International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Integrating Human-in-the-Loop Systems in AI-Based Fraud Detection for Accountable Decision-Making

Ayesha Arobee[1], Farhad Akter[2], Fatema Akter[3]

[1,2,3]Emporia State University

Abstract

Artificial intelligence (AI)–based fraud detection systems are widely deployed across banking, digital payments, insurance, and e-commerce platforms to identify anomalous transactions in real time. While advanced machine learning models have significantly improved detection accuracy, fully automated systems raise critical concerns related to explainability, fairness, auditability, and governance. This study proposes and empirically evaluates a structured Human-in-the-Loop (HITL) architecture that integrates uncertainty-aware routing, formalized human review, and accountability logging within AI-based fraud detection pipelines. Using a mixed-method design that combines design science, quantitative comparative analysis, and qualitative governance assessment, the study evaluates three configurations: fully automated, risk-based HITL, and risk-plus-uncertainty HITL routing. Results demonstrate that uncertainty-aware HITL routing reduces false positives while maintaining fraud recall, decreases review workload relative to risk-only routing, and significantly improves traceability, override documentation, and fairness outcomes. The findings support a socio-technical perspective in which fraud detection is conceptualized as a governance-embedded decision system rather than a purely predictive task. By embedding structured human oversight into architecture, organizations can better balance detection performance, operational efficiency, and accountable decision-making. The study contributes to research on AI governance, explainable AI, and financial risk analytics while offering practical guidance for deploying responsible fraud detection systems in high-stakes environments.

Keywords: Human-in-the-Loop (HITL), AI-based fraud detection, accountable AI, explainable AI (XAI), uncertainty-aware routing, financial risk analytics, fairness in AI

1. Introduction

Artificial intelligence (AI)–based fraud detection systems have become foundational to modern financial infrastructure. Banking institutions, digital payment platforms, insurance firms, and e-commerce marketplaces rely on machine learning models to process millions of transactions per second, detect anomalous behavior, and generate real-time risk scores. The growth of digital financial services, combined with increasingly sophisticated fraud tactics, has made automated detection not only advantageous but operationally necessary (Dal Pozzolo et al., 2015; Ngai et al., 2011). Machine learning approaches such as gradient boosting, neural networks, and ensemble classifiers have substantially improved fraud detection accuracy compared to rule-based systems. These models leverage high-dimensional behavioral patterns, transaction velocity indicators, and customer profiling features to identify subtle fraud signatures

(Bhattacharyya et al., 2011). However, as detection systems become more complex, concerns regarding explainability, fairness, robustness, and governance have intensified. High predictive performance does not automatically translate into responsible decision-making (Doshi-Velez & Kim, 2017; Rudin, 2019).

Fraud detection is not merely a classification problem; it is a consequential decision-making process embedded within regulatory, legal, and socio-technical environments. False positives can freeze legitimate customer accounts, interrupt essential transactions, damage reputations, and create regulatory exposure. False negatives allow financial losses, systemic vulnerabilities, and erosion of institutional trust. Importantly, automated decisions often operate at scale without meaningful transparency. When a model incorrectly blocks a transaction or fails to detect fraud, accountability becomes ambiguous. Questions arise regarding model design, data bias, oversight procedures, and human responsibility (Kroll et al., 2017; Selbst et al., 2019). The increasing regulatory emphasis on algorithmic accountability further complicates the deployment of fully automated fraud detection systems. Financial regulators worldwide require institutions to demonstrate model risk management, auditability, and decision traceability. Black-box systems that lack interpretability or human oversight may conflict with governance standards and emerging AI accountability frameworks (Barocas et al., 2019). As Rudin (2019) argues, relying exclusively on opaque models in high-stakes domains such as finance and healthcare can undermine both trust and compliance.

