those scenarios.

By combining a cross-sectional, scenario-based survey with quantitative analysis of inter-rater agreement and Spearman correlations, this research moves beyond descriptive case studies and regulatory reports to provide concrete, quantitative evidence on the human and systemic drivers of inconsistency in fintech fraud decision-making. The findings are intended to inform both academic literature on inter-rater reliability in non-clinical operational contexts and practitioner efforts to design more robust fraud and AML workflows, training programs, and supervisory frameworks.

2 LITERATURE REVIEW

2.1 Literature search and selection strategy:

To ensure a focused and relevant review of the academic and professional literature, a systematic yet pragmatic search strategy was adopted. The UK is a particularly relevant setting for this research because it combines a large and mature fintech ecosystem with active regulatory scrutiny, strong enforcement capacity, and repeated policy attention to financial crime and market integrity. *(UK FinTech - State of the Nation, 2019)* The main goal was to identify studies and reports concerning fraud and AML detection, human factors in automated systems-assisted decision-making, and regulatory perspectives model justifiability in financial crime contexts.

2.1.1 Search sources and keywords

The search was conducted using academic databases, institutional repositories, and conference proceedings relevant to information systems, and financial regulation (for example: IEEE Xplore, ACM Digital Library, ScienceDirect, Scopus, and the publisher platforms of journals such as the *International Journal of Intelligence Science*, *CHI Conference on Human Factors in Computing Systems*, and central banking / regulatory authority websites such as the Financial Conduct Authority (FCA), European Central Bank (ECB) and European Banking Authority (EBA)).

Search strings combined terms related to:

- fraud and AML (e.g., “fraud detection”, “anti-money laundering”, “financial crime compliance”, “payment fraud”, “SAR”), and
- human factors and regulation (e.g., “cognitive bias”, “automation bias”, “explainability”, “model justifiability”, “supervisor”, “regulator”, “human-in-the-loop”).
- Machine learning (e.g., “artificial intelligence”, “machine learning”, “hybrid intelligence systems”),

Where possible, Boolean operators (AND / OR) were used to combine keyword groups and refine the result set (for example: ("rule-based fraud detection" OR "hybrid intelligence systems") AND ("fintech" OR "banking")).

2.1.2 Filters and time scope

To keep the literature relevant to contemporary practice and regulatory developments, the search was limited to recent years (roughly 2018–2024/2025), reflecting the period during which fraud

detection and agentic/automated compliance systems have gained traction in the financial sector. Where available, the following filters were applied:

- Document type: peer-reviewed journal articles, conference papers, and technical / policy reports.
- Language: English-language publications only, to ensure accessibility and comparability.
- Domain: works explicitly referring to banking, fintech, AML, fraud detection, or financial crime compliance (excluding non-financial domains unless used for methodological inspiration).

2.1.3 Inclusion and exclusion criteria

A short set of inclusion and exclusion criteria was used to screen titles, abstracts, and, where necessary, full texts:

Inclusion criteria:

- Studies that explicitly address fraud/AML detection in financial or fintech settings.
- Works that discuss human factors, cognitive biases, trust, or "human-in-the-loop" roles in decision-making.
- Regulatory or supervisory analyses focus on model justifiability, explainability in highly regulated environments such as AML-CFT.
- Empirical studies (quantitative, qualitative, or mixed-methods) as well as conceptually grounded analyses that provide clear insights into the socio-technical context of financial-crime systems.

Exclusion criteria:

- Papers that focus only on non-financial domains (e.g., healthcare, transport, cybersecurity outside financial institutions) unless explicitly cited for methodological or theoretical transfer.
- Purely technical algorithm papers with no discussion of human roles, regulatory constraints, or operational workflows.
- Grey literature or vendor-produced white papers that lack clear methodological description, empirical evidence, or independent validation.

2.1.4 Selection and finalization of included studies

After an initial keyword-based search, results were reviewed in two stages:

1. Title and abstract screening to remove clearly irrelevant papers (e.g., those outside the financial domain or not addressing decision-making), and
2. full-text review of the remaining candidates to assess their relevance to the research questions on human consistency, system relied fraud decisions, and regulatory expectations.

Citations were also snowballed from reference lists of key papers (e.g., major AI-driven fraud-detection studies and supervisory-perspective works) to include foundational and methodologically relevant sources that initially did not appear in the keyword search.

Through this iterative process, a manageable set of core studies was selected for detailed discussion and integration into the literature review (see Table X). These studies span:

- AI-driven fraud detection and hybrid architectures (e.g., Abi, 2024; Sethupathy, 2025),
- Human factors and cognitive biases in rule-based decisions (e.g., Hagen et al., 2025),
- Regulatory and supervisory perspectives on model justifiability (e.g., Bertrand et al., 2024; Axelsen & Licht, 2025), and
- Macro-level fraud statistics and regulatory frameworks (e.g. FCA Reports, ECB & EBA, 2024).

This literature-search and selection strategy ensured that the chosen sources are methodologically diverse, contextually relevant, and aligned with the research focus on how systems, human analysts, and regulatory expectations interact in fintech fraud and AML decision-making.

2.2 Thematically organized literature Review:

Having outlined the literature search and selection strategy, the remainder of this chapter presents a systematic discussion of the selected studies, organized thematically rather than chronologically. Each paper is introduced with its core focus, methodology, and key findings, followed by a brief explanation of how it relates to the research questions on system generated fraud detection, human and system collaboration, and decision consistency in financial crime environments. By moving from search strategy to individual-paper analysis, this section builds a coherent narrative that situates the present study within the existing body of work and sets the stage for the identification of the research gap and motivation in the subsequent section.

(Vijayabaskar, 2020)) explore Hybrid Intelligence Systems (HIS) as a paradigm that integrates human expertise with artificial intelligence and robotic process automation to improve complex problem-solving in domains such as risk management and fraud detection. They argue that HIS enhances decision-making accuracy, efficiency, and innovation by combining human domain knowledge, contextual understanding, and ethical reasoning with data-driven insights and computational power. At the same time, the authors highlight significant challenges, including privacy concerns, data security risks, and algorithmic bias, and contend that future developments in explainable AI (XAI), edge AI, and neuro-symbolic AI will be crucial for building transparent, trustworthy, and human-centered HIS architectures in fintech and other regulated sectors.

(Axelsen et al., 2025) present an agentic AI system designed specifically for financial crime compliance in digitally native financial platforms. Developed through an Action Design Research process with a fintech and regulatory stakeholders, the system automates key FCC workflows such as onboarding, monitoring, investigation, and reporting, while emphasizing explainability, traceability, and compliance-by-design. Using an artifact-centric modelling approach, the architecture assigns clearly bounded roles to autonomous agents and supports task-specific model routing and audit logging. Their work contributes a reference architecture and a real-world prototype, illustrating how automation embedded within accountable governance structures can improve transparency and institutional trust in high-stakes, regulated environments.

The 2024 joint report by the European Central Bank (ECB) and the European Banking Authority (EBA) provides large-scale statistics on payment fraud across the European Economic Area, showing that the total value of fraud reached €4.3 billion in 2022 and €2.0 billion in the first half of 2023. The report indicates that most fraud by value is associated with credit transfers and card payments, while card transactions also account for the largest share of fraud volume. It further confirms that strong customer authentication (SCA) requirements under PSD2 are associated with lower fraud rates, especially for card payments within the EEA, and that fraud rates rise significantly when counterparties are located outside the EEA where SCA does not apply. These findings highlight both the scale of the fraud problem and the importance of regulatory-level technical controls in shaping the fraud environment in which fintech fraud analysts operate((EBA Report on Payment Fraud, 2024)

(Hagen et al., 2025) investigate human factors in cybersecurity, focusing on cognitive biases and trust issues that affect how professionals interact with AI tools. Through qualitative interviews with 19 cybersecurity analysts and a comparative analysis of system solutions from major vendors, they find that automation bias (47%) and confirmation bias (37%) strongly influence analysts' trust in system generated alerts, with many expressing skepticism and preferring hybrid human-system models over full automation. The authors argue that high false-positive rates and a lack of explainability amplify these biases, and they propose strategies such as bias-awareness training, and adaptive trust calibration to improve human-system collaboration. Their work is highly relevant to fintech fraud and AML settings, where analysts must interpret system-generated alerts under time pressure and uncertainty.

(Abi, 2025) examines the role of system-driven fraud detection systems in fintech, arguing that traditional rule-based approaches are increasingly inadequate due to high false-positive rates, limited adaptability, and poor scalability. The paper emphasizes hybrid architectures that combine supervised learning, which excels at identifying known fraud patterns, with unsupervised techniques such as clustering and

autoencoders, which detect anomalies and zero-day threats. By integrating these paradigms, fintech platforms can achieve more accurate, adaptive, and real-time fraud detection while reducing manual review burden. Abi also stresses the importance of explainability, privacy preservation, and human-in-the-loop frameworks as essential components for building scalable, ethical, and compliant fraud-detection systems in neobanks, payment gateways, and digital lending platforms.

((Sethupathy, 2025) proposes a scalable, risk-aware rule-based framework for financial fraud detection that integrates supervised learning, unsupervised anomaly detection, and domain-specific risk features within a modular, cloud-native architecture. The framework uses components of machine learning models for supervised classification and anomaly detection, supported by real-time data distributed inference (Kubernetes, ONNX), and explainability tools and counterfactual reasoning. On over 100 million anonymized transactions from multinational banks, the model achieves precision ≈0.94, recall ≈0.93, and F1-score ≈0.933, while reducing median inference latency to 126 ms and achieving a 92% fraud hit rate in the top-1% alerts. Sethupathy argues that such hybrid, explainable, and production-ready architecture offer a blueprint for banks and fintechs seeking to modernize fraud detection while meeting regulatory and operational constraints.

(Goldenfein, 2023) critically examines the regulatory concept of the "human in the loop," arguing that it functions less as a genuine safeguard and more as a flexible legal technology that can obscure distributions of accountability in automated systems. Using the example of an Australian immigration visa-processing platform, he shows how referring to a human officer as the final decision-maker can mask the extent to which automated workflows actually shape outcomes. The chapter contends that simply inserting a human reviewer into a system assisted process does not guarantee fairness, consistency, or meaningful oversight unless the design explicitly clarifies how human and machine roles interact. This critique is directly relevant to the research, as it motivates empirical investigation into how fintech fraud analysts actually interpret system generated alerts and whether their decisions are consistent across structurally similar cases.

(Bertrand et al., 2024) examine how system is entering regulated territory, focusing specifically on financial crime detection and the supervisors' perspective on model justifiability in anti-money laundering (AML). Through scenario-based workshops with 13 supervisors and 6 banking professionals, the study unpacks the socio-technical context and auditing practices of AML supervisors, highlighting how existing compliance requirements clash with system generated opacity. The authors analyse AML obligations and regulatory expectations to identify seven key needs supervisors have for model justifiability, arguing that explanations must serve as reliable, auditable evidence that support regulatory scrutiny rather than purely

technical performance. This work is directly relevant to the research because it foregrounds the regulatory demand for transparency and accountability in system-driven financial-crime tools, reinforcing the importance of studying how human analysts interpret and act on system-generated alerts in ways that remain consistent, defensible, and aligned with supervisory expectations.

