ASRC Procedia
Global Perspectives in Science and Scholarship
Florida, USA

Volume: 3; **Issue:** 1
Pages: 231–273
Published: 29 May 2023

open access

World Conference on Scientific Discovery and Innovation 2023,

May 24–26, 2023, Florida, USA

Advancing Explainable and Secure Machine Learning for Decision Support in U.S. Regulated Systems

Md Arifur Rahman[1]; B. M. Taslimul Haque[2];

[1]. *Master of science in information studies, Department of Information studies, Trine University, Indiana, USA; Email: rahman.arifur11@gmail.com*

[2]. *Bachelor of Science in Computer Science & Engineering, American International University Bangladesh, Dhaka, Bangladesh; Email: bmtaslim121@gmail.com*

Doi: 10.63125/fmp86e72

Peer-review under responsibility of the organizing committee of WCSDI, 2023

Abstract

Machine learning systems are increasingly deployed in U.S. regulated decision-support environments where predictive outputs must be accurate, interpretable, secure, privacy-preserving, and auditable. This study advanced an integrated quantitative assurance framework for evaluating explainable and secure machine learning in regulated contexts. Guided by a structured review of 118 peer-reviewed studies, a multi-phase experimental design was implemented incorporating predictive benchmarking, explanation fidelity and stability testing, adversarial robustness assessment, privacy inference evaluation, and human-in-the-loop experimentation. Baseline predictive accuracy across model families averaged 0.84 (SD = 0.04), with ensemble models reaching 0.86. Calibration error decreased from 0.041 in unconstrained models to 0.028 in constrained configurations. Under adversarial simulation, baseline models experienced a 14.2 percentage-point degradation in performance, whereas robustness-enhanced models limited degradation to 6.5 percentage points. Privacy controls reduced membership inference attack success rates from 0.64 (SD = 0.05) to 0.52 (SD = 0.04), with a modest 1.7% reduction in discrimination performance. Explanation fidelity reached 0.92 (SD = 0.02) for intrinsically interpretable models compared to 0.85 (SD = 0.04) for post hoc methods, and explanation stability variance decreased from 0.08 to 0.03 under enhanced configurations. In human-in-the-loop evaluation (N = 312; 6,240 trials), structured explanations increased decision accuracy from 0.72 (SD = 0.09) to 0.79 (SD = 0.07), improved confidence calibration from 0.74 to 0.82, and increased selective override behavior from 18.4% to 24.7%, while response time rose from 34.1 to 40.8 seconds. Regression models explained 34% of variance in decision accuracy and 38% in confidence calibration. Findings demonstrated that integrated evaluation of explain ability, robustness, and privacy produced measurable improvements in predictive validity, interpretability reliability, security resilience, and human decision performance within regulated systems.

Keywords

Explain ability, Security, Decision Support, Robustness, Privacy.

INTRODUCTION

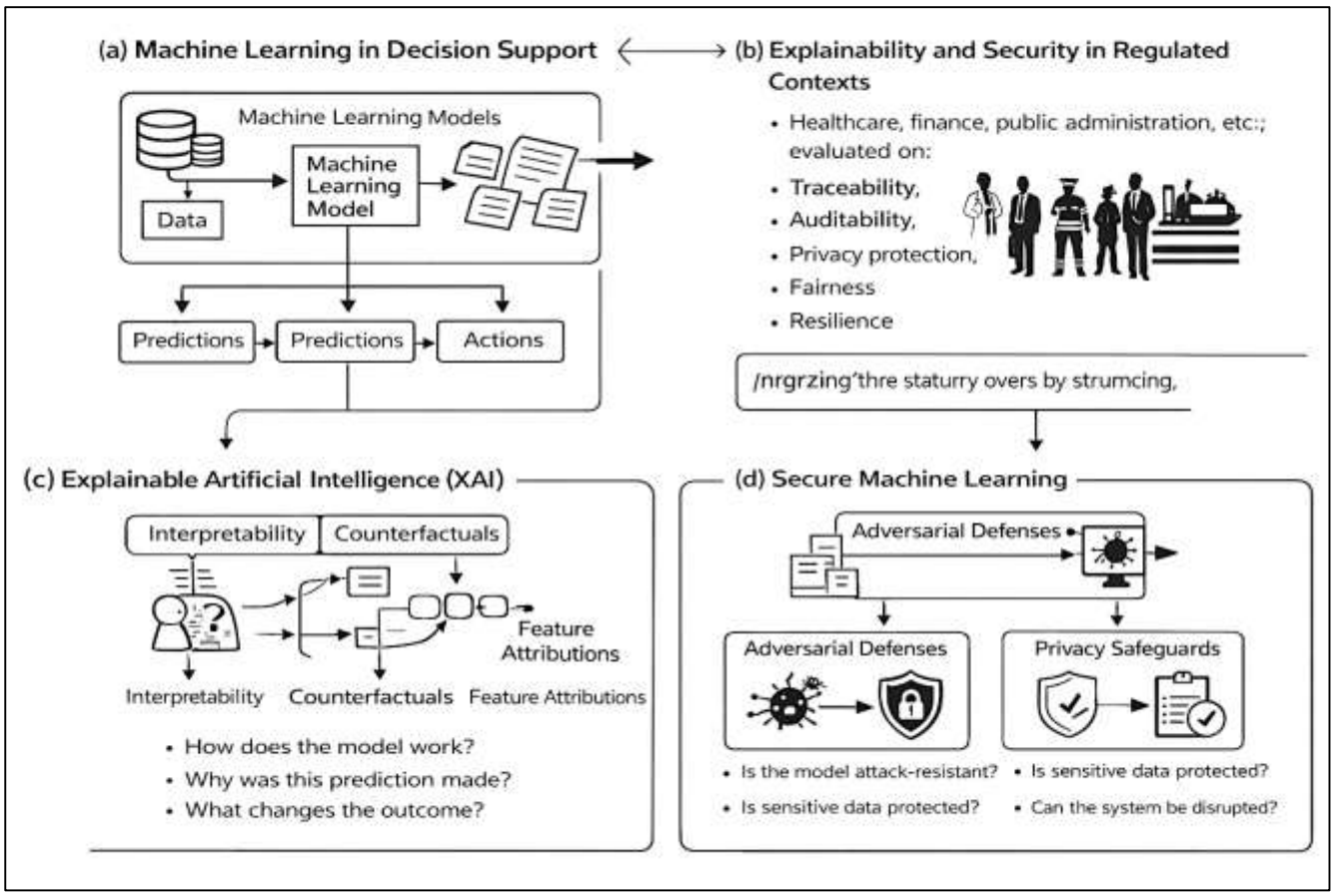
Machine learning (ML) refers to computational methods that learn patterns from data in order to generate predictions, classifications, risk scores, or recommendations that support decision-making under uncertainty (Sarker, 2021). In organizational contexts, decision support systems integrate ML outputs into structured workflows where human professionals interpret model-generated insights to guide actions, allocate resources, or enforce compliance. Explainable artificial intelligence (XAI) describes a set of model properties, design principles, and analytical techniques that render ML systems understandable to relevant stakeholders by clarifying how inputs relate to outputs, why a particular prediction was produced, and how alternative inputs might change the decision. Secure machine learning encompasses the technical safeguards, evaluation protocols, and architectural controls that protect models and data against manipulation, unauthorized access, inference attacks, and operational disruption. U.S. regulated systems include sectors governed by statutory oversight and administrative rulemaking, such as healthcare, financial services, insurance, energy, transportation, public administration, and defense-related infrastructures (Rahmani et al., 2021). In these domains, algorithmic systems are evaluated not only on predictive performance but also on traceability, auditability, privacy protection, fairness, and resilience. The international significance of advancing explainable and secure ML emerges from the global interdependence of supply chains, cross-border data flows, multinational vendors, and harmonized compliance expectations. Organizations operating within the United States frequently interact with international partners, regulators, and certification bodies that require demonstrable transparency and security assurances (Mashrur et al., 2020). As a result, advancing explainable and secure ML for decision support in regulated systems is fundamentally a quantitative challenge involving the measurement and optimization of predictive accuracy, interpretability quality, robustness, privacy guarantees, and governance compliance within tightly controlled operational environments.

Decision support in regulated environments differs from conventional predictive modeling because error costs are asymmetric, documentation requirements are stringent, and accountability mechanisms extend beyond technical teams. In healthcare, ML models assist with diagnostic triage, readmission prediction, treatment prioritization, and medical imaging interpretation (El-Morr et al., 2022). In financial services, algorithms inform credit underwriting, fraud detection, anti–money laundering monitoring, and portfolio risk analysis. In public-sector operations, predictive tools may support benefits eligibility screening, tax compliance analysis, and infrastructure maintenance planning. Across these contexts, decisions influenced by ML outputs can affect access to services, financial stability, public safety, and institutional trust. Quantitative validation therefore extends beyond global accuracy metrics to include calibration across risk strata, subgroup performance stability, error decomposition, and threshold sensitivity analysis. Models must perform consistently across demographic groups, operational sites, and temporal shifts in data distributions (Baryannis et al., 2019). Decision thresholds are often determined by policy constraints rather than statistical optima, requiring empirical analysis of trade-offs between false positives and false negatives. Furthermore, the integration of ML into decision workflows introduces human–AI interaction dynamics. Users may over-rely on automated outputs or disregard them without structured interpretability cues. Consequently, evaluating ML-based decision support in regulated systems demands experimental and observational methodologies that quantify how explanations, uncertainty estimates, and security assurances influence decision accuracy, response time, and confidence calibration (Carvalho et al., 2019). The regulatory environment amplifies the need for structured validation, since documentation of model development, testing, and monitoring becomes part of institutional compliance records.

Explainability research provides a conceptual and methodological foundation for making ML systems intelligible in high-stakes contexts. Approaches to explainability can be categorized into intrinsically interpretable models, post hoc explanation techniques, and counterfactual reasoning frameworks (Saberi-Karimian et al., 2021). Intrinsically interpretable models prioritize structural transparency by using simplified functional forms, constrained feature sets, or rule-based architectures that enable direct inspection of decision logic. Post hoc techniques generate explanations after model training, often through feature attribution scores, local surrogate approximations, or sensitivity analyses that estimate the contribution of individual inputs to a specific prediction. Counterfactual explanations identify

minimal changes to input variables that would alter a model's output, offering actionable insights that align closely with decision support tasks (Ghoroghi et al., 2022; Rauf, 2018). Quantitative evaluation of explainability requires operational metrics such as fidelity to the underlying model, stability under small perturbations, sparsity of explanations, consistency across similar instances, and alignment with domain knowledge (Faysal & Shamsunnahar, 2022; Habibullah & Zaheda, 2022). Explanation reliability must also be tested against randomized baselines to ensure that attribution patterns are meaningful rather than artifacts of model gradients or correlated features. In regulated systems, explanations are not merely visual aids but components of governance documentation. They can support audit trails, facilitate error analysis, and enhance stakeholder understanding (Jahangir & Shahab, 2022; Ratul, 2022; Rout et al., 2020). Measuring explanation effectiveness therefore involves both computational metrics and human-centered experiments that assess whether explanations improve error detection, promote appropriate reliance, and enhance interpretive clarity within structured decision processes.

