Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00242-4

Review Article

Allocating Human Judgment in Automated Blockchain Anti-Money Laundering: A Review of Detection, Attribution, Adjudication, and Reporting

Moshood Sorinola [1] , Kehinde Hamed [2,]

1. *Finance and Business Analytics, Georgia State University, Georgia, USA*
2. *Financial Crime and Compliance Management, Utica University, Utica, USA*

Received: June 23, 2026 | Review began: July 27, 2026 | Review ended: August 03, 2026 | Published: August 24, 2026

Abstract

Automated anti-money laundering (AML) on public blockchains is usually framed as a detection problem. Because the ledger is public and permanent, automated detection is feasible, but that record shows only that value moved, without showing who moved it or why. We review automated blockchain anti-money laundering as a system in which machine models and human analysts share each decision, following it through four stages: detection, attribution, adjudication, and reporting, and asking at each stage what automation does well, what it must leave to human judgment, and what goes wrong when that judgment is misplaced. The evidence shows a consistent asymmetry. Machine learning is strong at pattern-finding over the permanent public record, where models rank suspicion and clustering heuristics scale, but it weakens sharply as the task turns from finding a pattern to assigning meaning, identity, intent, or accountability. Drawing first on emerging AML-specific studies and then, where direct evidence remains insufficient, on human-factors research from adjacent high-stakes domains, we find an uneven evidence base. Automation is best supported for large-scale detection, alert prioritization, and parts of blockchain attribution, while the evidence becomes more limited as decisions require contextual interpretation, evidential judgment, accountability, and reporting. AML-specific studies identify explainability, flexibility, tool integration, supervisory justification, and human review as important operational requirements, but they do not yet establish how frequently analysts over-rely on, reject, or selectively follow automated recommendations. Evidence from aviation, healthcare, and public administration is therefore used to identify plausible failure mechanisms rather than to claim AML-specific effects. The result is an evidence-weighted allocation matrix that records, for each stage, the strength of the case for automation, the function retained by the analyst, the dominant failure mode, and the empirical question that remains unresolved. Support is strongest for detection, direct but context-dependent for attribution, and more provisional for adjudication and reporting.

Categories: Blockchain and Cryptocurrency, Machine Learning (ML), AI/ML-based decision support systems

Keywords: blockchain, anti-money laundering, human-ai decision systems, automation bias, meaningful human control, explainability, cryptocurrency forensics, learning to defer

Introduction And Background

A blockchain ledger is an unusual object for financial surveillance because it hides almost nothing. Nakamoto's design announces every transaction publicly and protects privacy only by allowing participants to keep their public keys anonymous, so an observer can watch value move between parties without learning who they are, much as a stock-exchange tape reports the time and size of trades but not the traders [1]. Pseudonymity is not anonymity. Europol,

reviewing its casework, concluded that cryptocurrencies are not anonymous and that the permanent public record gives investigators substantially more to work with than a comparable cash case [2]. Blockchain anti-money laundering (AML) therefore starts from a peculiar condition: the data is all there, and the difficulty lies in reading it.

The United Nations Office on Drugs and Crime estimates that 2% to 5% of global GDP, between $800 billion and $2 trillion, is laundered each year, although the clandestine nature of the activity makes any figure approximate. Laundering is conventionally described in three movements: placement, layering, and integration, which in practice may merge or repeat [3]. Cryptocurrency is a small but real part of this picture. Illicit activity involving bitcoin has been estimated at roughly $76 billion a year, with its proportional share falling as the market matured and more opaque alternatives drew activity elsewhere [4]. Europol's casework similarly characterizes criminal cryptocurrency use as genuine and expanding but still modest relative to cash [2].

Obligations built on top of this are global and concrete. Through its recommendations, the Financial Action Task Force (FATF) sets the international AML standard that member states implement through domestic legislation. Recommendation 1 requires a risk-based approach that matches controls to assessed exposure, while Recommendation 20 requires suspicious transactions to be reported to a national financial intelligence unit [5]. In the United States, the Bank Secrecy Act authorizes the Treasury to require reporting of suspicious transactions relevant to possible violations of law; covered institutions file suspicious activity reports under the applicable regulatory thresholds and suspicion criteria [6]. That report is the artifact the whole apparatus exists to produce.

At the scale of a live blockchain, no team can read the ledger by hand, and the work is necessarily automated. Automation does not remove the human from the loop so much as change where in the loop the human sits. An automated AML system is better understood as a human-AI decision system than as a classifier: a pipeline in which models rank suspicion and people decide what the rankings mean. Detection screens transactions for the risk profile of laundering; attribution attempts to connect an address to a controlling entity; adjudication is the analyst-mediated stage at which a flagged case is dismissed, escalated, investigated, or reported; and reporting turns the decision into a durable record for regulators, auditors, and courts. Each stage automates a different function and asks a different kind of judgment from the person downstream. Working through the pipeline one stage at a time, we ask how much judgment a model can carry and how much must remain with the analyst. The principal contribution is an evidence-weighted allocation matrix. For each stage, it identifies the strength of the evidence supporting automation, the function that remains with the analyst, the dominant way the human-machine arrangement can fail, and the specific empirical evidence still missing. The matrix therefore distinguishes recommendations supported by direct AML evidence from those that remain provisional or depend on findings from adjacent domains.

Search strategy and evidence selection

This narrative review was updated through July 27, 2026. Searches were conducted in Google Scholar, the ACM Digital Library, IEEE Xplore, ScienceDirect, SpringerLink, the NeurIPS proceedings archive, and relevant conference proceedings repositories. Official legal and regulatory materials were retrieved from FATF, the United Nations Office on Drugs and Crime, the US Financial Crimes Enforcement Network, US federal banking regulators, European Union legal repositories, and published court records.

