# Dynamic Trust Modulation and Human Oversight in AI-Driven AML Systems: A Conceptual Framework for Compliance

Julian Diaz[1], Abeer Alsadoon[2], Oday D. Jerew[3], Ahmed Hamza Osman[4]*,
Hani Moetque Aljahdali[5], Albaraa Abuobieda[6], Abubakar Elsafi[7]
Higher Education Leadership Institute (HELI), Melbourne, Australia[1, 2, 3]
Asia Pacific International College (APIC), Sydney, Australia[2, 3]
Department of Information Systems-Faculty of Computing and Information Technology in Rabigh (FCITR),
King Abdulaziz University, Jeddah 21911, Saudi Arabia[4, 5]
Department of Software Engineering-College of Computer Science and Engineering,
University of Hafr Al Batin, Hafr Al Batin 31991, Saudi Arabia[6]
Department of Software Engineering-College of Computer Science and Engineering,
University of Jeddah, Jeddah, 21959, Saudi Arabia[7]

***Abstract*—This literature review investigates how human trust, decision fatigue, explainability (XAI), and human oversight interrelate to influence analyst decision-making in AI-driven anti-money laundering (AML) systems. While prior research has predominantly emphasized algorithmic performance, detection accuracy, or regulatory compliance in isolation, a critical gap remains in understanding the human-centered dynamics that shape real-world operational outcomes. Addressing this gap, the review examines how financial institutions navigate compliance demands and operational constraints, drawing on the Australian regulatory environment as an illustrative governance reference, including expectations articulated by AUSTRAC. Building on this synthesis, the study identifies structural gaps in Trust Calibration and oversight practices. It introduces a Dynamic Trust Modulation (DTM) framework to conceptualize how trust evolves across AML workflows. The framework models trust as a fluid, context-dependent construct shaped by system behavior, analyst workload, explainability mechanisms, and regulatory pressure. By framing trust, explainability, and decision fatigue as interdependent components of human–AI collaboration, this review advances a more holistic perspective on socio-technical system design in financial crime detection. The proposed framework contributes theoretically by extending human–AI trust research into the AML domain and practically by offering actionable design principles to enhance system accountability, decision defensibility, and adaptive compliance in operational AML environments.**

***Keywords*—*Artificial intelligence; anti-money laundering (AML); Trust Calibration; Explainability; decision fatigue; human oversight; AUSTRAC Compliance; transaction monitoring; false positives; Analyst–System Interaction; Regulatory Technology (RegTech)***

## I. Introduction

Money laundering continues to pose a critical threat to financial systems worldwide, prompting financial institutions to adopt increasingly sophisticated anti-money laundering (AML) frameworks. Despite technological advancements, institutions continue to struggle with high false positive rates, escalating operational costs, and increasing regulatory pressure to ensure robust compliance [1].

Artificial intelligence (AI) and machine learning (ML) have emerged as promising tools to enhance AML transaction monitoring by uncovering anomalous behavior and enabling real-time detection. AI promises efficiency and improved detection precision; however, these systems often function as opaque "black boxes", raising concerns about explainability, regulatory defensibility, and human oversight [2]. The lack of transparency and adaptability to evolving compliance standards hinders institutional confidence and broader adoption.

While existing literature explores individual factors such as trust, decision fatigue, and explainability, it often treats them in isolation, overlooking their dynamic interplay in operational workflows. Critical questions remain unanswered: How do analysts rebuild trust after AI errors? How does human feedback shape system adaptation? And how can compliance frameworks incorporate meaningful oversight? These unresolved issues underscore the tension between technological performance and regulatory accountability in AI-supported AML environments.

This study addresses a critical gap in AI-driven AML research by developing a Dynamic Trust Modulation framework that conceptualizes how trust, explainability, and decision fatigue interact within transaction monitoring systems under conditions of uncertainty. Rather than treating these factors in isolation, the framework integrates them to capture the dynamic nature of analyst–system interaction and evolving oversight demands in AML workflows. By focusing on system-level decision processes such as confidence assessment, interpretability, and feedback loops, the framework provides a structured lens for understanding how human oversight can be supported and sustained in complex compliance environments. Drawing on human–AI collaboration and socio-technical systems literature, this study offers a generalizable conceptual foundation for advancing accountability, decision defensibility, and explainable decision-making in AI-supported AML systems.

*Corresponding author.

This study makes three key contributions to research on AI-supported anti-money laundering (AML) systems. First, it synthesizes trust calibration, decision fatigue, and explainability into an integrated conceptual perspective tailored to AML transaction monitoring, addressing their interdependence within high-stakes compliance environments. Second, it introduces the Dynamic Trust Modulation framework, which conceptualizes trust as a feedback-driven and context-dependent construct shaped by system performance, analyst workload, explainability mechanisms, and regulatory constraints. Third, the study offers practical and regulatory insights for the design of AI-enabled AML systems by highlighting design principles that support accountability, decision defensibility, and sustainable human oversight under regimes such as AUSTRAC.

The remainder of this study is organized as follows: Section II outlines background and related work; Section III presents case-based illustrations of implementation challenges; Section IV details the review methodology; Section V synthesizes thematic findings; Section VI discusses trust calibration dynamics and addresses oversight and governance mechanisms; Section VII discuss the regulatory frameworks and governance under AUSTRAC; Section VIII explores systemic risks; Section IX presents the structural ambiguity and compliance overload; Section X discusses fragmented oversight and risk diffusion; Section XI presents the global gaps and interoperability risks; Section XII summarizes limitations and future research; and Section XIII concludes the study.

## II. Background and Related Work

### A. AML and Compliance Failures

Lack of awareness and fragmented implementation of AML regulations have consistently led to enforcement failures and reputational risks. In Australia, AUSTRAC has imposed multimillion-dollar fines on institutions that failed to adopt a robust risk-based approach [3]. The cases of Commonwealth Bank of Australia (CBA), Westpac, and Tabcorp demonstrate that insufficient governance, underdeveloped compliance infrastructure, and inadequate technological adaptation remain key barriers to effective AML performance. For example, CBA failed to submit over 53,000 threshold transaction reports due to unmonitored Intelligent Deposit Machines, while Westpac committed over 23 million breaches due to poor internal oversight and delayed reporting [4]. International experiences echo similar challenges, such as Santander's outsourcing of AI-based AML systems, which resulted in substantial penalties due to vendor opacity and poor model explainability, whereas the Bunq vs. DNB case illustrates how deep regulatory understanding and system transparency enabled the bank to defend an AI-driven compliance approach [5]. Chainalysis and Elliptic offer additional insights into how institutions with strong AI and AML knowledge can support global regulatory goals and innovate proactively [6].

Prior studies indicate that higher AML regulatory ratings are associated with stronger financial outcomes [7]. Other research has identified persistent weaknesses in institutional coordination and in the empirical evaluation of illicit financial flows [8]. Additional analyses highlight embedded distrust and cognitive overload within compliance systems, particularly in the context of NLP-based monitoring tools [9]. Related work further suggests that traditional AML models tend to underperform in informal economies and across non-bank financial intermediaries [10].

### B. AI in AML Systems

AI is redefining AML monitoring through anomaly detection, natural language processing, and real-time predictive analytics. Machine learning and deep learning algorithms enable institutions to recognize sophisticated laundering strategies that evade rule-based models. Systems like those implemented by Chainalysis and Elliptic, or the NICE Actimize modules adopted by IDB Bank, show measurable improvements in detection and operational efficiency [6], [11]. Research highlights AI's scalability and capacity for continuous learning, with models adapting detection parameters as threats evolve [12]. Empirical studies further indicate that AI-based approaches outperform legacy systems in terms of precision and recall [13], and additional work demonstrates their viability even in underregulated markets [14]. However, challenges persist. AI requires high-quality data, domain-specific calibration, and contextual sensitivity. Overreliance on external vendors, as seen in Santander's case, limits transparency and introduces compliance risks [5]. Hybrid models, combining AI with human oversight, have emerged as practical compromises [3]. AI's success is not solely technical; it hinges on integration with compliance expectations and interpretability. As the Bunq case demonstrates, performance is insufficient without justification mechanisms acceptable to regulators. Institutions must embed auditability, adaptability, and role clarity into AI systems to optimize both compliance and risk mitigation.

### C. Trust and Explainability in AI

Trust and explainability are central to the adoption and effectiveness of AI in AML environments. Without interpretable models, analysts struggle to validate alerts, leading to inefficiencies and increased risk exposure. Decision fatigue arises from high false positive rates, undermining trust in AI systems and degrading regulatory defensibility [3]. Explainable AI (XAI) tools such as SHAP and LIME have been proposed to mitigate this issue, given that these techniques increase transparency by visualizing feature importance and supporting interpretability [15]. Recent studies emphasize that explainability is becoming a regulatory imperative, particularly under evolving AI governance standards [16]. Institutions that fail to provide traceable logic risk regulatory rejection, as evidenced in the initial pushback against Bunq's system. Despite growing literature on technical performance, few studies explore how AML professionals interact with AI or recover trust after errors. This research contributes by proposing a Dynamic Trust Modulation Loop, a framework that connects analyst feedback, system responsiveness, and explainability in a continuous cycle. This approach provides a pathway for aligning human-AI collaboration with institutional accountability and adaptive regulatory compliance.

Existing frameworks in AI-supported decision-making and human–AI collaboration have examined constructs such as trust, explainability, and human oversight, particularly in domains including healthcare and decision support systems. While these models offer valuable insights into how users interpret algorithmic outputs, they often treat trust and explainability as

static or isolated factors rather than as dynamically interacting elements within operational workflows. In anti-money laundering (AML) contexts, this limitation is especially critical. AML environments are characterized by persistent uncertainty, high false positive rates, regulatory accountability, and sustained cognitive load on analysts. Existing frameworks do not sufficiently capture how trust evolves in response to repeated system errors, how decision fatigue accumulates under continuous alert pressure, or how analyst feedback is operationally reintegrated into AI systems.

The Dynamic Trust Modulation framework proposed in this study addresses these gaps by conceptualizing trust as a feedback-driven and context-dependent construct. By explicitly modeling the interaction between trust calibration, decision fatigue, and explainability within AML transaction monitoring workflows, the framework offers a novel socio-technical perspective tailored to high-stakes compliance settings where accountability remains fundamentally human-centered.

## III. Case Studies

### A. Santander: Vendor Dependency and Oversight Gaps

*1) Context:* Santander, a major European bank, began integrating AI-powered transaction monitoring in 2020, outsourcing this function to the third-party vendor ThetaRay. The system was designed to detect laundering patterns in correspondent banking using machine learning.

*2) Event:* Despite this technological integration, Santander faced major penalties of $1 million in Norway (2019) and £107.7 million in the UK (2022) due to persistent compliance deficiencies [5]. The reliance on an external AI provider reduced internal transparency and limited control over system decisions. Under the EU AI Act, this division of responsibility complicated regulatory accountability, while high false positive rates overwhelmed internal analysts.

*3) Lesson learned:* Outsourcing AI without maintaining internal oversight and interpretability weakens institutional trust and regulatory defensibility. Banks must preserve visibility into AI decisions, even when compliance functions are delegated externally.

### B. Bunq: Challenging Regulatory Conservatism Through Explainability

*1) Context:* Bunq, a Dutch neobank, developed its own AI-based transaction monitoring system, arguing that it outperformed traditional rule-based methods.

*2) Event:* The Dutch Central Bank (DNB) rejected Bunq's model for lacking explainability and regulatory alignment [5]. Bunq challenged this stance in court, arguing for outcome-based rather than methodology-based evaluation. In October 2022, the Commercial and Industry Appeals Tribunal ruled in Bunq's favor, requiring regulators to assess performance rather than favoring manual approaches.

*3) Lesson learned:* Regulatory trust in AI depends not only on technical performance but also on transparency. Explainable systems are essential for regulatory acceptance, even when accuracy is high.

### C. Chainalysis and Elliptic: Blockchain Analytics and the Need for Explainability

*1) Context:* Chainalysis and Elliptic are leaders in AI-driven blockchain analytics for digital asset compliance. Both firms support cryptocurrency monitoring by detecting illicit financial activities through anomaly detection and network analysis.