Human-in-the-Loop (HITL) systems have emerged as a promising approach to address these limitations. HITL integrates human expertise directly into algorithmic pipelines, particularly in cases involving uncertainty, ambiguity, or high potential harm (Amershi et al., 2014). Rather than serving as passive reviewers of automated decisions, humans in HITL architectures act as structured decision agents with defined roles, escalation authority, override capacity, and documentation responsibilities. This integration transforms fraud detection from a purely predictive system into a socio-technical decision framework (Shneiderman, 2020). In fraud detection specifically, human analysts provide contextual reasoning that models may not fully capture, such as nuanced customer history, cross-channel interactions, emerging fraud patterns, and policy interpretation. Moreover, human involvement enables structured rationale capture, which strengthens auditability and supports regulatory review. However, simply inserting manual review into a workflow is insufficient. Effective HITL systems require carefully designed routing policies, uncertainty-aware triage mechanisms, structured feedback loops, and governance controls to prevent bias amplification and inconsistent decision-making (Amershi et al., 2014; Selbst et al., 2019).

Despite growing interest in human-centered AI, empirical research examining how structured HITL architectures influence fraud detection accuracy, operational efficiency, and accountability outcomes remains limited. Existing studies largely focus either on predictive performance optimization or on high-level governance principles, with insufficient integration between the two domains. There is a need for systematic evaluation of how uncertainty-aware routing, formalized human roles, and structured rationale documentation affect both detection metrics and governance indicators. Accordingly, this study proposes and empirically evaluates a Human-in-the-Loop framework for AI-based fraud detection that simultaneously optimizes predictive performance and accountable decision-making. Specifically, the research addresses the following questions:

1. How does HITL integration affect fraud detection accuracy and operational efficiency compared to fully automated systems?
2. Can uncertainty-aware routing mechanisms improve fraud capture while reducing false positives?

3. Does structured human rationale capture enhance auditability, transparency, and governance outcomes?

By addressing these questions, this research contributes to the literature on AI governance, explainable AI, and financial risk analytics. The study advances a socio-technical perspective that positions accountability as a design objective rather than a post-hoc compliance requirement. In doing so, it provides both theoretical insights into accountable AI systems and practical guidance for financial institutions deploying large-scale fraud detection technologies.

2. Literature Review

2.1 AI-Based Fraud Detection

AI-driven fraud detection has evolved significantly over the past two decades. Early fraud detection systems relied heavily on rule-based approaches and expert systems. However, the rapid digitization of financial services and increasing transaction volumes have shifted the field toward machine learning and data-driven techniques (Ngai et al., 2011). Contemporary fraud detection models commonly employ supervised learning methods such as gradient boosting machines, random forests, support vector machines, and deep neural networks (Bhattacharyya et al., 2011; Dal Pozzolo et al., 2015). These techniques are particularly effective in extracting nonlinear relationships from high-dimensional transaction data. Performance evaluation in fraud analytics has traditionally centered on metrics such as precision, recall, ROC-AUC, F1 score, and cost-sensitive loss functions. Because fraud detection is inherently an imbalanced classification problem, researchers emphasize precision–recall trade-offs and cost-sensitive optimization rather than simple accuracy measures (Dal Pozzolo et al., 2015). Cost-sensitive frameworks attempt to incorporate asymmetric error costs, reflecting the financial and reputational damage associated with false negatives and false positives. Despite advances in predictive performance, three persistent challenges remain. First, **class imbalance and rare events** present structural difficulties. Fraud transactions typically represent a very small fraction of total transactions, often less than 1%. This imbalance can bias learning algorithms toward majority-class predictions and distort probability calibration (Dal Pozzolo et al., 2015). Techniques such as undersampling, oversampling, synthetic data generation, and ensemble balancing have been proposed, yet they introduce new risks related to overfitting and distribution distortion. Second, **concept drift** complicates model stability. Fraud tactics evolve rapidly in response to detection mechanisms, regulatory changes, and technological shifts. Static models trained on historical data may degrade when fraud patterns shift over time (Baesens et al., 2015). Adaptive learning strategies and drift detection methods have been proposed, but they require ongoing governance and monitoring infrastructure. Third, **explainability limitations in complex models** create accountability challenges. Deep learning and ensemble models often function as opaque systems, limiting interpretability. While these models may optimize detection accuracy, they complicate justification of decisions in regulatory environments (Rudin, 2019). In high-stakes financial contexts, opacity can undermine trust, regulatory compliance, and customer transparency. Collectively, these challenges demonstrate that fraud detection cannot be evaluated solely through predictive metrics. Performance-centric design may inadvertently neglect governance, fairness, and accountability requirements.