The search was organized around the four stages examined in this review. Search combinations included: ("anti-money laundering" OR AML OR "financial crime" OR "transaction monitoring") AND (blockchain OR Bitcoin OR cryptocurrency) AND (detection OR attribution OR clustering OR adjudication OR reporting OR analyst OR "human judgment" OR "human-in-the-loop" OR explainability OR deferral). A second set combined AML or financial-crime terms with "automation bias," "algorithm aversion," "selective adherence," "human oversight," "meaningful human control," and "learning to defer." Because direct AML evidence remained limited, supplementary searches paired these human-factors concepts with aviation, healthcare, public administration, and finance.

Studies were eligible when they presented peer-reviewed empirical evidence, a peer-reviewed conceptual framework, or an official legal or regulatory source; addressed at least one stage of the proposed AML pipeline; and contributed evidence concerning automated performance, analyst judgment, human reliance, explainability, accountability, or

How to cite this article:
Sorinola M, Hamed K (August 24, 2026) Allocating Human Judgment in Automated Blockchain Anti-Money Laundering: A Review of Detection, Attribution, Adjudication, and Reporting. Cureus J Comput Sci 3 : es44389-026-00242-4. DOI https://doi.org/10.7759/s44389-026-00242-4

reporting. Foundational studies predating contemporary blockchain AML systems were retained when they established concepts directly used in the analysis, including automation bias, complacency, algorithm aversion, levels of automation, and meaningful human control. Duplicates, non-scholarly commentary, promotional material without independently reported methods, and sources unrelated to the allocation of functions between automated systems and human decision-makers were excluded.

Preprints were excluded from the revised evidence base when a final journal or conference publication was available. Conference and workshop papers were retained only when a formal venue and publication record could be identified. Titles and abstracts were assessed for relevance, followed by full-text review of potentially eligible sources. Evidence was prioritized in the following order: direct studies of AML professionals or AML systems; studies of financial-crime or fraud analysts performing comparable alert-review tasks; and evidence from other high-stakes domains. Adjacent-domain evidence was retained only when sufficiently direct AML evidence did not address the relevant human-factors question, and its analogical status is stated explicitly.

The evidence was synthesized narratively by pipeline stage. Statistical meta-analysis was not performed because the included studies differed substantially in populations, datasets, tasks, outcome measures, model architectures, and definitions of human involvement. Quantitative results were instead extracted and compared descriptively, including classification performance, label requirements, clustering error, traceability, analyst sample size, review capacity, and human-AI decision outcomes. AML-specific evidence concerning adjudication is limited but growing rather than absent; however, controlled studies have not yet estimated the prevalence or magnitude of automation bias, algorithm aversion, or selective adherence among practicing AML analysts.

Evidence hierarchy

The evidence was interpreted using a three-tier hierarchy. The first tier comprised direct studies of AML professionals, AML systems, financial supervisors, and banking practitioners working with financial-crime detection. These sources provide the closest evidence for the operational and human-judgment claims made in this review. The second tier comprised studies of financial-fraud detection, alert review, and human-AI deferral that involve comparable screening and escalation tasks but do not examine blockchain AML directly. The third tier comprised human-factors research from adjacent high-stakes domains, including aviation, healthcare, and public administration. Third-tier evidence was used only when the first two tiers did not address the relevant mechanism, and it was treated as evidence of a plausible risk rather than as an AML-specific effect estimate.

This hierarchy is applied throughout the review. Detection and attribution rely mainly on direct blockchain and financial-crime evidence. Adjudication combines limited direct AML evidence with financial-fraud deferral studies and carefully identified adjacent-domain evidence on automation bias, algorithm aversion, and selective adherence. Reporting draws on direct regulatory materials, financial-crime governance studies, and broader human-AI evidence concerning explanation and accountability.

Review

Automated blockchain AML is a sequence of decisions, and each stage hands the analyst a different kind of uncertainty.

Detection

Detection frames AML analytics as transaction screening, judging whether a payment, drawn from a stream dominated by legitimate activity, carries the risk profile of laundering. Alarab et al. [7] describe the Elliptic Data Set as a directed graph containing 203,769 transaction nodes, 234,355 payment-flow edges, and 166 features per node. Labels are sparse: 4,545 nodes (2%) are marked illicit, 42,019 (21%) licit, and the remainder unlabeled.

Later datasets deepened that lineage. Elmougy and Liu [8] extended the original with Elliptic++, adding more than 822,000 Bitcoin wallet addresses and shifting part of the analytical focus from transactions toward actors. Song et al. [9] used Elliptic2 to model address clusters controlled by one entity as nodes, transactions between entities as edges, and

laundering schemes as subgraphs rather than isolated suspicious payments. Altman et al. [10] addressed the scarcity of real financial data by releasing standardized synthetic AML datasets across different sizes and levels of difficulty, together with graph-neural-network baselines.

On the Elliptic benchmark, Alarab et al.'s [7] graph-convolutional model with linear layers achieved an illicit-class F1-score of 0.773, compared with 0.628 for the baseline graph convolutional network. Elmougy and Liu [8] reported 97.5% precision and 71.9% recall for their best transaction-level Random Forest detector. Alarab and Prakoonwit's [11] temporal graph convolutional network, combining a long short-term memory network with topology-adaptive convolution, reached an illicit-class F1-score of 0.806, compared with 0.720 for EvolveGCN and 0.705 for Skip-GCN. Table *1* summarizes these results.