*2) Event:* Despite strong operational success, such as Chainalysis's recognition by U.S. Homeland Security regulators like the Financial Stability Board continues to stress explainability [6]. Recent work cautions that even high-performing models may fail regulatory scrutiny if their underlying mechanisms remain opaque [16]. The 2025 acquisition of Alterya by Chainalysis highlights a broader shift toward deeper explainability and risk validation in financial crime analytics.

*3) Lesson learned:* Explainability remains critical even for top-performing AI systems. Without transparency and auditability, regulatory trust and long-term accountability are jeopardized.

### D. Westpac: Oversight Failures and Analyst Fatigue

*1) Context:* Westpac, one of Australia's major banks, deployed transaction monitoring systems for AML/CTF compliance but failed to ensure governance effectiveness.

*2) Event:* The institution was fined 1.3 billion AUD for over 23 million violations, including failure to monitor cross-border transfers and respond to alerts [3]. These failures were tied to weak board oversight, overburdened compliance teams, and fatigue caused by excessive alert volumes, leading to institutional breakdowns.

*3) Lesson learned:* Effective AML systems require governance structures that support human-AI workflows. Analyst fatigue and board-level disengagement can compromise trust and system effectiveness.

### E. Commonwealth Bank of Australia: Misaligned Trust Calibration

*1) Context:* The Commonwealth Bank of Australia relied on Intelligent Deposit Machines (IDMs) to support AML monitoring, automating transaction processing and reporting.

*2) Event:* In 2018, the bank was fined 700 million AUD for failing to report over 53,000 threshold transactions. Investigations revealed that analysts did not escalate anomalies, assuming system reliability despite lacking internal risk assessments or override mechanisms [3].

*3) Lesson learned:* Overreliance on automation without real-time human verification erodes compliance. AML systems must enable ongoing trust calibration and empower analysts to challenge system outputs.

### F. IDB Bank (New York): Successful Human-AI Collaboration

*1) Context:* IDB Bank implemented NICE Actimize modules to enhance AML operations, aiming to streamline alert processing and boost efficiency.

*2) Event:* The system led to a 60% reduction in manual post-alert work, allowing analysts to focus on high-risk investigations [12]. A structured feedback loop enhanced trust and explainability, improving understanding of AI behavior.

*3) Lesson learned:* AI adoption succeeds when it supports, not replaces, human judgment. Feedback-integrated, explainable systems reduce fatigue and build trust across compliance teams and regulators.

## IV. Review Methodology

This study adopts a systematic literature review methodology tailored for conceptual synthesis rather than empirical measurement. The purpose of this approach is to transparently map and integrate existing knowledge on AI integration within Anti-Money Laundering (AML) compliance, particularly in the areas of trust, explainability, oversight, and human-AI collaboration. The review adheres to academic standards for reproducibility, thematic coherence, and methodological transparency.

### A. Literature Search Strategy

The literature review spans the publication period from 2020 to 2025. The search was conducted across academic databases and authoritative industry sources, including Google Scholar, Scopus, and regulatory or policy-based repositories such as ACAMS, AUSTRAC, and the Attorney-General's Department of Australia. Relevant white papers from Deloitte, NICE Actimize, and Chainalysis were also included to ensure industry alignment.

Search terms were structured around the core themes of the study and included Boolean combinations such as: "Artificial Intelligence" AND "AML" OR "Anti-Money Laundering", "False positives" AND "transaction monitoring", "Human-AI collaboration" OR "trust in AI systems", "Explainability" OR "XAI" AND "compliance", "AUSTRAC" AND "regulatory expectations" OR "AML Australia", "Decision fatigue" OR "alert fatigue" AND "AML analysts". To ensure thematic alignment with the conceptual framework, literature sources were mapped according to their relevance to key constructs, as detailed in Table I. As shown in Table I, the reviewed sources span academic research, industry reports, and regulatory guidelines, providing a balanced foundation for both theoretical insights and practical relevance. This structured mapping demonstrates how each source informed key aspects of the framework, including the core analyst–system interaction constructs, along with human oversight and feedback loop integration. By explicitly aligning each paper with a core construct or operational gap, the review process ensures that the final conceptual framework is grounded in evidence and reflects the latest developments in AI-supported AML practice.

### B. Inclusion and Exclusion Criteria

Inclusion criteria focused on literature published in English between 2020 and 2025 that addressed financial crime detection, AI in compliance, regulatory frameworks, human-machine interaction, and AML case studies. Eligible documents included peer-reviewed journal articles, conference papers, government publications, and technical white papers. Exclusion criteria removed literature outside the financial sector (e.g., AI in medicine or agriculture), non-English publications, and theoretical papers with no relevance to regulatory frameworks or compliance technologies. Initial screening was based on titles and abstracts, followed by full-text review to ensure methodological rigor and relevance. The systematic selection process is visualized in Fig. 1, which outlines identification, screening, eligibility, and final inclusion of sources.

TABLE I. Literature Review Source Overview

| Theme | Key Authors (Type/Year) | Focus Summary | Framework Relevance* |
|---|---|---|---|
| Trust Calibration | Bertrand (Thesis, 2024), Wang (Thesis, 2024), Alkhalili (JA, 2021), Ghimire (JA, 2025), Kahur (Report, 2025) | Trust evolution, explanation paradox, ML filtering, analyst confidence | TC, EX |
| Explainability | Kahur (Report, 2025), Deloitte (Report, 2020), Rane (JA, 2023), Jensen (JA, 2023), Bello (JA, 2024) | SHAP/LIME, hybrid models, blockchain transparency, system performance | EX, FO |
| False Positives | Alahmadi (Conf, 2022), Oztas (JA, 2024), Mounika (JA, 2024), Ketenci (JA, 2021) | Analyst workload, anomaly detection, tuning thresholds, FP mitigation | DF, TC |
| Oversight | Bello & Bronitt (JA, 2024), Goldbarsht (BC, 2023), King (JA, 2020), Rennie (JA, 2021), Maxwell (Report, 2020) | AI governance, risk mitigation, compliance defensibility, human oversight | OV, TC |
| Feedback Loops | Deloitte (Report, 2020), Eddin (Preprint, 2021), Sausen (Report, 2020), Kumar (JA, 2024), Johnson (Report, 2025) | Human-in-the-loop design, AML adaptation, operational feedback systems | FL, TC |
| Regulatory Compliance | AUSTRAC (Reg, 2023/24), Quinn (Report, 2021), Momenta (IR, 2022), Goldbarsht & Sheedy (Report, 2024), Saputra (JA, 2021) | National risk assessment, policy frameworks, compliance alignment | RC, OV |
| Situational Awareness | Boudt (JA, 2025) | News monitoring for AML, risk signal integration | SA |

TC = Trust Calibration; EX = Explainability; DF = Decision Fatigue; OV = Oversight; FL = Feedback Loop; RC = Regulatory Compliance; SA = Situational Awareness.

## V. PRISMA Diagram

This review scoped academic, industry reports, and regulatory documents addressing AI trust, oversight, and anti-money laundering (AML) within both regulatory and technical domains. Emphasis was placed on studies that examine the interaction between machine learning systems and compliance practices, particularly in high-stakes financial environments. Thematic analysis focused on patterns related to explainability, analyst trust calibration, and institutional accountability in AI-augmented AML systems. The selected corpus synthesized

interdisciplinary perspectives across law, computer science, and financial governance.

The flow diagram outlines the identification, screening, eligibility assessment, and inclusion of sources in the systematic literature review, following PRISMA guidelines.

### A. Analytical Framework

Thematic coding was applied across seven domains: Trust Calibration, Explainability, False Positives, Oversight, Feedback Loops, Regulatory Compliance, and Situational Awareness. The distribution and classification of sources across these themes are synthesized in Table III.

Furthermore, to clarify methodological depth and diversity, Table II summarizes the methodological types, data sources, and limitations of the reviewed literature. This enables a comparative understanding of the balance between technical modeling, policy alignment, and human-centered design in the current AML literature.

As shown in Table II, the classification of methodologies reveals irregular scholarly attention where technical modelling approaches prioritize algorithmic performance but rarely consider analyst workload or regulatory defensibility. In contrast, policy-driven studies emphasize compliance and ethical safeguards yet lack empirical validation. Only a small number of industry reports attempt to incorporate practitioner perspectives, but these remain descriptive rather than evaluative. Taken together, this shows fragmented evidence base where rigor and applicability rarely converge, underscoring the need for the trust–feedback framework proposed in this study.

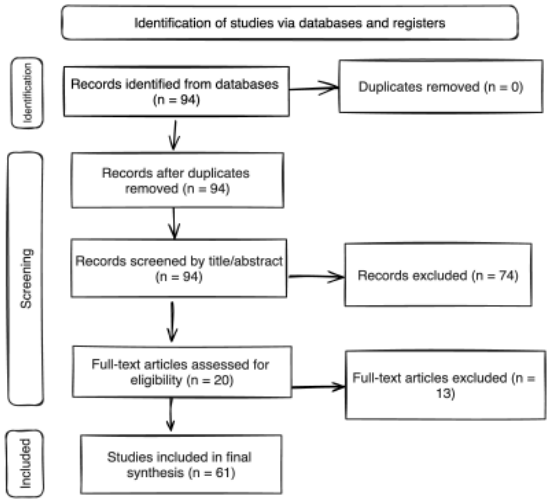


Fig. 1. PRISMA flow diagram of the literature selection process.

TABLE II. METHODOLOGIES IN REVIEWED STUDIES

| Study | Methodology | Data Type | Strengths | Limitations |
|---|---|---|---|---|
| Ghimire (2025) | Simulation + Case Study | Secondary | Focused on precision-recall metrics and system design. | Not practitioner-based; lacks real-world validation. |
| Ketenci et al. (2021) | Quantitative - ML Models | Experimental | Robust testing of false positive reduction strategies. | Lacks human-AI interaction component. |
| Oztas et al. (2024) | Quantitative - ML Clustering | Experimental | Uses adaptive algorithms for anomaly detection. | Does not address human oversight or compliance validation. |
| Bertrand (2024) | Theoretical Analysis | Conceptual | Explores trust, fatigue, and decision frameworks in AI settings. | Lacks empirical or applied industry context. |
| AUSTRAC (2024) | Risk Assessment + Regulatory Review | Policy and Compliance Documents | Grounds research in national compliance standards. | Not focused on technical implementation of AI. |
| Australian Government (2024) | Policy Framework Analysis | Conceptual | Establishes ethical and assurance principles for AI. | Applies broadly to government, not AML-specific. |
| Kute et al. (2021) | Experimental + XAI Evaluation | Quantitative | Demonstrates explainability using SHAP and LIME. | Focuses more on technical output than end-user integration. |
| Deloitte & UOB (2021) | Case Study + Interview-Based Review | Industry Report | Real-world application of AI-human hybrid models | Case-specific and not generalisable across all banks. |
| AUSTRAC (2023–24) | regulatory Summary and Sector Review | Government Report | High-level operational and compliance insights | Limited detail on model evaluation or metrics |

## VI. THEMATIC SYNTHESIS AND FINDINGS

A detailed review of the literature reveals a constellation of interrelated themes that shape the discourse on AI-supported Anti-Money Laundering (AML) systems. Rather than treating technological capabilities in isolation, the literature emphasizes the entanglement of AI transparency, analyst cognitive dynamics, and regulatory alignment within evolving financial compliance ecosystems. This section synthesizes the findings using the interaction constructs introduced in Section III: trust calibration, oversight, feedback loops, and explainability, and maps these onto practical tensions observed in industry and empirical studies.

While Table III categorizes the literature into thematic domains, its real value lies in revealing the disciplinary requirements that fragment AI-AML research. A key pattern emerges: technical contributions largely optimize model performance metrics, such as precision or false positive (FP) reduction, without engaging with the socio-institutional challenges of operational trust and oversight. In contrast, regulatory and policy-focused literature emphasizes compliance defensibility but lacks guidance on integrating AI tools into analyst workflows or trust repair mechanisms following an AI error. This divide reflects an unresolved tension between algorithmic sophistication and institutional interpretability, which is central to real-world AI deployment in AML systems.