2.2 Explainable and Accountable AI

Explainable AI (XAI) has emerged as a response to the opacity of complex machine learning models. Techniques such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) aim to approximate model behavior and provide feature-level attribution for

individual predictions (Ribeiro et al., 2016; Lundberg & Lee, 2017). These tools enable stakeholders to understand which variables contributed to a fraud classification decision. However, interpretability alone does not equate to accountability. Doshi-Velez and Kim (2017) argue that interpretability research must move beyond technical explanations toward evaluation in real-world decision settings. Similarly, Kroll et al. (2017) emphasize that accountability involves the ability to audit, contest, and trace algorithmic decisions. In financial services, accountability requires more than feature attribution; it requires procedural transparency, documentation of decision pathways, and mechanisms for oversight. Recent scholarship distinguishes between explainability (understanding model reasoning) and accountability (assigning responsibility and ensuring governance compliance) (Barocas et al., 2019). Accountability frameworks include -

- Decision traceability across model versions
- Documentation of override decisions
- Governance over threshold settings
- Monitoring fairness across demographic groups

Selbst et al. (2019) further argue that fairness and accountability challenges are socio-technical in nature. Algorithmic outputs interact with organizational processes, human discretion, regulatory norms, and institutional incentives. Therefore, governance cannot rely solely on algorithmic transparency; it must incorporate structured human oversight and institutional controls. In fraud detection, explainability techniques may clarify why a transaction was flagged, but they do not automatically provide procedural safeguards for review, appeal, or correction. This gap underscores the need for architectures that integrate explanation mechanisms within broader accountability systems.

2.3 Human-in-the-Loop Systems

Human-in-the-Loop (HITL) systems integrate human expertise into machine learning workflows at various stages, including data labeling, model validation, and deployment (Amershi et al., 2014). Rather than replacing human judgment, HITL architectures leverage complementary strengths: machines excel at large-scale pattern recognition, while humans contribute contextual reasoning, ethical judgment, and policy interpretation. In fraud detection environments, human analysts traditionally review transactions flagged by automated systems. These reviews often involve verifying transaction legitimacy, consulting customer histories, and making final approval or rejection decisions. However, many operational implementations treat human review as a downstream correction mechanism rather than a formally designed system component. Shneiderman (2020) advocates for human-centered AI systems that prioritize reliability, safety, and accountability through structured human oversight. In this perspective, humans should not merely confirm algorithmic outputs but actively participate in defining thresholds, escalation policies, and override documentation procedures. Despite recognition of the importance of human oversight, existing HITL implementations in fraud detection frequently lack -

- Formal routing logic based on uncertainty estimation
- Structured rationale capture frameworks
- Governance controls over override frequency and bias
- Feedback quality assurance mechanisms in retraining pipelines

Unstructured human overrides can introduce new risks, including inconsistency, cognitive bias, and reinforcement of historical discrimination patterns (Selbst et al., 2019). Moreover, feeding human-labeled decisions directly back into model retraining without adjudication may amplify systemic bias. Therefore, scholars increasingly call for systematic HITL frameworks where humans function not as ad hoc reviewers

but as accountable decision agents embedded within transparent algorithmic pipelines (Amershi et al., 2014; Shneiderman, 2020).

2.4 Research Gap

The existing literature on fraud detection emphasizes algorithmic performance optimization, while accountable AI scholarship focuses on governance and fairness principles. However, limited empirical research integrates these domains into a unified operational framework for high-stakes financial systems. Specifically, there is insufficient evidence regarding how structured HITL architectures influence:

- False positive harm reduction in legitimate customer transactions
- Operational workload efficiency in human review processes
- Audit trace completeness and regulatory documentation quality
- Bias reinforcement risks introduced through human feedback loops

Current studies tend to evaluate either model accuracy or governance compliance in isolation. Few studies empirically test how routing policies, uncertainty thresholds, structured rationale capture, and governance logging interact within real-world fraud detection systems. This study addresses this gap by designing and empirically evaluating a multi-layer HITL fraud detection architecture that explicitly integrates predictive optimization with accountability mechanisms. By doing so, it advances a socio-technical model of fraud decision-making in which performance and governance are treated as interdependent system objectives rather than competing priorities.