Study (dataset)	Best-performing method	Illicit-class performance
Alarab et al. [7] (Elliptic)	GCN with linear layers	F1-score 0.773, compared with 0.628 for the baseline graph convolutional network [7]
Elmougy and Liu [8] (Elliptic++)	Random Forest	Precision 97.5%; recall 71.9% [8]
Alarab and Prakoonwit [11] (Elliptic, temporal)	Temporal GCN	F1-score 0.806, compared with 0.720 for EvolveGCN and 0.705 for Skip-GCN [11]

TABLE 1: Reported illicit-class detection performance on the Elliptic dataset family

GCN, Graph Convolutional Network

These F1-score, precision, and recall values were obtained exclusively from the Elliptic dataset family. Live exchange, wallet, and mixer environments may differ in label noise, class imbalance, transaction behavior, and rates of concept drift. The reported figures should therefore be interpreted as benchmark results rather than estimates of expected performance in production systems.

The benchmark results above assume abundant labels, a condition that live AML data rarely meets. Lorenz et al. [12] found that label scarcity limits the practical use of conventional supervised methods and that unsupervised anomaly detection performed substantially below the supervised baseline. Their active-learning approach recovered the fully supervised baseline using approximately 5% of the labels by directing analyst review toward cases about which the model was most uncertain; the contamination rate represented the fixed alert budget available to the compliance team. Alarab and Prakoonwit [11] reached a similar result using a Bayesian active learning by disagreement acquisition function, which matched the fully supervised model using only a fraction of the labels. The analyst is therefore already inside the detector as the source of labels that sustain its performance.

Federated learning offers another possible response to institutional data silos in financial-fraud detection. Participating institutions can train a shared model by exchanging model updates rather than centralizing their raw transaction records. This may broaden the information available for identifying patterns that do not appear within one institution's data alone, while reducing the need for direct sharing of sensitive records. The approach is not yet established for blockchain AML, however, and it introduces its own challenges, including differences between institutional datasets, governance of the shared model, privacy leakage from model updates, and exposure to adversarial manipulation [13].

Detection performance is also sensitive to changes over time. Studies using temporal and label-scarce versions of Elliptic report uneven performance across time steps and show that models trained on earlier transaction patterns may not transfer cleanly when the underlying distribution changes [11,12]. A change that an analyst would interpret as a shift in the operating environment reaches the model as degraded or unstable performance.

Cost pulls against precision. Elmougy and Liu [8] argue that blockchain AML detectors may need to favor recall because a missed illicit transaction can be more consequential than an additional alert. Lorenz et al. [12] frame analyst review as a constrained alert budget, making the operating threshold a workflow decision rather than a purely statistical one. The practical objective is therefore not maximum model accuracy in isolation, but a screening system that controls missed cases without overwhelming the analysts who must investigate what remains.

Two boundary conditions shape everything downstream. Not all outliers are illicit, and not all illicit transactions are outliers, because sophisticated laundering is designed to blend in; a flagged anomaly is therefore a question rather than a verdict [12]. Conversely, the public and persistent nature of blockchain records allows later investigative methods to be applied to earlier transactions, as the Monero traceability evidence demonstrates [14]. Detection produces ranked suspicion over a transparent but ambiguous ledger and passes that suspicion to attribution.

Attribution

Attribution involves linking a blockchain address to the person or entity that controls it. Reid and Harrigan [15] framed the problem as building the one-to-many mapping between one user and the many public keys that user controls. They showed that transaction and user networks can be reconstructed from public data and that, when this structure is combined with external information and flow analysis, particular holdings and movements can be traced.

Address clustering is the lever that makes attribution scale. Meiklejohn et al. [16] formalized the multi-input heuristic, under which addresses spent together as inputs to one transaction are treated as commonly controlled. They paired this with label propagation: identify one public key belonging to a known service, and the label can spread across the associated cluster. Hand-tagging 1,070 addresses through 344 transactions and propagating those labels enabled the authors to identify large flows, although the goal was targeted attribution rather than total de-anonymization [16].

That power rests on fragile ground. The change-address heuristic infers which output returns funds to the spender from contingent wallet software behavior rather than a guarantee of the Bitcoin protocol. Meiklejohn et al. [16] therefore cautioned that incorrect change-address links can merge addresses belonging to different users. Möser and Narayanan [17] subsequently showed that clustering heuristics have rarely been evaluated or calibrated rigorously: their effectiveness changes as protocol use evolves, the multi-input heuristic produces false positives in the presence of CoinJoin, and naive clustering of predicted change outputs can cause cluster collapse. Longitudinal estimates that failed to account adequately for change behavior were inaccurate by at least 11% to 14% [17].

Privacy technologies sit on the same edge between concealment and exposure. In an empirical analysis of Monero, Möser et al. [14] found that approximately 62% of transaction inputs containing one or more mixins were vulnerable to the evaluated chain-reaction traceability method. The weakness arose partly because decoys did not sufficiently resemble genuine spending, allowing the newest input to be inferred as the real one. Although a mandatory two-mixin minimum substantially reduced the weakness, the permanent ledger meant that later analytical advances could still be applied retrospectively to earlier transactions [14].

CoinJoin illustrates a related tension. It combines multiple senders in one transaction to weaken linkability and deliberately frustrates heuristics such as multi-input clustering. Nevertheless, Stütz et al. [18] found that CoinJoin transactions generated through the Wasabi and Samourai wallets could themselves be identified on-chain with high precision. Tracing transactions entering and leaving the mix could narrow the effective anonymity set, and a prior CoinJoin could later trigger compliance review when the funds reached an exchange.

The attribution literature carries two lessons downstream. Even the strongest results assume more than the chain itself, since practical tracing may depend on exchange records and other off-chain information [14]. In addition, on-chain heuristics remain consequential but unsafe: a mislinked change address can fuse two unrelated users, and the methods

themselves drift with the ecosystem [16,17]. Attribution narrows the field of suspects; it rarely closes a case alone, and what it yields is a candidate identity that someone must adjudicate.