TABLE III. LITERATURE CLASSIFICATION BY THEMES

| Theme | Key Authors (Year) | Key Insights | Gaps | Gap Relevance* |
|---|---|---|---|---|
| AI in AML | Ketenci (2021), Oztas (2024), Kahur (2025) | ML & features reduce FP; new patterns | Few human-AI workflow studies | TC, FP |
| Human-AI Collaboration | Bertrand (2024), Wang (2024) | Trust, sensemaking, explanation paradox | No empirical AML banks data | TC, EX |
| Trust & Oversight | AUSTRAC (2024), Australian Gov (2024), King (2020) | Oversight & compliance defensibility | Generic, no dynamic trust focus | OV, TC |
| False Positives & Fatigue | Alahmadi (2022), Mounika (2024) | High FP rates cause decision fatigue | Limited AML-specific FP research | DF, TC |
| Explainability | Eddin (2021), Bertrand (2024), Jensen (2023) | SHAP/LIME aid sensemaking | Not embedded in daily workflows | EX, FO |
| Feedback Loops | Deloitte (2020), NICE Actimize (2020), Kumar (2024) | HITL & loops cut FP | Case-specific; lacks generalization | FL, TC |
| Regulatory Compliance | AUSTRAC (2024), Gov. Australia (2024) | Regulatory guidance & assurance | Broad, not tailored to AML | RC, OV |
| Situational Awareness | Boudt (2025), Rane (2023) | News monitoring boosts AML relevance | No human feedback integration | SA, OV |

TC = Trust Calibration; FP = False Positives; DF = Decision Fatigue; EX = Explainability; FO = Follow-through; OV = Oversight; FL = Feedback Loop; RC = Regulatory Compliance; SA = Situational Awareness.

Moreover, the table surfaces a notable epistemic gap: few studies offer empirical insight into how AML analysts interact with AI tools in practice. While theoretical models of human-AI collaboration, for example [17] and [18], conceptualize trust dynamics and explanation paradoxes, they remain untested in operational settings. This undermines our ability to assess decision fatigue, escalation behaviors, or feedback efficacy. Similarly, while explainability tools such as SHAP and LIME are technically validated, their real-world impact on analyst cognition and institutional defensibility remains speculative. These oversights suggest a need for workflow-embedded research that moves beyond algorithmic validation toward situated understanding of analyst behavior under compliance constraints.

Taken together, the classification in Table III does more than organize the literature; it reveals a systemic blind spot in current research: the lack of integrated, cross-disciplinary models that capture how trust, fatigue, explainability, and oversight interact in real-time AML operations.

Organizing the literature into these synthesized themes and acknowledging both convergence and contradiction, this section sets the stage for deeper thematic unpacking, detailed in Section VI(A) to Section VI(D). These themes are not merely descriptive categories; they are operational leverage points within the trust-feedback loop central to the conceptual framework.

### A. Trust Calibration

This theme highlights how fluctuations in analysts' trust, as discussed by [17] and [18], directly inform the design of the Dynamic Trust Modulation Loop (see Fig. 3), which operationalizes how trust is depleted after repeated false positives and restored through feedback-driven model adjustments.

### B. Theoretical Perspective: Human-AI Collaboration and Sensemaking in AML

To strengthen the understanding of how AML professionals collaborate with AI systems in transaction monitoring, this study draws on key concepts from Human-AI interaction theory: cognitive trust, situational awareness, and decision fatigue. These constructs are essential in analyzing how human analysts interpret, respond to, and calibrate their oversight of AI-generated alerts in high-stakes AML environments. This theme aligns with the conceptual framework in Fig. 2, where high false positives and alert ambiguity are shown to contribute to trust calibration challenges.

Cognitive trust reflects the degree to which users perceive AI systems as competent, reliable, and predictable. Prior research shows that trust in automation is shaped by tangibility, transparency, and task characteristics, especially relevant in compliance settings where legal accountability is still held by human actors [17]. When trust in AI drops due to unexplained false positives, analysts may disengage or ignore alerts entirely, creating operational blind spots.

Situational awareness involves understanding the transaction flagged by AI as well as the broader customer behavior, laundering typologies, and threat patterns. Overly complex or "black box" systems risk eroding this awareness [19]. Decision fatigue arises when analysts are overwhelmed by high alert volumes, often due to false positives. Bertrand introduces the "cry wolf effect", where repetitive false alerts reduce trust and increase the chance of missing legitimate threats [17].

Further research provides insight into this issue, indicating that security analysts working in similar high-alert environments often lack the contextual data needed to assess alert relevance [20].

This parallels AML settings, reinforcing the need for systems that enable explainability and reinforce human oversight. Graph-based machine learning triage models show promise; Feedzai reports an 80% reduction in false positives with over 90% detection of true positives [21]. The real-world relevance of such models emphasizes the value of applying theoretical insights to institutional settings.

Table IV reveals contrasting institutional approaches to trust and oversight in AI-supported AML systems. Cases like Westpac and CBA illustrate how poor governance and absent feedback loops lead to trust erosion, especially under high alert fatigue and low system transparency. In contrast, IDB Bank and

Chainalysis show that trust can be reinforced when automation is paired with high explainability and real-time feedback. Their success lies in workflows that support analyst understanding and continuously refine detection models.

Also in Table IV, the Santander case highlights the risks of relying on external AI vendors, but without full transparency or internal adaptability, institutions may lose control over explainability and oversight, weakening both analyst trust and regulatory assurance. Bunq's legal challenge underscores the role of regulatory alignment in trust calibration. Even high-performing models may face resistance if explainability standards are unclear or contested.

Overall, the mapping suggests that sustainable AML performance depends less on model complexity and more on system designs that embed transparency, human oversight, and adaptive feedback.

Table V draws on Alahmadi's thematic coding of analyst feedback to explore how cognitive trust, situational awareness, and oversight are shaped by real-world interactions with AI-generated alerts. These insights reveal how high-pressure environments strain the balance between automation and human judgment in AML contexts.

The coded responses illustrate a core tension: automation often falls short of supporting analysts in fast-paced environments. Instead of reducing workload, systems frequently overwhelm users with unclear or irrelevant alerts, triggering distrust and cognitive strain.

Situational awareness is another weak point. Analysts report difficulty in understanding system behavior or interpreting alerts without contextual clues, undermining their ability to act decisively and accurately. Trust calibration emerges as an emotional as well as procedural challenge. Analysts feel personally responsible for missed threats, especially when system limitations are known but uncorrected. This underscores the importance of transparency and oversight.

Importantly, the responses show consensus around the need for human-in-the-loop processes. Validation by analysts not only improves compliance defensibility but also restores confidence in system recommendations. So, the limited feedback from analysts back into AI models reflects a broader weakness in adaptive learning. Without structured error reporting, systems cannot evolve meaningfully, reinforcing inefficiencies and trust gaps over time.

TABLE IV. MAPPING OF CASE STUDIES TO FRAMEWORK COMPONENTS

| Case Study | Trust Calibration | Explainability | Situational Awareness | Decision Fatigue | Feedback Loop Integration |
|---|---|---|---|---|---|
| Westpac (AU) | Trust erosion due to systemic oversight failures | The low absence of clear rationale for alerts | Limited by poor governance structures | High, driven by excessive unresolved alerts | Absence of no mechanism to feed analyst observations back into the system |
| Commonwealth Bank of Australia (AU) | Over-reliance on partially automated systems without adaptive override | Moderation of some transaction data is available, but insufficient for risk context | Compromised in weak link between alerts and broader risk picture | Medium to High, given sustained operational strain | Minimal limited escalation learning |
| IDB Bank (US) | Strengthened trust through automation of post-alert processes | High and integrated case management provides contextual clarity | Enhanced of the analysts focus on high-risk escalations | Reduced 60% lower manual review workload | Strong feedback directly informs system tuning |
| Santander (EU) | Mixed efficiency gains but reliance on external AI reduces internal trust | Moderate vendor models lack full transparency | Variable dependent on third-party reporting | Medium, due to persistent false positives | Weak limited capacity to adapt vendor models |
| Bunq vs. DNB (NL) | Initially questioned by regulator; restored through legal vindication | Contested regulator deemed insufficient, court found adequate | Adequate AI model contextualized transactions effectively | Low automation reduced manual screening load | Moderate to internal testing but limited regulator feedback |
| Chainalysis & Elliptic (Global) | Generally high, supported by law enforcement use | High anomaly detection is well-documented | Strong in real-time blockchain monitoring | Low AI filters high-volume low-value alerts | Strong investigative feedback improves detection models |

TABLE V. THEMATIC ASSERTIONS FROM PARTICIPANT RESPONSES ALIGNED WITH AML THEORY

| Theme | Codes | Summary Result & AML Relevance |
|---|---|---|
| Cognitive Trust | A1, A8 | Analysts rely on personal judgment and feel strong responsibility when threats are missed, reflecting trust calibration and the emotional burden of oversight. |
| Situational Awareness | A2, A6, A10 | Lack of context, unclear alerts, and evolving system behavior hinder rapid, accurate decisions, aligning with situational awareness theory and explainability gaps. |
| Trust Erosion from False Positives | A3, A4 | Repetitive irrelevant alerts reduce trust, increase cognitive strain, and risk underreporting, supporting decision fatigue theory. |
| Oversight & Human-in-the-Loop | A5 | High consensus that human validation is essential for AI oversight, reinforcing analyst confidence and compliance defensibility. |
| Performance Pressure | A7 | Speed-focused performance metrics undermine investigative depth, linking to decision fatigue and compliance risks. |
| Feedback Loop Weakness | A9 | Limited reporting of errors back to systems hinders AI adaptation, evidencing feedback loop failures. |

### C. Understanding Accountability in Human-AI Decision-Making for AML Monitoring

AI's black-box nature complicates accountability in transaction monitoring, as prior research stresses that analysts must be able to justify AI-driven decisions through interpretability and effective oversight [18]. Related studies also warn against over-reliance on automated systems, noting that excessive dependence can erode critical thinking and increase regulatory risk [20]. Accountability-sharing frameworks are emerging, such as human-in-the-loop models and AI review committees, which formalize roles and support ethical oversight [22] and [23]. These mechanisms must be embedded in governance structures to uphold compliance defensibility and public trust.

### D. Post-Alert Investigation Processes

High false positive rates delay investigations and burden analysts; for example, NICE Actimize has improved workflows by centralizing alerts, SAR generation, and case management, as seen in IDB Bank's reported 60% manual workload reduction [12]. External tools like Belgium's NEMO integrate media signals into risk assessments, showcasing dynamic post-alert adaptations [24]. Meanwhile, predictive triage models further improve investigative accuracy [21].

Table VI summarizes the root causes of false positives in AML systems and links each to targeted mitigation strategies. These findings show how technical and governance gaps contribute to analyst fatigue and eroded trust.

Table VI outlines primary causes of false positives and mitigation strategies, reinforcing that explainability and feedback loops are critical to trust and efficiency.

Note 1: Several risk-based adaptive thresholds have been shown to significantly reduce false positive rates in AML models, as mentioned by Ketenci.

Note 2: Referring to SHAP values and other explainability tools, these provide transparency in model decisions, thereby increasing analyst confidence and reducing investigation time, as mentioned by Kahur.

Note 3: Feedback cycles allow AML professionals to improve model performance over time by incorporating real-world alert fixes, as mentioned by Sausen and Liegel.

As shown in Table VI, false positives are not just technical glitches; they are trust failures since the alerts lack context or cannot be explained, and analysts disengage, leading to oversight gaps and underreporting. The table highlights that improving accuracy alone is not enough. The most effective mitigation strategies, like SHAP-based explainability or human-in-the-loop retraining, address cognitive and procedural dimensions of trust.

Importantly, explainability functions as a trust calibration mechanism, so this one enables analysts to reconcile system output with professional judgment, reducing reliance on blind acceptance or rejection of alerts. However, transparency without action is useless, which is why feedback loops remain underdeveloped, forcing analysts to revalidate the same errors without influencing model behavior.