Having discussed the selected studies in narrative form, the following table provides a synthesized overview of the key literature used in this research. The table consolidates information on each paper's author(s) and year, research focus, methodology or data source, main findings, and its specific relevance to the present study on system-assisted fraud decision-making, human-analyst consistency, and regulatory expectations. By presenting this information in a structured format, the table enables the reader to quickly compare the contributions of different works, while the preceding literature-review paragraphs provide the detailed contextual explanation. This combination of narrative discussion and tabular summary ensures both depth and clarity in the positioning of the research within the existing body of knowledge.

Table 1, which is shown below summarizes the key findings from all the reviewed research papers for this research article.

Author	**Focus**	**Methodology**	**Key Findings**	**Relation to This Research**
(Vijayabaskar, 2020)	Hybrid Intelligence Systems (HIS) for risk management and fraud detection	Conceptual review of HIS integrating human expertise with AI and RPA	HIS can improve decision-making accuracy, efficiency, and innovation by combining human domain knowledge and ethical reasoning with AI's data processing and automation, but raises concerns about privacy, security, and algorithmic bias.	Supports the broader context of human-system collaboration in fraud detection and highlights the need for explainable and human-centered design, which this research addresses by examining how human analysts interpret system-generated alerts.

(Axelsen et al., 2024)	Agentic AI for financial crime compliance (FCC)	Action Design Research (ADR) with a fintech and regulatory stakeholders; prototype implementation	An agentic AI system automates onboarding, monitoring, investigation, and reporting under an explainable, traceable, compliance-by-design architecture, improving transparency and institutional trust when embedded in accountable governance structures.	Provides a concrete example of how system can reconfigure FCC workflows under regulatory constraints, reinforcing the need to study how human analysts respond consistently (or not) to system-assisted decisions.
(2024 REPORT ON PAYMENT FRAUD, 2024)	Payment fraud statistics in the EEA	Semi-annual reporting by payment service providers; aggregated fraud data across payment instruments	Total fraud value was €4.3 billion in 2022 and €2.0 billion in the first half of 2023; SCA-authenticated transactions show lower fraud rates, especially for card payments within the EEA.	Offers macro-level evidence of the scale and distribution of payment fraud, setting the economic and regulatory context in which your study on fraud-analyst decision consistency is situated.
(Hagen et al., 2025)	Human factors in AI-driven cybersecurity (cognitive biases and trust)	Qualitative interviews with 19 cybersecurity professionals; comparative analysis of AI tools from Microsoft, CrowdStrike, Darktrace, and IBM	Automation bias (47%) and confirmation bias (37%) strongly affect trust in tools; many analysts prefer hybrid human-system models and express skepticism about fully automated alerts.	Directly supports this research; focus on cognitive biases and trust in system-assisted fraud decisions, motivating empirical investigation of how such biases affect consistency among fintech fraud analysts.
(Abi, 2025)	AI-driven fraud detection systems in fintech using hybrid supervised - unsupervised learning	Conceptual and technical overview of hybrid AI architectures for fraud detection	Hybrid systems combining supervised learning and unsupervised anomaly detection improve detection accuracy, reduce false positives, and adapt to new fraud typologies, but require explainability and human-in-the-loop designs.	Grounds this research in the technical literature on high-performance fraud-detection models, while leaving open the empirical question of how human analysts respond consistently to system-generated alerts.

(Sethupathy, 2025)	Risk-aware models for financial fraud detection from big transaction al data	Empirical evaluation of a modular, five-layer framework on >100 million anonymized transactions from multinational banks	The hybrid AI framework achieves precision ≈0.94, recall ≈0.93, and F1-score ≈0.933 with low latency and high fraud hit rate in top-1% alerts; SHAP-based explanations support regulatory transparency.	Demonstrates that high-performing, explainable systems can reduce analyst workload and improve detection, which underscores the importance of studying human decision-consistency in this sample of fintech fraud analysts.
(Goldenfein, 2023)	The "human in the loop" as a regulatory concept in automated decision-ma king	Conceptual and legal analysis using the Australian visa-processing case	The "human in the loop" often functions as a legal technology that redistributes accountability without ensuring genuine human oversight or consistency, potentially masking the real influence of AI.	Critically questions the assumption that human reviewers automatically guarantee consistent, fair, or accountable decisions, thereby motivating our empirical measurement of inter-analyst agreement in fintech fraud cases.
(Bertrand et al., 2024)	Supervisors ' perspective on model justifiability in AML-CFT	Scenario-based workshops with 13 supervisors and 6 banking professionals; analysis of AML obligations	The study identifies seven needs supervisors have for model justifiability and argues that explanations must serve as reliable, auditable evidence for regulatory scrutiny in AML-CFT.	Directly links this research to the regulatory demand for transparent, defensible system-driven decision-making, highlighting why consistent, explainable human decisions are essential for meeting supervisors' expectations.

Table 1: Summary of reviewed Literature

2.3 Research gap and motivation

Existing research demonstrates that system-driven and hybrid fraud-detection systems can achieve high technical performance (e.g., ≈0.94 precision and ≈0.93 recall in risk-aware AI frameworks) and reduce analyst workload, while human-factors and HCI studies highlight the role of cognitive biases, trust, and "human-in-the-loop" designs in AI-assisted decision-making. At the same time, regulatory and supervisory work, such as (Axelsen et al., 2025; Bertrand et al., 2024) stresses the need for model justifiability, transparency, and accountability in financial-crime detection under AML-CFT regimes.

However, despite this progress, a critical gap remains: literature has not systematically measured how consistently human fraud analysts interpret and act on the same SOP-defined signals in fintech

environments. Most studies treat fraud detection as a purely technical problem (optimizing recall, reducing false positives), or only describe human factors qualitatively, without quantifying inter-analyst agreement or linking it to training, workload, SOP clarity, and social influence. Yet regulatory enforcement actions and industry benchmarks show that identical red flags often lead to different outcomes and that only a fraction of investigated cases are escalated to SARs, indicating substantial hidden inconsistency at the human level.

This research addresses that gap by presenting the first primary empirical study that:

1. measures inter-analyst agreement using Fleiss' kappa on a set of standardized fraud case scenarios in a sample of UK-regulated fintech analysts; and
2. examines the relationships between training, workload, SOP clarity, and social influence (team precedent, manager instructions) and individual decision-consistency indices derived from those scenarios.

The motivation for this study is threefold:

- Operational: To help fintechs and AML teams understand and mitigate the drivers of inconsistency in high-volume fraud-review workflows.
- Regulatory: To address expectations for consistent and auditable decision-making by identifying how human and organizational factors influence variability in fraud analysis outcomes.
- Theoretical: To extend the literature on inter-rater reliability and human-system collaboration by moving beyond descriptive case studies and technical performance metrics to quantitative evidence on the human and systemic factors that shape fraud-analyst decisions.

By combining a scenario-based survey, inter-rater agreement metrics, and correlation analysis, this research provides a concrete, data-driven foundation for more robust fraud and AML frameworks, training programmers, and supervisory expectations in today's financial crime landscape.

3 RESEARCH DESIGN, DATA COLLECTION, AND ANALYTICAL MODEL

This chapter outlines the research design, data collection procedures, and analytical model used in this study. The research adopts a mixed-methods, cross-sectional design combining a quantitative survey with qualitative semi-structured interviews. This study focuses on the UK context because UK-regulated financial institutions operate within a system where the FCA, UKFIU, and industry bodies all treat fraud detection, AML reporting, and escalation quality as core governance concerns.(*Financial Conduct Authority Annual Report and Accounts*, 2025) The quantitative component uses a scenario-based judgement task embedded in an online questionnaire to measure decision consistency among fraud/AML analysts in UK-regulated fintech firms. The qualitative component complements the survey data with in-depth interviews to explore how human and systemic factors such as SOP clarity, workload, and team dynamics influence fraud decision-making.

3.1 Research philosophy and design

The study is grounded in an interpretivist-pragmatic philosophy, aiming both to understand the lived experiences of fraud analysts and to identify empirically measurable patterns in decision consistency. Methodologically, it follows a convergent parallel mixed-methods approach, with the survey providing quantitative data on inter-analyst agreement and individual-level consistency, and the interviews offering qualitative/contextual insight into how analysts interpret automated alerts, apply SOPs, and experience workload and social influence. The overall design is cross-sectional, capturing data at a single point in time rather than tracking changes longitudinally.

The research addresses three core questions:

1. How consistently do fraud analysts classify identical case scenarios as suspicious and decide on escalation;
2. To what extent do training, workload, SOP clarity, and social influence correlate with decision consistency; and
3. How do practitioners themselves describe the interplay between automated systems, SOPs, and personal judgment in high-pressure environments.

3.2 Study population and sampling

For this research, participants were selectively approached from fintech organizations that operate within the UK regulatory environment. The inclusion of UK-focused regulatory reports is also justified because

the FCA's enforcement data and annual reports provide an external benchmark for the scale of supervision, cancellations, interventions, and financial-crime enforcement in the environment in which the sampled firms operate. *(FCA Enforcement Data, 2025)* Additionally, respondents were encouraged, where appropriate, to share the survey within their teams to increase participation among relevant professionals. To address the sensitivity of the industry and encourage candid responses, it was explicitly communicated that the study was fully anonymous, with no collection or disclosure of personal or organizational identifiers, thereby ensuring confidentiality and building participant trust. While not all respondents were physically based in the UK, they were employed in companies serving a UK user base and routinely handling UK-regulated fraud and compliance cases. The sampling focused on fintechs with significant operations in the UK, including both firms exclusively operating within the UK and those with broader international presence. Using professional networking platforms and existing industry contacts, the survey was distributed specifically to individuals working in fraud prevention, financial crime, and compliance-related roles. The participants represented a diverse range of experience levels to ensure a varied and balanced sample.

The target population consists of fraud/AML/financial crime analysts and compliance officers working in UK-regulated or UK-serving fintech firms whose roles involve reviewing fraud/AML alerts and making decisions on escalation or SAR/STR submission. Inclusion criteria required that participants be actively engaged in reviewing alerts or cases, have at least some exposures to fraud/AML systems. Participants who did not meet these criteria were excluded from the final analysis.

A convenience sampling strategy was employed, recruiting participants through the researcher's industry contacts and professional networks. This approach enabled access to a targeted group of practitioners in a niche and sensitive domain but limits the statistical generalizability of the findings beyond the sample. The survey collected 33 complete responses, and four semi-structured interviews were conducted with experienced fraud/AML professionals, purposively selected to reflect variation in role, firm size, years of experience, and organizational context.

3.3 Data collection: Survey design and implementation

The primary data collection instrument was an online questionnaire hosted on Qualtrics and designed to capture both descriptive information about participants' works environments and their decision consistency across standardized fraud case scenarios. The survey began with a consent and screening block, in which respondents were informed that the study was part of academic research examining how human and systemic factors influence fraud decision-making and operational efficiency within

fraud/financial crime teams in FinTech companies. The consent text stressed that the survey was anonymous, that no personal names, company identities, or directly identifiable information would be collected or reported, and that responses would be analyzed in aggregated form only. Respondents who did not consent were directed to exit the survey, and their responses were not recorded.

3.4 Survey structure and content

The questionnaire comprised the following blocks, the detail with each survey question is given in Annexure A.