Adjudication

Adjudication hands a flagged case to a person. By this point, the detector has ranked suspicion and attribution has proposed an identity; an analyst must determine whether to dismiss the alert, obtain additional evidence, escalate the matter, open an investigation, or file a report. Direct empirical evidence on this decision remains limited, but it is no longer accurate to describe the AML-specific evidence base as absent.

Oztas et al. [19] interviewed eight AML specialists for a combined 480 minutes about existing transaction-monitoring practices and requirements for improved methods. The participants identified explainability, flexibility, scalability, regulatory compliance, and the capacity to identify emerging risks and hidden relationships as important requirements. These findings support retaining analysts in the interpretation and contextual review of automated alerts because a useful system must expose enough information to be challenged and adapted as risks change. The study did not, however, compare analysts with models or measure systematic over-reliance on automated recommendations.

Two further studies extend the direct evidence. Bertrand et al. [20] conducted scenario-based workshops with 13 financial supervisors and six banking professionals. Their findings emphasize that AI used in financial-crime detection must provide evidence that supports supervisory challenge and legal justification rather than merely producing a persuasive explanation. Pyper and Wiese [21] conducted 15 semi-structured interviews with AML professionals and described investigation as work performed through a constellation of software tools whose design and integration affect both investigative effectiveness and the effort required. Together, these studies place human judgment within a broader sociotechnical workflow of evidence gathering, tool coordination, escalation, documentation, and accountability.

A related financial-fraud benchmark offers a computational account of how cases might be allocated between automated systems and human reviewers, but it is not direct blockchain-AML evidence. Alves et al.'s [22] published Financial Fraud Alert Review dataset contains predictions from 50 synthetic fraud analysts for 30,000 alert-review instances and models differences in behavior and work-capacity constraints. The benchmark shows that the relative performance of learning-to-defer methods changes according to which experts are available and how much work they can absorb. It supports capacity-aware allocation as a research direction but does not establish the proper division of authority between real AML analysts and blockchain AML models.

The AML-specific studies above are qualitative, participatory, or synthetic rather than controlled evaluations of analyst reliance on model advice. They do not establish the prevalence or magnitude of automation bias among AML analysts. The human-factors evidence that follows is therefore used to identify mechanisms that could plausibly affect AML adjudication under similar conditions of imperfect automation, workload, and accountability; it should not be interpreted as evidence that AML analysts exhibit those effects at the same frequency or magnitude as participants in other domains.

The reason the hand-off cannot simply be automated away traces to Bainbridge's [23] observation that automation creates new human difficulties. People are retained to intervene precisely when a system encounters an abnormal situation requiring unusual judgment, yet automation of routine work deprives them of the regular practice needed to maintain that judgment. In a simulated flight task, Skitka et al. [24] found that participants without an automated aid outperformed those using a highly reliable but imperfect one. Participants using the aid committed omission errors by missing events the system failed to flag and commission errors by following incorrect automated recommendations, a pattern the authors described as automation bias. A systematic review by Goddard et al. [25], which screened 13,821 records and included 74 studies, found comparable reliance failures across decision-support settings.

Why these effects resist obvious correction is especially relevant to system design. Parasuraman and Manzey [26] treated complacency and automation bias as related attentional phenomena and concluded that complacency can emerge under multitask load in both novices and experts. They also found that imperfect automation can produce omission and commission errors that are not reliably eliminated through practice, instruction, or training alone. Goddard et al. [25]

identified mediating factors such as cognitive style, experience with the system, and task-specific expertise, as well as possible mitigations including training and designs that make users accountable for the outcome. The implication for AML adjudication is not that training has no value, but that accountability and active review must also be embedded in the workflow.

Over-reliance is only half the failure surface because people may also discount accurate automated advice. Dietvorst et al. [27] documented algorithm aversion: participants preferred human judgment even when an evidence-based algorithm performed better, and confidence in the algorithm declined more sharply than confidence in a human after comparable errors. In two survey experiments involving 605 and 904 participants, Alon-Barkat and Busuioc [28] did not find a general tendency to defer to algorithmic advice. Instead, they found selective adherence, in which participants were more likely to follow an algorithm when its recommendation confirmed an existing expectation. They also cautioned that selective reliance may reinforce disadvantage when automated systems operate on already marginalized populations. An AML workflow must therefore guard against both deference to an incorrect model and rejection of a correct one.

Meaningful human control supplies the normative basis for retaining responsibility at the adjudication stage. Santoni de Sio and van den Hoven [29] propose two conditions: tracking, under which a system responds to the relevant human reasons in its environment, and tracing, under which an outcome can be connected to at least one person who understands the system and its moral significance. When neither condition is met, an automated system can create a responsibility gap in which no person can be held adequately accountable for the decision [29]. Adjudication is intended to prevent that gap from opening, while reporting preserves the evidence needed to trace the decision afterward.

Reporting

Reporting turns a decision into a record that others can inspect, including a suspicious activity report, audit trail, model-risk document, or account-closure record. A central technical question is what an explanation is expected to accomplish. Rudin [30] argues that using a black-box model and attaching an explanation afterward is an unsuitable default for high-stakes decisions because post-hoc explanations can obscure or misrepresent the reasoning they are intended to reveal. She instead advocates interpretable models where their use is feasible.

Complex models may nevertheless outperform simpler alternatives on some large datasets, creating demand for post-hoc methods. Shapley additive explanations assign additive feature-importance values to individual predictions and satisfy defined consistency properties within their model class [31]. Local interpretable model-agnostic explanations approximate a classifier near a particular prediction using a simpler local surrogate, allowing users to inspect the features influencing that result [32]. These methods produce different forms of explanation, but neither independently establishes that the underlying decision is correct or that the user has understood it.

A catch connects reporting directly back to adjudication. Across three human-AI decision datasets, Bansal et al. [33] found that explanations did not reliably improve complementary team performance; instead, they increased the rate at which people accepted model recommendations whether or not the recommendations were correct. An explanation can therefore buy compliance rather than understanding, allowing an artifact intended to support oversight to weaken it.