This lack of adaptation directly contributes to decision fatigue. Table VI illustrates that reducing false positives requires integrated solutions that align model design with analyst workflows, situational awareness, and regulatory defensibility.

*1) Key findings:* Prior research shows that trust in AI-driven systems is dynamic and influenced by factors such as false positives, limited explainability, and the degree of analyst control [18], [19], [25].

*2) Contradictions/trends:* Some perspectives conceptualize trust as a dynamic construct recalibrated through feedback mechanisms, while other accounts frame trust as primarily compliance driven [26]. Industry evidence further indicates that integrating structured feedback into AML systems can improve trust stability, as demonstrated in reports by NICE Actimize [12].

*3) Limitations/gaps:* There is limited empirical evidence on how trust is rebuilt. This highlights the value of the Dynamic Trust Modulation Loop (see Fig. 3), which offers a visual model to capture the trust-feedback interaction over time.

TABLE VI. SUMMARY OF FALSE POSITIVE CAUSES AND MITIGATION STRATEGIES

| Cause of False Positive | Description | Mitigation Strategy | Supporting Study |
|---|---|---|---|
| Rule-based thresholds | Static thresholds trigger alerts on legitimate transactions | Move to dynamic, risk-based profiling with adaptive AI | Ketenci et al. (2021) |
| Poor or incomplete data | Missing or outdated KYC data affects transaction interpretation | Improve data governance; refresh KYC frequently | AUSTRAC (2023); Jensen & Iosifidis (2023) |
| Model bias | Outdated or narrow training data flags benign behavior | Retrain models with updated, diverse data; integrate typologies | Tsapa (2023); Kahur (2025) |
| Lack of contextual awareness | Alerts ignore behavioral patterns, customer background, or geography | Incorporate contextual variables into feature sets | Deloitte (2020); Feedzai (2022) |
| Black-box outputs | Analysts can't understand why the system flagged the alert | Use explainable AI (e.g., SHAP values); implement model transparency features | Momenta (2022); Kahur (2025) |
| Feedback not integrated | Analyst input on false positives is not used to improve the system | Establish feedback loops and human-in-the-loop retraining protocols | NICE Actimize (2020); Sausen & Liegel (2020) |

### E. Explainability in AI Alerts

*1) Limitations of AI in AML monitoring: the false positive problem:* One of the most pressing challenges in AI-enabled AML systems is the persistently high rate of false positives, legitimate transactions incorrectly flagged as suspicious. As shown in Table II, methodological approaches vary across empirical and conceptual studies. Although AI and machine

learning (ML) have been praised for modernizing transaction monitoring, the practical reality is more complex. False positive rates frequently exceed 90% in both traditional rule-based and some AI-based models [27], [28]. This over-detection places a heavy burden on AML practitioners, who must manually review thousands of flagged transactions, most of which turn out to be harmless. Excessive alert volume not only leads to alert fatigue and increased operational costs but can also desensitize analysts, making it difficult to detect truly suspicious activity when it arises [28].

The effectiveness of AML systems lies in their capacity to direct human attention toward genuinely high-risk events, as noted in prior research [25]. When analysts spend most of their time reviewing false alerts, the fundamental objective of AML, preventing financial crime, is compromised. The psychological toll of continuous exposure to irrelevant alerts may also reduce an analyst's sensitivity to true red flags over time. These limitations are especially acute in smaller institutions with fewer resources, leading to greater reliance on manual reviews and further straining compliance operations [29].

The variability of AML-related data, which may involve hundreds or thousands of accounts, transactions, and behavioral features, adds to the challenge. Designing effective models requires both high-quality data and deep domain expertise to identify meaningful feature interactions. Consequently, many implementations suffer from performance limitations or require extensive manual configuration, which reduces scalability in high-volume environments [28].

To mitigate these issues, recent approaches have focused on integrating iterative, confidence-based learning cycles. In these models, analyst feedback is not only used to validate alerts but is also quantified and incorporated into retraining processes. Confidence indicators such as analyst override rates, decision certainty, and SHAP-based interpretability scores are used to assign weighted values to human corrections and feature relevance. This allows the system to prioritize learning from high-confidence analyst inputs, which are more likely to represent reliable domain knowledge.

What is particularly novel about this approach is that the model's learning is shaped around human decision-making. It does not merely enhance technical accuracy but also improves its ability to align with how analysts interpret and assess risk. This approach is aligned with AUSTRAC's expectations: AI systems must provide explainable decisions, incorporate risk considerations, and place human oversight at the core of the process. Moreover, it introduces the potential to use human trust itself as a feature in training AI systems for regulatory compliance contexts. For example, prior work has introduced novel feature sets based on time–frequency analysis, transforming transactional data into two-dimensional signal representations [30]. When implemented with Random Forest classifiers, this method reduced the false positive rate to 11.85% and improved the F1 score to 74.06%, demonstrating a more balanced trade-off between precision and recall.

As this study adopts a qualitative approach to understanding how AML practitioners manage false positives in AI systems, the proposed framework creates a foundation for future empirical validation. Variables such as perceived trust in AI, decision time, cognitive load, and analyst satisfaction with explainability could be measured through surveys or controlled experiments. These variables are closely tied to core components of the model, including trust calibration, feedback loops, and interpretability support tools. Accordingly, future research could adopt mixed methods designs to evaluate how these variables influence escalation outcomes and trust dynamics, strengthening both the empirical application and theoretical robustness of the proposed framework.

Table VII compares leading AI techniques used in AML transaction monitoring, evaluating their performance across false positive reduction, explainability, and real-world deployment. These comparisons reveal adjustments between model transparency and operational effectiveness.

The comparison in Table VII shows that high performance doesn't always mean high transparency, given that graph-based models reduce false positives effectively but lack full explainability, which may hinder regulatory trust. On the other hand, SHAP values, while highly interpretable, lack published data on their real-world impact. This reflects a broader tension between model transparency and measurable efficiency.

Eventually, techniques that align with analyst workflows by aiding prioritization and enabling oversight offer more than just accuracy. They support trust, usability, and compliance defensibility. Table VII supports the need for hybrid AI designs that balance performance, explainability, and human feedback integration.

TABLE VII. CROSS-COMPARISON OF AI TECHNIQUES IN AML

| AI Technique | Use Case | False Positive Reduction | Explainability | Real-World Application |
|---|---|---|---|---|
| Random Forest + Time-Frequency | Transaction pattern detection | Reduced FPR to 11.85% | Moderate | Mounika et al. (2024) |
| SHAP Values | Model interpretability | Data on false positive reduction are not disclosed in public sources. | High | Momenta (2022); Kahur (2025) |
| Graph-Based Triage Model | Risk prioritization via transaction networks | Reduced FPR by up to 80% | Moderate | Feedzai (2022); Eddin et al. (2022) |

Practical deployments have begun to address long-standing limitations; the best example might be "NICE Actimize, which has demonstrated a 30–50% reduction in false positives, improving the efficiency of compliance teams" [12]. As summarized in Table III, literature themes converge around trust, explainability, and regulatory compliance. These findings suggest that advanced feature engineering, including time frequency features such as kurtosis and skewness, can lead to more accurate and operationally viable AML systems [31].

Despite these improvements, false positives remain a persistent challenge; until AI models can reliably distinguish

between genuinely suspicious transactions and benign anomalies, human analysts will remain essential to the investigation process, and the vision of full automation will remain unattainable. As emphasized throughout the literature, improving model explainability, refining input features, and enhancing human–AI collaboration are critical for ensuring that AI augments rather than overwhelms the AML function. This reinforces the relevance of the core analyst–system interaction constructs articulated in human–AI collaboration theory. Analysts continue to experience decision fatigue when static thresholds consistently flag non-suspicious behavior.

- Key findings: Prior research emphasizes that explainable AI (XAI) tools, such as SHAP and LIME, can help analysts better understand system outputs and increase confidence [21], [31], [32]. However, practical adoption of these tools remains uneven.
- Contradictions/Trends: Although a degree of consensus exists regarding the utility of technical explainability tools, prior studies emphasize that these tools are often disregarded when they are not seamlessly integrated into analysts' daily workflows, particularly through accessible visualizations [32], [33]. Several studies highlight the tension between explanation complexity and the speed required for real-time decision-making.
- Limitation/Gap: There is a lack of empirical evidence demonstrating whether explainability features directly reduce false positives or analyst fatigue in operational settings. The inclusion of "explainability satisfaction" as a construct within this study's framework aims to address this measurement gap by capturing the analyst's perception of the clarity and usefulness of AI-generated outputs.

### F. Oversight and Governance

*1) Human-AI collaboration in AML compliance:* The integration of artificial intelligence (AI) into anti-money laundering (AML) systems has introduced a dual-layered structure of operational support and cognitive challenge. While AI enhances scalability and anomaly detection, it also generates persistent governance issues, particularly concerning false positives, opaque decision-making, and compliance defensibility. As such, oversight in AI-AML systems must be understood not only as a regulatory function but also as a dynamic interface between automated pattern recognition and human interpretive judgment [19], [34].

Despite the adoption of advanced AI tools, studies continue to report false positive rates as high as 90–95% [33], leading to alert fatigue and cognitive overload among analysts [4]. These issues are compounded by the reliance on historical data that may encode legacy biases, leading to detection errors and redundant investigations. In response, human analysts play a critical oversight role: validating AI outputs, applying contextual knowledge, and modulating escalation behavior functions that current AI systems are poorly equipped to replicate.

The following Table VIII clearly summarizes the practical response strategies employed by AML analysts when addressing AI-generated alerts, highlighting the range of responses that effectively mitigate alert fatigue and enhance compliance outcomes.

The interview responses in Table VIII reveal that alert fatigue, lack of feedback loops, and limited explainability significantly undermine trust in AI-assisted AML systems. Analysts report defaulting to intuition over system guidance and express frustration when feedback isn't acknowledged, indicating weak cognitive trust and disengagement from human-in-the-loop processes. While tools like SHAP offer some interpretability, they are not sufficient on their own; without actionable transparency and feedback integration, AI systems risk being sidelined despite their technical performance. These practitioner themes reinforce that trust calibration must be embedded in system design to ensure adoption and compliance defensibility.

TABLE VIII. PRACTITIONER THEMES FROM INTERVIEWS

| Theme | Sample Quote | Frequency | Interpretation |
|---|---|---|---|
| Alert Fatigue | I catch myself rushing after 20 false alerts in a row. | Q3, Q15 | Decision Fatigue |
| Trust in AI | Sometimes I ignore the system and trust my instincts. | Q7, Q8 | Cognitive Trust |
| Lack of Feedback Loops | I've submitted feedback, but I don't know if it's used. | Q10 | Human-in-the-loop |
| Need for Explainability | SHAP helps me understand why the alert was triggered. | Q9, Q16 | Interpretability |

## VII. REGULATORY FRAMEWORKS AND GOVERNANCE UNDER AUSTRAC

In the Australian context, AUSTRAC operates both as a financial intelligence unit and the chief regulator responsible for enforcing the AML/CTF Act 2006. Although AUSTRAC mandates that all transaction monitoring programs, manual or automated, be risk-based and support "reasonable grounds for suspicion" in SMR submissions, the agency has yet to issue dedicated AI-specific guidance. This omission has created a regulatory blind spot in which institutions must interpret traditional compliance obligations in the context of complex, often opaque, AI systems [1], [35].

The operational consequences of this regulatory ambiguity are evident in Table IX, which highlights the misalignment between AUSTRAC expectations and common AI system features. For instance, while human validation and feedback loops are increasingly present in advanced setups, critical areas such as transparency, accountability, and real-time contextual awareness remain underdeveloped. High-profile failures at institutions like Westpac and CBA have illustrated the dangers of opaque models and unmanaged alert volumes, while newer hybrid models from Bunq to Elliptic showcase the value of explainability and real-time feedback integration.