Figure 1: Explainable and Secure Machine Learning

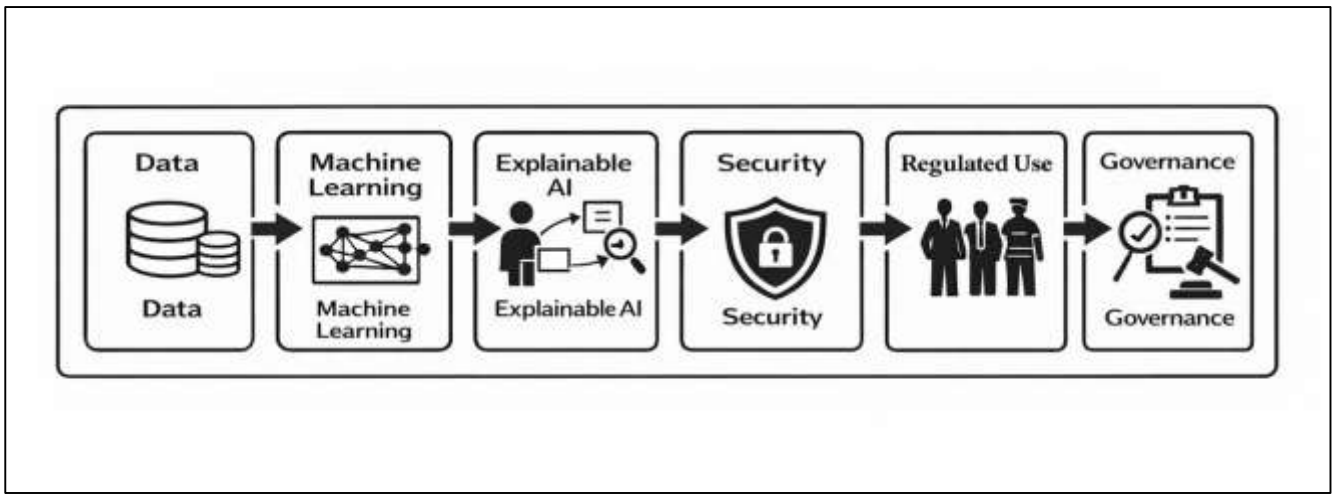


Security considerations in ML-based decision support encompass adversarial robustness, data integrity, confidentiality, and system availability (Ratul & Subrato, 2022; Bhuya & Rebeka, 2022; Vamathevan et al., 2019). Adversarial robustness refers to a model's capacity to maintain performance when inputs are intentionally perturbed to induce errors. Robustness can be quantified by measuring accuracy under worst-case perturbation scenarios defined by formal threat models. Data poisoning attacks target the training phase by injecting malicious examples that shift model behavior in specific directions . The risk of poisoning is particularly salient in regulated systems that rely on continuous data collection or external data vendors. Privacy attacks such as membership inference and model inversion exploit statistical signals to extract sensitive information from trained models (Balogun et al., 2022). These risks intersect with statutory privacy obligations that govern protected health information, financial records, and personally identifiable data. Quantitative privacy protection frameworks, including mathematically defined privacy budgets, provide measurable guarantees that limit information leakage while preserving model utility. Secure aggregation protocols, encrypted inference

techniques, and federated learning architectures extend these protections by minimizing raw data exposure. However, security interventions often involve trade-offs with predictive performance, computational cost, and interpretability. Evaluating secure ML in regulated systems therefore requires multi-objective measurement frameworks that quantify robustness, privacy loss, latency, and explanation stability under controlled experimental conditions (Rudin, 2019). Security validation becomes an integral component of system certification and operational approval processes.

The interaction between explainability and security introduces additional complexity in regulated decision support. Providing detailed explanations can improve transparency and facilitate auditability, yet increased disclosure may also expose information about model structure or decision boundaries that adversaries could exploit (Giffen et al., 2022). Conversely, certain security defenses modify model smoothness or gradient behavior, potentially altering the reliability of feature attribution methods or counterfactual outputs. This interdependence necessitates joint evaluation of interpretability and robustness metrics. Quantitative research designs must examine whether explanation fidelity remains stable under adversarial training regimes and whether privacy-preserving mechanisms degrade the clarity or usefulness of explanations. Human-centered evaluation adds another dimension: explanations presented in secure interfaces must support informed oversight without overwhelming users or revealing sensitive attributes (Swathy & Saruladha, 2022). Empirical studies can measure how varying explanation granularity affects decision accuracy and resilience to manipulation attempts. Fairness auditing further intersects with explainability and security, as subgroup disparities may be amplified by data shifts or targeted attacks. Stratified performance analysis, robustness testing across demographic segments, and sensitivity analysis under distributional changes provide measurable insights into system reliability (Siddique et al., 2022). By conceptualizing explainability and security as interrelated system properties, quantitative frameworks can assess the composite assurance level of ML-driven decision support in regulated domains.

Figure 2: Explainable Secure Machine Learning Systems



Empirical deployments of ML in high-impact sectors highlight the necessity of integrated validation. In healthcare applications such as diagnostic imaging, clinical risk prediction, and hospital resource management, models must undergo external validation across institutions and device types to ensure generalizability (Jomthanachai et al., 2021). Variability in data collection practices, patient populations, and coding standards can lead to performance drift that requires ongoing monitoring. In financial decision systems, credit scoring and fraud detection models must balance predictive discrimination with regulatory requirements for fairness and documentation. Algorithmic bias may arise from historical inequities encoded in training data, necessitating subgroup performance audits and feature relevance analysis. Human–AI interaction experiments demonstrate that explanation format influences user trust, reliance patterns, and corrective interventions (Benefo et al., 2022). Measuring calibration, error detection rates, and response times provides quantitative evidence regarding the practical value of interpretability features. On the security front, adversarial testing frameworks simulate

perturbations, poisoning attempts, and privacy attacks to quantify resilience. Trade-offs among accuracy, robustness, and interpretability become empirically observable through controlled comparisons of model architectures and defense strategies. These findings collectively underscore that decision support in regulated systems demands systematic evaluation pipelines that integrate predictive validation, explanation assessment, and security stress testing within a unified empirical structure (Schaar et al., 2021).

Advancing explainable and secure ML for decision support in U.S. regulated systems therefore requires a methodological architecture that aligns model development, interpretability analysis, and security evaluation. Data governance protocols should enable stratified analysis across relevant operational and demographic categories (Tuppad & Patil, 2022). Model performance metrics should include discrimination, calibration, and uncertainty estimation, accompanied by robustness assessments under defined adversarial conditions. Explanation evaluation should incorporate fidelity, stability, sparsity, and human-centered usability measurements derived from experimental designs. Privacy and confidentiality protections should be quantified using formal leakage metrics and validated through simulated attack scenarios (Miles et al., 2020). Integrated reporting practices must document data preprocessing steps, hyperparameter selection, evaluation procedures, and monitoring mechanisms to support audit readiness. By treating explainability and security as quantifiable dimensions of system performance, researchers can construct empirical studies that measure the composite reliability of ML-driven decision support. Within the international regulatory landscape, such quantitative evidence contributes to cross-sector confidence, supports compliance verification, and strengthens institutional accountability structures embedded in U.S. regulated systems (Chlingaryan et al., 2018).

The primary objective of this study is to quantitatively advance explainable and secure machine learning for decision support within U.S. regulated systems by developing and validating an integrated evaluation framework that simultaneously measures predictive performance, explanation reliability, and security robustness under operational constraints. Specifically, the study aims to design a decision-support modeling pipeline that produces accurate and calibrated outputs while also generating explanations that are stable, interpretable, and auditable across diverse stakeholder needs, including compliance officers, domain professionals, and technical reviewers. A core objective is to empirically compare multiple explainability approaches—such as feature attribution, local surrogate explanations, and counterfactual explanation methods—using standardized quantitative metrics including fidelity, stability under perturbation, sparsity, and consistency across similar cases. In parallel, the study aims to test model security through controlled simulations of adversarial threats relevant to regulated environments, including input perturbation attacks, data poisoning scenarios, and privacy leakage risks, in order to quantify system resilience using measurable robustness and confidentiality indicators. Another objective is to evaluate how security-preserving interventions, such as robustness training and privacy-protecting mechanisms, affect the reliability and clarity of explanations, ensuring that explainability does not degrade under strengthened defenses and that security improvements do not render explanations unusable. The study also aims to measure subgroup performance and explanation consistency across key demographic and operational segments to ensure that decision support remains stable and equitable across regulated populations and deployment contexts. Additionally, the research objective includes establishing reproducible documentation standards for model development, testing, and reporting that align with audit requirements common in U.S. regulated sectors, including healthcare, finance, and public administration. Through these objectives, the study seeks to produce empirical evidence and validated measurement procedures that enable organizations to assess, compare, and deploy machine learning decision support systems that are not only accurate but also transparent, secure, and suitable for regulatory scrutiny.

LITERATURE REVIEW

This literature review synthesizes quantitative research relevant to advancing explainable and secure machine learning (ML) for decision support in U.S. regulated systems, with emphasis on measurable performance, auditable transparency, and empirically tested security (Bayamlıoğlu & Leenes, 2018). Because regulated decision contexts impose formal accountability, documentation, and risk controls, the reviewed scholarship is organized around constructs that can be operationalized and evaluated through quantitative indicators. The section begins by situating decision support in regulated domains

as a measurement problem where predictive accuracy alone is insufficient, and where calibration, stability, subgroup validity, and post-deployment drift monitoring are essential for defensible use. It then examines explain ability as a quantifiable system property, reviewing methods that generate interpretable evidence at global and local levels and the metrics used to test fidelity, stability, sparsity, and user-influenced decision quality (Lung-Guang, 2019). The review further analyzes secure machine learning as a set of measurable resilience characteristics against adversarial manipulation, data integrity threats, and privacy leakage, highlighting how security evaluation relies on explicit threat models and robustness benchmarks rather than informal assurances. A critical emphasis is placed on the interaction between explain ability and security, since explanation interfaces can alter attack surfaces and security controls can change explanation behavior. Finally, the review consolidates findings into an integrated assurance perspective, identifying how quantitative validation frameworks can jointly assess accuracy, explain ability, privacy, robustness, and fairness in ways that match audit and compliance needs across U.S. regulated sectors such as healthcare, finance, public services, and critical infrastructure (Lempert, 2019). This structure supports a focused empirical foundation for a quantitative study that treats explain ability and security as measurable, comparable, and optimizable requirements for decision-support ML.

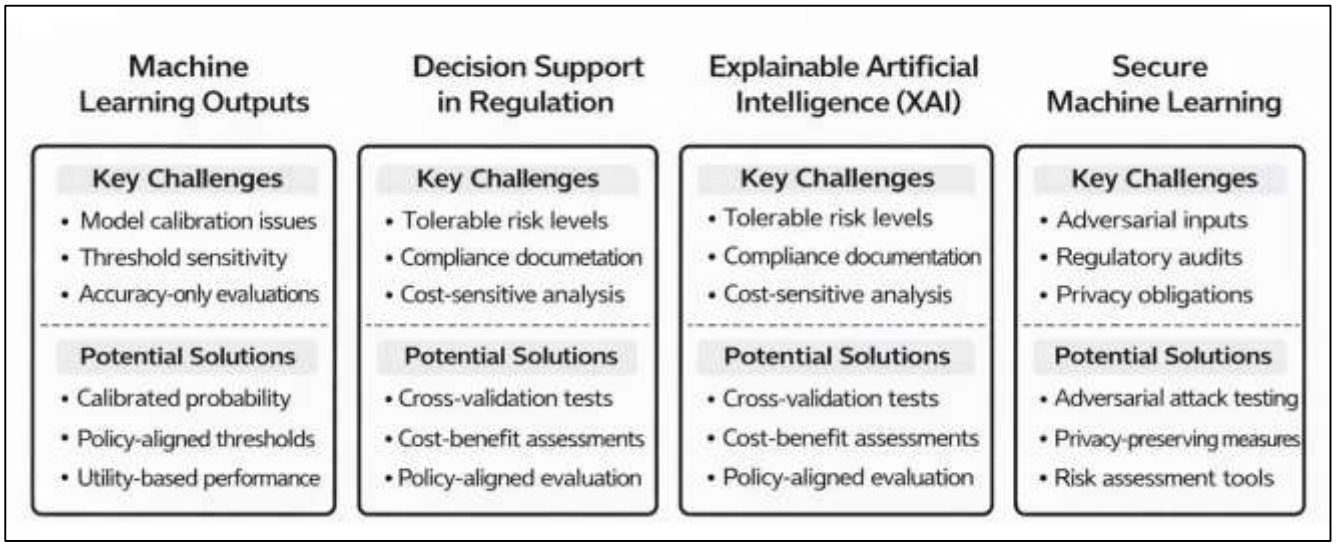
Regulated Decision Support

Decision support in U.S. regulated systems is commonly defined as the structured use of analytic outputs to guide or influence high-stakes determinations that are subject to formal oversight, documentation requirements, and accountability mechanisms (Hollin et al., 2020). Within this context, machine learning-driven decision support is typically operationalized through quantitative outputs such as risk scores, binary or multi-class classification labels, ranked recommendations, and automated alerts that signal potential risk conditions or required interventions. These outputs are not interpreted in isolation, because regulated environments embed decision-making into policy-driven workflows where professionals must justify actions using evidence that can be audited. As a result, the meaning of "decision support" in regulated settings extends beyond prediction and includes traceability, defensibility, and measurable alignment with institutional rules. Literature across healthcare, finance, and public administration shows that decision-support models are frequently deployed as risk stratification tools, prioritization engines, and compliance screening systems rather than as autonomous decision makers, emphasizing the human-in-the-loop structure that dominates regulated operations. Studies examining clinical prediction systems highlight how risk scoring can influence triage, resource allocation, and diagnostic escalation, while also revealing that model-driven prioritization can shift workload and decision patterns in ways that must be empirically measured. In financial decision support, algorithmic scoring systems shape creditworthiness evaluations, fraud detection pipelines, and regulatory monitoring processes, where documentation and auditability are fundamental (Trunk et al., 2020). Public-sector systems similarly apply predictive models to guide eligibility screening, service targeting, and compliance enforcement, and the literature indicates that transparency requirements become especially important when decisions affect rights, access, or legal outcomes. Across these domains, the quantitative foundation of decision support involves decision thresholding, which determines how model scores are converted into actions, and this conversion is frequently governed by policy rather than statistical optimality. In regulated settings, decision thresholds are typically tied to tolerable risk levels, safety margins, or compliance rules, making evaluation dependent on decision outcomes rather than raw prediction accuracy (Settembre-Blundo et al., 2021). Researchers have also emphasized the relevance of utility-based evaluation in which the cost of false positives and false negatives is not symmetric, meaning that an error type can carry disproportionate harm or regulatory consequence. This has led to widespread use of cost-sensitive learning approaches and policy-aligned performance analysis, particularly in clinical and financial risk modeling. The literature therefore conceptualizes regulated decision support as a measurable system in which ML outputs, thresholding rules, and audit requirements jointly determine whether a model is operationally valid.