Governance frameworks are beginning to encode these concerns, although they do so unevenly. In US banking supervision, Supervisory Letter SR 26-2 was issued jointly by the Federal Reserve, the Office of the Comptroller of the Currency, and the Federal Deposit Insurance Corporation on April 17, 2026. It superseded SR 11-7 and the SR 21-8 Bank Secrecy Act/AML statement and is expected to be most relevant to banking organizations with more than $30 billion in assets. The guidance adopts a risk-based approach scaled to the institution and emphasizes that validation quality depends on the rigor and effectiveness of the challenge applied to a model. It is expressly non-binding and does not establish an independently enforceable legal standard [34].

Reporting obligations themselves are concrete and growing. FATF extended its AML and counter-terrorist-financing standards to virtual assets and virtual asset service providers in 2019 and applied the Travel Rule, under which originator and beneficiary information must be collected, held, and transmitted with qualifying transfers. Virtual asset service

providers across 19 reporting jurisdictions filed 134,500 suspicious transaction reports between 2018 and March 2020 [35]. The European Union's AI Act adds a parallel governance layer aimed at systems rather than individual transactions, requiring risk management, technical documentation, logging, disclosure to deployers, and effective human oversight for systems classified as high risk [36]. Whether a particular AML application falls within a high-risk category depends on its specific purpose and legal context; the Act nonetheless reinforces the broader design expectation that oversight must be operational rather than nominal.

The hardest case for any reporting regime is one in which no accountable human exists to report to or about. In Van Loon, the Fifth Circuit held in November 2024 that the immutable smart contracts of the Tornado Cash mixer were not property that the government could sanction because, once deployed, they were unownable, uncontrollable, and unchangeable even by their developers. A 2020 setup ceremony involving more than 1,100 participants had locked at least 20 of those contracts beyond anyone's control. The court did not dispute the underlying harm but found no owner to hold within the statutory definition [37]. That is the responsibility gap of the adjudication literature made literal, and it marks an outer limit of what reporting can reach.

Reporting failures also propagate backward into who gets banked. FATF treats the wholesale severing of entire customer categories without individual assessment as improper de-risking and warns that it can push activity into less visible channels, deepen financial exclusion, and reduce the transparency on which AML depends. Its 2025 financial-inclusion guidance asks countries to align inclusion and AML objectives rather than trade one against the other [37]. A reporting regime tuned only to avoid missed reports, and never to the cost of over-exclusion, minimizes one error while amplifying another.

Discussion

Automation does not suit the four stages equally, and the reason is structural rather than incidental. Parasuraman et al. [38] treat automation as something that can be applied to four classes of function: acquiring information, analyzing it, selecting a decision or action, and implementing it, each along a continuum from fully manual to fully automatic. Their framework provides a principled basis for deciding which functions to automate and how far, while warning that automation does not simply replace a human task; it reshapes the task and loads new coordination demands onto whoever remains [39].

Read across the pipeline, the evidence sorts into a clear asymmetry. Automation is strongest where the work is pattern-finding over the public record. Detection models rank suspicion effectively when labels exist, while later analyses can revisit the same ledger as methods and external information improve [2,14]. Automation weakens as the work turns from spotting patterns to settling what they mean. Attribution is the hinge: clustering heuristics scale and then fail in contingent, drifting ways that require review [17]. By adjudication, the system can prioritize or defer, but the decision to dismiss, escalate, investigate, or report remains accountable human work [19-22]. Reporting runs partly back toward automation because explanations, logs, and audit trails can be generated at scale, yet its hardest demands, including locating responsibility, judging evidential sufficiency, and weighing the cost of exclusion, are not resolved by generation alone.

Putting a human in the loop is necessary but not sufficient, because the human can fail in either direction. Human-AI teaming is attractive when full autonomy carries too much risk, but explanations do not automatically produce complementarity. Bansal et al. [33] found that explanations increased acceptance without consistently improving team accuracy. Complementarity materializes only when the human contribution is genuine scrutiny rather than a signature, and that condition is difficult to maintain under workload, automation bias, algorithm aversion, and selective adherence [25-28]. Allocation must therefore be engineered into the workflow. AML-specific studies point to explanation quality, tool integration, supervisory challenge, and access to supporting evidence as conditions for meaningful review [19-21].

Quantitative Synthesis and Empirical Basis for Allocation

The included studies were too heterogeneous for a pooled meta-analysis. They evaluated different units of analysis, including transactions, wallet addresses, entity clusters, synthetic analyst decisions, and human-AI classification tasks. Their outcome measures also differed, including F1-score, precision, recall, label efficiency, clustering error, traceability, recommendation acceptance, and review capacity. A descriptive quantitative synthesis is therefore used to identify where automation has demonstrated operational value and where human judgment remains necessary because the evidence is unstable, context-dependent, or incomplete.

Detection has the strongest quantitative basis for substantial automation. Across the Elliptic dataset family, reported illicit-class F1-scores reached 0.773 and 0.806, while the Elliptic++ Random Forest model achieved 97.5% precision and 71.9% recall [7,8,11]. Active-learning approaches approached the fully supervised baseline using approximately 5% of the labels, indicating that analyst attention can be concentrated on strategically selected cases rather than applied uniformly [11,12]. These results support allocating initial screening, ranking, and label-efficient case selection to automation. They do not support autonomous final decisions because performance deteriorated under the distribution shift and remained dependent on sparse and imperfect labels.