3.4.1 Section A (Demographics and role):

Questions 1–5 captured the respondent's primary role (e.g., fraud analyst, AML analyst, financial crime analyst, compliance officer), years of professional experience in fraud/AML-related roles, approximate number of active customers or users served by their firm, and whether their work involved UK-regulated or UK-located customer cases. These items help contextualize the sample in terms of professional background and organizational scale.

3.4.2 Section B (Workload and operational practices):

Questions 6–13 measured the respondent's workload and operational environment. Items included: typical weekly hours spent reviewing fraud/AML alerts or cases; the number of fraud/AML alerts/cases reviewed per month; average time per "simple" case and per "complex" case; the number of cases handled simultaneously during busy periods; hours of fraud/AML/compliance training received in the last 12 months; the number of SARs/STRs personally submitted in that period; and the extent to which workload pressure was perceived to affect decision quality. These variables serve as key predictors related to workload and capacity in the analytical model.

3.4.3 Section C (Decision protocols, trust, and biases):

Questions 14–24 assessed the influence of formal and informal decision-making factors, as well as cognitive biases and trust in the organization's automated fraud system. Items covered the clarity of analysis/escalation SOPs (Q14), the degree of "interpretation room" left by SOPs (Q15), and the frequency with which written SOPs, team precedent, personal judgment, and manager instructions influenced respondents' decisions (Q16). The block also included items on trust in the automated fraud system (e.g., perceived reliability, predictability, understandability, and comfort relying on the system) and behaviors indicative of automation bias, confirmation bias, fatigue-induced inconsistency, and

reliance on intuition. These items were generally measured on 5-point Likert-type scales (e.g., 1 = Strongly Disagree / Never to 5 = Strongly Agree / Very Often), coded consistently for analysis.

3.4.4 Section D (Red-flag patterns and email identifiers):

Section D focused on how common red-flag patterns are treated in practice. Respondents rated the typical treatment of patterns such as velocity checking, high-risk destinations, sanctions nexus, structuring, and round-tripping on a Likert-type scale (e.g., "rarely", "sometimes", "often"), providing insight into the application of risk-based typologies in daily operations. A separate set of questions asked whether specific "gibberish" style email patterns would be treated as high-risk and whether they would be escalated for further review. These items allow examination of how technical signals such as email structure and domain choice influence escalation decisions.

3.4.5 Section E (Scenario-based judgement tasks):

Section E contained eight standardized fraud case scenarios (Q27–Q34), each presenting a brief customer or account profile and associated transaction or behavioral history. For each scenario, respondents answered three core questions:

- "Suspicious per SOP?" (Yes/No)
- "Escalate to SAR/Compliance?" (Yes/No)
- "Action?" (e.g., release, request info(RFI), Escalate, Immediate block, N/A – I have no prior experience with this type of fraud case).

These scenario-based items form the core input for the primary analytical model, as they enable measurement of both inter-analyst agreement (Fleiss' kappa) and per-analyst consistency against the modal decision across identical cases.

3.5 Technical and administrative details

The Qualtrics deployment recorded standard administration variables such as Start Date, End Date, Response Type, Progress, finished status, Recorded Date, Response ID, and distribution channel for quality control. The survey was distributed in English, and all consent, instructions, and scenario text were written in clear, professional language suitable for an academic research context.

3.6 Data collection: Qualitative interviews

To complement the survey data, four semi-structured interviews were conducted with experienced fraud/AML practitioners who had also completed the questionnaire. The interviews were conducted online (via video-conference platforms) and lasted approximately 45-60 minutes each. Participants were

selected purposively to reflect diversity in years of experience, firm size, and specific focus (e.g., card-based fraud, account-opening fraud, or AML-only roles).

An interview guide structured the conversations around key themes, including:

- How analysts interpret and act on system-generated alerts;
- How written SOPs are applied in practice versus "unwritten" team norms;
- How workload and time pressure influence decision-making and escalation thresholds;
- How managers' instructions and team precedent shape individual decisions; and
- Specific examples of cases where similar-looking alerts were treated differently.

Interviews were transcribed verbatim (with consent). A thematic analysis approach was used to code the transcripts inductively, identifying recurring patterns related to trust in system, reliance on SOPs versus intuition, and the role of workload and social influence. These qualitative insights were used to contextualize the quantitative findings, particularly in interpreting why certain patterns of inconsistency emerged and how practitioners themselves rationalize variations in decision-making.

3.7 Interview Questions and their Link with the study

Table 2 shows all the interview question designed and their references from literature.

No.	Interview Theme	Question text	Theoretical Ground	Relevant Reference
1	Core demographics	What is your primary role? (Fraud Analyst/Specialist, AML/CTF Analyst/Specialist, Compliance Officer, or something else?)	Professional role and domain expertise in fraud/AML.	(Axelsen et al., 2025).
2	Experience	How many years of professional experience do you have in Fraud, AML, or related financial crime roles?	Experience as a proxy for decision-making maturity and pattern recognition.	(*PSFA Standards*, 2025)
3	Firm size / customer base	Approximately how many active customers/users does your firm serve?	Firm scale as a driver of workload and systemic risk exposure.	(Axelsen et al., 2025)
4	Regulatory exposure (UK-focused)	Does your role involve UK-regulated or UK-located customer cases?	Regulatory context as a constraint on operational design.	(*PSFA Standards*, 2025)

5	Workload (monthly alerts)	In a typical month, how many Fraud/AML alerts or cases do you personally review?	Alert volume as a proxy for cognitive load and alert fatigue.	(Hagen et al., 2025)
6	Time per case	On average, how long do you spend on: (a) a simple case; (b) a complex case?	Time allocation as a marker of depth of review and decision quality.	(Axelsen et al., 2025)
7	Concurrent cases	At busy times, how many cases do you handle simultaneously?	Multitasking as a source of cognitive load and automation complacency.	(Parasuraman & Manzey, 2010)
8	Training	In the last 12 months, roughly how many hours of Fraud/AML/Compliance training have you received?	Training as a mediator of capability, consistency, and bias awareness.	(*PSFA Standards*, 2025)
9	SAR/STR activity	In the last 12 months, how many SARs/STRs have you personally submitted? If not, how do you handle cases flagged to SAR/Compliance?	SAR volume as a key FCC output metric and decision-threshold indicator.	(Axelsen et al., 2025)
10	Workload effect on decision quality	How does workload pressure affect your decision quality? Can you give an example of when workload negatively affected a decision?	Time pressure and environmental factors as amplifiers of inconsistency.	(Umemuro Hiroyuki, 2011)
11	Decision-influencing factors	In practice, how often do each of these influence your decisions on a case? (Written SOPs, Team precedent, Personal judgment, Manager instructions)	Negotiated decision-making shaped by formal and informal norms.	(*PSFA Standards*, 2025)
12	Trust in automated system	Overall, how much do you trust your company's AI-driven Fraud/AML system? Why?	Trust and skepticism toward AI-assisted tools.	(Hagen et al., 2025)

13	Perceived system qualities	Do you see the system as reliable, predictable, understandable, and comfortable to rely on?	Dimensions of trust and explainability in AI-assisted decision-making.	Schaefer et al. (2021), automation-trust literature.
14	SOP clarity and subjectivity	How clear are your SOPs? And how much room for personal interpretation do they leave?	Ambiguous rules and organisational culture as drivers of inconsistency.	(Umemuro Hiroyuki, 2011)
15	Low-risk-acceptance behaviour	How often do you accept a "Low Risk" recommendation without a detailed manual check?	Automation-bias-like behaviour and reliance on low-risk flags.	Goddard et al. automation-bias review.
16	Disagreement with system	If your own judgment disagrees with the system's recommendation, do you usually re-check the evidence carefully?	Active monitoring vs. complacency toward AI outputs.	(Parasuraman & Manzey, 2010)
17	Cognitive bias 1 (initial impression)	How often have you stuck with your initial impression even when new minor evidence appeared?	Confirmation-bias-like tendency to maintain first impressions.	(Umemuro Hiroyuki, 2011)
18	Cognitive bias 2 (selective search)	How often do you mainly look for information that supports your first assessment?	Confirmation bias in evidence search.	(*AML Watcher*, 2025)
19	Cognitive bias 3 (stress-induced inconsistency)	How often have you made different decisions on similar cases when tired, under time pressure, or stressed?	Emotional and cognitive-load factors amplifying inconsistency.	(Umemuro Hiroyuki, 2011)
20	Cognitive bias 4 (intuition use)	In borderline cases, how often do you rely more on intuition or "gut feel" than on formal rules or scoring?	Heuristic-driven ("System 1") thinking under stress.	(Hagen et al., 2025)
21	Universal red-flags	How are velocity, high-risk destinations, sanctions nexus, structuring, and round-tripping typically treated in practice?	Established AML/Fraud typologies and red-flag recognition.	AML typology reports, e.g., AML Watcher.

22	SOP coverage gaps (email patterns)	Do your SOPs define "gibberish" or higher-risk email patterns? How would you handle a suspicious email under SOPs?	Situational ambiguity and edge-case coverage.	(*PSFA Standards*, 2025)
23	Scenario-based decision	For each selected scenario: based on SOPs, would this be treated as suspicious? Would you escalate, to whom? What final action would you take?	Scenario-based judgement under SOP-guided conditions.	own typology-based scenario design (aligned with UK Finance, AML typology sources).
24	Scenario-detail drivers	What specific detail in this scenario most influences your decision?	Feature-salience in human decision-making.	Human-factors and behavioral-finance literature.
25	Scenario under stress	If you were more tired or under heavier workload, would your decision in this scenario change?	Cognitive-load implications on threshold-setting.	(Umemuro Hiroyuki, 2011)
26	Experience with similar cases	Have you personally handled a similar case before?	Experience and familiarity as mediators of decision confidence.	(*PSFA Standards*, 2025)
27	Source of inconsistency	What, in your view, is the biggest source of decision-making inconsistency between analysts on similar cases in your team?	Perceived organizational and systemic drivers of inconsistency.	(Umemuro Hiroyuki, 2011)
28	Wishlist for improvement	If you could change one thing (tools, SOPs, training, workload, team structure) to improve decision quality and consistency, what would it be?	Prescriptive suggestions for organizational design.	General literature on organizational learning and safety-culture.

Table 2: Summary of Interview Questionnaire

3.8 Analytical model and data analysis methods

This section details the analytical model and the statistical methods used to address the research questions. The main analyses are based on the 33 complete survey responses, with the interview data analyzed separately to provide qualitative context.

3.8.1 Inter-analyst agreement (Fleiss' kappa)

To quantify decision consistency across analysts, Fleiss' kappa was computed for each of the eight scenarios separately. For each scenario, the following binary outcomes were extracted from the survey data:

- "Suspicious per SOP?" (Yes/No)
- "Escalate to SAR/Compliance?" (Yes/No)

One Fleiss' kappa value was calculated per scenario for each outcome, representing the degree of inter-rater agreement beyond chance across the respondent panel. Kappa values were interpreted using conventional thresholds (e.g., slight, fair, moderate, substantial, almost perfect) to characterize the overall level of consistency in fraud classification and escalation decisions

Per-analyst consistency metrics

Alongside group-level agreement metrics were computed. For each scenario, the modal (majority) decision across all analysts was identified for each outcome (suspicious vs not suspicious, escalate vs not escalate, and action type).