A central finding across quantitative studies is that accuracy alone is insufficient for evaluating machine learning models in regulated contexts, because accuracy fails to capture critical dimensions of reliability, safety, and decision appropriateness (Kuziemski & Misuraca, 2020). Research in medical AI

has repeatedly shown that models may achieve strong discrimination while still producing poorly calibrated risk estimates, meaning predicted probabilities do not align with real-world event rates. This is significant in regulated decision support because decision-makers often interpret risk scores as actionable probability estimates that justify interventions. For example, in healthcare, mis calibrated predictions can lead to inappropriate escalation or under-treatment, which can create safety risks and compliance concerns. Similarly, in financial decision support, miscalibration may distort risk-based pricing, default probability estimation, or fraud alert thresholds, potentially leading to inequitable outcomes and regulatory exposure. The literature has therefore emphasized calibration as a regulatory-relevant property, because it supports consistent interpretation of risk across populations and operational contexts. Studies in applied ML also show that models may appear accurate in development environments but degrade substantially when deployed across different institutions, geographic regions, or time periods (Daniel & Daniel, 2018). This has motivated the use of stability and reproducibility measures, including performance variance across cross-validation folds, performance drift across time windows, and sensitivity to sampling changes. The evidence indicates that high-stakes systems must be evaluated for consistency, not only peak performance, because regulated decision-making requires dependable outputs under operational variability. A major theme in the literature is external validity, which refers to the ability of a model to generalize across settings, and studies have shown that cross-institution evaluation often reveals domain shift problems that are not visible in single-site validation. Domain shift occurs when data distributions differ due to population differences, measurement practices, coding standards, or operational workflows. In regulated systems, domain shift is particularly important because the model may be used across multiple sites, vendors, or agencies, and performance degradation can produce systemic risks. Research also highlights that stability and generalization are closely connected to fairness concerns, because performance instability often affects subgroups differently (Leicht-Deobald et al., 2022). This means that evaluation must incorporate not only aggregate metrics but also stratified performance analysis across demographic and operational categories. Overall, the literature positions regulated ML decision support as a measurement problem in which calibration, stability, and external validity are essential complements to accuracy, because they align model behavior with the evidentiary and accountability expectations of regulated environments.

Figure 3: Explainable Secure Machine Learning Framework



Quantitative performance evaluation in regulated decision support typically relies on three interrelated metric families: discrimination metrics, calibration metrics, and decision utility metrics. Discrimination metrics measure how well a model separates positive from negative outcomes, supporting ranking quality and triage performance in decision pipelines. Such metrics are widely used in healthcare prediction studies, including readmission risk, mortality prediction, and diagnostic classification, and

also in financial modeling such as default prediction and fraud detection (Hoegen et al., 2018). However, the literature indicates that discrimination metrics are often insufficient to support policy decisions because they do not capture whether predicted probabilities are meaningful or whether performance is consistent across subgroups. Calibration metrics address this gap by measuring whether predicted risks correspond to observed outcomes, which is critical when models are used to justify interventions, allocate resources, or determine compliance actions. Calibration evaluation is particularly emphasized in medical AI research where decision thresholds may correspond to clinical guidelines, and in finance where risk estimates can influence pricing and regulatory reporting. Beyond discrimination and calibration, decision utility metrics are increasingly central in regulated decision support because they incorporate the consequences of decisions into evaluation. In this approach, performance is assessed based on the expected costs or benefits associated with different actions, reflecting real-world trade-offs that are often mandated by policy (Sanders & Crozier, 2018). In healthcare, decision utility frameworks have been used to evaluate whether models improve triage efficiency, reduce unnecessary interventions, or support earlier detection. In financial services, similar frameworks quantify the trade-off between false alarms and missed fraud, or between credit denial errors and default losses. Decision curve analysis and net benefit approaches have become common in clinical literature to evaluate whether using a model provides measurable benefit relative to standard practice, and these approaches translate naturally to regulated settings where decision-making is governed by policy thresholds. Another theme in the literature is that metric selection must match operational use, because a model used for ranking may prioritize discrimination, while a model used for risk communication requires calibration (Čartolovni et al., 2022). Researchers have also emphasized that performance metrics must be reported with uncertainty estimates and reproducibility checks, because regulated environments require defensible evidence rather than single-point performance claims. As a result, the quantitative literature supports a multi-metric evaluation structure that combines discrimination, calibration, and decision utility to reflect the practical requirements of regulated decision support.

The literature also emphasizes that regulated decision support must be evaluated as an end-to-end system rather than as a standalone predictive model, because operational validity depends on how outputs are thresholder, interpreted, and acted upon. Studies across high-stakes ML domains show that decision thresholds are rarely chosen to maximize global accuracy and are instead selected based on institutional risk tolerance, safety margins, and regulatory expectations. This means that evaluation must quantify how different thresholds affect error rates, resource burden, and subgroup impacts, since a small threshold shift can dramatically change who receives interventions or scrutiny (Bateman & Mace, 2020). Researchers have argued that policy-aligned thresholding requires cost-sensitive analysis, where different error types are weighted according to their consequences. In regulated systems, false positives may trigger unnecessary investigations, interventions, or service denial, while false negatives may permit harm, fraud, or safety failures. The literature demonstrates that these trade-offs are not merely technical, because they directly reflect governance choices that must be documented and justified. Another critical dimension is auditability, which requires that performance evidence be reproducible across time and evaluators. As a result, researchers increasingly recommend validation designs that include cross-site testing, temporal splits, and repeated evaluation cycles to quantify stability. The literature also shows that model evaluation must incorporate operational context, including differences in data collection practices, coding standards, and workflow integration (Feng et al., 2021). Cross-institution studies reveal that models trained in one environment may not transfer effectively to another, even within the same sector, and this has led to emphasis on external validation and transportability analysis. Furthermore, research indicates that performance instability can be amplified when models interact with human decision-makers, because users may change behavior in response to model outputs, altering the data distribution over time. This reinforces the importance of monitoring and reproducibility measures. Overall, the literature establishes that conceptual and quantitative foundations for regulated decision support require multi-dimensional evaluation: models must be assessed for discrimination and ranking ability, calibrated risk estimation, stability across time and settings, and decision utility under policy thresholds (Parish et al., 2020). This foundation provides

the basis for subsequent literature on explain ability and security, because interpretability and robustness must ultimately be evaluated within the same regulated decision-support framework rather than as isolated technical enhancements.

Data Governance in Regulated ML

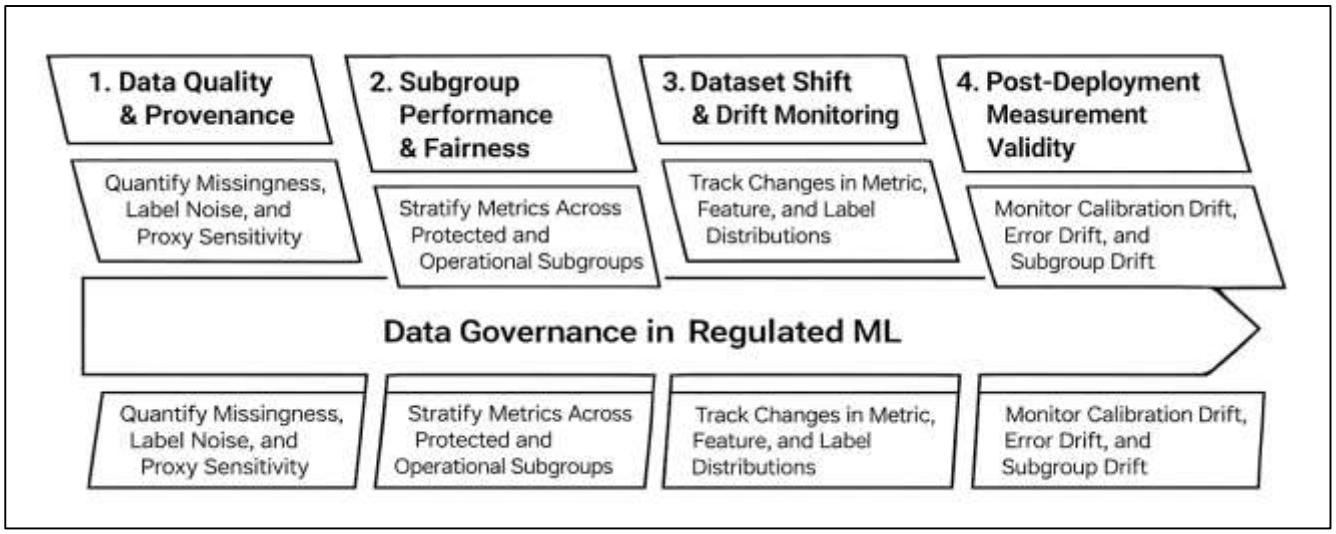
Data governance in regulated machine learning is treated in the literature as a quantitative risk-control function because the quality, structure, and traceability of data directly determine whether model outputs are reliable enough for audited decision support. Studies across healthcare, finance, and public-sector analytics consistently show that missingness is not merely a nuisance but a measurable signal of process variation, access inequities, and workflow artifacts that can systematically bias estimates (Benfeldt et al., 2020). Missingness patterns can be monotone, intermittent, or feature-dependent, and each pattern influences error profiles differently when imputation or case-wise deletion is applied. Empirical research also highlights that label noise is a frequent and under-measured driver of model risk in regulated settings because outcomes are often derived from administrative coding, claims submissions, clinician documentation, or compliance determinations rather than direct measurement. Label noise rates can vary by institution, time period, and subgroup, producing inflated performance during development and unstable generalization at deployment. A related concern is the use of proxy variables, where features correlated with protected characteristics or sensitive attributes indirectly shape predictions in ways that are hard to detect through aggregate metrics (Müller et al., 2022). The literature emphasizes that proxies emerge naturally in regulated data environments because operational variables capture access, location, billing practices, or service utilization patterns, which are structured by policy and resource availability. Quantitative governance approaches therefore prioritize data provenance and documentation as measurable audit artifacts: logging how data were collected, transformed, and linked; recording variable definitions and code sets; tracking inclusion and exclusion rules; and maintaining versioned datasets and feature engineering pipelines. Research on model documentation and auditability underscores that reproducibility is itself a quantitative property of governance, because models cannot be reliably validated or challenged if the underlying data lineage cannot be reconstructed (Van De Sande et al., 2021). As a result, the literature frames regulated ML data governance as an end-to-end measurement discipline, where risk is reduced by quantifying missingness, label uncertainty, and proxy sensitivity, while strengthening audit readiness through systematic provenance records and transparent variable documentation.

Subgroup performance and fairness measurement form a second pillar of regulated ML validity because regulated decisions must be defensible across protected populations and operationally relevant groups. The literature consistently reports that aggregate performance can conceal large disparities, particularly when class prevalence differs between groups or when measurement processes vary by setting (de Hond et al., 2022). For this reason, studies recommend performance stratification across demographic categories as well as across operational segments such as geography, facility type, payer category, device type, and workflow pathway. Quantitative subgroup evaluation typically extends beyond one metric, because different fairness indicators capture distinct forms of disparity in error allocation and decision impact. Some studies focus on differences in error rates, examining how false positive and false negative patterns differ across groups when a single threshold is applied. Other work emphasizes parity in positive classification rates, which can be important in screening and eligibility contexts where resource allocation is triggered by model outputs. Another stream evaluates whether predicted risk scores mean the same thing across groups by assessing calibration within groups, since calibrated scores support consistent interpretation and can reduce the risk of systematically overestimating or underestimating risk for specific populations (Weissler et al., 2021). Research also shows that fairness measurement is tightly coupled to decision thresholds and base rates, so fairness cannot be inferred from model architecture alone. Small threshold changes can shift subgroup impacts substantially, and studies therefore recommend reporting subgroup performance across a range of thresholds aligned to policy and operational constraints. The literature additionally highlights that fairness indicators can conflict, so rigorous assessment involves comparing multiple indicators and explicitly connecting them to the decision context. In regulated systems, fairness auditing is also treated as a monitoring requirement rather than a one-time test because changes in population composition, service access, and coding practices can alter subgroup behavior even when

the model remains unchanged (Wang et al., 2022). Consequently, quantitative fairness work emphasizes repeatable subgroup dashboards, uncertainty intervals for subgroup metrics, and clear definitions of protected and operational groups that reflect both legal requirements and real-world workflow segmentation.