Attribution has a moderate but less stable empirical basis for automation. Failure to model change-address behavior adequately produced longitudinal errors of at least 11% to 14%, while historical analysis of Monero found that approximately 62% of evaluated inputs with mixins were vulnerable to the traceability method studied [14,17]. These findings demonstrate both the power and instability of automated attribution. Heuristics can narrow the field substantially, but their reliability depends on wallet behavior, protocol design, privacy techniques, and off-chain information. Human validation is retained not because analysts can reproduce graph analysis manually, but because automated links may be operationally consequential despite resting on contingent assumptions.

The empirical basis for adjudicative allocation is more limited. Direct studies involved eight AML specialists, 13 supervisors, six banking professionals, and 15 AML professionals [19-21]. These studies establish the practical importance of explainability, flexibility, tool integration, supervisory justification, and human review, but they do not compare model and analyst accuracy or quantify reliance errors. FiFAR adds a quantitative benchmark of 30,000 instances and 50 synthetic fraud analysts, showing that deferral performance changes with analyst behavior, availability, and capacity [22]. Because the experts are synthetic and the task concerns financial fraud rather than blockchain AML, the benchmark supports capacity-aware design rather than a fixed allocation rule.

Reporting and explanation likewise lack evidence supporting full automation. Across three human-AI datasets, explanations increased acceptance of model recommendations without consistently improving team performance [33]. Explanation generation, logging, and evidence summarization can be automated, but the resulting artifact does not establish that meaningful review occurred. Human responsibility remains necessary for judging whether the available evidence justifies escalation or reporting and for ensuring that the record accurately reflects the basis of the decision.

Automation is most useful before a case turns on interpretation and accountability. Models can screen large datasets, rank suspicion, direct scarce labeling effort, trace network relationships, and generate records. Analysts remain necessary when a decision depends on validating unstable links, bringing in off-chain evidence, interpreting intent and context, or deciding whether the legal and evidential threshold for escalation or reporting has been met. Table *2* converts these findings into an allocation matrix. For each stage, it separates the empirical basis for automation, the judgment retained by the analyst, the dominant failure mode, and the unanswered empirical question.

How to cite this article:
Sorinola M, Hamed K (August 24, 2026) Allocating Human Judgment in Automated Blockchain Anti-Money Laundering: A Review of Detection, Attribution, Adjudication, and Reporting. Cureus J Comput Sci 3 : es44389-026-00242-4. DOI https://doi.org/10.7759/s44389-026-00242-4

Stage	Automation strength and empirical basis	Human function retained	Dominant failure mode	Remaining empirical gap
Detection	High, but benchmark-dependent. Illicit-class F1-scores of 0.773–0.806; precision of 97.5% and recall of 71.9%; active learning approached supervised performance using approximately 5% of labels [7,8,11,12].	Audit labels, identify distribution shifts, set operating thresholds, and review uncertain or high-risk alerts.	Model confidence persists despite concept drift, sparse labels, or biased training data.	External validation using live institutional data, realistic alert volumes, and production label noise.
Attribution	Moderate and context-dependent. Inadequate treatment of change behavior produced errors of at least 11%–14%; approximately 62% of historically evaluated Monero inputs with mixins were vulnerable to the studied tracing method [14,17].	Validate address links, incorporate off-chain evidence, and determine whether attribution is sufficiently reliable for further action.	Fragile clustering or outdated heuristics assign one person's activity to another person or entity.	Prospective evaluation against verified identities across changing wallet designs, mixers, and privacy technologies.
Adjudication	Limited. Published studies involved eight AML specialists, 13 supervisors and six banking professionals, and 15 AML professionals; FiFAR contains 30,000 cases and 50 synthetic fraud analysts [19-22].	Interpret context, challenge model recommendations, decide whether to dismiss, escalate, investigate, or report, and remain accountable for the decision.	Over-reliance, algorithm aversion, or selective adherence distorts the analyst's judgment under workload and uncertainty.	Controlled studies with practicing AML analysts measuring acceptance, override, error, confidence, workload, and review time.

Reporting	Moderate for record generation, limited for judgment. Across three human-AI datasets, explanations increased acceptance without consistently improving team performance [33].	Determine evidential sufficiency, preserve traceability, approve reporting, and consider the consequences of the decision.	A persuasive explanation is mistaken for understanding, or the audit record fails to identify an accountable decision-maker.	AML-specific evaluation of explanation formats, audit trails, reporting quality, and downstream regulatory outcomes.

TABLE 2: Evidence-weighted allocation matrix for automated blockchain AML

AML, Anti-Money Laundering

Table *2* does not establish that every case should follow an identical allocation. It identifies the functions for which the available evidence is strongest. Detection can be substantially automated because performance can be measured against labeled outcomes, although concept drift and sparse labels remain important constraints. Attribution requires greater human validation because the underlying heuristics can drift or produce consequential false links. Adjudication and reporting remain human-led, not because direct experiments have established universal human superiority, but because the available automated evidence does not independently resolve meaning, intent, evidential sufficiency, or accountability.

The two most important points to instrument are the deferral boundary in adjudication and the traceability of the reporting record. These are where a model recommendation becomes a human decision and where the quality of that judgment can later be evaluated. A system tuned only to minimize missed reports may optimize the wrong objective, because the cost of over-exclusion falls on people who do not appear in conventional detection metrics.

The largest gap is not the complete absence of AML evidence but the absence of controlled AML-specific evidence on analyst reliance. Existing studies involving AML specialists, professionals, supervisors, and banking practitioners establish the practical importance of explainability, system integration, supervisory justification, flexibility, and human review [18-20]. They do not quantify how frequently analysts accept an incorrect automated recommendation, reject a correct one, or change their behavior under different levels of workload, uncertainty, explanation, and accountability. The next empirical step should be a field or controlled study using realistic AML alerts and practicing analysts, measuring override behavior, false-positive and false-negative decisions, review time, confidence, workload, and the effects of different explanation and accountability designs.