TABLE IX. AI FEATURES VS. AUSTRAC PRINCIPLES: AN OPERATIONAL ALIGNMENT

| AUSTRAC Expectation (Grouped) | Typical AI Feature | Compliance Gap / Risk + Construct* | Real Case Example (Year) |
|---|---|---|---|
| Risk-based monitoring & SMRs | ML detects patterns; real-time anomaly detection | Over-flagging & alert fatigue (DF); delayed SMRs (OV) | Santander fine (2025); CBA SMR failure (2018) |
| Auditability & Explainable decisions | Black-box AI; complex algorithms | Limited traceability (EX); need human validation (SA, TC) | Westpac breaches; Bunq interpretable AI (2021) |
| Ongoing CDD/KYC & Oversight | Profile risk scoring; semi-autonomous operation | Misses' real-time context (SA); oversight needed (OV, FL) | Chainalysis blockchain data; Elliptic hybrid alerts (2025) |

Legend: DF = Decision Fatigue; OV = Oversight; EX = Explainability; SA = Situational Awareness; TC = Trust Calibration; FL = Feedback Loop.

Table IX illustrates critical misalignments between AUSTRAC's regulatory expectations and current AI practices in AML. While AI tools offer real-time monitoring and pattern detection, they often trigger excessive false positives and lack transparency, leading to delayed reporting and reduced auditability. These gaps highlight the persistent need for human oversight to contextualize system outputs and maintain regulatory compliance. Moreover, inconsistent implementation and poor documentation practices, such as reliance on ad hoc risk assessments, undermine both accountability and trust. The analysis reinforces that AI must be integrated not only as a technical solution but within a governance structure that supports explainability, oversight, and traceability.

Moreover, AUSTRAC's regulatory reach does not yet extend to high-risk sectors such as legal, accounting, and real estate services (DNFBPs), creating systemic gaps that sophisticated laundering networks continue to exploit. Proposed "Tranche 2" reforms aim to address these omissions, but until implemented, oversight remains fragmented, undermining both national security and FATF compliance rankings [36].

## VIII. GOVERNANCE AND THE FUTURE OF OVERSIGHT

Traditional compliance models, which rely on linear thresholds and rule-based control points, are increasingly unfit for AI-enhanced AML environments. Instead, a new paradigm of adaptive governance is emerging, one that integrates explainability tools such as SHAP and LIME with real-time trust metrics and analyst confidence scores to modulate compliance thresholds dynamically. This approach aligns with AUSTRAC's emphasis on risk-based monitoring while operationalizing transparency, traceability, and human accountability as interdependent elements of effective oversight.

AUSTRAC's alignment with the Australian Government's AI Assurance Framework [37] further reinforces this shift. The framework's eight principles explainability, contestability, accountability, transparency, fairness, reliability, privacy, and human-centered values echo AUSTRAC's stated priorities but go further by offering specific policy tools such as lifecycle risk assessments and AI transparency statements. These tools provide institutions with a governance scaffolding that not only supports compliance but also anticipates the ethical and legal challenges posed by increasingly autonomous AI systems.

The implication is clear: oversight in AI-AML systems must evolve from static rule adherence to dynamic, feedback-integrated governance. This transition demands that institutions embed structured feedback loops, document decision pathways, and ensure that human analysts remain central in validating and contextualizing automated outputs. Without these mechanisms, the promise of AI in AML remains undermined by its own operational opacity and regulatory uncertainty.

### A. Systemic Risks

Systemic risks in artificial intelligence-driven anti-money laundering (AI-AML) frameworks emerge not solely from algorithmic inefficiencies but from misalignments between institutional practices, regulatory expectations, and the operational realities of compliance systems. Within the Australian context, the AUSTRAC regulatory environment, while comprehensive in principle, reveals structural and procedural gaps that, if unaddressed, may scale into broader systemic vulnerabilities.

## IX. STRUCTURAL AMBIGUITY AND COMPLIANCE OVERLOAD

AUSTRAC's AML/CTF Act 2006 does not distinguish between manual and AI-based transaction monitoring systems, applying uniform expectations around Suspicious Matter Report (SMR) submissions based on "reasonable grounds for suspicion". However, this regulatory equivalence masks a key risk: AI systems often generate large volumes of alerts, many of which are opaque and lack interpretability. Institutions are legally obligated to process these alerts with the same rigor as human-generated suspicions, potentially overwhelming compliance workflows and increasing the likelihood of delayed or inappropriate filings, a systemic failure point [1].

The absence of formal thresholds for false positive or false negative rates introduces additional ambiguity. Without clear benchmarks, institutions are left to self-assess AI system performance, resulting in inconsistent standards of compliance across the sector. Table X illustrates how these misalignments manifest: while "effectiveness" and "ongoing oversight" are broadly achieved, critical areas such as transparency, accountability, and traceability remain structurally deficient. This incomplete alignment reflects a broader systemic fragility where critical compliance decisions may rely on black-box outputs with limited auditable pathways.

Table X assesses how current AI systems align with AUSTRAC's AML/CTF regulatory principles, based on system characteristics and observed operational practices.

As shown in Table X, it illustrates how AUSTRAC and its principles for compliance systems align, or do not align, with the typical characteristics of AI-based transaction monitoring. It is understood that AI can improve efficiency and facilitate oversight when implemented in feedback mechanisms, which

often present deficiencies in areas such as transparency, accountability, and traceability. These weaknesses reinforce the need for continuous human oversight, explainable models, and improved auditability to ensure compliance with AUSTRAC standards.

TABLE X. OPERATIONAL ALIGNMENT: AUSTRAC PRINCIPLES AND AI FEATURE ANALYSIS

| AUSTRAC Principle | AI Feature | Operational Scenario | Alignment |
|---|---|---|---|
| **Transparency** | Deep learning model | Analyst receives an alert but cannot explain why it was triggered | It's not aligned |
| **Accountability** | Automated alert generation without review | Alerts are escalated without human validation | It's not aligned |
| **Traceability** | No logging of analyst actions or model decisions | Supervisor cannot trace why an alert was closed or escalated | It's not aligned |
| **Proportionality** | Rule-based scoring applied uniformly | Low-risk customers flagged with the same thresholds as high-risk clients | Its partially aligned but need improvements |
| **Effectiveness** | AI flagging paired with human validation | Analyst reviews a flagged transaction, confirms suspicion, files SMR | Aligned |
| **Ongoing Oversight** | Feedback loops and the model retraining | Analyst labels false positives: system learns and improves over time | Aligned |

## X. FRAGMENTED OVERSIGHT AND RISK DIFFUSION

As highlighted in AUSTRAC's expectations and reflected in the Australian Government's AI Assurance Framework, human oversight is essential not only for ethical compliance but for the iterative improvement of AI systems. However, most AML frameworks fail to operationalize this principle. Feedback from analysts is often informal, undocumented, or siloed. Consequently, AI models are rarely retrained based on field-level insights, allowing detection errors to propagate unchecked. This oversight vacuum introduces not only model stagnation but also a false sense of algorithmic reliability, amplifying systemic exposure to compliance failure.

Moreover, AUSTRAC's current scope excludes several high-risk professions such as lawyers and real estate agents (DNFBPs), leaving exploitable gaps in the regulatory perimeter. These omissions enable criminal networks to route illicit flows through sectors with reduced AI scrutiny, creating shadow pathways that weaken the broader AML ecosystem. As described in international comparisons (see Table XI), other jurisdictions such as the UK's FCA or the US FinCEN offer more robust mechanisms for AI explainability, sandbox testing, and auditability, reflecting a maturity in systemic risk mitigation that AUSTRAC has yet to formalize.

TABLE XI. GLOBAL AI-AML REGULATORY COMPARISON

| Regulatory Body | AI & AML Focus | Key Compliance Expectation | Implication for AUSTRAC |
|---|---|---|---|
| AUSTRAC (Australia) | Risk-based, flexible, minimal AI-specific rules | Outcome-focused SMR accountability | AUSTRAC could benefit from transparency requirements by providing new, clearer metrics for measuring AI traceability and SAR justification. |
| FinCEN (Financial Crimes Enforcement Network) (US) | Supports AI-driven SARs, stresses auditability | Explainable, testable AI systems | AUSTRAC could benefit transparency requirements by providing new, clearer metrics for measuring AI traceability and SAR justification. |
| FCA (Financial Conduct Authority) (UK) | High transparency and fairness in AI models | Clear algorithmic accountability, sandboxing | AUSTRAC could benefit from incorporating new FCA model explainability standards, promoting tools like SHAP and new decision tree interfaces. |
| FATF (Financial Action Task Force) (Global) | Global oversight principles: explainability, proportionality | Human oversight + innovation risk control | AUSTRAC can lead implementation by FATF members by proposing national AI audit protocols and assisting with participation in FATF AML working groups. |

Table XI highlights key differences in how major regulatory bodies approach AI governance in AML, revealing a spectrum from general principles to detailed algorithmic standards. While AUSTRAC maintains a flexible, risk-based approach, it lacks AI-specific guidance compared to FinCEN and the FCA, which emphasize explainability, auditability, and accountability. The FCA's sandboxing and model transparency initiatives, along with FinCEN's support for testable AI systems, suggest a more proactive stance that AUSTRAC could emulate. Meanwhile, FATF's global guidance underscores human oversight and proportionality but stops short of prescribing technical norms. The comparison signals a clear opportunity for AUSTRAC to lead regionally by adopting clearer traceability metrics, formal AI audit protocols, and aligning with emerging global standards on explainable machine learning in financial compliance.

## XI. GLOBAL GAPS AND INTEROPERABILITY RISKS

On a global level, the divergence in regulatory maturity across jurisdictions creates further systemic risk through regulatory arbitrage. Institutions operating in multiple countries must navigate heterogeneous expectations for AI transparency and accountability, which can lead to fragmented governance strategies. As shown in Table XI, AUSTRAC's outcome-based model lacks the explicit interpretability and audit trails mandated by bodies like FinCEN or the FCA. This fragmentation may incentivize minimal compliance in lower-regulation jurisdictions, while overburdening analysts in stricter environments, distorting the effectiveness of AI-AML implementation globally.

### A. Implications for Framework Development

The cumulative effect of these systemic factors' regulatory ambiguity, deficient oversight, fragmented feedback loops, and global interoperability gaps poses a risk not only to institutional compliance but also to national and transnational financial security. The proposed framework responds to these challenges by embedding structured human-in-the-loop mechanisms, adaptive threshold modulation, and explainability tools that align with both AUSTRAC's principles and international standards. By explicitly linking trust calibration to analyst feedback and system transparency, the framework mitigates systemic fragility and enhances the defensibility of SMR decisions.

## XII. Discussion

The proposed framework was developed following a Design Science Research (DSR) approach, integrating patterns, themes, and operational gaps identified in 61 systematically selected studies published between 2020 and 2025, alongside regulatory reports and industry guidelines. This structured process ensured that each component of the model is logically derived from both academic evidence and practitioner insights, aligning with AUSTRAC's compliance principles.

Unlike previous AML frameworks, which often treat human review as a final, static checkpoint, this model introduces two novel elements: 1) a dynamic trust modulation cycle, capturing how analyst confidence fluctuates and recalibrates in response to false positives; and 2) explicit feedback loops that feed analyst input directly into AI model refinement.

While the present study focuses on theory building, future research will empirically validate this model through quantitative surveys and statistical analysis to test the hypothesized pathways between explainability, trust, fatigue, and escalation decisions.

Fig. 2 illustrates this journey and highlights where key theoretical concepts, such as decision fatigue, confidence calibration, and explainability, are activated in practice.

The proposed human-centered framework focuses on a design that specifically addresses key shortcomings in both the academic literature on AI-AML systems and practitioner rules, such as those of AUSTRAC.