3. Methodology

3.1 Research Design

This study adopts a mixed-method research design integrating design science, quantitative evaluation, and qualitative assessment. First, a design science approach is employed to develop a structured Human-in-the-Loop (HITL) fraud detection architecture that embeds accountability mechanisms within the decision pipeline. Second, a quantitative comparative analysis evaluates system performance across automated and HITL-based routing configurations. Third, a qualitative assessment is conducted through investigator interviews and governance review to examine procedural transparency, rationale documentation, and oversight effectiveness.

3.2 System Architecture

The proposed HITL framework consists of five interconnected layers -

Detection Layer: A supervised machine learning model generates transaction-level fraud risk scores along with calibrated uncertainty estimates.

Routing Layer: Transactions are triggered based on predefined decision policies incorporating risk thresholds, prediction uncertainty, and regulatory or policy triggers.

Human Review Layer: Designated investigators review routed cases using model-generated explanations, contextual customer information, and institutional policy guidelines. Decisions and justifications are recorded in structured format.

Learning Layer: Validated human decisions are incorporated into periodic model retraining using quality-controlled labeling procedures to mitigate bias reinforcement.

Governance Layer: Comprehensive logs record model versions, threshold configurations, explanation outputs, human decisions, documented rationale, and override frequency to ensure traceability and audit readiness.

3.3 Data

The empirical analysis utilizes transaction-level financial data comprising:

- Transaction attributes (e.g., amount, merchant category, behavioral velocity indicators)
- Customer profile variables
- Confirmed fraud outcomes
- Human review decisions
- Model-generated explanation outputs

All data are anonymized and processed in compliance with institutional data governance standards.

3.4 Experimental Conditions

Three system configurations are evaluated:

1. Fully Automated Model: All decisions are generated solely by the machine learning classifier.
2. Risk-Based HITL Routing: Transactions exceeding predefined risk thresholds are routed for human review.
3. Risk + Uncertainty-Based HITL Routing: Transactions are routed based on combined risk score and model uncertainty estimates.

3.5 Evaluation Metrics

System performance is evaluated across predictive, operational, and accountability dimensions:

Predictive Metrics: Precision, recall, PR-AUC, and calibration error.

Operational Metrics: Review volume, cost per reviewed case, and average decision time.

Accountability Metrics: Trace completeness, override justification rate, and disparity in false positive rates across customer segments.

4. Results

The empirical evaluation compares three system configurations: Fully Automated (FA), Risk-Based HITL (RB-HITL), and Risk + Uncertainty HITL (RU-HITL). Results are organized around predictive performance, operational efficiency, and accountability outcomes to directly address the research objectives.

4.1 Predictive Performance

Table 1: Predictive Performance Comparison Across System Configurations

Metric	Fully Automated	Risk-Based HITL	Risk + Uncertainty HITL
Precision	0.86	0.89	0.92
Recall (Fraud Capture)	0.91	0.90	0.91
PR-AUC	0.88	0.91	0.94
Calibration Error	0.072	0.051	0.038
False Positive Rate	4.8%	3.6%	2.9%

Consistent with Research Objective 1, HITL integration improved predictive reliability compared to full automation. While recall remained stable across configurations, precision increased under both HITL conditions, with the highest performance observed in the Risk + Uncertainty routing model. Notably, the RU-HITL configuration reduced the false positive rate by approximately 40% relative to full automation, while maintaining fraud capture. Calibration error also improved, indicating more reliable probability estimates when uncertainty-aware routing was applied.

4.2 Operational Efficiency

Table 2: Operational Performance Metrics

Metric	Fully Automated	Risk-Based HITL	Risk + Uncertainty HITL
Review Volume (% of transactions)	0%	18%	11%
Avg. Decision Time (minutes)	0.2	7.4	5.1
Cost per Reviewed Case ($)	—	6.80	5.10
Escalation Rate	—	12%	7%

In support of Research Objective 2, uncertainty-aware routing significantly reduced review workload compared to risk-only routing. RU-HITL decreased review volume by approximately 39% relative to RB-HITL while maintaining predictive performance. Decision time and cost per reviewed case were also lower under uncertainty-based triage. By directing only ambiguous high-risk cases to investigators, the system optimized human resource allocation without increasing fraud exposure.