Limitations of the Present Review

Several limits should be kept in view when interpreting this review. The qualitative and participatory studies were not subjected to a formal risk-of-bias assessment. The discussion of adjudication also draws partly on evidence from aviation, healthcare, and public administration, and it remains uncertain whether the same effects occur with similar frequency or strength among AML analysts. The review does not estimate the operational costs associated with false positives, false negatives, analyst workload, delayed investigations, or financial exclusion. Its narrative design was intended to identify evidence relevant to the four-stage framework rather than to provide an exhaustive systematic synthesis. The allocation matrix should therefore be read as a structured guide for design and evaluation, not as a validated operational standard.

Conclusions

Evidence for automation is not evenly distributed across the AML pipeline. Detection has the clearest quantitative support, although its performance remains benchmark-dependent and vulnerable to sparse labels and distribution shift. Attribution benefits from scalable tracing and clustering methods, but those methods rest on assumptions that change across wallets, protocols, and privacy tools. Direct AML evidence becomes much thinner at adjudication and reporting, where context, evidential sufficiency, and accountability still sit with the analyst. The allocation matrix reflects this evidence gradient rather than prescribing a universal division of labor.

The next test should be a controlled field study involving practicing AML analysts working with realistic alert queues. Such a study should record when analysts accept, reject, or override model recommendations; compare error rates, review times, and confidence under realistic alert volumes; and vary explanation formats and accountability requirements. Linking those decisions to false positives, false negatives, escalation quality, and reporting outcomes would show where human review improves the system and where it adds delay or another source of error.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Kehinde Hamed, Moshood Sorinola

Acquisition, analysis, or interpretation of data: Kehinde Hamed, Moshood Sorinola

Drafting of the manuscript: Kehinde Hamed, Moshood Sorinola

Supervision: Kehinde Hamed, Moshood Sorinola

Critical review of the manuscript for important intellectual content: Moshood Sorinola

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

Data sharing is not applicable. This work is theoretical; no datasets were generated or analyzed.

References

1. Nakamoto S: Bitcoin: a peer-to-peer electronic cash system. (2008). Accessed: June 22, 2026: https://bitcoin.org/bitcoin.pdf.
2. Europol: Cryptocurrencies: tracing the evolution of criminal finances. (2022). Accessed: June 22, 2026: https://www.europol.europa.eu/publications-events/publications/cryptocurrencies-tracing-evolution-of-criminal-finances.
3. United Nations Office on Drugs and Crime: Money laundering. (2024). Accessed: June 22, 2026: https://www.unodc.org/unodc/en/money-laundering/overview.html.
4. Foley S, Karlsen JR, Putniņš TJ: Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies?. The Review of Financial Studies. 2019, 32:1798-1853. 10.1093/rfs/hhz015

How to cite this article:
Sorinola M, Hamed K (August 24, 2026) Allocating Human Judgment in Automated Blockchain Anti-Money Laundering: A Review of Detection, Attribution, Adjudication, and Reporting. Cureus J Comput Sci 3 : es44389-026-00242-4. DOI https://doi.org/10.7759/s44389-026-00242-4

5. Financial Action Task Force: International standards on combating money laundering and the financing of terrorism and proliferation: The FATF Recommendations. (2012). Accessed: June 22, 2026: https://www.fatf-gafi.org/en/publications/Fatfrecommendations/Fatf-recommendations.html.

6. Financial Crimes Enforcement Network: Interpretation of suspicious activity reporting requirements to permit the unitary filing of suspicious activity and blocking reports. (2004). Accessed: June 22, 2026: https://www.fincen.gov/resources/statutes-regulations/guidance/interpretation-suspicious-activity-reporting-requirements.

7. Alarab I, Prakoonwit S, Nacer MI: Competence of graph convolutional networks for anti-money laundering in Bitcoin blockchain. Proceedings of the 2020 5th International Conference on Machine Learning Technologies. 2020, 23-27. 10.1145/3409073.3409080

8. Elmougy Y, Liu L: Demystifying fraudulent transactions and illicit nodes in the Bitcoin network for financial forensics. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023, 3979-3990. 10.1145/3580305.3599803

9. Song K, Dhraief MA, Xu M, Cai L, Chen X, Mithal A, Chen J: Identifying money laundering subgraphs on the blockchain. Proceedings of the 5th ACM International Conference on AI in Finance. 2024, 195-203. 10.1145/3677052.3698635

10. Altman E, Blanuša J, von Niederhäusern L, Egressy B, Anghel A, Atasu K: Realistic synthetic financial transactions for anti-money laundering models. Advances in Neural Information Processing Systems. 2023, 36:29851-29874. 10.52202/075280-1300

11. Alarab I, Prakoonwit S: Graph-based LSTM for anti-money laundering: experimenting temporal graph convolutional network with Bitcoin data. Neural Processing Letters. 2023, 55:689-707. 10.1007/s11063-022-10904-8

12. Lorenz J, Silva MI, Aparício D, Ascensão JT, Bizarro P: Machine learning methods to detect money laundering in the Bitcoin blockchain in the presence of label scarcity. Proceedings of the First ACM International Conference on AI in Finance. 2020, 1-8. 10.1145/3383455.3422549

13. Madem S, Margana S, Manda VK, Tarnanidis T, Sousa BB: Federated learning for financial fraud detection. Strategic Frameworks for Managing Risk in FinTech Ecosystems. Faisal SM, Khan AK (ed): IGI Global Scientific Publishing, Hershey, PA; 2026. 127-156. 10.4018/979-8-3373-2980-2.ch005

14. Möser M, Soska K, Heilman E, et al.: An empirical analysis of traceability in the Monero blockchain. Proceedings on Privacy Enhancing Technologies. 2018, 2018:143-163. 10.1515/popets-2018-0025