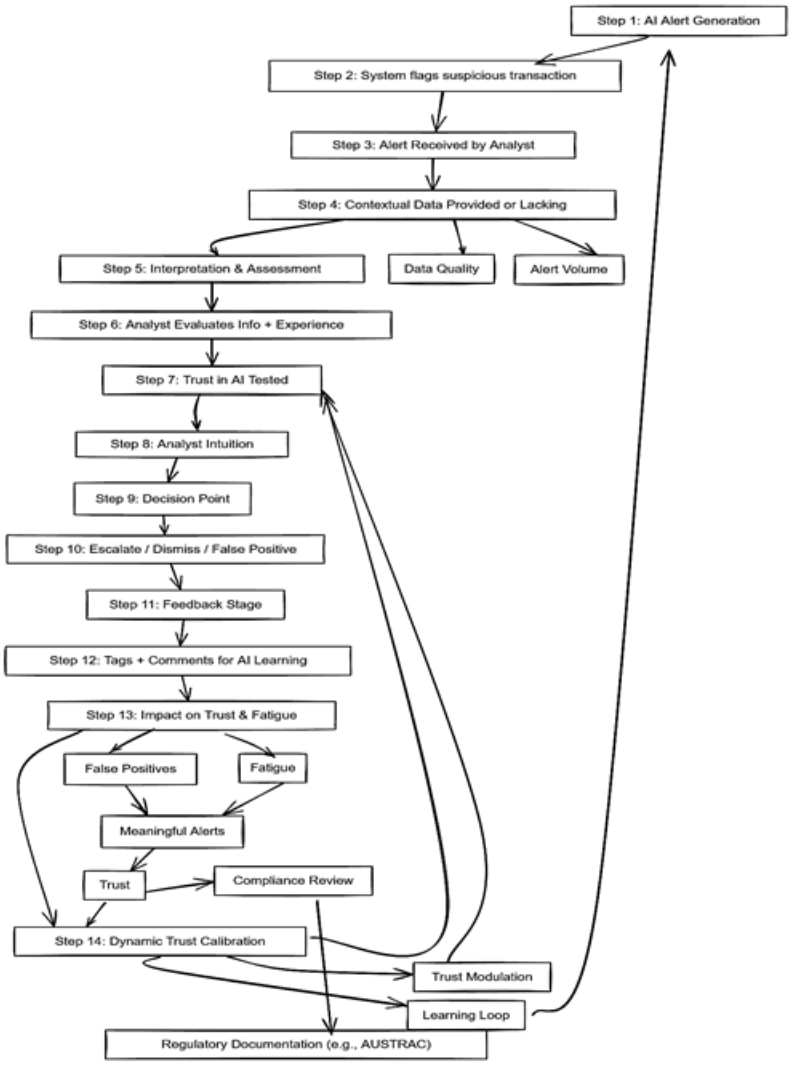


Fig. 2. Conceptual framework: AML analyst journey.

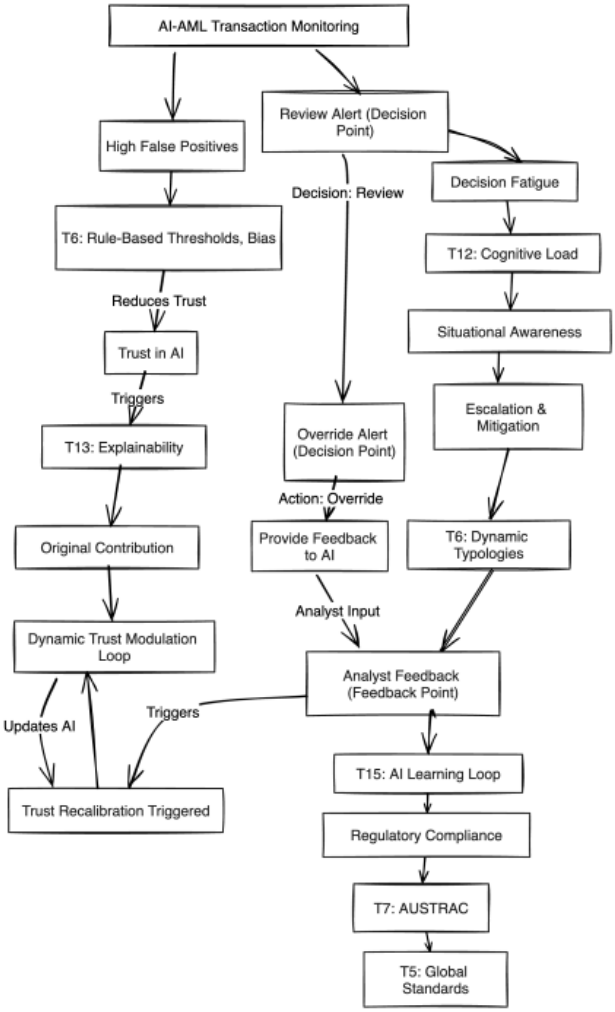


Fig. 3. Analyst workflow model for AI-supported AML decision-making.

The previous discussion on practical strategies, alert fatigue, and regulatory frameworks laid the foundation for the conceptual model. This model integrates theoretical constructs such as human-AI trust, sensemaking, and regulatory alignment, and is illustrated in Fig. 3.

This framework seeks to improve the interaction between human and artificial intelligence decisions in anti-money laundering (AML) settings by dynamically calibrating trust, decision fatigue, and continuously enhancing situational awareness during the alert review process. The goal is to shorten the time between static compliance checkpoints and real-world analytical understanding by integrating explicit feedback loops, trust-based escalation thresholds, and adaptive compliance monitoring.

As shown in Fig. 3, the feedback loop mechanism illustrates this integration. The framework introduces two novel contributions absent in most current AML models:

- A dynamic trust modulation cycle, in which trust in AI declines in response to repeated false positives and is gradually restored through feedback-driven improvements; and
- Explicit feedback loops, whereby analysts' responses and insights are integrated into the model's learning process, thereby narrowing the gap between human decision execution and system adaptation. This process is illustrated in Fig. 4.

### *A. Contribution to Anti-Money Laundering (AML) Research and Practice*

This study contributes to the AML field by introducing a dynamic, human-centered conceptual framework that operationalizes critical analyst–AI interaction variables such as trust calibration, decision fatigue, explainability, and regulatory alignment. Unlike previous models that treat human review as a static checkpoint, the framework demonstrates how analyst trust evolves in response to system outputs and feedback integration, and how this process impacts escalation decisions and regulatory defensibility.

The Framework Analytical Layer Mapping (Table XII) ensures systematic traceability from literature synthesis to conceptual modeling. This transparent development process strengthens the framework's credibility and auditability for practical application in real-world AML settings. It also bridges the gap between theoretical constructs and operational tools,

offering a structure that aligns with AUSTRAC principles and regulatory trends in high-risk financial environments.

Furthermore, the framework provides practical value for compliance officers and AML teams by embedding explainability, analyst confidence, and oversight into daily operations. Through trusted nodes, feedback loops, and threshold calibration, it addresses key gaps identified in AUSTRAC's expectations and supports more defensible Suspicious Matter Reports (SMR).

TABLE XII. FRAMEWORK ANALYTICAL LAYER MAPPING

| Table/Figure | Title | Framework Stage | Key Construct(s) |
|---|---|---|---|
| **Table I** | Literature Review Source Overview | Included in all stages | All Constructs |
| **Table II** | Methodologies in Reviewed Studies | Included in all stages | All Constructs |
| **Table III** | Literature Classification by Themes | Stage 1: Supporting Theme Extraction | Trust, Fatigue, Situational Awareness |
| **Table IV** | Mapping of Case Studies to Framework Components | Stage 1: Theme-to-Framework Alignment | Trust, Explainability, Feedback Loops |
| **Table V** | Thematic Assertions from Participant Responses Aligned with AML Theory | Stage 1: Practitioner Theme Extraction | Trust, Fatigue, Situational Awareness |
| **Table VI** | Summary of False Positive Causes and Mitigation Strategies | Stage 2–3: Problem–Solution Mapping | Decision Fatigue, Trust, Feedback Loops |
| **Table VII** | Cross-Comparison of AI Techniques in AML | Stage 2: Technical System Evaluation | Explainability, Trust |
| **Table VIII** | Practitioner Themes from Interviews | Stage 1–2: Theme Validation | Trust, Fatigue, Explainability |
| **Table IX** | AI Features vs. AUSTRAC Principles: An Operational Alignment | Stage 3: Regulatory Gap Alignment | Compliance Defensibility, Oversight |
| **Table X** | Operational Alignment: AUSTRAC Principles and AI Features Analysis | Stage 3: Regulatory Analysis | Situational Awareness, Oversight |
| **Table XI** | Global AI-AML Regulatory Comparison | Stage 3: Regulatory Comparison | Oversight, Trust Calibration |
| **Table XII** | Framework Analytical Layer Mapping | Stage 4: Analytical Impl. | All Constructs |
| **Table XIII** | Contribution to Explainable AI in AML | Stage 3–4: Novelty Justif. | Explainability, Trust |
| **Table XIV** | AI Error Types and Human Responses | Stage 2: Problem Id. | Trust, Decision Fatigue |
| **Table XV** | Linking Theory–Objective Mapping | Included in all stages | Trust, Fatigue, Situational Awareness |
| **Table XVI** | Summary of Identified Gaps | Stage 3: Gap Synthesis | Trust, Oversight, Feedback, Explainability |
| **Table XVII** | Alignment of Regulatory Expectations with AML System Capabilities | Stage 3: Regulatory Fit–Gap Closure | Compliance Defensibility, Oversight |
| **Table XVIII** | Mapping of Framework Constructs to Potential Measurable Indicators | Stage 4: Operational Metrics Layer | Trust, Feedback, Decision Fatigue, Explainability |
| **Table XIX** | Appendix A – Likert-Scale Survey for AI-AML Constructs | Future Empirical Layer | All Constructs |
| **Table XX** | Appendix C: Constructs and Measurement for Empirical Validation | Future Empirical Layer | Trust, Feedback, Decision Fatigue, Explainability |
| **Table XXI** | Appendix C: Mapping of Interview Questions to Research Objectives and Theoretical Constructs | Future Empirical Layer | Trust, Fatigue, Situational Awareness |
| **Figure 1** | PRISMA Flow Diagram of Literature Selection Process | Stage 1: Literature Selection Justification | All Constructs |
| **Figure 2** | Conceptual Framework: AML Analyst Journey | Stage 4: Framework Synthesis | Trust, Explainability, Feedback, Compliance |
| **Figure 3** | Analyst Workflow Model for AI-Supported AML Decision-Making | Stage 4: Process Operationalization | Trust, Feedback Loops, Situational Awareness |
| **Figure 4** | Dynamic Trust Modulation Loop | Stage 4: Novel Component Highlight | Trust Depletion/Restoration |
| **Figure 5** | Feedback Loop Mechanism | Stage 4: Feedback Operationalization | Analyst Feedback, System Learning |

### *B. Contribution to AI Trust and Human–AI Collaboration*

The study advances cognitive trust theory in regulated AI contexts by introducing the Dynamic Trust Modulation Cycle, a core innovation that captures how repeated false positives erode analyst trust and how structured feedback and improved explainability restore it. This contribution builds upon, but extends beyond, models like Chhetri's A²C framework by explicitly incorporating decision fatigue, situational awareness, and escalation confidence into the trust-feedback dynamic.

This process is illustrated in Fig. 4, as analyst confidence declines in response to repeated false positives, but it also demonstrates that confidence can be recalibrated through structured feedback and model improvements. Visualizing this dynamic cycle of trust modulation, human oversight, and feedback integration emphasizes that it functions as a continuous process rather than remaining static as a final checkpoint.

To understand the dynamic process of trust calibration, the framework also incorporates explicit feedback loops that connect analyst decisions to AI. Since traditional AML systems typically consider human review as a final step in the process, this design demonstrates that analyst input, such as the confirmation of false positives or the justification for escalating

an alert, is systematically captured and used to refine the model's thresholds, just like feedback data.

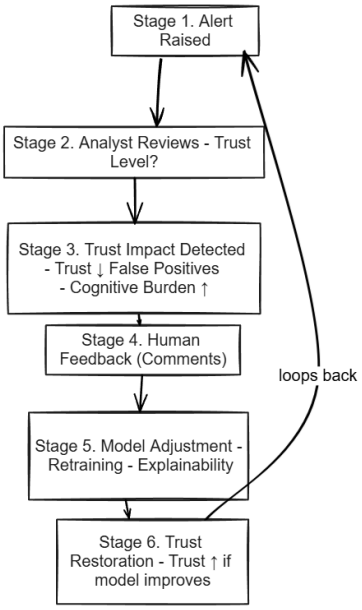


Fig. 4. Dynamic trust modulation loop.