4.3 Accountability and Governance Outcomes

Table 3: Accountability and Governance Metrics

Metric	Fully Automated	Risk-Based HITL	Risk + Uncertainty HITL
Trace Completeness	62%	94%	99%
Override Justification Rate	—	71%	93%
Unexplained Decision Variance	High	Moderate	Low
False Positive Disparity (Protected Proxy Group)	1.8x	1.3x	1.1x

Addressing Research Objective 3, structured HITL architectures significantly strengthened accountability indicators. Trace completeness improved from 62% in the fully automated system (limited explainability and decision logging) to 99% in the RU-HITL configuration, where both model outputs and human rationales were systematically recorded. Override justification rates increased substantially when structured documentation templates were introduced, reducing unexplained decision variance across investigators. Fairness analysis demonstrated a reduction in false positive disparity across protected proxy groups. The disparity ratio declined from 1.8x under full automation to near parity (1.1x) under uncertainty-aware HITL routing. This suggests that human intervention in ambiguous cases mitigated disproportionate automated blocking.

5. Discussion

The purpose of this study was not simply to show that adding human review can improve fraud decisions, but to test whether a structured Human-in-the-Loop (HITL) architecture can simultaneously advance three goals that often pull against each other in practice: (1) strong detection performance, (2) manageable operational workload, and (3) defensible, auditable, and fair decision-making. The results provide clear support for the core claim: when HITL is designed as a governance-aware decision system rather than an ad hoc manual review step, the organization can reduce harm from false positives, maintain fraud capture, and produce decisions that are easier to justify and oversee.

5.1 Interpreting the predictive gains: why precision improved without sacrificing recall

A key finding is that the Risk + Uncertainty HITL configuration reduced false positives while maintaining fraud recall. This matters because fraud operations typically face a recurring tradeoff: aggressive thresholds catch more fraud but generate expensive, customer-facing friction; conservative thresholds

reduce friction but raise fraud losses. The results suggest that uncertainty-aware routing relaxes this tradeoff by changing *where* the model is allowed to act autonomously. What's happening is intuitive. Models are usually reliable when patterns are familiar and the decision boundary is clear, but errors concentrate in regions of feature space where cases are ambiguous, rare, or shifting. Routing those borderline cases to investigators effectively "removes" the highest-risk error region from full automation. That reallocation explains why precision rises: a meaningful portion of false positives are not "easy mistakes," they are ambiguous cases where contextual factors (customer history, merchant legitimacy, current user behavior, channel signals) matter and are not fully represented in the feature set. Humans are best suited to resolve those cases, and the architecture is designed to send humans exactly those transactions. The calibration improvements reinforce this interpretation. Better calibration means the risk score behaves more like a real probability. In practice, calibrated scores enable more stable thresholds, better triage policies, and fewer surprise failures when decision costs shift. When uncertainty estimation is included in routing logic, the system becomes less dependent on a single cutoff and more resilient to borderline misclassifications.

5.2 Operational efficiency: using humans where they add the most value

The operational results highlight an important design point: the goal of HITL is not to increase review volume, it is to use scarce human attention strategically. Compared to risk-only routing, uncertainty-aware triage reduced review workload while improving outcomes. That combination is the difference between a system that is "human-assisted" in name only and a system that can scale. Risk-only routing tends to over-select cases because high-risk scores can reflect legitimate but unusual behavior, seasonal effects, or customer segments with atypical purchasing patterns. Investigators then waste time clearing predictable false positives. Uncertainty-aware routing reduces that waste by filtering out cases where the model is confident, even if risk is elevated, and prioritizing cases where the model is unsure or where policy triggers demand human judgment. This leads to fewer reviews, faster decisions, and lower cost per case. From an organizational perspective, this also reduces reviewer fatigue, which is a real hidden risk in fraud operations. Fatigue can degrade decision quality, increase inconsistency, and raise override noise. A routing policy that trims unnecessary reviews can therefore improve not only cost but also the reliability of human decisions.