15. Reid F, Harrigan M: An analysis of anonymity in the Bitcoin system. Security and Privacy in Social Networks. Altshuler Y, Elovici Y, Cremers A, Aharony N, Pentland A (ed): Springer, New York, NY; 2013. 197-223. 10.1007/978-1-4614-4139-7_10

16. Meiklejohn S, Pomarole M, Jordan G, Levchenko K, McCoy D, Voelker GM, Savage S: A fistful of Bitcoins: characterizing payments among men with no names. Proceedings of the 2013 Conference on Internet Measurement Conference. 2013, 127-140. 10.1145/2504730.2504747

17. Möser M, Narayanan A: Resurrecting address clustering in Bitcoin. Financial Cryptography and Data Security. Eyal I, Garay J (ed): Springer, Cham; 2022. 13411:386-403. 10.1007/978-3-031-18283-9_19

18. Stütz R, Stockinger J, Moreno-Sanchez P, Haslhofer B, Maffei M: Adoption and actual privacy of decentralized CoinJoin implementations in Bitcoin. Proceedings of the 4th ACM Conference on Advances in Financial Technologies. 2022, 254-267. 10.1145/3558535.3559782

19. Oztas B, Cetinkaya D, Adedoyin F, Budka M, Aksu G, Dogan H: Transaction monitoring in anti-money laundering: a qualitative analysis and points of view from industry. Future Generation Computer Systems. 2024, 159:161-171. 10.1016/j.future.2024.05.027

20. Bertrand A, Eagan JR, Maxwell W, Brand J: AI is entering regulated territory: understanding the supervisors' perspective for model justifiability in financial crime detection. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 2024, 1-21. 10.1145/3613904.3642326

21. Pyper JD, Wiese J: An exploratory study of sociotechnical issues for anti-money laundering workers. Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. 2026, 1-16. 10.1145/3772318.3791584

22. Alves JV, Leitão D, Jesus S, et al.: A benchmarking framework and dataset for learning to defer in human-AI decision-making. Scientific Data. 2025, 12:506. 10.1038/s41597-025-04664-y

23. Bainbridge L: Ironies of automation. Automatica. 1983, 19:775-779. 10.1016/0005-1098(83)90046-8

24. Skitka LJ, Mosier KL, Burdick M: Does automation bias decision-making?. International Journal of Human-Computer Studies. 1999, 51:991-1006. 10.1006/ijhc.1999.0252

How to cite this article:
Sorinola M, Hamed K (August 24, 2026) Allocating Human Judgment in Automated Blockchain Anti-Money Laundering: A Review of Detection, Attribution, Adjudication, and Reporting. Cureus J Comput Sci 3 : es44389-026-00242-4. DOI https://doi.org/10.7759/s44389-026-00242-4

25. Goddard K, Roudsari A, Wyatt JC: Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association. 2012, 19:121-127. 10.1136/amiajnl-2011-000089

26. Parasuraman R, Manzey DH: Complacency and bias in human use of automation: an attentional integration. Human Factors. 2010, 52:381-410. 10.1177/0018720810376055

27. Dietvorst BJ, Simmons JP, Massey C: Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General. 2015, 144:114-126. 10.1037/xge0000033

28. Alon-Barkat S, Busuioc M: Human-AI interactions in public sector decision making: "automation bias" and "selective adherence" to algorithmic advice. Journal of Public Administration Research and Theory. 2023, 33:153-169. 10.1093/jopart/muac007

29. Santoni de Sio F, van den Hoven J: Meaningful human control over autonomous systems: a philosophical account. Frontiers in Robotics and AI. 2018, 5:15. 10.3389/frobt.2018.00015

30. Rudin C: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019, 1:206-215. 10.1038/s42256-019-0048-x

31. Lundberg SM, Lee S-I: A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, 30:4768-4777.

32. Ribeiro MT, Singh S, Guestrin C: "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016, 1135-1144. 10.1145/2939672.2939778

33. Bansal G, Wu T, Zhou J, et al.: Does the whole exceed its parts? The effect of AI explanations on complementary team performance. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021, 1-16. 10.1145/3411764.3445717

34. Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency, Federal Deposit Insurance Corporation: Revised guidance on model risk management (Supervisory Letter SR 26-2). (2026). Accessed: July 27, 2026: https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm.

35. European Parliament and Council of the European Union: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending regulations. (2024). Accessed: June 22, 2026: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689.

36. United States Court of Appeals for the Fifth Circuit: Van Loon v. Department of the Treasury, 122 F.4th 549 (5th Cir. 2024). (2024). Accessed: June 22, 2026: https://www.ca5.uscourts.gov/opinions/pub/23/23-50669-CV0.pdf.

37. Financial Action Task Force: Drivers and actions to tackle de-risking; and guidance on anti-money laundering and terrorist financing measures and financial inclusion, 2025 update. (2025). Accessed: June 22, 2026: https://www.fatf-gafi.org/en/publications/Fatfrecommendations/Fatf-action-to-tackle-de-risking.html.

38. Parasuraman R, Sheridan TB, Wickens CD: A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans. 2000, 30:286-297. 10.1109/3468.844354

39. Financial Action Task Force: Updated guidance for a risk-based approach to virtual assets and virtual asset service providers; and 12-month review of the revised FATF standards on virtual assets and VASPs. (2020). Accessed: June 22, 2026: https://www.fatf-gafi.org/content/dam/fatf-gafi/guidance/RBA-VA-VASPs.pdf.

How to cite this article:
Sorinola M, Hamed K (August 24, 2026) Allocating Human Judgment in Automated Blockchain Anti-Money Laundering: A Review of Detection, Attribution, Adjudication, and Reporting. Cureus J Comput Sci 3 : es44389-026-00242-4. DOI https://doi.org/10.7759/s44389-026-00242-4