Fig. 5 provides an overview of this mechanism, demonstrating how human judgment continuously shapes system behavior over time and strengthens the compliance case.

Fig. 5 shows the explicit feedback loop for model refinement. It illustrates how analyst feedback on alerts triggers system retraining and threshold adjustments, closing the loop between human judgment and AI model evolution. This cycle enhances the iterative, human-centered approach that distinguishes this framework from static compliance models and ensures that analysts have integrated expertise into the continuous improvement of AI systems.

In the following Table XIII, there are studies that have contributed to the development or application of explainable AI in AML monitoring. It highlights tools such as SHAP values, decision trees, and visual interfaces, and links these contributions to analyst confidence, transparency, and the veracity of effective processes.

As shown in Table XIII, the different AI tools and their explainability, as well as SHAP, LIME, and visual dashboards, directly contribute to analyst efficiency in clarifying model logic and improving alert interpretation. This is especially important in AML/CFT environments, when analysts must justify their case escalation and compliance reporting, following AUSTRAC regulatory standards. When AI decisions are more transparent and interpretable, it reduces decision fatigue, builds cognitive trust, and facilitates human validation, improving operational efficiency and accountability in compliance workflows.

In addition to explicit feedback loops, this framework positions itself as a bridge between algorithmic decision-making and human accountability, in line with AUSTRAC's regulatory expectations. These elements directly support the research objectives, specifically the investigation of analyst behavior in handling alerts, the strategies they adopt to manage false positives, and the regulatory compliance they must follow.

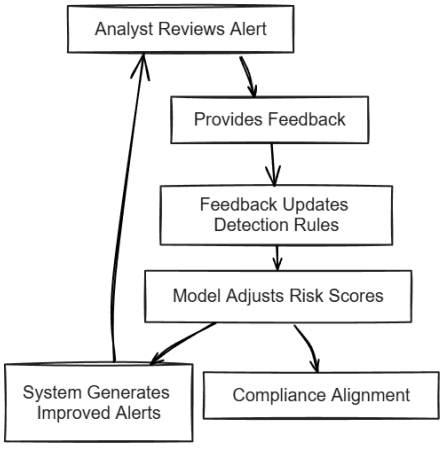


Fig. 5. Feedback loop mechanism.

TABLE XIII. CONTRIBUTION TO EXPLAINABLE AI IN AML

| Study | XAI Techniques Used | Contribution to Transparency | Compliance Impact |
|---|---|---|---|
| Kute et al. (2021) | SHAP, LIME, Decision Trees | Improved interpretability of AI-generated alerts for compliance analysts | Supports justifiable SMR decisions; enhances trust in AI recommendations |
| Bertrand (2024) | Conceptual - Explainability as an ethical requirement | Highlights explainability as key to human trust and accountability | Aligns with AUSTRAC's 'reasonable grounds' requirement for SMRs |
| Australian Government (2024) | Framework Guidelines (Transparency Standards) | Calls for formal documentation of model logic and intended use | Reinforces AUSTRAC-aligned auditability and oversight expectations |
| Ghimire (2025) | Precision-recall, threshold tuning | Optimizes model performance visibility | Helps reduce false positives, supports model defensibility |

TABLE XIV. AI ERROR TYPES AND HUMAN RESPONSES

| Study | Error Type Focused | Human Response Strategy | Tools/Models Used |
|---|---|---|---|
| Mounika et al. (2024) | False Positives | Time-frequency features to reduce alert overload | Random Forest + Skewness/Kurtosis |
| Eddin et al. (2021) | False Positives | AI enhancement with human validation loop | Machine learning classifier ensemble |
| Ghimire (2025) | False Positives | Precision-recall optimization to reduce noise | Rule-based + simulated AI outputs |
| Ketenci et al. (2021) | False Positives | Data integration to reduce alert redundancy | Decision Trees, Logistic Regression |
| Bertrand (2024) | Cognitive Overload / Alert Fatigue | Human-AI trust calibration and workload awareness | Theoretical (no models) |
| Deloitte & UOB (2021) | Alert overload | Human-in-the-loop feedback for dynamic tuning | Anomaly Detection + Feedback Loops |

Table XIV shows the main types of errors generated by AI in AML systems, such as false positives and the errors it can make when misclassifying them. It also summarizes how analysts respond in practice, highlighting the importance of explainability and a structured process to prevent overload and improve compliance outcomes.

Table XIV shows a comparative summary of studies highlighting the types of artificial intelligence errors, particularly false positives and cognitive overload in analysts, and the human response mechanisms used to mitigate their impact in AML contexts. For a better understanding, each row links a specific error type, such as false positives and alert fatigue, with corresponding strategies, such as feedback gaps, data integrations, or confidence calibration. This comparison between the reference error type and the response mechanism reveals a need to balance technical optimization with human-centered design. Importantly, the table reveals that since most studies focus on technical refinement, not many include confidence as a dynamic and modifiable variable.

The Theory Objective Mapping (Table XV) demonstrates how constructs such as trust, fatigue, and situational awareness are not abstract but are directly linked to AML professionals' lived experiences. Each theoretical construct is aligned with specific research objectives and supports the framework's relevance for both practitioners and scholars seeking to understand human–AI interaction in compliance-driven domains.

As shown in Table XV, a structured mapping between the study's research objectives and their corresponding theoretical foundations ensures the conceptual integrity of the proposed framework. Each objective, ranging from managing false positives to strengthening compliance defensibility, is explicitly linked to established theories such as cognitive trust, human–AI collaboration, and explainability.

The framework transforms conceptual elements of trust and human oversight into measurable, actionable tools for AML professionals. By integrating analyst trust levels, real-time confidence scores, and feedback loops into system retraining protocols, it offers a nuanced mechanism for ensuring compliance that is both adaptive and explainable.

### C. Practical and Policy Implications

From a practical standpoint, the architecture offers AML practitioners a clear mechanism for managing false positives, establishing trust levels, and directly incorporating analyst feedback to retrain the system. Through trusted nodes, explainability tools, and feedback loops in daily operations, the framework facilitates real-world compliance defenses based on AUSTRAC's "reasonable grounds" principle. Institutions can demonstrate to regulators that human judgment is not only present but consistently guiding the evolution of AI systems over time.

The framework also contributes policy insights by highlighting the absence of specific regulatory thresholds and explainability standards in existing AUSTRAC guidance. By embedding trust calibration and analyst feedback into the compliance lifecycle, it sets the stage for regulatory advancements aligned with global best practices.

Tables such as the Summary of Identified Gaps (Table XVI), Framework Analytical Layer Mapping (Table XII), and Alignment with Regulatory Expectations (Table XVII) provide empirical support for how the framework addresses persistent challenges in AML operations and policymaking.

Table XVI summarizes these critical gaps and demonstrates their direct relevance to the study's framework.

TABLE XV. LINKING THEORY-OBJECTIVE MAPPING

| Theory / Construct | Description | Linked Research Objectives |
|---|---|---|
| Trust Theory | Trust in AI systems and cognitive trust determine how analysts accept that decision, override the alert, or otherwise escalate the AI alert. This includes the responsibility for monitoring and calibration after each alert. | RO1: Understand how professionals detect FPs<br>RO2: Understand how professionals evaluate AI alerts<br>RO3: Assess confidence in AI-based alerts |
| Decision Fatigue Theory | The cognitive load and mental fatigue caused by frequent false positives affect analysts' concentration, willingness to investigate, and the accuracy with which they escalate each alert or case. | RO1: Understand impact of FPs on workload<br>RO2: Understand how professionals evaluate AI alerts<br>RO4: Improve human-AI collaboration and effectiveness |
| Situational Awareness Theory | When we talk about analysts' ability to perceive, interpret, and act on contextual information, it can impact judgment, confidence, escalation confidence, and compliance defensibility. | RO2: Understand how professionals evaluate AI alerts<br>RO3: Assess confidence in AI-based alerts<br>RO4: Improve human-AI collaboration and effectiveness |

TABLE XVI. SUMMARY OF IDENTIFIED GAPS

| Framework Component | Gap in Literature |
|---|---|
| Feedback loops to AI systems | Most studies addressing AI-AML view trust as something static, something that doesn't improve or makes little difference, which means that how analysts' trust evolves isn't explored (Fig. 4). |
| Analyst fatigue | Cognitive fatigue is often underexplored in AML research, especially in the context of AI false positives (Q5). |
| Explainability | How analyst feedback changes AI results are rarely investigated in the literature; for example, AUSTRAC lacks feedback norms or standards. |
| Regulatory compliance | Some technical articles describe the different tools that exist and are included in some research, but not how practitioners use or interpret them (Table XIII). |
| Human decision after the alert | Academic works usually lack a link to the regulatory context; a clear example can be seen in Table XII on AUSTRAC's obligations. |
| Trust calibration | In the previous models, a limitation to the generation of alerts is observed, this change, this framework includes a complete human escalation process (Fig. 3) |

Table XVI functions as a conceptual tracing map, linking the "why" of the framework to the gaps in the literature, and the "what" of the components, thereby providing a foundation for future empirical testing. Table XVI shows how each component of the proposed conceptual framework directly addresses unanswered questions in the existing AML and AI literature, linking features such as the trust modulation cycle, adaptive compliance thresholds, and feedback mechanisms to specific scholarly and operational gaps.

Table XVII illustrates how each component of the framework addresses specific knowledge gaps in the literature and regulation, highlighting the originality of the model and its contribution to both theory and practice in the field of regulatory compliance within the area of anti-money laundering.

TABLE XVII. ALIGNMENT OF REGULATORY EXPECTATIONS WITH AML SYSTEM CAPABILITIES

| AUSTRAC Principle | System Capability | Compliance Gap | Example |
|---|---|---|---|
| Explainability | Black-box alerts limit transparency | Trust and auditability issues | Westpac |
| Timely SMRs | Alerts are not always escalated in time | Regulatory risk in SMR timing | CBA |
| Ongoing Oversight | Feedback loops are not always implemented | Lack of analyst feedback loop | Elliptic |
| Accountability | Responsibility unclear between the system and analyst | Ambiguity in fault ownership | Tabcorp |

As shown in Table XVII, each component of the proposed framework addresses the gaps specifically identified through the literature review and analysis of regulatory practices. When these gaps are combined with important elements of the framework, such as adaptive trust calibration, feedback loops, and explainability, this table demonstrates how the framework not only aligns with current regulatory expectations in AML prevention but also offers new contributions to the system.

### *D. Limitations and Future Work*

While the framework provides a robust foundation for AML-AI system design, it remains conceptual and context-specific. Its development is grounded in 61 literature sources and Australian regulatory guidelines, but its generalizability across jurisdictions still needs validation. Moreover, the framework has not yet been tested with primary empirical data, which currently limits real-world applicability.

Future work should focus on operationalizing the framework's constructs into measurable indicators to support empirical evaluation. These indicators, spanning trust calibration, explainability, decision fatigue, and oversight, are designed to reflect actual AML workflows and can inform both qualitative studies, for example, practitioner interviews, and quantitative assessments using operational data.

While prior studies have identified issues such as trust erosion or analyst fatigue (Alahmadi, 2022), few offer mechanisms for real-time trust recalibration or system adaptivity. This framework uniquely incorporates feedback loops as functional components that not only address cognitive strain but also support oversight continuity. In effect, it fills a critical gap in the intersection between compliance design and human-AI interaction.

TABLE XVIII. MAPPING OF FRAMEWORK CONSTRUCTS TO POTENTIAL MEASURABLE INDICATORS

| Framework Construct | Proposed Measurable Indicator | Example Data Source | Potential Use Case |
|---|---|---|---|
| Trust Calibration | Percentage of AI-generated alerts overridden or escalated by analysts | Internal AML system logs | Assessing how analyst confidence in AI decisions evolves over time |
| Explainability | Average time analysts take to resolve an AI-generated alert | Case management system | Evaluating impact of explainability tools on decision speed |
| Decision Fatigue | Number of alerts processed per analyst per shift and error rates | Analyst workload reports | Identifying workload thresholds that degrade decision quality |
| Feedback Loop Integration | Frequency and type of analyst feedback incorporated into model retraining | Model governance records | Measuring the effectiveness of human-in-the-loop system updates |
| Regulatory Alignment | Proportion of escalations meeting AUSTRAC reporting criteria | Compliance audit reports | Tracking alignment of AI decisions with statutory obligations |

Table XVIII explains that once the indicators are defined, the framework becomes testable in operational AML environments. It is suggested that future research could employ a mixed-methods design, combining practitioner surveys on trust and fatigue with quantitative analysis of system logs to help validate the relationships between the constructs and help refine the model for broad implementation.