Dataset shift and drift monitoring are repeatedly identified in the literature as core operational risks in regulated ML because decision support systems operate in environments where measurement processes, population characteristics, and institutional policies evolve. Studies across applied ML demonstrate that models validated on a static dataset can degrade after deployment when the distribution of inputs changes or when the relationship between inputs and outcomes shifts (Alhassan et al., 2019). In regulated settings, such shifts are common because guidelines change, coding standards are updated, new devices or software systems are introduced, and service patterns respond to policy incentives. The literature describes drift as multifaceted: feature drift occurs when input distributions move; label drift occurs when outcome prevalence changes; and concept drift occurs when the mapping from features to outcomes changes due to new practices or interventions. Quantitative drift detection methods are used to measure these changes by comparing current data to a baseline reference window, often at regular intervals (Maleki et al., 2020). Research also emphasizes the need for stratified drift detection because distribution changes can be localized to specific facilities, regions, or subgroups, creating a hidden failure mode when global monitoring appears stable. Drift detection is not treated as a purely statistical exercise; it is linked to governance and auditability because monitoring alerts require documented investigation procedures, thresholds for action, and a traceable record of responses. Empirical work highlights that drift signals can also reflect beneficial changes, such as improved clinical processes or reduced fraud, so monitoring programs must interpret drift within operational context rather than treating all change as degradation. The literature therefore recommends multi-signal monitoring that tracks shift in input distributions, changes in predicted score distributions, and changes in outcome-linked performance indicators, ensuring that monitoring remains tied to decision support validity (Q. Zhou et al., 2021). In regulated environments, drift monitoring is also connected to change management because model updates, feature changes, and data pipeline modifications can produce artificial drift, which must be distinguished from real-world population changes through careful versioning and provenance control.

Figure 4: Regulated Machine Learning Data Governance



Post-deployment measurement validity is increasingly framed in the literature as a continuous assurance problem that integrates data governance, subgroup fairness, and drift monitoring into a single operational evaluation layer. Studies show that monitoring only aggregate accuracy-like statistics is insufficient because it can miss calibration deterioration, changes in error composition, and subgroup-specific failures that emerge under evolving conditions (Liu & Panagiotakos, 2022). As a result, post-deployment monitoring signals described in the literature include calibration drift, where

predicted risk no longer aligns with observed outcomes; error-rate drift, where sensitivity and specificity change in ways that alter decision consequences; and subgroup drift, where disparities widen or shift even when overall performance appears stable. Quantitative monitoring programs often combine statistical change detection with operational key performance indicators tied to workflow outcomes, such as investigation burden, escalation rates, or resource utilization triggered by model outputs. The literature also emphasizes the importance of governance controls that define when a drift signal requires human review, when thresholds should be recalibrated, and when model retraining or rollback is warranted, all while maintaining traceable documentation for auditors (Varona & Suárez, 2022). Research on regulated ML operations stresses that monitoring must include data integrity checks, such as changes in missingness rates, unusual feature value frequencies, unexpected category growth, and pipeline break indicators, because these issues can create silent failures. Subgroup monitoring is treated as a priority because changes in service access, demographic composition, or institutional practices can alter both prevalence and measurement quality across groups, affecting fairness indicators and calibration within groups. The literature further emphasizes that measurement validity depends on consistent outcome definitions and stable label generation processes, so governance teams often monitor label latency, coding updates, and documentation shifts as part of the same assurance system (Ma et al., 2020). Overall, the research base positions data governance, bias measurement, and drift monitoring as quantitative foundations for regulated ML validity, requiring systematic measurement of data quality risks, transparent provenance documentation, stratified fairness evaluation, and multi-signal post-deployment monitoring that preserves auditability and decision defensibility.

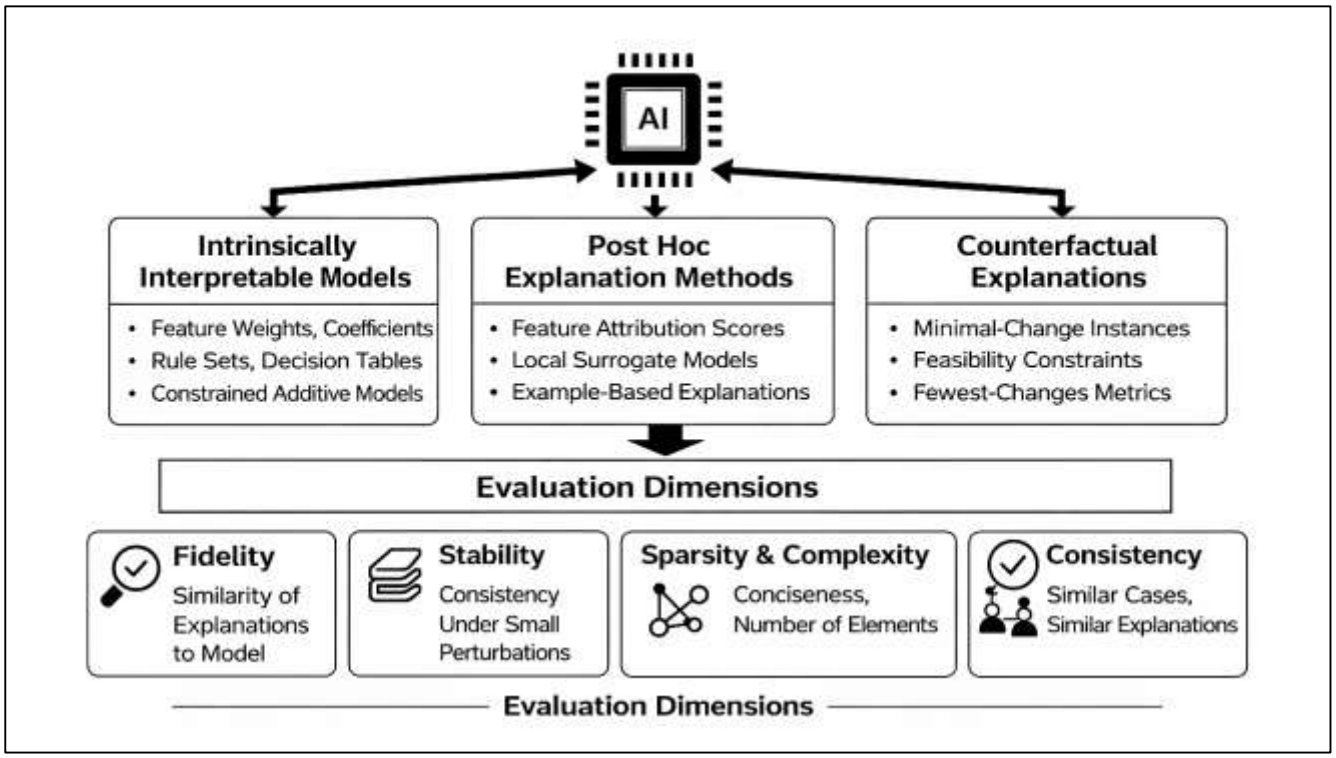
Explainable Machine Learning

Explainable machine learning is widely treated as a quantitative assurance component in high-stakes decision environments because it transforms opaque prediction mechanisms into measurable interpretive artifacts that can be evaluated for reliability, traceability, and audit readiness. The literature typically distinguishes three major categories of explain ability that map directly to different kinds of measurable outputs (Jia et al., 2022). Intrinsically interpretable models generate transparency through their structure, producing outputs such as coefficients, explicit feature weights, rule lists, decision sets, scoring tables, or constrained additive components that can be inspected without additional explanation layers. This category is often linked to regulated decision support because transparency is embedded by design, enabling analysts to document which variables drive predictions and how decision logic is formed. Post hoc explain ability methods, in contrast, create measurable explanation artifacts after training a complex model. These artifacts include feature attribution scores, local surrogate decision approximations, or example-based explanation summaries that attempt to approximate the behavior of a black-box model in a human-interpretable format. A third category, counterfactual explanations, produces minimal-change instances that represent how input conditions could be altered to change a model decision, yielding measurable outputs such as the number of changes required, the magnitude of changes, and the feasibility constraints that define realistic counterfactuals (Asaadi, Denney, et al., 2020). Across these categories, the literature emphasizes that explain ability is not a single "look and feel" property but a set of measurable system behaviors that can be tested in the same way as predictive performance. Explanations become quantitative objects: coefficients can be analyzed for sign consistency and magnitude stability; rules can be measured for length, coverage, and contradiction rates; attributions can be measured for ranking stability and sensitivity; and counterfactuals can be measured for proximity and feasibility. In regulated contexts, this measurement orientation is central because explanations support documentation requirements, audit trails, and structured review by oversight bodies. As a result, explainable machine learning is framed as part of the evidence package that supports defensible use of models, where the explanation outputs must be reproducible and systematically evaluated across repeated tests, different datasets, and operational segments (Asaadi et al., 2019). This framing positions explain ability as an assurance layer that complements, rather than replaces, predictive validation by contributing measurable transparency and supporting risk controls associated with accountability.

Quantitative evaluation of explanation quality is consistently presented in the literature as necessary because explanation artifacts can appear plausible while being inaccurate, unstable, or inconsistent

with the underlying model. One core dimension is fidelity, which refers to how closely an explanation reflects the actual behavior of the predictive model it describes. Fidelity can be assessed by comparing the explanation's implied decision logic or feature contributions with the model's actual responses across representative inputs (J. Zhou et al., 2021). A second major dimension is stability, which reflects whether explanations remain similar when inputs are minimally perturbed, when resampling is applied, or when equivalent model training runs are repeated. Stability matters in decision support because explanations that change dramatically for nearly identical cases can reduce trust and hinder auditability. A third dimension is sparsity and complexity, which captures how concise an explanation is and how many elements it uses to express reasoning. For interpretable models, complexity may be expressed through the number of terms or rules; for post hoc explanations, complexity may be expressed through how many features are highlighted or how dense a surrogate decision summary becomes. The literature commonly notes that more concise explanations are easier to interpret, yet overly compressed explanations may omit important drivers of the model, which means clarity and fidelity can trade off in measurable ways. A fourth dimension is consistency, typically evaluated by examining whether similar cases produce similar explanations (Studer et al., 2021). Consistency is crucial for regulated decision support because comparable individuals or events should not receive materially different rationales without defensible reasons. Quantitative approaches to consistency evaluate similarity in explanation rankings, similarity in identified rules, or similarity in counterfactual directions among cases that share relevant characteristics. Across the literature, these evaluation dimensions are treated as multi-metric rather than singular because strong performance in one dimension does not guarantee strong performance in another. For example, an explanation may be sparse but low-fidelity, or high-fidelity but unstable. This has led to systematic evaluation practices that compare explanation methods across datasets, across repeated runs, and across operational segments in order to quantify reliability (Roscher et al., 2020). In regulated settings, this multidimensional measurement approach is essential because explanations often become part of compliance evidence, meaning they must withstand scrutiny as repeatable artifacts rather than serving as illustrative narratives. Consequently, the literature treats explanation evaluation as a structured validation program that parallels and extends model performance validation.

Figure 5: Explainable Machine Learning Assurance Framework



Local explanation methods are particularly emphasized in the literature because regulated decision support frequently requires case-level justification, such as explaining why a specific alert was triggered or why a particular individual was assigned a high-risk score. However, extensive empirical work identifies measurable failure modes that can undermine local explanation reliability. A common failure mode involves brittleness under correlated features, where multiple variables encode similar information and the explanation method may arbitrarily attribute importance to one variable over another (Bücker et al., 2022). This leads to instability in feature rankings and can produce different rationales for nearly identical cases, even when the predicted output remains unchanged. Another failure mode is sensitivity to baseline choice, which occurs when an explanation depends on a reference condition used to measure feature contribution. Different baselines can lead to different importance distributions, producing inconsistency across studies and difficulty in standardizing explanations for audit review. Sampling neighborhoods also affects local surrogate explanations because these methods approximate the model in a local region around a target instance, and the composition of that region determines what the surrogate learns. If the neighborhood sampling includes unrealistic combinations of feature values or fails to reflect operational data constraints, the resulting explanation can misrepresent how the model behaves in real cases. Noise sensitivity is another measurable limitation, because small perturbations in inputs or slight measurement error can shift which features are highlighted, particularly in high-dimensional spaces or when models rely on subtle patterns (Klauschen et al., 2018). In response to these issues, the literature emphasizes quantitative sanity checks that test whether explanations behave as expected when model structure is disrupted. For example, when model parameters are randomized or labels are permuted, explanation outputs should degrade in meaningful ways; if explanations remain similar under such disruptions, this suggests the method may not reflect learned relationships. These sanity checks have strengthened evaluation rigor by providing measurable diagnostics for explanation validity. The literature also stresses that local explanations should be assessed for their ability to detect spurious correlations, proxy effects, and brittle dependencies, especially in regulated data where administrative patterns and access-related variables can dominate predictions (Kailkhura et al., 2019). Overall, local explain ability is treated as a high-value capability for accountability, yet one that requires systematic measurement, stress testing, and validation against known failure modes to ensure explanations reflect actual model behavior rather than presenting persuasive but unreliable rationales.