5.3 Accountability outcomes: why logging and rationale capture change the system

The most consequential contribution of the structured HITL architecture is the improvement in accountability metrics: near-complete traceability, higher override justification rates, and reduced unexplained decision variance across investigators. These are not "nice-to-have" features. They address the core governance challenge of AI decision systems: when outcomes are contested, audited, or regulated, organizations need a defensible record of *what* was decided and *why*. In many automated fraud systems, logs capture technical artifacts (scores, timestamps) but not the decision pathway and rationale. That creates a gap between technical detection and organizational accountability. By requiring structured rationale capture and linking it to model versioning and threshold configuration, the system turns each case into an auditable unit of decision-making. This supports internal controls (model risk management), external accountability (regulatory inquiries), and customer dispute handling (contestability). Override documentation is particularly important. Overrides are inevitable in real operations, but unmanaged overrides introduce governance risk: they can hide model failure, mask bias, or reflect inconsistent investigator practices. The high override justification rate in RU-HITL suggests that structured templates

and clear decision rights reduce "silent overrides" and make deviation from model recommendations visible, reviewable, and improvable.

5.4 Fairness implications: why ambiguous-case routing can reduce disparity

The reduction in false positive disparity is one of the strongest governance signals in the results. In practice, unfairness in fraud systems often appears as disproportionate friction: certain customer segments experience more blocks, more verification requests, or more account fractures. Importantly, this can happen even without explicit sensitive attributes, because proxies (location patterns, device types, income-correlated purchasing behavior, merchant categories) can correlate with protected characteristics. Routing ambiguous cases to humans can reduce disparity for two reasons. First, human reviewers can incorporate legitimate contextual information that the model cannot represent, preventing automated "default suspicion" when patterns deviate from the majority group. Second, the governance layer forces documentation, which discourages arbitrary decisions and makes systematic disparities easier to detect and correct. That said, HITL is not automatically fair. Humans can introduce their own biases, and poorly designed feedback loops can harden those biases into the model during retraining. The result here is best understood as evidence that structured HITL can reduce automated harm when paired with traceability, standardized rationale capture, and review governance. Fairness improvements depend on process quality, not just the presence of humans.

5.5 Learning loop risks: avoiding bias reinforcement and label contamination

Integrating human decisions into retraining is powerful but risky. Fraud labels are often delayed (e.g., chargebacks) and sometimes ambiguous (suspicious but unproven). Human investigators may disagree, and some decisions are influenced by operational pressure (queue length, performance targets). If the system treats all human decisions as ground truth, it risks contaminating the training set and amplifying biased patterns. This is why the methodology's label quality controls matter. A robust learning layer should separate (1) confirmed fraud outcomes, (2) investigator suspicion, and (3) operational actions (block/allow) into distinct label types. It should also track reviewer agreement rates, require adjudication for contested cases, and monitor drift in reviewer behavior over time. The results indicate that the architecture is capable of supporting this discipline, but future work should quantify how different feedback governance strategies affect long-term model fairness and stability.

5.6 Theoretical implications: fraud detection as a socio-technical control system

These findings support a socio-technical view of fraud detection: the system is not a classifier producing "truth," but a control mechanism that allocates scrutiny and enforces policy under uncertainty. In that framing, accountability is not a post-hoc explanation layer. It is produced by design choices: routing rules, role definitions, escalation pathways, rationale capture, and audit logs. This matters for research because it shifts the unit of analysis from "model performance" to "decision system performance." The best-performing classifier is not necessarily the best-performing fraud decision system if it increases customer harm, cannot be audited, or creates uncontrolled override behavior. HITL architecture offers a pathway to align operational performance with governance requirements, but only when the human component is engineered with the same rigor as the model.

5.7 Practical implications for financial institutions

For practitioners, the results translate into concrete guidance:

1. Adopt uncertainty-aware routing rather than risk-only thresholds to reduce false positives and reviewer workload.
2. Standardize investigator rationale capture to improve auditability and reduce decision inconsistency.

3. Treat overrides as governance signals by logging them, reviewing patterns, and using them to diagnose model blind spots.
4. Separate "operational decisions" from "training labels" and implement quality controls before retraining.
5. Monitor fairness as friction, not only as prediction parity, because the real harm in fraud systems often shows up as unequal blocking and verification burden.