These indices allow comparison between analysts and serve as outcome variables in correlation analyses, helping to link individual decision-making tendencies with background and contextual factors.

3.8.2 Correlation analysis with human and systemic factors

To examine associations between human and systemic factors and decision consistency, Spearman rank correlation coefficients were computed, given the ordinal nature of many predictors and the relatively small sample size ($N \approx 33$). Spearman's ρ was chosen because it is appropriate for Likert-type and interval-like ordinal data and does not assume normality.

The following variables were extracted from the survey and used as predictors:

- Training: Q11 ("Hours of Fraud/AML/Compliance training received last 12 months").
- Workload: Q6–Q7 and Q10 (weekly hours spent reviewing alerts, number of alerts reviewed per month, and number of concurrent cases handled during busy periods).
- SOP clarity and interpretation room: Q14 ("How CLEAR are analysis/escalation SOPs within your company?") and Q15 ("How much room for personal interpretation do SOPs leave?").

- Social influence: Q16_2 ("How often do team precedent influence your decisions?") and Q16_4 ("How often do manager instructions influence your decisions?").
- Bias-related items: Q19–Q24 (items on automation bias, confirmation bias, fatigue-induced inconsistency, and reliance on intuition).

Correlations were computed between each predictor, as well as (secondarily) with suspicious_rate and escalation_rate. Significant correlations are reported with their Spearman ρ and p-values and interpreted in terms of direction and practical implications for fraud team design and policy.

3.8.3 Software and visualization

All data wrangling, analysis, and visualization were conducted in Python. The main libraries used included:

- pandas and numpy for data cleaning and transformation;
- scipy.stats.spearmanr for computing Spearman rank correlation coefficients;
- statsmodels.stats.inter_rater.fleiss_kappa for calculating Fleiss' kappa; and
- seaborn and matplotlib for generating visualizations such as correlation heatmap of spearman correlation..

3.9 Ethical considerations

The study was conducted in accordance with university ethical guidelines and data-protection principles like those under GDPR and the UK Data Protection Act. All participants provided informed consent at the start of the survey, and interview participants were similarly briefed about the purpose of the research, anonymity, and data usage. Survey and interview data were anonymized; no personally identifiable information (e.g., names, company names) was reported in the study. Data was stored in password-protected systems accessible only to the researcher, and all findings are presented in aggregate form without reference to individual respondents.

4 CHAPTER 4: PRESENTATION OF RESULTS

4.1 Summary of Key Findings

This research investigated the extent of decision inconsistency among fraud analysts in UK-regulated fintech firms and examined how human and systemic factors contribute to this inconsistency. These findings are consistent with the wider UK regulatory environment, where the FCA and UKFIU both emphasise early intervention, good-quality reporting, and the need to move from detection to action without delay *(FCA Enforcement Data, 2025; UKFIU, SARs Annual Report, 2024)*

Using a scenario-based survey with 33 respondents, the study measured inter-analyst agreement using Fleiss' kappa and explored relationships between decision consistency and factors such as training, workload, SOP clarity, and reliance on team/manager influence.

4.2 Respondent Profile (Context for Findings)

Demographic	**Key Finding**	**Percentage**
Primary Role (Q2)	Fraud Analyst/Specialist	45.45% (15/33)
	Financial Crime Analyst/Specialist	27.27% (9/33)
	AML/CTF Analyst/Specialist	21.21% (7/33)
	Compliance Officer	6.06% (2/33)
Experience (Q3)	1y1m-3 years	42.42% (14/33)
	3y1m-5 years	39.39% (13/33)
	5y1m-9y11m	15.15% (5/33)
	0-1 year	3.03% (1/33)
Firm Size (Q4)	100k-999k	39.39% (13/33)
	≥5M	30.30% (10/33)
	1M-4.9M	21.21% (7/33)
	0-99k	9.09% (3/33)
Weekly Hours (Q6)	≥40	51.52% (17/33)
	20-29	24.24% (8/33)

	30-39	15.15% (5/33)
	10-19	9.09% (3/33)
Monthly Cases (Q7)	50-199	30.30% (10/33)
	200-499	21.21% (7/33)
	500-999	21.21% (7/33)
	1k-2k	18.18% (6/33)
	<50	9.09% (3/33)
Training Hours (Q11)	0-5	66.67% (22/33)
	6-10	21.21% (7/33)
	≥41	9.09% (3/33)
	21-40	3.03% (1/33)

Table 2: Summary of Demographics of Respondents

The sample represents experienced frontline fraud professionals working in relatively large firms. Nearly half of the respondents reported working 40 or more hours per week reviewing fraud and AML alerts, indicating a substantial workload and sustained exposure to high-pressure decision-making. At the same time, 66.67% of the sample received only 0-5 hours of formal fraud/AML or compliance training over the past 12 months. This combination of intensive workload and minimal structured training suggests that many analysts are expected to operate at the front line of complex financial crime decisions with limited opportunities for guided skill development, policy refreshment, or bias-awareness support. The details of each participant on their demographics can be found in annexure B.

In such an environment, fatigue, time pressure, and cognitive overload are likely to increase, raising the risk of reliance on heuristics, automation bias, and intuitive judgments rather than consistent, rule-based reasoning. The lack of regular training also limits exposure to updated typologies, regulatory expectations, and case-based best practices, which can further contribute to variability in how similar red-flag patterns are interpreted and escalated. Overall, this high-workload, low-training context creates fertile ground for decision inconsistency, both between analysts and within the same analyst over time and underscores the importance of examining how organizational factors such as training investment and workload management shape operational outcomes in automated system-assisted fraud workflows (UK Finance Annual Fraud Report, 2024; UK Finance Annual Fraud Report, 2025)

4.3 Three Core Findings

1. Overall decision consistency is low.

The aggregated Fleiss' kappa across all eight scenarios indicates only slight agreement among analysts:

- Suspicion (SOP): $\kappa = 0.161$
- Escalation: $\kappa = 0.195$
- Action: $\kappa = 0.141$

2. These values fall well below the threshold for "moderate" agreement ($\kappa > 0.40$), confirming that analysts do not consistently reach the same conclusions on identical cases.
3. Inconsistency is most severe in the link between suspicion and action. Decision consistency taken from Escalation to Action declined, indicating worse than chance alignment. Even when analysts agree a case is suspicious, they do not consistently choose the same action.
4. This proves that identifying risk does not translate to consistent action.

4.4 Key Descriptive Findings: The Context of Inconsistency

4.4.1 Systemic Evidence: SOP Clarity and Interpretation

Question	Finding	Percentage
Q14: SOP Clarity	Somewhat clear	48.48% (16/33)
	Somewhat unclear	36.36% (12/33)
	Extremely unclear	6.06% (2/33)
	Extremely clear	6.06% (2/33)
	Neutral	3.03% (1/33)
Q15: Room for Interpretation	A lot	33.33% (11/33)
	A little	30.30% (10/33)
	Moderate	27.27% (9/33)
	None	6.06% (2/33)
	A great deal	3.03% (1/33)

Table 3: SOP Clarity and Interpretation

A key finding from the survey data is that only 48.48% of respondents considered their analysis and escalation SOPs to be at least "somewhat clear", while 42.42% explicitly described them as unclear. This indicates that nearly one-third of analysts struggle to interpret the core rules that are supposed to guide their daily decisions. Moreover, 63% of respondents reported that their SOPs leave at least a moderate amount of room for personal interpretation.

This pattern reveals a systemic level of ambiguity embedded directly into the operational framework: policies are present and documented, but they are not sufficiently specific or contextualized to constrain decision-making consistently. As a result, analysts are effectively expected to fill in the gaps using their own experience, intuition, and situational context, which can vary significantly from person to person and from day to day. In high-pressure, high-volume environments, this ambiguity magnifies the risk of divergent decisions on structurally similar cases, for example, one analyst may escalate a pattern of velocity or structuring while another deems it acceptable because the SOPs do not explicitly prohibit or mandate such behavior.

In combination with workload and limited training, this SOP, level ambiguity becomes a structural driver of inconsistency rather than a neutral background condition. It shifts the burden of consistency from the institution's policy and guidance to the individual analyst, making it difficult to align operational practice with regulatory expectations or promised risk-based frameworks. Recognizing and measuring this degree of systemic ambiguity is therefore central to understanding why identical red-flag patterns can lead to very different outcomes in practice.(UK Finance Annual Fraud Report, 2024)

4.4.2 Human Evidence: Reliance on Informal Decision Protocols

Decision Influence (Q16)	**Often + Very Often**	**Percentage**
Team Precedent (Q16_2)	Often 27.27% + Very often 54.55%	81.82%
Manager Instructions (Q16_4)	Often 42.42% + Very often 15.15%	57.58%
Past Experience (Q16_3)	Often 9.09% + Very often 6.06%	15.15%
Written SOPs (Q16_1)	Often 45.45% + Very often 27.27%	72.73%

Table 4: Reliance on informal decision protocols

A critical finding from the survey data is that, although 72.73% of respondents report relying on written SOPs as a primary decision-making resource, a substantial proportion also depend heavily on manager instructions (57.58%) and team precedent (81.82%) when handling fraud and AML cases. This pattern reveals that formal policies do not operate in isolation; instead, they coexist with powerful informal norms and supervisory expectations that actively shape how alerts are interpreted and escalated. Many analysts

explicitly indicate that manager-issued guidance or established team practices supersede or at least heavily influence their application of the written SOPs, especially in borderline or novel situations.

This interplay between formal and informal decision drivers produces a phenomenon that can be described as "local consistency" within teams but "global inconsistency" across the organization. Within any given team, analysts tend to align their decisions because they share the same manager's instructions and common working routines, leading to relatively uniform behavior around certain red-flag patterns. However, when these team-specific norms differ from one group to another, due to varying managers, risk appetites, or historical cases, similar alerts can be treated very differently across teams, even within the same firm(UK Finance Annual Fraud Report, 2024)

The coexistence of formal SOPs and informal influences thus creates a dual-track system of decision-making, where the organization's official rules struggle to fully standardize practice. In a high-volume, automated system-assisted environment, this can amplify inconsistency because the same automated alert may be escalated by one team while another downgrades it based on unwritten norms rather than explicit policy. Recognizing this dynamic is essential for understanding how organizational culture, managerial guidance, and team-level patterns contribute to the observed variation in fraud-analyst decisions, despite the presence of a formal control framework.

4.4.3 Human Amplifier: Workload Pressure and Cognitive Bias

Question	Finding	Percentage
Q13: Workload Impact on Decision Quality	Very negatively	54.55% (18/33)
	Somewhat negatively	24.24% (8/33)
Q24: Rely on Gut Feel in Borderline Cases	Often	15.15% (5/33)
	Always	6.06% (2/33)
	Sometimes	18.18% (6/33)
Q21: Anchoring Bias (stuck with initial impression)	Often	12.12% (4/33)
Q22: Confirmation Bias (sought supporting info only)	Often	27.27% (9/33)
Q23: Stress-Induced Inconsistency	Often	33.33% (11/33)

Table 5: Workload Pressure and Cognitive Bias

A key insight from the data is that 78.78% of respondents feel that workload pressure negatively affects the quality of their decisions, either “somewhat” or “very much.” This reflects a context in which analysts must review large volumes of alerts under tight time constraints, increasing the likelihood of fatigue, heuristic-driven shortcuts, and reduced attention to detail. In such conditions, the need for speed can outweigh the desire for thoroughness, leading to more minimal review behaviors and a higher chance of both missed risks and unnecessary escalations.