Table XIX presents a draft of a Likert-scale survey developed to measure the key theoretical constructs explored in this study, such as trust in AI, decision fatigue, explainability, and perceptions of supervision. Each item of this scale survey is designed to capture practitioner attitudes using a 5-point scale (1 = Strongly Disagree to 5 = Strongly Agree). This survey aims to support future quantitative extensions of the current research framework.

Future research should address these limitations by:

- Deploying the Likert-scale survey instruments (Table XIX) to gather practitioner data across diverse financial institutions.
- Conducting scenario-based testing and interviews to validate trust calibration, escalation confidence, and decision fatigue in practice.
- Testing the model in other regulated industries, such as healthcare, cybersecurity, or insurance, to examine cross-domain relevance.

- Integrating findings from quantitative log data and qualitative feedback into a mixed-methods study design for broader validation.

The measurement mapping (Table XX) confirms that each construct in the model can be tested using well-established methods and tools. This ensures that the framework is not only theoretically sound but also empirically actionable, a rare feature among current AI-AML models.

As shown in Table XX, the components of the framework are directly linked to a measurable construct, supported by established variables and practical methods that reflect practices in AML and XAI research. This mapping demonstrates that the framework is not only conceptually sound but also empirically testable by asking clear survey questions and scenario-based tasks. When constructs such as confidence calibration, analyst fatigue, and explainability are aligned with validation methods by [18], [31], [32], among others, the study reinforces its commitment to evidence-based design and provides a solid foundation for future quantitative validations. This clarity ensures the framework's relevance to both academic and practitioner settings.

When there is a clear definition of what is to be measured and how, the study provides a roadmap for future researchers. While this study was developed for AML transaction monitoring, the framework's core principles are broadly applicable across sectors where AI and human oversight intersect in high-stakes decision-making environments.

TABLE XIX. LIKERT-SCALE SURVEY FOR AI-AML CONSTRUCTS

| Construct | Questionnaire Item (5-point Likert Scale) |
|---|---|
| Trust in AI | I trust the AI alerting system to correctly identify suspicious transactions. |
| Trust in AI | I believe the AI system supports my decision-making process. |
| Decision Fatigue | I feel mentally exhausted after reviewing numerous AI-generated alerts. |
| Decision Fatigue | I rush through alerts when I know most are false positives. |
| Explainability | I understand why the AI system flags certain transactions. |
| Explainability | The AI system provides clear and understandable explanations. |
| Oversight & Responsibility | I feel personally responsible for verifying the accuracy of AI alerts. |
| Oversight & Responsibility | Responsibility for AI alert errors lies with the analyst more than the system. |

TABLE XX. CONSTRUCTS AND MEASUREMENT FOR EMPIRICAL VALIDATION

| Framework | Construct | Example Measurement Variable | Method |
|---|---|---|---|
| Feedback Loops to AI Systems | Feedback quality and frequency | Analyst perceptions of how feedback is integrated into model updates. | Likert-scale items or Frequency scale<br>Example: How often do you provide feedback? |
| Analyst Fatigue | Decision fatigue and Cognitive load | Perceptions of the amount of workload due to the constant repetition of false positives. | Cognitive Load Scale<br>Example: Rate the mental effort required for daily alert review |
| Explainability | Explainability Satisfaction | Greater clarity and interpretation of alerts generated by the AI. | Likert-scale items<br>Example: I understand how the AI reached its decision for each alert |
| Regulatory Compliance | Compliance | Measurement of analysts' confidence in the decisions they make and whether they align with internal or AUSTRAC policies. | Likert-scale agreement items<br>Example: I believe my decisions would be defensible in a compliance audit. |
| Human Decision After the Alert | Escalation confidence | Measurement of analyst confidence in deciding whether to escalate or dismiss the alert. | Scenario-based questions + Likert items<br>Example: Given this scenario, how confident are you in escalating this alert? |
| Trust Calibration | Trust in AI Systems | Perceptions of system reliability and analyst confidence levels | Likert-scale items<br>Example: I trust the AI system's output when explanations are clear and consistent. |

## XIII. CONCLUSION

This study examined how trust, decision fatigue, and explainability interact within AI-driven anti-money laundering (AML) transaction monitoring systems, addressing a critical gap in the literature that has traditionally examined these constructs in isolation. Through a structured synthesis of academic research, regulatory guidance, and industry practice, the study highlighted how high false positive rates, sustained alert pressure, and opaque system behavior undermine both analyst performance and compliance defensibility.

To address these challenges, the study introduced the Dynamic Trust Modulation framework, which conceptualizes trust as a feedback-driven and context-dependent construct shaped by system performance, analyst workload, explainability mechanisms, and regulatory constraints. By framing trust calibration, decision fatigue, and explainability as interdependent elements of AML workflows, the framework provides a socio-technical lens for understanding human–AI collaboration in high-stakes compliance environments.

The contributions of this work are both theoretical and design-oriented, with implications for practice. The framework extends human-AI collaboration research into the AML domain and offers design-relevant insights for developing AI systems that support accountability, decision defensibility, and sustainable human oversight. While this study is conceptual in

nature, it establishes a foundation for future empirical validation and mixed-methods research aimed at operationalizing trust dynamics and feedback mechanisms in real-world AML settings.

## ACKNOWLEDGMENT

This project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, under grant no. (GPIP: 1807-830-2024). The authors, therefore, acknowledge with thanks DSR for technical and financial support.

## REFERENCES

[1] AUSTRAC, AUSTRAC Money Laundering in Australia National Risk Assessment, 2024. [Online]. Available: https://www.pmc.gov.au/government/its-honour

[2] A. Bello and K. Olufemi, "Artificial intelligence in fraud prevention: Exploring techniques, applications, challenges and opportunities," Computer Science & IT Research Journal, vol. 5, no. 6, pp. 1505–1520, Jun. 2024, doi: 10.51594/csitrj.v5i6.1252.

[3] D Goldbarsht and E. Sheedy, "Money laundering and the risk in the risk-based approach: The Australian context," 2024. [Online]. Available: https://www.cfatf-gafic.org

[4] D Goldbarsht, "Leveraging AI to mitigate money laundering risks in the banking system," in Money, Power, and AI. Cambridge University Press, 2023, pp. 51–69, doi: 10.1017/9781009334297.006.

[5] L Leontjeva and Z. Oubari, "Maximizing anti-money laundering compliance through AI: Assessing the obligations and responsibilities of financial institutions under the proposed EU AI Act," May 2024. [Online]. Available: https://www.diva-portal.org

[6] A J. Ajayi et al., "The impact of artificial intelligence on cyber security in digital currency transactions," Archives of Current Research International, vol. 25, no. 2, pp. 329–351, Feb. 2025, doi: 10.9734/acri/2025/v25i21090.

[7] A Alqudah et al., "Mitigating financial crimes: How anti-money laundering mechanisms shape bank outcomes," 2025. [Online]. Available: https://papers.ssrn.com

[8] N. Sawangnetr, "The implementation and effects of international AML/CTF standards in Thailand: A case study of banking institutions," May 2025. [Online]. Available: https://sussex.figshare.com

[9] M. E. Lokanan, "Unveiling sentiments of the Cullen Commission: Exploring AML compliance and regulation through deep learning," Journal of Economic Criminology, vol. 7, p. 100138, Mar. 2025, doi: 10.1016/j.jeconc.2025.100138.

[10] J. S. Dote-Pardo and P. Severino-González, "Money laundering in emerging countries: Patterns, trends, and knowledge gaps," Journal of Money Laundering Control, 2025, doi: 10.1108/JMLC-01-2025-0002.

[11] S. K. Vududala, "Enhancing anti-money laundering compliance using NICE Actimize: A data-driven approach," International Journal of Computer Techniques, vol. 11, no. 2, 2024.

[12] H. Shah and H. Sadiya, "Predictive analytics and AI integration: Revolutionizing AML and fraud detection," 2024, doi: 10.13140/RG.2.2.10702.27205.

[13] V Boateng, "Harnessing artificial intelligence for combating money laundering and fraud," Finance & Accounting Research Journal, vol. 7, no. 1, pp. 37–49, Feb. 2025, doi: 10.51594/farj.v7i1.1814.

[14] M Celestin and J. Asamoah, A Study on Fintech–RegTech Integration for Combating Money Laundering and Terrorist Financing, 2020.

[15] P Weber, K. V. Carl, and O. Hinz, "Applications of explainable artificial intelligence in finance: A systematic review," Management Review Quarterly, vol. 74, no. 2, pp. 867–907, Jun. 2024, doi: 10.1007/s11301-023-00320-0.

[16] P Lakkarasu, "Advancing explainable AI for AI-driven security and compliance in financial transactions," 2024.

[17] A Bertrand, "Misplaced trust in AI: The explanation paradox and human-centric paths," Ph.D. dissertation, 2024.

[18] H Wang, "Trust in human–AI interaction: How AI's identity affects judicial decisions," master's thesis, 2024.

[19] C Agorbia-Atta and I. Atalor, "Enhancing AML capabilities: Strategic use of AI and cloud technologies," World Journal of Advanced Research and Reviews, vol. 23, no. 2, pp. 2035–2047, Aug. 2024, doi: 10.30574/wjarr.2024.23.2.2508.

[20] B A. Alahmadi et al., "99% false positives: A qualitative study of SOC analysts' perspectives," USENIX Security Symposium, 2022.

[21] A N. Eddin et al., "Anti-money laundering alert optimization using machine learning with graphs," arXiv:2112.07508, 2021.

[22] T C. King et al., "Artificial intelligence crime: An interdisciplinary analysis," Science and Engineering Ethics, vol. 26, no. 1, pp. 89–120, 2020, doi: 10.1007/s11948-018-00081-0.

[23] D A. Zetzsche, "Artificial intelligence in finance: Putting the human in the loop," 2020.

[24] K. Boudt et al., "A news monitoring system for AML supervision," Risk Sciences, p. 100018, Apr. 2025, doi: 10.1016/j.risk.2025.100018.

[25] A Ghimire, "AI-powered anomaly detection for AML compliance in U.S. banking," Global Trends in Science and Technology, vol. 1, no. 1, pp. 95–120, Feb. 2025.

[26] W Maxwell, A. Bertrand, and X. Vamparys, "Are AI-based AML systems compatible with European fundamental rights?" Vienna, 2020.

[27] M Alkhalili et al., "Machine learning for watch-list filtering in AML," IEEE Access, vol. 9, pp. 18481–18496, 2021, doi: 10.1109/ACCESS.2021.3052313.

[28] V Mounika et al., "Detecting AML using time–frequency and transaction abstraction," 2024.

[29] S Aidoo, "Evaluating the effectiveness of AML regulations: A critical review," May 2025.

[30] U G. Ketenci et al., "Time–frequency-based suspicious activity detection for AML," IEEE Access, vol. 9, pp. 59957–59967, 2021, doi: 10.1109/ACCESS.2021.3072114.

[31] Deloitte, Advanced Analytics and Innovation in Financial Crime Compliance, 2020.

[32] B Oztas et al., "Transaction monitoring in AML: Industry perspectives," Future Generation Computer Systems, vol. 159, pp. 161–171, Oct. 2024, doi: 10.1016/j.future.2024.05.027.

[33] K Powelson and A. Capstone, "The impact of artificial intelligence on AML programs," 2022.

[34] Momenta Group, "What will AML Tranche 2 mean for your business?" 2022.

[35] Australian Government, National Framework for the Assurance of Artificial Intelligence in Government, 2024.

[36] B Scott and M. Webster, "AML/CTF Tranche 2 reform in Australia," International Journal of Contemporary Intelligence Issues, 2024.