Human-in-the-Loop Decision Support

Human-in-the-loop decision support is described in the literature as a joint system in which human judgment and machine learning outputs combine to determine decision quality, particularly in regulated environments where accountability cannot be fully delegated to automation. Within this perspective, explanations are evaluated not as decorative transparency features but as measurable interventions that can change how people decide, how quickly they decide, and how accurately they decide under policy constraints (Farhood et al., 2021). A common quantitative approach compares the accuracy of human decisions without algorithmic assistance to the accuracy of decisions made with model outputs and explanations available. This comparison is often conducted at both the aggregate level and the case level, allowing researchers to estimate whether explanations improve decision correctness when models are accurate and whether they help humans detect or correct model mistakes when models fail. The literature also places strong emphasis on operational decision metrics such as time-to-decision and intervention rates because regulated workflows are resource-constrained, and explanation interfaces can either reduce cognitive effort or increase deliberation time depending on their design and complexity (Walsh & Feigh, 2021). Override rates are frequently analyzed as a behavioral signal indicating whether users appropriately accept or reject algorithmic guidance; high override may reflect distrust or confusion, while low override may reflect automation bias when users defer to model outputs even under uncertainty. In high-stakes screening tasks, the frequency of interventions triggered by model outputs is measured to evaluate whether explanations support more selective, policy-aligned actions rather than indiscriminate escalation. Another key construct is confidence calibration, which measures whether users' confidence levels align with their correctness after viewing explanations. Calibration is important because overconfident users may act on incorrect model-supported decisions, and underconfident users may ignore correct recommendations. The

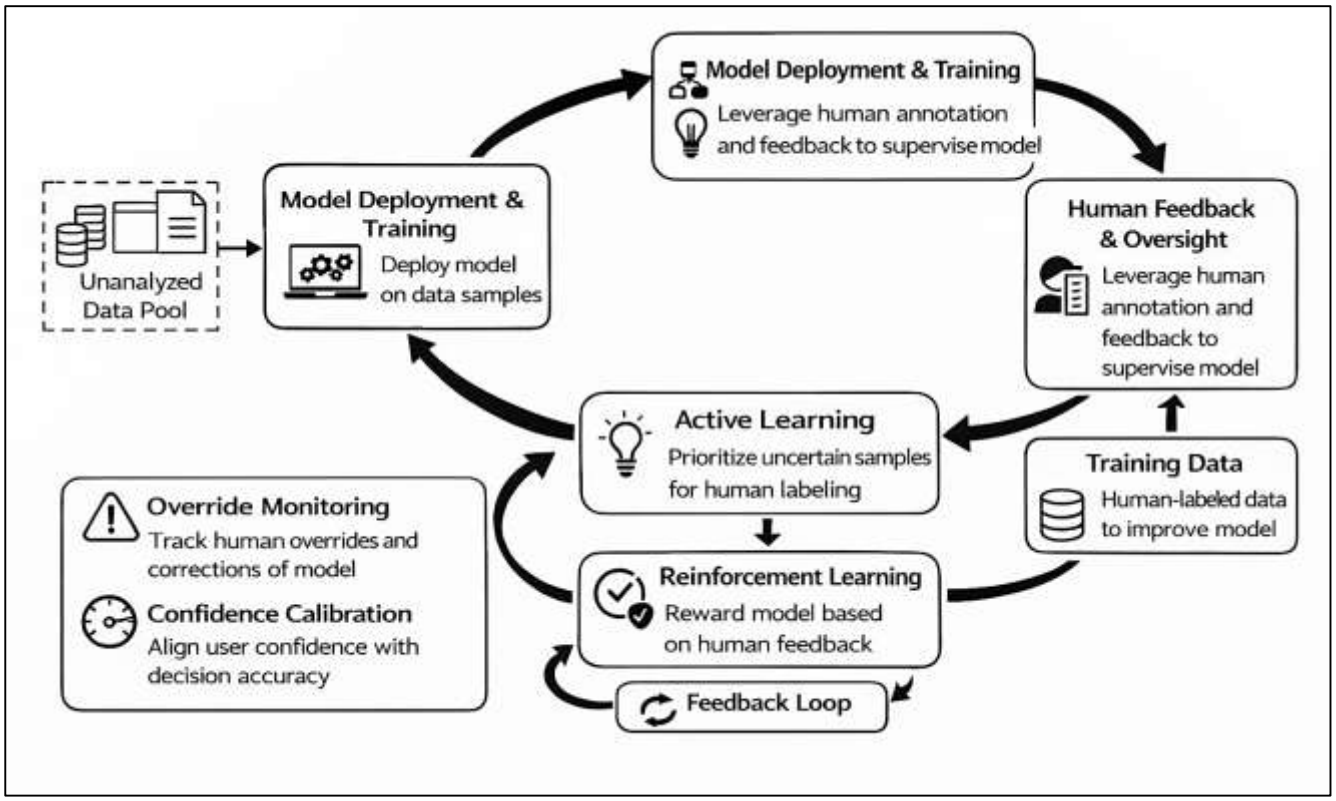
literature indicates that explanations can shift confidence and reliance in complex ways, making calibration a central outcome variable in human–AI evaluation (Estivill-Castro et al., 2022b). In regulated decision support, these metrics are often interpreted through the lens of accountability: explanation interfaces are valuable when they improve accuracy, maintain manageable decision time, support appropriate override behavior, and calibrate confidence so that users neither blindly trust nor reflexively reject the system. This has led to the view that decision quality must be measured as a combination of correctness, efficiency, and reliance behavior rather than as a single accuracy statistic.

Experimental designs in explain ability research commonly treat explanation type and presentation format as independent variables that can be manipulated to observe changes in human decision behavior. Randomized controlled experiments are frequently used to compare explanation conditions such as no explanation, simple feature highlights, structured rationales, example-based explanations, or counterfactual-style information. Participants are typically assigned to conditions, and their performance is evaluated across standardized decision tasks that mirror real-world classification, ranking, or triage decisions (Thrun, 2022). This design supports causal claims about how explanation formats affect accuracy, time, and reliance patterns. A/B testing is another widely used design, often implemented in simulated workflows or controlled interface environments to compare alternative explanation layouts, levels of detail, or uncertainty displays. Researchers measure outcomes such as decision accuracy, response latency, acceptance and override behavior, and interaction patterns with the explanation interface. Because human decision-making varies across individuals and contexts, explain ability studies commonly apply statistical models that can account for repeated measures and clustered data, such as mixed-effects approaches, as well as models that examine relationships between explanation exposure and reliance behaviors. Researchers also use analysis-of-variance styles of comparisons when conditions are balanced and tasks are standardized, and regression-based approaches are common for estimating how explanation features interact with user expertise, task difficulty, or model correctness. The literature emphasizes that human–AI evaluation requires careful control for confounding factors such as learning effects, fatigue, and ordering effects, which can influence decision behavior across trials (Zeng et al., 2021). Many studies therefore randomize case order, balance difficulty across conditions, and include attention checks or comprehension checks to ensure that outcomes reflect explanation influence rather than participant disengagement. Another theme is that explanation effects may differ based on whether the model is right or wrong, leading researchers to analyze results separately for correct and incorrect model outputs. This decomposition enables measurement of whether explanations help users detect model errors or merely increase trust in the system. In regulated settings, experimental rigor is especially important because organizations need evidence that explanation interfaces improve decision quality in reproducible ways (Treiss et al., 2020). The literature therefore frames experimental design and statistical modeling as central tools for quantifying the causal utility of explain ability in decision support rather than relying on subjective impressions of transparency.

Cognitive load and usability measurement are widely treated as essential in explain ability research because explanations can increase understanding while simultaneously increasing mental effort, which can undermine operational feasibility in time-sensitive or high-volume regulated workflows. The literature often quantifies cognitive load using standardized workload instruments that capture perceived mental demand, effort, frustration, and task complexity, alongside behavioral proxies such as response latency and interaction time (Zhao et al., 2021). Response latency is used as a practical indicator of effort, especially in decision pipelines where throughput and timeliness are critical. Researchers also employ comprehension assessments, where participants answer questions about why the model predicted a certain outcome, what factors were most influential, or what changes might alter the result. These tests provide quantitative evidence of whether explanations actually increase understanding rather than merely providing a sense of transparency. Usability is commonly assessed through structured questionnaires and performance-based indicators, including error detection rates, the ability to identify irrelevant features, and the ability to recognize uncertainty. The literature shows that explanation interfaces that are overly detailed may increase cognitive load without improving accuracy, particularly for non-expert users. Conversely, explanations that are too simplified may

reduce comprehension in complex cases, encouraging shallow reliance rather than informed oversight. These findings support the use of multi-dimensional evaluation where cognitive load and usability are measured alongside decision accuracy and override behavior (Enarsson et al., 2022). Another theme is that cognitive load effects vary by user expertise: domain experts may benefit from richer detail, while novices may be overwhelmed by complex attribution summaries. Explain ability studies therefore often segment participants by experience and measure interaction effects, such as whether detailed explanations improve expert performance while decreasing novice performance. In regulated systems, these distinctions matter because user populations include both highly trained professionals and administrative staff with different roles and training. The literature also emphasizes that usability must be evaluated under realistic constraints, including time pressure and information overload, because explanation performance in relaxed laboratory conditions may not translate to operational environments (Estivill-Castro et al., 2022a). Overall, cognitive load metrics, response time measures, and comprehension tests provide a quantitative basis for determining whether explain ability improves decision support in practice, not only in theory, by balancing interpretive clarity against mental workload.

Figure 6: Human-Centred Machine Learning Workflow



The literature also evaluates explanation preference versus performance trade-offs, demonstrating that what users like is not always what improves decision quality. Preference is commonly measured through surveys, ranking tasks, or forced-choice comparisons among explanation formats, while performance is measured through accuracy, calibration of confidence, and detection of model errors. Many studies find that users may prefer explanations that are simpler, more narrative, or more visually intuitive, even when those explanations do not increase correctness or do not help users recognize when the model is wrong (Zahabi & Kaber, 2019). This disconnect has led researchers to emphasize the importance of measuring both subjective satisfaction and objective performance, treating preference as a separate outcome rather than a proxy for utility. Another documented pattern is that explanations can increase trust without increasing accuracy, which can be operationally risky in regulated decision support because inflated trust can reduce vigilance. Quantitative analysis of reliance often includes metrics such as acceptance rates, override frequencies, and changes in decision thresholds when explanations are presented. These measures help distinguish between explanations that improve

informed decision-making and explanations that simply increase compliance with algorithmic recommendations. Researchers also assess whether explanations help users calibrate their trust, meaning that users rely more when the model is likely correct and rely less when the model is likely incorrect (Lee et al., 2022). When explanations improve this selective reliance pattern, they are considered more effective for human-in-the-loop governance. Performance trade-offs are also measured across task types, with evidence that explanation benefits may be stronger in some tasks than others, such as anomaly detection versus routine classification. In regulated workflows, these task differences are important because decision support includes both high-frequency routine decisions and low-frequency high-impact decisions. The literature suggests that explanation evaluation should therefore include multiple task scenarios and measure how explanation design interacts with risk level, complexity, and time constraints. By quantifying preference, cognitive load, decision accuracy, time-to-decision, override behavior, and confidence calibration together, the literature establishes a comprehensive approach to evaluating the utility of explanations in human-in-the-loop decision support (Grønsund & Aanestad, 2020). This approach reflects the central argument that explain ability is valuable when it measurably improves decision outcomes and oversight behaviors within the constraints of regulated operational environments.