Adding to this pressure, nearly 21.21% of respondents report that they rely on gut feel “often” or “always” when dealing with borderline cases, particularly where SOPs are vague or silent. This reliance on intuition represents a compensatory strategy: when formal rules are unclear or incomplete, analysts turn to experience and implicit knowledge to fill the gaps. While this can sometimes be effective, it also introduces substantial subjectivity into decision-making, as different analysts’ intuitions will not necessarily lead to the same outcomes for structurally similar cases.

Further, about 12.12% of respondents acknowledge that they often experience confirmation bias, sticking with their initial impression of a case and seeking evidence that supports it rather than challenges it, and that their decisions change under stress or fatigue. These human factors are not isolated; they are amplified in environments where SOPs are ambiguous and training is limited. In such settings, workload pressure, reliance on intuition, and cognitive biases interact to increase the probability of inconsistent decisions on comparable red-flag patterns, undermining both operational reliability and regulatory defensibility. This pattern underscores that the observed inconsistency in fraud-analyst behavior is not simply a matter of individual variability, but a systemic issue shaped by the interaction of human factors and weak formal guidance.

4.4.4 *Spearman Correlations Result*

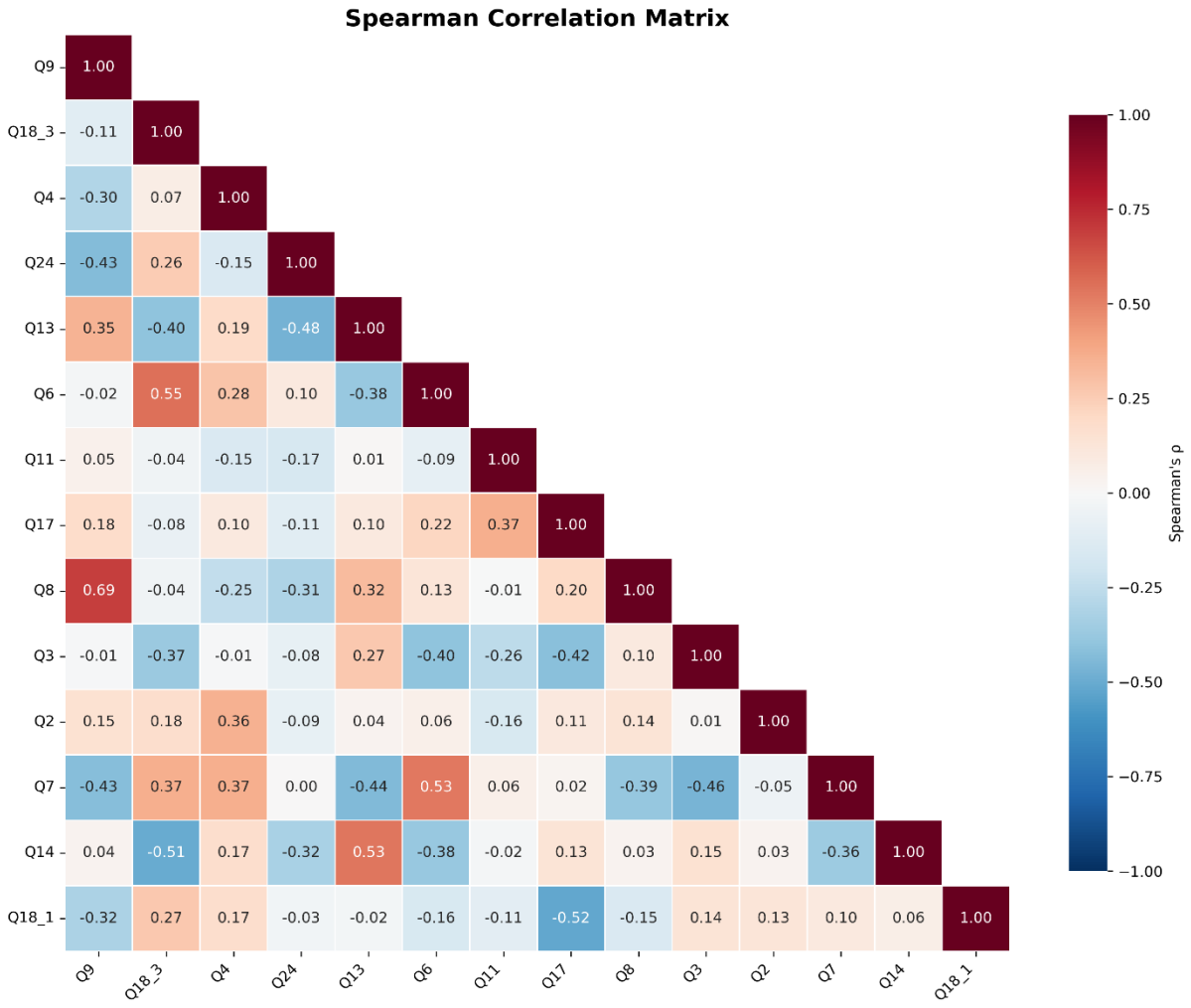


Figure 1: Spearman Correlation Heatmap of 14 Variables

The correlation matrix for these 14 variables shows both strong positive and clear negative relationships between key survey items related to workload, training, experience, decision-making patterns, and SOPs. Overall, many variables are only weakly correlated, which suggests that the selected features capture somewhat distinct aspects of analysts' behaviors and context. However, several moderates to strong correlations stand out, indicating that certain pairs of questions tend to move together (or in opposite directions) across respondents.

Main positive and negative patterns

One of the strongest positive correlations is between Q8 (average time spent on *simple* cases) and Q9 (average time spent on *complex* cases), with a correlation of 0.69. This indicates that analysts who tend to spend more time on complex cases also tend to spend more time on simple ones, suggesting a generally thorough, time-intensive working style across the board. Alongside this, Q6 (number of hours spent reviewing cases weekly) shows a moderate positive correlation with Q7 (monthly fraud/AML cases

reviewed), also around 0.53, which reflects that respondents who report higher working hour per week also report reviewing more fraud/AML cases in total.

On the negative side, Q7 (monthly fraud/AML cases reviewed) and Q3 (years of experience) are negatively correlated with a value of −0.46, implying that more experienced analysts tend to handle fewer fraud/AML cases per month, possibly because of role changes, seniority, or delegation. Similarly, Q7 is negatively correlated with Q14 (SOP clarity) at −0.36, suggesting that analysts with heavier workloads may perceive SOPs as less clear or less supportive, perhaps because they have less time to interpret and rely on them. Another notable negative pair is Q14 (SOP clarity) and Q18_1 (system is reliable), with a correlation of −0.52, indicating that respondents who find SOPs clearer, tend to find that system is not that reliable, and vice versa.

These patterns suggest that workload, experience, and SOP clarity are intertwined in analysts' responses: High case volume is associated with less experience and perceived less clarity, while greater experience is linked to higher clarity. The strong positive link between simple-case and complex-case time also shows that effort and scrutiny are not confined to complex scenarios but spread across case types. At the same time, the presence of moderate correlations but also many near-zero values implies that the feature set is not fully redundant, which is beneficial for modelling: it suggests that the variables retain distinct informational value even though some are clearly related. This balance supports the use of these features in further analysis while still requiring careful interpretation of highly correlated pairs.

Pair	Variable_1	Variable_2	Spearman	P_value	Interpretation	Significance
Avg time per complex case vs Workload Pressure	Q9	Q13	0.346001	0.0486	Weak positive	*
Avg time per simple case vs Workload Pressure	Q8	Q13	0.320269	0.0692	Weak positive	+
Q3 vs Training Hours	Q3	Q11	-0.26077	0.1427	Weak negative	ns
System understandable vs Gut Feel Reliance	Q18_3	Q24	0.255111	0.1519	Weak positive	ns
Q4 vs SOP Clarity	Q4	Q14	0.172052	0.3384	Very weak positive	ns
Trust automated system vs SOP Clarity	Q17	Q14	0.133996	0.4572	Very weak positive	ns
Hours per week reviewing cases vs Training Hours	Q6	Q11	-0.09091	0.6148	Very weak negative	ns

System reliable vs SOP Clarity	Q18_1	Q14	0.063755	0.7245	Very weak positive	ns
Cases per month vs Training Hours	Q7	Q11	0.056557	0.7546	Very weak positive	ns
Q2 (categorical) vs SOP Clarity	Q2	Q14	0.032413	0.8579	Very weak positive	ns

Table 3: Spearman Correlation and P values of highly influential pairs

Table 3 presents the Spearman correlation analysis between key operational variables. The strongest observed association is between average time per complex case and workload pressure ($\rho = 0.346$, $p < 0.05$), indicating a weak but statistically significant positive relationship, suggesting that higher workload pressure is associated with increased time spent on complex cases. Similarly, average time per simple case and workload pressure shows a weak positive trend ($\rho = 0.320$, $p < 0.10$), reflecting marginal significance. While Spearman correlation analysis was conducted to explore relationships between decision consistency and human/systemic factors, most correlations were found to be statistically non-significant ($p > 0.05$).

This lack of statistical significance should be interpreted in the context of the study design rather than as evidence of no relationship. Several factors explain this outcome:

First, the small sample size (N = 33) limits statistical power, making it difficult to detect moderate relationships even when they may exist in practice. In operational and behavioral research, particularly within niche professional populations such as fraud analysts, small samples are common and often constrain inferential testing.

Second, the findings suggest that decision inconsistency is not driven by single isolated factors but rather emerges from the interaction of multiple human and systemic influences, including workload pressure, ambiguity in SOPs, and social decision norms. Such multi-factor interactions are not fully captured through pairwise correlation analysis.

Importantly, although most correlations were not statistically significant, the direction and magnitude of several relationships remain theoretically meaningful. For example, weak positive associations between workload pressure and time spent per case, and negative trends between experience and workload, are consistent with both the descriptive findings and qualitative interview insights.

Therefore, rather than relying solely on statistical significance, this study adopts a triangulation approach, where correlation results are interpreted alongside:

- Fleiss' kappa (low agreement),
- Descriptive statistics (clear patterns),
- Qualitative evidence (strong explanations).

Together, these provide a more robust understanding of the drivers of decision inconsistency in fintech fraud teams. Overall, the results suggest that while workload pressure has some measurable influence on case handling time, most other factors exhibit negligible or statistically insignificant relationships, pointing toward a complex and potentially non-linear interaction structure among the studied variables.

4.4.5 Scenario-Specific Agreement Patterns

The per-scenario data illustrates where inconsistency is most severe:

Scenario	Case Type	SOP Agreement	Escalation Agreement	Action Agreement	Key Observation
Q27	Student + Crypto	78.79	51.52	45.45	SOP highest agreement
Q28	Sanctioned region IP	69.7	45.45	33.33	Low action agreement
Q29	EU hosting provider	54.55	75.76	48.48	Escalation highest
Q30	New merchant (6 weeks)	63.64	60.61	36.36	Low action agreement
Q31	Cash structuring	66.67	63.64	51.52	Moderate consistency
Q32	Round-tripping	57.58	60.61	39.39	Action lowest
Q33	Scam	81.82	54.55	57.58	SOP very high
Q34	Charity Organization	84.85	57.58	57.58	SOP highest overall

Table 6: Scenario-Specific Agreement Patterns

The data reveals a clear and consistent pattern across all eight scenarios: SOP (Suspicion) agreement is consistently the highest among the three decision types, ranging from 54.55% (Q29) to 84.85% (Q34). This suggests that the company's written procedures are relatively effective at guiding analysts to a common conclusion about whether a transaction pattern is *technically* suspicious according to policy. Scenarios involving clearly defined risk indicators, such as UK Professional (Q34) with 84.85% agreement and the software engineer (Q33) with 81.82% show the strongest alignment, indicating that SOPs work well for straightforward, lower-risk profiles.