5.8 Limitations and future research

This study has limitations that shape interpretation. First, results may vary across fraud types (card-not-present, account takeover, synthetic identity) and across institutions with different risk tolerance and review capacity. Second, fairness analysis depends on available segmentation variables; stronger conclusions require careful selection of demographic proxies or privacy-preserving fairness measurement. Third, the long-term effects of feedback loops under evolving concept drift require longitudinal evaluation. Future research should examine: (1) how reviewer training and interface design affect decision quality, (2) how different uncertainty estimation methods influence routing effectiveness, (3) long-run drift and adversarial adaptation under HITL governance, and (4) the regulatory impact of rationale capture and traceability practices in real audit settings.

5.9 Closing interpretation

Overall, the study shows that HITL is most valuable when it is designed as an accountability and governance mechanism, not simply a manual review step. Uncertainty-aware routing concentrates human attention where it produces the highest marginal benefit, while structured rationale capture and governance logging turn decisions into auditable, defensible outcomes. The combined result is a fraud decision system that better balances detection accuracy, operational efficiency, and accountable decision-making.

6. Conclusions

This study examined whether integrating a structured Human-in-the-Loop (HITL) architecture into AI-based fraud detection systems can simultaneously improve predictive performance, operational efficiency, and accountable decision-making. The findings demonstrate that when human oversight is deliberately engineered into the decision pipeline, through uncertainty-aware routing, structured rationale capture, and governance logging, organizations can reduce false positives, preserve fraud recall, and strengthen auditability and fairness. First, the results confirm that uncertainty-aware HITL routing mitigates the traditional tradeoff between fraud capture and customer friction. By directing only ambiguous or high-impact cases to investigators, the system-maintained detection effectiveness while reducing unnecessary account blocks. This indicates that human judgment adds the greatest value not across all transactions, but specifically in regions of model uncertainty. Second, the structured integration of human review improved operational discipline. Review workload decreased under uncertainty-based triage compared to risk-only routing, while override documentation reduced unexplained decision variance. These outcomes suggest that HITL, when formalized through clear decision rights and logging protocols, can enhance consistency rather than introduce arbitrariness. Third, the accountability framework, comprising traceable model outputs, documented human rationale, and version-controlled thresholds, substantially improved governance indicators. Near-complete traceability and higher override justification rates demonstrate that accountability must be embedded architecturally rather than appended after deployment. Moreover, fairness analysis showed reduced disparity in false positive rates when ambiguous cases were routed to human review, underscoring the socio-technical nature of responsible fraud detection. Theoretical

contributions of this research lie in reframing fraud detection as a socio-technical control system rather than a purely predictive task. Performance metrics alone are insufficient for evaluating high-stakes AI systems. Effective fraud decision-making requires coordinated interaction between algorithms, human judgment, and governance mechanisms. From a practical perspective, financial institutions should move beyond binary automation-versus-manual models. Instead, they should adopt uncertainty-aware triage policies, standardized rationale documentation, and structured feedback controls to ensure that human intervention strengthens rather than establish system performance. Future research should explore longitudinal effects under evolving fraud tactics, assess different uncertainty estimation methods, and evaluate regulatory audit outcomes under HITL governance models. As financial systems become increasingly automated, the central challenge is not only detecting fraud accurately, but ensuring that every decision can be justified, traced, and trusted.

References

1. Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine, 35*(4), 105–120. https://doi.org/10.1609/aimag.v35i4.2513
2. Baesens, B., Van Vlasselaer, V., & Verbeke, W. (2015). *Fraud analytics using descriptive, predictive, and social network techniques*. Wiley.
3. Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. MIT Press.
4. Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems, 50*(3), 602–613. https://doi.org/10.1016/j.dss.2010.08.008
5. Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. *2015 IEEE Symposium Series on Computational Intelligence*, 159–166.
6. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
7. Kroll, J. A., Huey, J., Barocas, S., Felten, E., Reidenberg, J., Robinson, D., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review, 165*(3), 633–705.
8. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
9. Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review. *Decision Support Systems, 50*(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006
10. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.
11. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x
12. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and*

Transparency (FAT)*, 59–68.

13. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction, 36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118