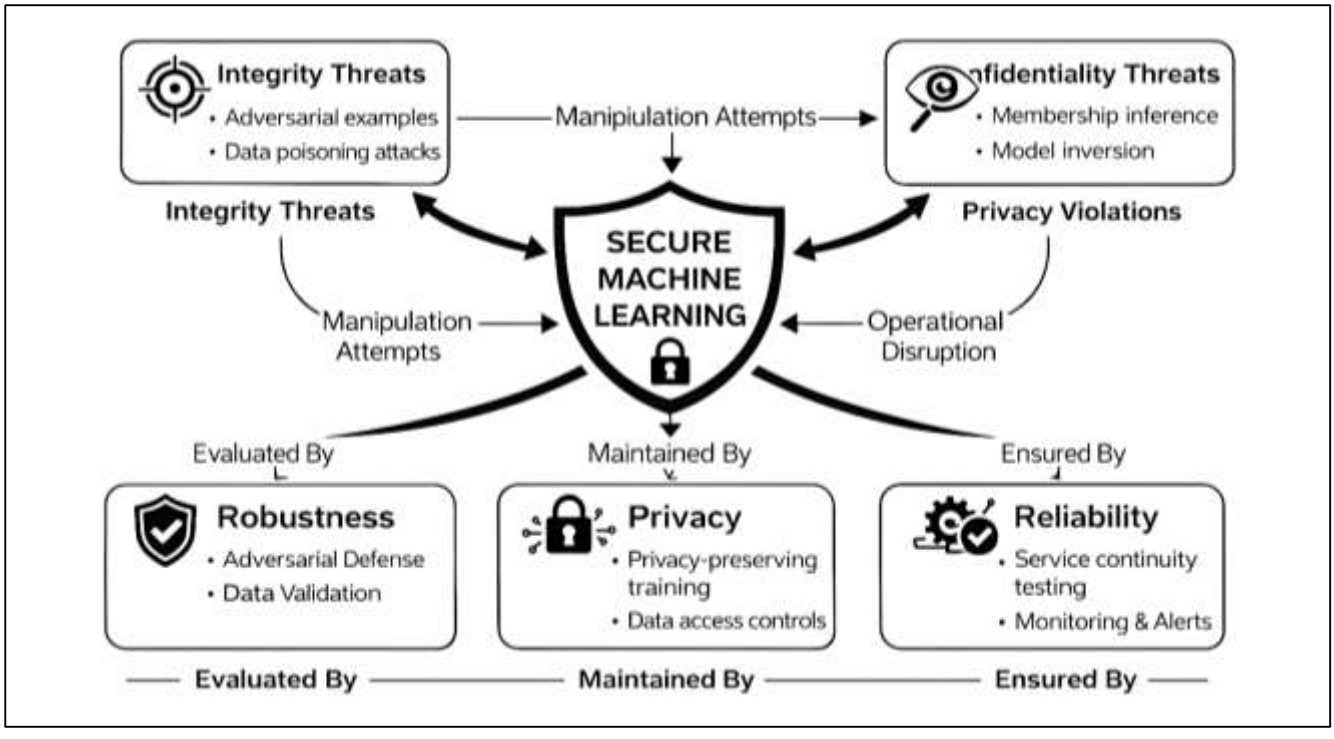
Secure Machine Learning for Regulated Decision Support

Secure machine learning for regulated decision support is treated in the literature as a system-level assurance requirement because high-stakes workflows rely on model outputs that must remain trustworthy under deliberate attack, accidental corruption, and operational disruption. Research typically frames security through threat modeling that separates integrity, confidentiality, and availability risks, each tied to measurable failure outcomes. Integrity threats involve attempts to manipulate model behavior so that predictions become wrong in targeted or broad ways (Bettaieb et al., 2019). This includes adversarial examples that alter inputs in subtle or constrained manners to trigger incorrect classifications, rankings, or alerts, as well as data poisoning attacks that contaminate training data to implant systematic errors. In regulated systems, integrity threats are viewed as especially salient because even small shifts in decision outputs can translate into clinical harm, compliance failures, wrongful denials, or missed detection events. Confidentiality threats are treated as violations of privacy and protected data obligations, where adversaries infer sensitive information from the model or its outputs. Membership inference attacks attempt to determine whether a particular individual's record was part of training data, while model inversion attempts to reconstruct sensitive attributes or representative training examples from model behavior. The literature emphasizes that confidentiality threats are not limited to raw data leakage; they also arise from overly informative prediction interfaces, repeated querying, and exposure of explanation artifacts that reveal decision boundaries (Bettaieb et al., 2020). Availability threats focus on disruption of the decision pipeline, including denial-of-service against inference endpoints, resource exhaustion through query flooding, and degradation of model serving infrastructure. In regulated settings, availability is treated as a safety and continuity issue because workflows depend on timely scoring, alerting, and triage decisions. Threat modeling in this literature is quantitative in its orientation because it requires explicit assumptions about attacker knowledge, attacker capabilities, access pathways, query limits, and attack objectives. These assumptions then map to measurable outcomes such as attack success rates, degradation in decision accuracy under attack, leakage indicators for privacy compromise, and service-level reliability measures for availability (Imran et al., 2022). Across domains, the literature positions threat modeling as the foundation that links security risks to evaluation protocols, governance documentation, and audit-ready reporting requirements.

Robustness evaluation is discussed as a central quantitative pillar for secure machine learning because it measures how well decision-support models maintain acceptable performance when inputs or operational conditions are manipulated or perturbed. In the literature, robustness is typically assessed under controlled perturbation regimes that reflect plausible attacker constraints or realistic measurement noise, enabling comparison across models and defense strategies (Papernot et al., 2018). A common theme is the distinction between average-case robustness, which reflects typical perturbations or naturally occurring noise, and worst-case robustness, which reflects the most damaging perturbations an attacker can craft under a defined constraint set. This distinction is critical

in regulated decision support because rare worst-case failures can carry disproportionate consequences compared to frequent minor degradations. Studies also differentiate empirical robustness from certified robustness. Empirical robustness is measured through attack simulations and stress tests that attempt to find successful manipulations; certified robustness refers to provable guarantees that the model's prediction remains unchanged within certain bounded regions of the input space. The literature treats certified approaches as attractive for high-assurance contexts, while also acknowledging that certifications often apply under restrictive assumptions and may not cover the full range of real-world manipulation strategies. Robustness benchmarks are frequently used to standardize evaluation, enabling consistent measurement of robustness across model families, training procedures, and defense mechanisms (Olowononi et al., 2020). In regulated settings, benchmarking is emphasized because oversight requires comparable evidence rather than ad hoc claims. Another recurring point is that robustness is not a single metric but a profile that varies by class, subgroup, feature region, and operational condition. For example, robustness can degrade disproportionately for rare categories or for subgroups with distinct measurement patterns, which is especially important in regulated systems where fairness and consistency are audited. The literature also highlights that robustness evaluation must be tied to decision thresholds and workflow consequences. A small robustness loss near a threshold can create a large change in action rates, such as unnecessary escalations or missed detections. As a result, secure decision support evaluation often extends beyond overall robustness summaries to include threshold sensitivity analysis, stratified robustness testing, and stability checks across temporal and site-specific data segments (Pugliese et al., 2021). This robustness-centered view supports a measurable assurance posture in which security claims are grounded in repeatable stress testing and clearly defined threat assumptions rather than informal assurances of safety.

Figure 7: Secure Machine Learning Assurance Framework

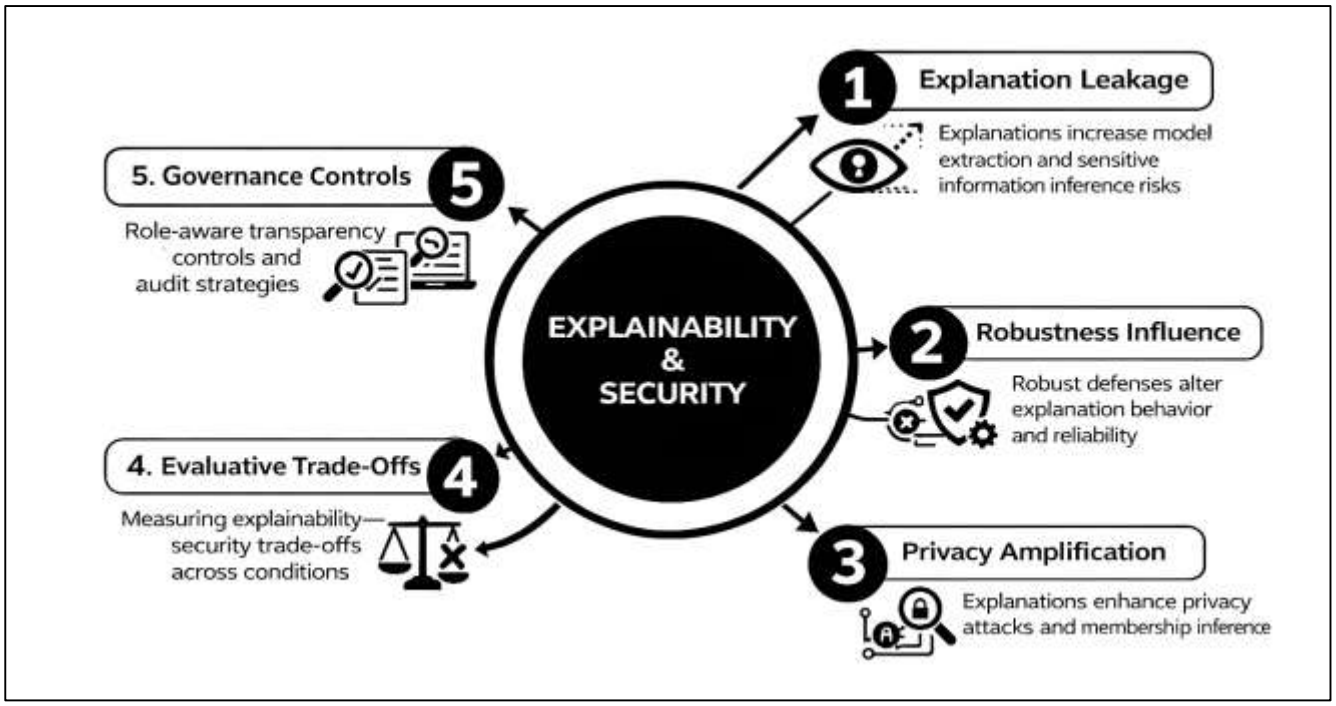


Interactions Between Explain ability

Interactions between explain ability and security are treated in the literature as a central assurance challenge in regulated machine learning because transparency mechanisms can simultaneously improve accountability and increase exposure to adversarial misuse. Explanation leakage risk is discussed as a measurable phenomenon where explanation outputs disclose information that helps an

attacker infer model structure, feature reliance, decision boundaries, or sensitive training characteristics (Vigano & Magazzeni, 2020). Research on model extraction and black-box exploitation emphasizes that repeated querying of a prediction interface can approximate the model, and explanation outputs can accelerate this process by revealing which variables matter most and how they influence outcomes. In regulated decision support, this risk is amplified because explanations are often made available to multiple stakeholders, potentially expanding the attack surface beyond tightly controlled technical teams. The literature describes information disclosure indicators that can be quantified by analyzing explanation outputs for their specificity, consistency, and granularity. For example, detailed local explanations that provide ranked feature contributions for individual cases can reveal which features dominate decisions in specific regions of the input space, enabling targeted perturbations. Counterfactual explanations, while valuable for interpretability, can disclose actionable directionality of decision changes, effectively providing an attacker with a blueprint for how to alter inputs to achieve a desired classification outcome (Roque & Damodaran, 2022). Global explanations can similarly expose model reliance patterns, allowing adversaries to focus on highly influential features. As a result, explain ability is framed not only as a user-facing interpretability feature but also as a potential side channel that can leak operational intelligence about the decision system. The literature therefore emphasizes measuring attack performance with and without explanation access, treating explanation availability as an independent variable in adversarial evaluation. Empirical studies compare attack success rates, query efficiency, and boundary estimation accuracy under different transparency conditions, showing that certain explanation interfaces can materially increase adversarial capability. In regulated systems, this raises a governance question that is treated quantitatively: how much explanatory detail is necessary for auditability and oversight, and how much increases exposure beyond acceptable risk thresholds (Klobas et al., 2019). This framing establishes explanation leakage as a measurable security dimension that must be evaluated alongside interpretability benefits.

Figure 8: Explain ability and Security Integration Framework



A second major theme in the literature is the impact of robustness defenses on explanation behavior, where security interventions change model properties in ways that alter the reliability, stability, and interpretive meaning of explanations. Adversarial training and other robustness-focused methods often modify decision boundaries, feature sensitivities, and gradient structures, which directly affects post hoc explanation methods that rely on local sensitivity patterns (Warnecke et al., 2020). Studies note that models trained for robustness can exhibit different feature reliance profiles compared to standard

models, which can shift attribution rankings and change the apparent rationale for predictions. This creates a measurable phenomenon of attribution drift, where explanation outputs change under defense configurations even when predictive performance remains comparable. The literature treats attribution drift as an important assurance problem because regulated systems depend on stable explanations for documentation and audit narratives. If the explanation for a given type of case changes substantially after a robustness intervention, oversight bodies may interpret the change as inconsistency or reduced transparency, even if the model is objectively more secure. Researchers therefore evaluate explanation stability before and after defense application using repeated sampling, perturbation tests, and similarity measures across explanation outputs (Kim et al., 2020). Another documented issue is that some defenses can increase explanation smoothness, potentially improving stability, while simultaneously reducing fidelity if explanations become less aligned with the exact decision logic of the defended model. In contrast, other defenses can introduce non-intuitive reliance patterns, causing explanations to highlight features that appear unrelated to the domain context. This is especially problematic in regulated settings where domain alignment is scrutinized. The literature therefore recommends joint testing protocols that measure robustness gains alongside explanation fidelity and stability under the same evaluation dataset, rather than treating explain ability evaluation and security evaluation as separate activities (Chen et al., 2018). This integrated approach quantifies whether the security improvements are achieved at the expense of explanation reliability, whether explanation artifacts remain consistent under perturbation, and whether interpretive summaries remain valid for governance and audit purposes. In this view, robustness defenses are not purely technical improvements; they alter the interpretability landscape and must be evaluated as interventions that change both risk and transparency properties.