Despite moderate to good alignment on suspicion, Escalation agreement drops noticeably (ranging from 45.45% to 75.76%), and Action agreement is the lowest across all scenarios (ranging from just 33.33% to 57.58%). This gap is critical. While analysts may agree that a case is suspicious (e.g., Q28 with 69.70% SOP agreement), they disagree on what to actually do about it, only 45.45% agreed on escalation, and just 33.33% agreed on the specific action. The widest gap appears in Q28 (returns/refund pattern), where SOP agreement (69.70%) is nearly double the action agreement (33.33%). This suggests that while policies define "what looks wrong," they leave excessive room for personal judgment on "what to do next," leading to inconsistent case outcomes.

The consistently low action agreement (averaging ~47% across scenarios) is a red flag for operational consistency and regulatory risk. Two analysts reviewing the identical case could arrive at completely different dispositions. The highest action agreement (57.58% for Q33 and Q34) still means over 40% of analysts would take a different action on the same scenario. To improve this, SOPs should be enhanced beyond identifying suspicion to include clear, conditional escalation matrices (e.g., "If X and Y are present, action must be Z"). Without this, bias, workload pressure, and individual intuition will continue to drive final decisions, undermining both fairness and auditability.

4.5 Interpretation: Human vs. Systemic Factors

The central research question was whether human factors (bias, intuition, workload) or systemic factors (SOP clarity, training, organizational structure) are the primary drivers of inconsistency. The results reveal a causal chain:

4.5.1 Primary Driver: Systemic Factors (The Root Cause)

Evidence for Systemic Gaps:

Evidence	Finding	Interpretation
Q14 (SOP Clarity)	42.42% find SOPs unclear	Written guidance is ambiguous
Q15 (Interpretation)	63.63% report moderate+ room for interpretation	SOPs are not prescriptive
Aggregated Fleiss' κ	SOP=0.096, Escalation=0.166, Action=0.105	Low inter-analyst agreement overall
Q29 SOP Agreement	51.52% (lowest)	Grey areas expose SOP failure

Table 7: Evidence for Systematic/ Suspicious-Action Gaps

The " Suspicion-Action Gap ": The decline in percentage of agreement between "Suspicious per SOP" and "Action". It proves that SOPs define what constitutes risk but fail to specify what to do next.

For example, in Q27 (student receiving multiple small payments then crypto):

- 81.82% flagged as suspicious per SOP
- 48.48% chose to escalate
- Only 45.45% escalated, other Actions are split: Request Info 30.33%, Release 12.12%, Block 3.03%, NA / No experience 9.09%

A critical confounder in the "Action" agreement results is the organizational variability in how analysts are authorized to act on suspicious cases. In some firms, fraud analysts have the power to directly block transactions (e.g., freeze accounts, decline payments), while in others they are restricted to escalating cases to a separate team (e.g., compliance or SAR-specialist units etc.). The "Action" variable in the survey therefore combines two distinct roles: block (direct action by the analyst) and escalate (recommendation to another team). When comparing across a mixed sample of both types of environments, this structural difference artificially lowers the observed agreement on "Action," because some analysts are structurally prevented from choosing "Block," even if they would do so under different organizational rules.

However, the core finding of inconsistency holds even when abstracting away this confounder. By focusing only on the universal escalation decision, that is, whether analysts mark a case as "Escalate to SAR/Compliance?" across all eight scenarios, the analysis still yields a low inter-rater agreement (Fleiss' kappa = 0.195). This value indicates only slight agreement beyond chance, demonstrating that analysts disagree on when to escalate suspicious cases, regardless of whether their firm allows them to block or only to escalate. In other words, the inconsistency exists at the level of decision threshold, not merely at the level of firm-specific action authority. Analysts diverge on how suspicious a case must be before it warrants escalation, even within the same scenario and under the same regulatory framework.

This structural-role confounder is therefore important for contextualizing the Action-level results, but it does not invalidate them. Instead, it strengthens the conclusion that the observed inconsistency is rooted in human and systemic drivers, such as workload, SOP ambiguity, and reliance on social norms and intuition, not simply in organizational differences in who executes the final action. By disentangling the escalation decision from the blocking decision, the analysis isolates the point at which judgment, rather than structure, produces variability, underscoring the need for more explicit, shared escalation thresholds and clearer SOPs across the organization.

4.5.2 Mechanism: Human Factors (The Amplifier)

Once systemic gaps exist, human factors become the mechanism through which inconsistency manifests:

Human Factor	**Evidence**	**Percentage**
Team Precedent	Q16_2: Often + Very often	81.82%
Manager Instructions	Q16_4: Often + Very often	57.58%
Gut Feel (Borderline)	Q24: Often + Always	21.21%
Confirmation Bias	Q22: Often	27.27%
Stress-Induced Variation	Q23: Often	33.33%
Workload Impact (Negative)	Q13: Very + Somewhat negatively	78.79%

Table 8: Evidence for Human Factor

A key insight from the data is that, in the absence of clear and tightly scoped SOPs, analysts systematically default to social learning and intuition as their primary decision-making anchors. Many respondents report relying heavily on team precedent and manager instructions, indicating that local norms and supervisory expectations become the de facto guide for handling borderline or ambiguous cases. At the same time, 21.21% of respondents admit to relying on "gut feel" often or always in such situations, showing that personal experience and implicit heuristics fill the gaps left by under-specified policies. This combination of social influence and intuition produces "silos" of consistency within individual teams, where analysts align around shared practices, while generating fragmentation across the broader organization, as different teams and managers develop divergent standards for the same types of alerts. But more importantly more fragmentation in the industry following same guidelines from same authority.

Further evidence for this pattern comes from the high inter-item correlations in the scenario-based responses ($r = 0.69–0.88$), which indicate that analysts apply stable personal thresholds when evaluating cases. A "strict" analyst may escalate a case like Q29 at around 50% perceived suspicion, while a "lenient" analyst may require closer to 80% suspicion before doing the same. Both types of analysts can be highly internally consistent, applying the same standard across multiple scenarios, yet the organization and industry as a whole lacks a shared, explicit threshold that would ensure comparable behavior across analysts.

This reveals a crucial structural issue: the system is strong on individual consistency, but weak on systemic uniformity. The result is that fraud-decision outcomes depend as much on which analyst, team, or manager or firm reviews a case as on the underlying risk pattern itself, creating a hidden source of operational and regulatory risk beneath the surface of high-performing automated system-driven detection systems.

4.5.3 Training: The Missing Systemic Lever

Q11: Training Hours (Last 12 Months)	**Count**	**Percentage**
0-5 hours	21	66.67%
6-10 hours	7	21.21%
21-40 hours	1	3.03%
≥41 hours	3	9.09%

Table 9: Amount of Training hours

The data suggest a clear structural weakness in the current operating environment: 66.67% of respondents received only 0-5 hours of training in the last 12 months, while 51.52% reported working 40 or more hours per week reviewing fraud cases. This combination indicates that many analysts are expected to make high-stakes decisions with limited formal development and under sustained time pressure. In practice, that leaves little opportunity to build shared judgment, refresh knowledge of emerging typologies, or calibrate decision thresholds across analysts.

This matters because training is one of the few levels an organization can use to reduce inconsistency in fraud and AML decisions. When training is sparse, analysts are more likely to rely on personal experience, intuition, or local team norms to fill gaps in SOPs. If heavy workload is added to that, decisions may become faster but not necessarily more consistent, because analysts have less time to compare cases, reflect on edge situations, or align their judgments with colleagues. Over time, this can produce uneven escalation behavior even when the same red flags appear across cases.

Taken together, the findings suggest that training is a potential systemic fix, but current levels are inadequate to standardize decision-making. A more robust training framework would likely need to be continuous rather than occasional, with regular case calibration, scenario-based exercises, and explicit discussion of ambiguous cases. Without that kind of investment, analysts remain dependent on informal learning and individual judgment, which makes consistency difficult to achieve in a high-volume fraud environment.

4.6 Qualitative results: Interview data structure

Interview data for four fraud/AML analysts were captured in a standardized interview-data table, mapping each interview question to four columns:

- the question,
- the final cleaned answer,
- additional context from the transcript, and

This structure allowed systematic comparison across participants while preserving the nuance of the spoken responses.

Across the four interviews, several common patterns emerge. First, participants consistently reported high-risk treatment of patterns such as velocity, high-risk jurisdictions, sanctions nexus, and round-tripping, showing that key red-flag typologies are well-entrenched in practice. Second, SOPs are described as well-document busted at the edges, with ownership structures and borderline cases requiring substantial interpretation and frequent reliance on manager instructions or team precedent. Third, workload and time pressure are associated with more minimal review behaviors, especially in past roles, while trust in the automated system is generally moderate and has improved as case-assignment and review processes have become more structured. These qualitative patterns align with the quantitative finding that human decision-consistency is influenced by training, SOP clarity, and social influence, and they help contextualize the sources of variation seen in the scenario-based survey data.

5 CONCLUSION, LIMITATIONS AND FUTURE RESEARCH

5.1 Conclusion

This study set out to examine how human and systemic factors influence decision consistency in fraud and financial crime teams operating within UK-regulated fintech environments. The findings show that, although analysts often work within formal SOP frameworks and supported by automated fraud systems, decision-making remains highly variable once cases move beyond the initial suspicion stage. Agreement is generally stronger on whether a case is suspicious than on whether to escalate it or what action to take, indicating that ambiguity increases as decisions become more operationally consequential.

The results also show that inconsistency is not caused by a single factor, but by the interaction of several pressures. Heavy workloads, limited training, unclear SOPs, and reliance on manager instructions or team precedent all contribute to uneven decision-making. In many cases, analysts compensate for ambiguity by using intuition or local team norms, which can create internal consistency within teams but inconsistent practice across the wider organization and industry. This suggests that the problem is structural as much as individual, shaped by how work is organized, guided, and supervised.

The scenario-based findings further demonstrate that even when analysts recognize the same red flags, they do not always agree with the correct response. This is especially important because it shows that the issue is not only whether suspicious behavior is detected, but whether the organization has a shared and clearly defined threshold for escalation and action. Where action authority varies across firms, the comparison becomes more complex, yet the low agreement on escalation itself confirms that inconsistency exists before firm-specific action choices are even reached.

The qualitative interviews support this conclusion by showing how practitioners navigate ambiguity in practice. Participants described SOPs as useful but incomplete, especially in edge cases, and many indicated that managerial guidance and experience often fill the gaps left by formal policy. This confirms that the final decision is often shaped by a combination of policy, hierarchy, and individual judgement rather than by a single uniform rule set. As a result, fraud and AML operations may appear standardized on paper while remaining uneven in practice.

From a managerial and organizational perspective, these findings have several critical implications:

First, financial institutions must prioritize SOP clarity and standardization. Policies should move beyond high-level guidance and provide case-based, scenario-driven examples that reduce ambiguity and constrain interpretation.

Second, there is a clear need to strengthen training frameworks. With many respondents receiving minimal formal training, organizations risk relying on informal knowledge transfer and intuition. Regular, structured training programs, particularly those incorporating real-case simulations, can improve both consistency and regulatory defensibility.

Third, firms should address workload management and cognitive pressure. High alert volumes and time constraints encourage heuristic decision-making and increase reliance on gut feeling. Introducing workload balancing mechanisms, decision-support tools, and review buffers can reduce these effects.

Fourth, organizations must recognize and actively manage the role of informal decision systems. While team precedent and managerial guidance can improve local efficiency, they can also create inconsistencies across teams. Aligning these informal practices with formal SOPs is essential for achieving organization-wide consistency.

Finally, the findings highlight the importance of designing fraud operations as socio-technical systems, where human decision-making is supported, not replaced, by automated systems, clear policies, and structured governance.

Overall, this research contributes empirical evidence to the literature on fraud detection, human-automated system collaboration, and inter-rater agreement by showing that inconsistency in fintech financial crime decisions is a measurable and meaningful organizational and industrial issue. The study demonstrates that improving technical fraud detection alone is not sufficient if human decision pathways remain ambiguous. Stronger SOP design, more consistent and ongoing trainings over time, clearer escalation thresholds, and better alignment between teams, supervisors and organizations are likely necessary to reduce variability and improve defensibility in fraud operations.

5.2 Limitations of the Study

This study provides useful empirical evidence on decision consistency in fraud and financial crime teams, but it is subject to several limitations that should be acknowledged when interpreting the findings. First, the study is based on a relatively small sample of 33 complete survey responses, supplemented by four semi-structured interviews. Although this sample is adequate for exploratory analysis and for identifying patterns in a specialized professional population, it limits the extent to which the findings can be generalized to the wider population of fraud, AML, and financial crime analysts working in fintech or banking environments. The small number of respondents also constrains the statistical power of the analysis, particularly when examining relationships between multiple organizational and human-factor variables.

Second, the sampling approach was based on convenience and professional network recruitment. This approach was necessary given the highly specialized and access-sensitive nature of the population, but it may have introduced selection bias. Individuals who were more accessible, more engaged with the topic, or more willing to discuss their work may be overrepresented in the sample. As a result, the findings may reflect the practices and perspectives of a specific subset of analysts rather than the full range of decision-making behaviors across the sector. In addition, because the sample was drawn from UK-regulated or UK-serving firms, the results are most directly applicable to the UK context and may not be transferred directly to other regulatory environments, organizational structures, or fraud operating models.

Third, the study relies heavily on self-reported survey data. Measures such as workload, training, reliance on SOPs, trust in automated systems, and the influence of bias-related behaviors were reported directly by participants. While this provides valuable insight into how analysts perceive their own working conditions, self-reported data are inherently vulnerable to recall bias, exaggeration, underreporting, and social desirability bias. Some respondents may present their practices in a more favorable light, particularly when describing judgment, escalation behavior, or reliance on intuition. In addition, some questions required participants to estimate training hours, case volumes, or decision influences over a past period, which may have introduced imprecision into the responses.

Fourth, the scenario-based survey design captures how analysts say they would respond to standardized fraud cases, rather than observing their behavior in live operational settings. Hypothetical scenarios are valuable because they allow direct comparison across respondents and enable the measurement of inter-analyst agreement under common conditions. However, they may not fully replicate the time pressure, emotional load, competing priorities, and organizational constraints present in real fraud review environments. In practice, analysts may behave differently when facing live alerts, customer pressure, system interruptions, managerial oversight, or account-specific commercial considerations. Therefore, the scenario-based findings should be interpreted as a structured approximation of operational decision-making rather than a complete representation of real-world behavior.

Fifth, organizational differences across firms may have affected the interpretation of some variables, particularly the “Action” responses. In some organizations, analysts have the authority to directly block or stop transactions, whereas in others they are restricted to escalating cases to a separate team such as compliance or financial crime operations etc. This means that the “Action” variable does not represent a perfectly uniform decision across all participants, because the available response options were shaped by the respondent’s organizational role and authority.

Sixth, the qualitative component of the study is limited by the small number of interviews conducted. Four interviews were sufficient to provide contextual depth and to illustrate common themes such as workload pressure, SOP ambiguity, trust in systems, and reliance on manager instructions or team precedent. However, this number does not allow for full thematic saturation across all possible fraud operating models or analyst experiences. The interview findings should therefore be viewed as complementary and illustrative rather than exhaustive. They help explain the survey results, but they do not claim to represent every fraud analyst perspective in the sector.

Seventh, the study uses a cross-sectional design, meaning that all data were collected at a single point in time. This allows the researcher to identify associations between decision consistency and factors such as training, workload, and SOP clarity, but it does not permit causal inference. For example, while the analysis may show that lower training or higher workload is associated with lower consistency, it cannot prove that these factors directly cause inconsistency. It is also possible that organizational culture, team structure, managerial style, or prior role experience influences both workload perception and decision behavior in ways not fully captured by the study.

Finally, although the study examines several important human and systemic factors, it cannot fully capture the complexity of fraud decision-making in live environments. Variables such as customer segment, product type, firm maturity, system design, escalation thresholds, and local risk appetite may all shape decision outcomes, but they are difficult to measure comprehensively in a short survey. Similarly, some of the most important influences on analyst behavior may be implicit or tacit, making them difficult to observe directly through questionnaire items alone. For this reason, the findings should be interpreted as evidence of meaningful patterns rather than as a complete map of all sources of inconsistency in fraud and financial crime operations.

Taken together, these limitations do not undermine the value of the study, but they do define the scope within which the findings should be interpreted. The research is best understood as an exploratory and context-specific investigation into decision consistency in UK fintech fraud and AML teams. Its main contribution lies in providing early empirical evidence on how analysts interpret identical scenarios differently, and on how organizational ambiguity, workload, and informal norms may shape those differences. Future research with larger samples, broader geographic coverage, and live operational data would be useful in extending and validating these findings. Future research would benefit from linking human decision-making directly to operational outcomes such as reporting quality, escalation speed, and intervention success, because the regulatory reports show that these outcomes matter materially in practice. *(FCA Enforcement Data, 2025; UKFIU, SARs Annual Report, 2024)*

5.3 Future Research

The present study opens several avenues for future research. First, larger samples are needed to test whether the patterns observed here hold across a broader range of fraud, AML, and financial crime professionals. A larger and more diverse sample would improve statistical power and allow researchers to examine whether decision consistency differs by role, seniority, firm size, product type, or regulatory environment. Future studies could also compare UK fintechs with banks, payment institutions, and firms operating in other jurisdictions to assess whether the same sources of inconsistency appear across different compliance regimes.

Second, future research should move beyond cross-sectional self-report data and examine decision-making in live operational settings. While scenario-based surveys are useful for measuring inter-analyst agreement under controlled conditions, they cannot fully capture the speed, pressure, and uncertainty of real fraud operations. Longitudinal or observational studies could track how the same analysts behave over time, especially as workloads fluctuate, policies change, or new fraud typologies emerge. Such designs would help establish whether inconsistency is stable, context-dependent, or shaped by organizational change.

Third, there is scope for deeper analysis of the interaction between human judgment and automated systems. The current study shows that analysts do not simply follow system generated outputs or SOPs mechanically; instead, they interpret, adjust, and sometimes override them. Future work could examine how different system designs, such as explainable AI, triage automation, or agentic AI workflows, affect analyst consistency, trust, and escalation behavior. This would be especially useful considering recent work suggesting that hybrid human and system generated models can improve fraud detection while still requiring human accountability and contextual judgment.

Fourth, future studies should investigate the organizational sources of inconsistency in greater depth. The current findings suggest that SOP ambiguity, manager instructions, team precedent, and workload pressure all shape decision-making, but the mechanisms behind these effects remain only partially understood. Qualitative case studies or comparative organizational research could explore how different teams develop local norms, how managers influence escalation thresholds, and how informal practices interact with formal policies. This would help explain why some firms achieve stronger internal alignment than others even when they operate under similar regulatory rules.

Fifth, further research should consider separating suspicion, escalation, and action more explicitly in both surveys and operational datasets. In this study, escalation agreement remained low even after accounting for organizational differences in who can block or escalate transactions, suggesting that inconsistency arises before firm-specific action authority comes into play. Future work could design more fine-grained measures of decision thresholds, such as confidence scoring, escalation triggers, or action hierarchies, to identify precisely where divergence begins and how it propagates through the case-handling process.

Finally, future research should examine how training, feedback loops, and supervisory review affect decision consistency over time. Recent discussions of automated system-led AML emphasize the need for analysts to interrogate machine outputs, identify bias, and continuously recalibrate both human and system performance. Studies that test targeted interventions, such as structured calibration sessions, decision-review workshops, or SOP redesign, would be valuable for determining whether inconsistency can be reduced through organizational learning. Such work would not only extend the present study but also provide practical evidence for firms seeking to build more transparent, defensible, and reliable fraud- and AML-decision processes.

Taking together, these future research directions would help move the field from exploratory measurement of inconsistency toward a more complete understanding of how human judgment and organizational structure rule-based systems interact in financial crime decision-making.

6 REFERENCES

Abi, R. (2025). AI-Driven Fraud Detection Systems in Fintech Using Hybrid Supervised and Unsupervised Learning Architectures. *International Journal of Research Publication and Reviews*, *6*(6), 4375–4394. https://doi.org/10.55248/gengpi.6.0625.2161

AML Watcher. (2025).

Axelsen, H., Licht, V., & Damsgaard, J. (2025). *Agentic AI for Financial Crime Compliance*. https://doi.org/https://doi.org/10.48550/arXiv.2509.13137

Bertrand, A., Eagan, J. R., Maxwell, W., & Brand, J. (2024, May 11). AI is Entering Regulated Territory: Understanding the Supervisors' Perspective on Model Justifiability in Financial Crime Detection. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3613904.3642326

EBA Report on Payment fraud. (2024). https://www.eba.europa.eu/sites/default/files/2024-08/465e3044-4773-4e9d-8ca8-b1cd031295fc/EBA_ECB%202024%20Report%20on%20Payment%20Fraud.pdf

FCA Enforcement data. (2025). https://www.fca.org.uk/data/fca-operating-service-metrics-2024-25/enforcement-data

Financial Conduct Authority Annual Report and Accounts. (2025). https://www.fca.org.uk/publications/annual-reports/annual-report-2024-2025

Goldenfein, J. (2023). *Lost in the Loop-Who is the "human" of the Human in the Loop?* https://doi.org/10.2139/ssrn.4750634

Hagen, R. A., Øverlier, L., & Helkala, K. (2025). Human Factors in AI-Driven Cybersecurity: Cognitive Biases and Trust Issues. *Digital Threats: Research and Practice*, *6*(4). https://doi.org/10.1145/3759260

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. *Human Factors*. https://pubmed.ncbi.nlm.nih.gov/21077562/

PSFA Standards. (2025, May). PSFA Standards - Standard for Fraud Detection Practitioners (HTML).