The interaction between explain ability and privacy is also emphasized in the literature because explanation outputs can increase the risk of sensitive information inference even when direct access to training data is restricted. Privacy attacks often rely on subtle signals in model outputs, and explanations can enrich these signals by revealing how features contribute to predictions and how decision boundaries behave around specific inputs (Li et al., 2021). This creates measurable differences in inference attack performance when explanations are available, particularly in scenarios where attackers can issue repeated queries. In regulated decision support, privacy obligations are strict, and the literature treats privacy risk as an empirical property that depends on model capacity, output interfaces, and transparency mechanisms. Explanations can function as amplifiers of confidence information: feature attributions and counterfactual directions can reveal whether a model is strongly reliant on certain sensitive proxies or whether membership-specific patterns are present. As a result, researchers treat explanation design choices—such as the number of features shown, the level of numerical detail, and whether counterfactuals include actionable thresholds—as privacy-relevant configuration variables. The literature also connects these concerns to governance models in regulated systems, where different stakeholders require different levels of transparency (Zwilling et al., 2022). Auditors and technical evaluators may require deep explanation access for verification, while operational users' may need only limited interpretive cues. This role-based transparency concept is framed as a measurable risk-control strategy because limiting explanation exposure reduces the number of queries and the richness of information available to potential adversaries. Quantitative evaluation approaches compare privacy leakage indicators under different explanation policies, measuring how attack success changes when explanations are restricted, coarsened, or aggregated. The literature also recognizes that some interpretability approaches can be adapted to reduce leakage, such as using global summaries rather than individualized attributions, or providing categorical rather than numeric contributions. However, these adjustments create measurable trade-offs in interpretive usefulness that must be balanced against privacy protection requirements (Gui et al., 2020). In regulated contexts, the literature presents explanation–privacy interaction as a multi-dimensional evaluation issue where transparency is necessary for accountability but must be engineered to avoid creating an uncontrolled leakage pathway.

A central synthesis across the literature is that explain ability and security form a joint optimization problem in regulated assurance, requiring simultaneous measurement of predictive performance,

calibration, robustness, privacy risk, and explanation quality. Rather than optimizing one dimension independently, researchers increasingly frame evaluation as multi-objective, where improvements in one property can degrade another (Markus et al., 2021). For example, stronger robustness defenses can alter explanation behavior, and richer explanations can increase leakage and attack efficiency. This has led to the use of comparative evaluation across model families and control configurations, where each configuration is assessed across a shared set of metrics and then compared as a trade-off profile rather than a single score. Studies emphasize that regulated assurance demands transparent reporting of these trade-offs because governance bodies must justify why a chosen model configuration is acceptable given policy constraints. This approach treats model selection as a structured evidence exercise: performance metrics demonstrate predictive validity; calibration measures support interpretability of risk outputs; robustness measures support integrity under adversarial or noisy conditions; privacy indicators quantify confidentiality risk; and explanation fidelity and stability measures support auditability (Ismagilova et al., 2022). The literature also notes that trade-offs vary by context, such as differences between screening systems and diagnostic support systems, or between batch scoring and real-time inference pipelines. This variation reinforces the need for comprehensive measurement rather than single-metric optimization. Comparative frameworks are frequently described in terms of identifying configurations that are not dominated across objectives, meaning they offer balanced trade-offs relative to alternatives. Such comparisons support governance by clarifying which configurations provide strong robustness without unacceptable explanation drift, or which explanation interfaces provide interpretive value without materially increasing attack success (Minh et al., 2022). Overall, the literature frames the explain ability–security interaction as a quantifiable design space where regulated decision support demands integrated evaluation, role-aware transparency controls, and evidence-based selection of model and interface configurations grounded in measurable trade-offs rather than informal assumptions.

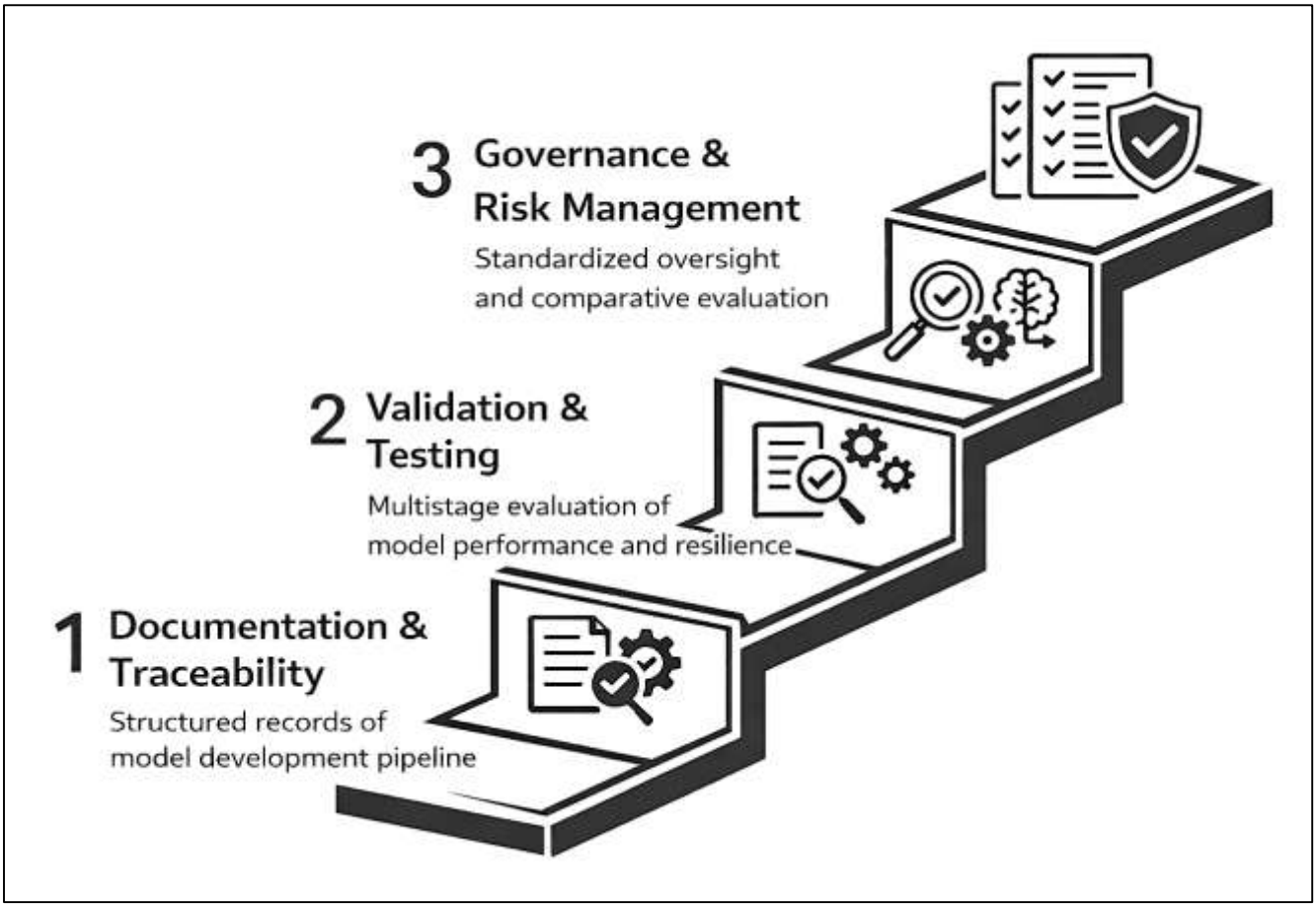
Integrated Assurance Frameworks

Integrated assurance frameworks for regulated machine learning decision support are described in the literature as structured governance systems that transform technical model development into auditable, measurable, and repeatable organizational processes (Asaadi, Beland, et al., 2020). In regulated sectors, the central concern is not simply whether a model produces accurate predictions, but whether the organization can demonstrate—through documented evidence—that the model is valid, controlled, and accountable across its lifecycle. For this reason, assurance frameworks emphasize audit-ready evidence packages, which consolidate model objectives, dataset characteristics, evaluation results, risk controls, and limitations into standardized reporting structures. Model cards and similar documentation templates are frequently discussed as a practical mechanism for turning technical validation into a measurable artifact. These documentation formats typically include explicit sections that require reporting of predictive performance, calibration quality, subgroup reliability, security testing results, and explain ability validation outcomes. The literature treats the completeness of these documents as a measurable governance indicator, because missing metrics or unclear evaluation procedures reduce auditability and increase institutional risk (Poth et al., 2020). Another major element in evidence packages is reproducibility, which is positioned as a foundational assurance requirement. Reproducibility indicators include clear recording of dataset versions, preprocessing steps, feature engineering procedures, training parameters, and evaluation configurations. This ensures that performance claims can be independently verified and that future audits can reconstruct the exact conditions under which a model was approved. The literature also emphasizes pipeline traceability as a critical component, meaning that the entire model development workflow—from raw data ingestion to model deployment—must be documented with version control and change logs. Traceability is presented not as an optional technical preference but as an assurance control that supports accountability when discrepancies occur, when complaints arise, or when regulatory review requires evidence of due diligence (Ward & Habli, 2020). Collectively, the literature frames audit-ready evidence packages as structured quantitative reporting systems that enable regulated organizations to treat machine learning models as governed decision instruments rather than experimental analytics.

Validation protocols aligned to regulated deployment represent a second core dimension of integrated assurance, and the literature repeatedly emphasizes that these protocols must reflect real-world

operational risk rather than laboratory-level performance. Pre-deployment validation is commonly described as a multi-stage process that includes internal testing, external validation, and stress testing under plausible failure conditions. External validation is treated as a critical quantitative step because regulated systems often deploy models across multiple sites, agencies, institutions, or geographic regions, where data distributions and workflow conditions differ (Mökander et al., 2021). The literature highlights that single-site validation can produce inflated performance estimates that do not generalize, which creates operational risk once the model is deployed in new environments. Stress testing is also discussed as a required assurance practice, particularly in regulated contexts where worst-case outcomes carry high consequences. Stress testing may include robustness evaluation against input perturbations, security testing against simulated attacks, subgroup fairness auditing, and calibration stability assessment under distribution shift. The literature frames these tests as measurable resilience evaluations, intended to quantify not only whether the model performs well in ideal conditions but whether it remains reliable under operational disturbances (Jarabek & Hines, 2019). Post-deployment validation is positioned as equally important because regulated environments are dynamic: policies change, measurement systems evolve, coding practices shift, and population characteristics fluctuate. The literature emphasizes continuous monitoring as a quantitative assurance mechanism, where organizations track changes in input distributions, prediction distributions, calibration behavior, error rates, and subgroup disparities over time. Periodic re-validation is discussed as a structured process in which models are reassessed at scheduled intervals or after significant operational changes, ensuring that performance remains aligned with the conditions under which the model was originally approved. This lifecycle-oriented validation structure is presented as a governance requirement, because regulated decision support cannot rely on one-time approval when model behavior can drift (Kovalchuk et al., 2022). The literature therefore conceptualizes regulated ML validation as an ongoing quantitative control program that integrates pre-deployment generalization testing, resilience stress testing, and post-deployment monitoring into a single assurance pipeline.

Figure 9: Integrated Machine Learning Assurance Framework



Comparative evaluation frameworks for regulated procurement are also emphasized in the literature because many regulated organizations acquire machine learning systems from vendors or must choose among competing internal model designs. In such cases, assurance is not only about validating a single model but about selecting the most defensible option among alternatives under compliance constraints. The literature describes the use of standardized scorecards as a method for comparing models using a shared set of measurable criteria (Amann et al., 2020). These scorecards typically combine predictive performance metrics with explain ability quality measures and security robustness indicators, enabling decision makers to evaluate trade-offs across multiple assurance dimensions. This approach reflects the recognition that the "best" model in regulated decision support is not necessarily the one with the highest discrimination performance, but the one that offers the most balanced profile of accuracy, calibration, fairness stability, interpretability reliability, privacy protection, and adversarial resilience. The literature also discusses weighted scoring systems as a governance mechanism for aligning model selection with institutional risk tiers and policy thresholds. For example, a model used for safety-critical clinical triage may require stronger calibration and robustness weighting, while a model used for administrative prioritization may prioritize throughput and interpretability (Valente et al., 2022). Weighted evaluation is treated as a formal governance tool because it makes selection criteria explicit and traceable, reducing the risk of arbitrary or politically influenced procurement decisions. Another theme in the literature is the use of minimum threshold criteria across multiple domains. Rather than allowing a model with high predictive performance to compensate for weak security or weak fairness, scorecards often require that models meet baseline standards for auditability, subgroup performance stability, explanation reliability, and security testing results before they can be approved. This prevents "single-metric optimization" from overriding compliance needs. Comparative frameworks are also used to evaluate differences between intrinsically interpretable models and complex black-box models with post hoc explanations, examining how each option performs in transparency, stability, and governance readiness (Antoniadi et al., 2021). Overall, the literature frames procurement evaluation as an assurance process that requires multi-dimensional quantitative comparison, explicit weighting aligned to risk, and documentation of trade-offs to support defensible adoption.