Sethupathy, U. K. A. (2025). Risk-Aware AI Models for Financial Fraud Detection: Scalable Inference from Big Transactional Data. *International Journal of Intelligence Science*, *15*(04), 162–183. https://doi.org/10.4236/ijis.2025.154009

UK Finance Annual Fraud Report. (2024). https://www.ukfinance.org.uk/system/files/2025-10/Annual%20Report%20for%20the%20year%20ended%2031%20December%202024.pdf

UK finance Annual Fraud Report. (2025). https://www.ukfinance.org.uk/system/files/2025-05/UK%20Finance%20Annual%20Fraud%20report%202025.pdf

UK FinTech - State of the Nation. (2019). https://assets.publishing.service.gov.uk/media/5cd9511740f0b62d81cad294/UK-fintech-state-of-the-nation.pdf

UKFIU, SARs Annual Report. (2024). https://cilexregulation.org.uk/2026/01/05/ukfiu-sars-annual-report-2025/

Umemuro Hiroyuki. (2011). General Factors That Elicit Human Affect across Countries. *Universal Access in Human-Computer Interaction. Users Diversity - 6th International Conference, UAHCI 2011, Held as Part of HCI International 2011, Orlando, FL, USA, July 9-14, 2011, Proceedings, Part II*. https://www.researchgate.net/publication/221095859_General_Factors_That_Elicit_Human_Affect_across_Countries

Vijayabaskar, S. (2020). Harnessing Generative AI for Risk Management and Fraud Detection in Fintech: A New Era of Human-Machine Collaboration. *International Journal of Scientific Research and Management (IJSRM)*, *8*(04), 369–379. https://doi.org/10.18535/ijsrm/v8i04.ec01

Annexure A

Block 1 – Consent (Q1)

1. Do you consent to participate in this anonymous academic study on fraud decision-making? Data will be analysed in aggregate and you may withdraw at any time before submitting.

   □ Yes, I consent

   □ No (screen-out)

Block 2 – Section A: Demographics & Role (Q2–Q5)

2. Which option best describes your current primary role?

   □ Fraud Analyst/Specialist

   □ AML/CTF Analyst

   □ Compliance Officer

   □ Financial Crime Analyst (generalist)

   □ Risk/Internal Audit (fraud-focused)

   □ Other (specify):

3. How many years of professional experience do you have in fraud, AML, or related financial-crime roles?

   □ 0–1 year

   □ 1–3 years

   □ 3–5 years

   □ 5–10 years

   □ >10 years

4. Approximately how many active customers/users does your firm serve?

   □ <100k

   □ 100k–999k

   □ 1M–4.9M

   □ ≥5M

   □ Not sure

5. Does your role involve UK-regulated or UK-located customer cases?

   □ Yes

   □ No

   □ Not sure

Block 3 – Section B: Workload, Training & SOPs (Q6–Q15)

6. In a typical week, how many hours do you spend reviewing fraud/AML alerts/cases?

   □ 0–9

   □ 10–19

   □ 20–29

   □ 30–39

   □ ≥40

7. How many fraud/AML alerts/cases do you personally review monthly?

   □ <50

   □ 50–199

   □ 200–499

   □ 500–999

   □ 1k–2k

   □ ≥2k

   □ Not sure

8. Average time per SIMPLE case:

   □ <5 min

   □ 5–9 min

   □ 10–14 min

   □ 15–19 min

   □ 20–29 min

   □ ≥30 min

9. Average time per COMPLEX case:

   □ <30 min

   □ 30–59 min

   □ 1–2 h

   □ 2–4 h

   □ ≥4 h

10. At busy times, how many cases do you handle simultaneously?

    □ 1–3

    □ 4–7

    □ 8–12

□ 13–20

□ >20

11. Hours of fraud/AML/compliance training in the last 12 months:

□ 0–5

□ 6–10

□ 11–20

□ 21–40

□ >40

12. How many SARs/STRs have you personally submitted in the last 12 months?

□ 0–5

□ 6–15

□ 16–30

□ 31–60

□ >60

□ Don't file SARs

13. How does workload pressure affect your decision quality?

□ Very negatively ⟶ □ Very positively (5-point scale)

14. How CLEAR are escalation SOPs?

□ Extremely unclear ⟶ □ Extremely clear (5-point scale)

15. How much room for personal interpretation do SOPs leave?

□ Not at all ⟶ □ A great deal (5-point scale)

Block 4 – Decision sources (Q16)

How often do each of the following influence your decisions?

- Written SOPs
- Team precedent
- Own judgment
- Manager instructions

(Scale: Never ⟶ Very often, 5-point)

Block 5 – Trust & Automation (Q17–Q20)

17. How much do you trust the automated fraud/AML system?

□ No trust ⟶ □ Complete trust (5-point)

18. Please indicate how much you agree/disagree with each of the following about the system:

- The system is reliable
- The system is predictable
- The system is understandable
- The system is comfortable to rely on

(Scale: Strongly disagree → Strongly agree, 5-point)

19. How often do you accept the system's "Low Risk" recommendation without a detailed manual check?

    □ Never → □ Always (5-point)

20. If your own judgment disagrees with the system's recommendation, do you usually re-check the evidence carefully before deciding?

    □ Strongly disagree → □ Strongly agree (5-point; reverse-coded)

Block 6 – Cognitive biases (Q21–Q24)

21. To what extent do you tend to stick with your initial impression of a case even when new, minor evidence appears?

    □ Strongly disagree → □ Strongly agree (5-point)

22. To what extent do you mainly look for information that supports your first assessment rather than information that might overturn it?

    □ Strongly disagree → □ Strongly agree (5-point)

23. To what extent would you make different decisions on similar cases when you are tired, under time pressure, or stressed?

    □ Strongly disagree → □ Strongly agree (5-point)

24. To what extent do you rely on intuition or "gut feel" rather than formal rules or scoring in borderline cases?

□ Strongly disagree → □ Strongly agree (5-point)

Block 7 – Red-flag patterns (Q25–Q26)

25. For each of the following patterns, please indicate how SEVERE or RISKY they are treated in your firm:

- Velocity / rapidly escalating activity
- High-risk destinations (countries/sectors/channels)
- Sanctions nexus
- Structuring around thresholds
- Round-tripping (funds out and back to the same owner chain)

(Scale: Not at all risky → Extremely risky, 5-point)

26. Do your SOPs define "gibberish" or higher-risk patterns for customer email addresses or identifiers?

    If yes, how would you typically treat an address that looks random or suspicious under your SOPs?

    If no, how would you handle such an address in practice?

Block 8 – Scenario-based questions (Q27–Q34)

For each of the following scenarios, please indicate:

- Based on your current SOPs, would this typically be treated as SUSPICIOUS?
- Would you ESCALATE this? If yes, to whom or which team (e.g. SAR/Compliance, senior analyst, manager)?
- What final ACTION would you take personally?
  Options may include: Release, Request more information (RFI), Monitor, Escalate, Block/Exit, Close relationship, etc.
- What specific detail in this scenario most influences your decision?
- If you were more tired or under heavier workload, do you think your decision in this scenario might change? If so, how?
- Have you personally handled a similar case before, or is this a type of case where you have limited or no prior experience?

Scenario 1 – Q27: Student-to-crypto cash-out pattern (typology-based crypto-mule-like scenario)

Scenario 2 – Q28: Tehran-/Istanbul-travel-related business-payment pattern (sanctions-related geographic risk)

Scenario 3 – Q29: Structuring-below-thresholds business-payment pattern

Scenario 4 – Q30: "Swiss Prestige Timepieces"-style scam-risk business pattern

Scenario 5 – Q31: Fake charity / sanctions-like NGO-type scenario

Scenario 6 – Q32: Multi-small-to-large-crypto-cash-out pattern

Scenario 7 – Q33: Velocity-driven retail-payment-top-up pattern

Scenario 8 – Q34: Round-tripping-between-entities pattern

Annexure B

Detailed Demographic Table for each participant

S.No	Primary role? - Selected Choice	How many years of professional experience do you have in Fraud, AML, or related Financial crime roles?	Approximately how many active customers/users does your firm serve?	Does your role involve UK-regulated or UK-located customer cases?
1	AML/CTF Analyst/Specialist	3 year 1 month-5 years	1M-4.9M	Yes
2	Financial Crime Analyst/Specialist	1 year 1 month-3 years	5M	Yes
3	Fraud Analyst/Specialist	3 year 1 month-5 years	100k-999k	Yes
4	Fraud Analyst/Specialist	1 year 1 month-3 years	5M	Yes
5	AML/CTF Analyst/Specialist	3 year 1 month-5 years	100k-999k	Yes
6	Financial Crime Analyst/Specialist	1 year 1 month-3 years	5M	Yes
7	Fraud Analyst/Specialist	1 year 1 month-3 years	100k-999k	Yes
8	Compliance Officer	3 year 1 month-5 years	100k-999k	Yes
9	Financial Crime Analyst/Specialist	3 year 1 month-5 years	1M-4.9M	Yes
10	Financial Crime Analyst/Specialist	3 year 1 month-5 years	100k-999k	Yes
11	Fraud Analyst/Specialist	3 year 1 month-5 years	0-99k	Yes
12	Financial Crime Analyst/Specialist	3 year 1 month-5 years	5M	Yes
13	Fraud Analyst/Specialist	5 year 1 month-9 years 11 months	1M-4.9M	Yes
14	Financial Crime Analyst/Specialist	1 year 1 month-3 years	5M	Yes
15	Financial Crime Analyst/Specialist	1 year 1 month-3 years	100k-999k	Yes
16	AML/CTF Analyst/Specialist	5 year 1 month-9 years 11 months	1M-4.9M	Yes
17	Fraud Analyst/Specialist	1 year 1 month-3 years	1M-4.9M	Yes
18	Fraud Analyst/Specialist	3 year 1 month-5 years	100k-999k	Yes
19	AML/CTF Analyst/Specialist	5 year 1 month-9 years 11 months	5M	Yes
20	Compliance Officer	0-1 year	0-99k	Yes
21	Fraud Analyst/Specialist	3 year 1 month-5 years	100k-999k	Yes
22	Fraud Analyst/Specialist	3 year 1 month-5 years	100k-999k	Yes
23	Fraud Analyst/Specialist	1 year 1 month-3 years	1M-4.9M	Yes
24	Fraud Analyst/Specialist	1 year 1 month-3 years	5M	Yes
25	Fraud Analyst/Specialist	1 year 1 month-3 years	100k-999k	Yes
26	Financial Crime Analyst/Specialist	1 year 1 month-3 years	5M	Yes
27	Financial Crime Analyst/Specialist	1 year 1 month-3 years	5M	Yes
28	Fraud Analyst/Specialist	3 year 1 month-5 years	100k-999k	Yes
29	AML/CTF Analyst/Specialist	3 year 1 month-5 years	100k-999k	Yes

30	AML/CTF Analyst/Specialist	5 year 1 month-9 years 11 months	1M-4.9M	Yes
31	Fraud Analyst/Specialist	1 year 1 month-3 years	0-99k	Yes
32	AML/CTF Analyst/Specialist	5 year 1 month-9 years 11 months	5M	Yes
33	Fraud Analyst/Specialist	1 year 1 month-3 years	100k-999k	Yes