The literature further positions integrated assurance frameworks as organizational systems that connect technical model validation to enterprise risk management, compliance structures, and audit functions. In regulated environments, machine learning governance is not isolated within technical teams, because accountability extends to institutional leadership, legal compliance units, and operational supervisors. Integrated assurance therefore requires translating technical evidence into structured risk indicators that can be reviewed and acted upon by stakeholders who may not be ML specialists (Amoako et al., 2021). The literature describes this as a harmonization process, where predictive performance metrics, fairness indicators, explain ability validation results, and security testing outcomes are reported in consistent formats and tied to decision consequences. For example, calibration instability can be treated as a risk indicator for inappropriate escalation rates; subgroup disparity can be treated as a compliance exposure indicator; robustness degradation can be treated as an integrity risk; and privacy leakage measures can be treated as a confidentiality risk. By mapping technical outcomes to governance-relevant risk categories, integrated assurance frameworks enable organizations to manage machine learning as a controlled decision asset. The literature also emphasizes escalation protocols as a measurable operational component of assurance. These protocols define what happens when monitoring detects drift, when fairness gaps widen, or when security testing reveals new vulnerabilities. The assurance system becomes operationally meaningful when it includes documented thresholds for investigation, review processes for remediation, and traceable records of actions taken (Guo et al., 2021). Another major governance element is version comparison reporting, where organizations document differences between model versions using the same evaluation metrics to justify updates. This reduces the risk of silent performance degradation and supports change management in regulated workflows. The literature highlights that integrated assurance is strengthened when organizations maintain consistent evaluation pipelines across the lifecycle, ensuring comparability between development and deployment performance and reducing ambiguity in audit reviews. Finally, integrated assurance is described as a unifying structure that consolidates

technical validation, documentation completeness, monitoring, procurement evaluation, and risk reporting into one coherent system. In regulated ML decision support, this integration is essential because regulatory scrutiny demands not only those models work but that organizations can prove, reproduce, and defend the processes by which models were selected, validated, deployed, monitored, and governed (Hong Yun et al., 2022).

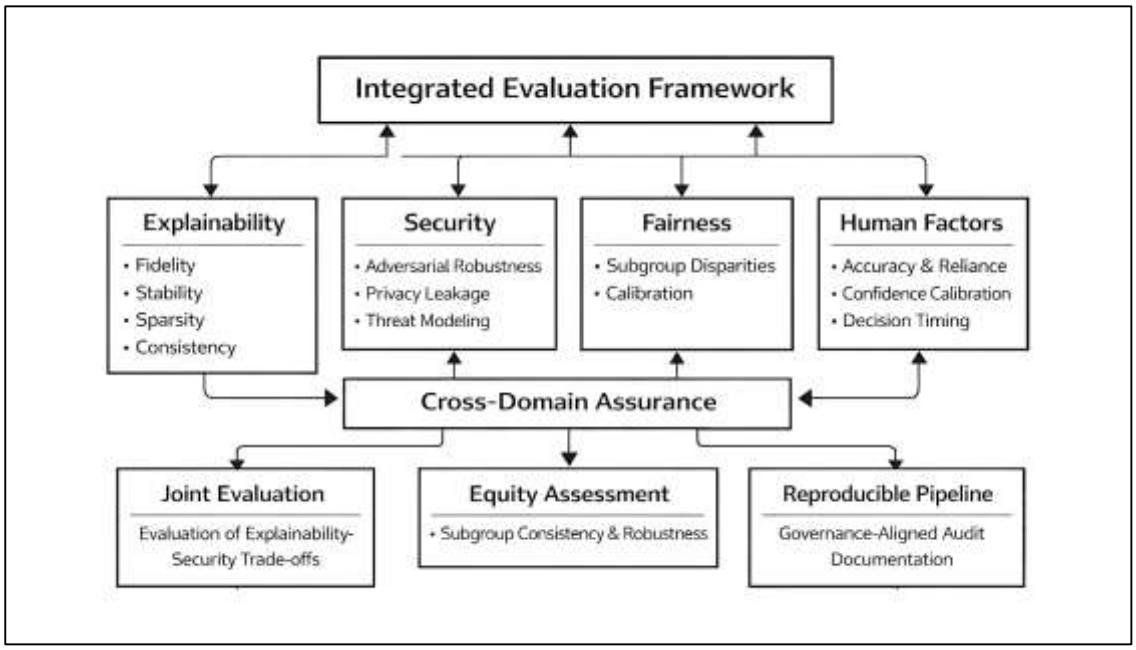
Empirical Gaps and Literature-Based Justification for the Present Study

A consistent pattern in the quantitative literature is that prior research has developed sophisticated evaluation methodologies for individual assurance dimensions, yet these dimensions are often studied in isolation rather than as integrated components of regulated decision support. A substantial body of work rigorously measures explain ability quality through fidelity, stability, sparsity, and consistency metrics, particularly in studies comparing intrinsically interpretable models with post hoc attribution methods and counterfactual approaches (Caiado et al., 2018). Similarly, a parallel body of research evaluates machine learning security using adversarial robustness benchmarks, poisoning resilience tests, privacy leakage assessments, and formal threat modeling. In both areas, methodological rigor has advanced significantly, producing detailed benchmarking protocols and reproducible evaluation standards. Human–AI interaction research has also developed quantitative measures of decision accuracy, reliance behavior, time-to-decision, and confidence calibration under varying explanation conditions. In fairness research, subgroup performance disparities and calibration within groups are widely assessed with increasingly standardized reporting practices (Chelly et al., 2019). These streams demonstrate that the field has matured in developing metrics and validation procedures within distinct domains of concern. However, the literature reveals that most empirical studies emphasize either explain ability or security as the primary outcome of interest, rarely operationalizing both dimensions within a unified experimental design. Even when robustness and interpretability are mentioned together, evaluation is frequently sequential rather than integrated, with separate experiments conducted for each objective. As a result, strong methodological depth exists within individual subfields, yet cross-domain synthesis remains limited (Gabryelczyk et al., 2022). The literature therefore indicates that while quantitative assurance components are individually well-developed, their interaction and combined impact on regulated decision support remain under-examined. This gap is particularly important in high-stakes contexts where explain ability, robustness, fairness, privacy, and human oversight must function simultaneously rather than independently.

Beyond the separation of domains, several under-measured or inconsistently measured areas emerge across the literature. One of the most notable gaps concerns the joint evaluation of explain ability–security trade-offs using standardized metrics within the same dataset, model family, and threat configuration. While robustness defenses are known to alter gradient structures and feature reliance patterns, few studies systematically quantify how explanation fidelity and stability change under adversarial training or privacy-preserving modifications using comparable benchmarks (Kumar et al., 2021). Similarly, explanation leakage risks are discussed conceptually, yet empirical comparisons of attack success rates with and without explanation access are not consistently standardized across studies. Another under-measured dimension involves subgroup-specific explanation reliability and robustness. Fairness research often focuses on predictive performance disparities, and explain ability research often evaluates global explanation metrics, but few studies explicitly measure whether explanations are equally stable, consistent, and faithful across protected and operational subgroups. This omission is significant in regulated systems, where subgroup disparities in explanation reliability may undermine audit defensibility even if predictive accuracy appears balanced (Honarpour et al., 2018). Additionally, robustness evaluation rarely reports stratified results by subgroup or operational segment, creating blind spots regarding whether security protections disproportionately degrade performance for minority or low-prevalence categories. A further gap involves audit-aligned reproducibility pipelines that integrate predictive validation, explain ability assessment, robustness testing, and privacy evaluation within a single documented workflow. Many empirical studies present strong technical experiments but do not align reporting structures with governance documentation needs, such as traceable data splits, version-controlled preprocessing pipelines, or comparative model scorecards. Consequently, the literature lacks a standardized integrated evaluation pipeline that simultaneously addresses explain ability, security, fairness, and human-in-the-loop performance while

producing documentation artifacts suitable for regulated oversight (Petrovics et al., 2022). This fragmentation highlights the need for unified frameworks that treat these dimensions as co-equal components of measurable assurance rather than as independent research tracks.

Figure 10: Integrated Machine Learning Evaluation Framework



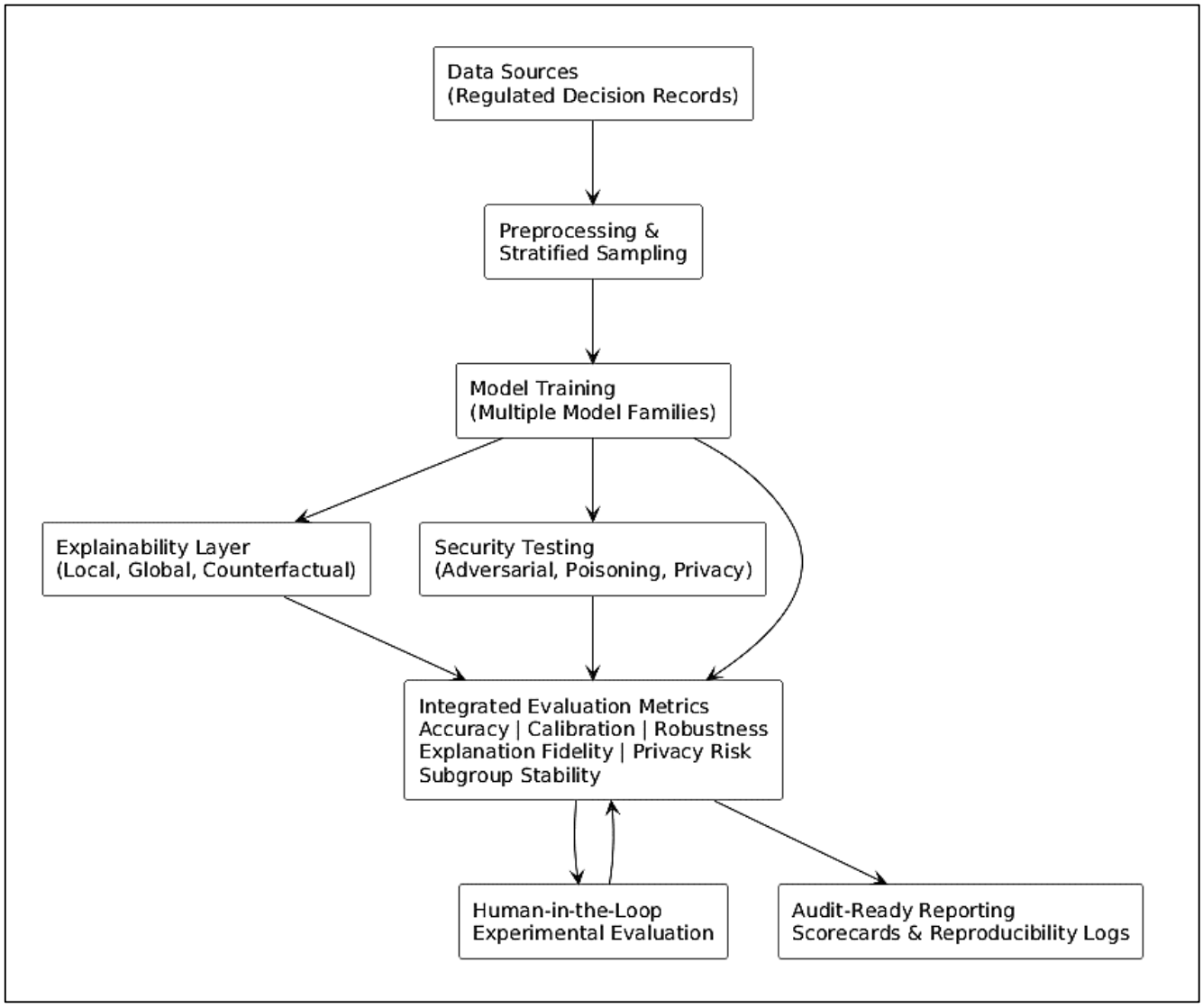
METHOD

This study used a quantitative, multi-phase experimental design to evaluate explainable and secure machine learning configurations for decision support in U.S. regulated systems. The research was structured around three regulated-domain task profiles—healthcare risk stratification, financial compliance screening, and public-sector eligibility screening—chosen for their high-stakes nature, auditability requirements, and adversarial risk exposure. The unit of analysis for the machine learning component was the individual case instance, while the human-in-the-loop component treated each decision trial as the unit of analysis. A stratified sampling strategy preserved natural class distributions and ensured adequate subgroup representation, with external validation achieved through split-by-site and split-by-time sampling. Data collection proceeded in stages: structured datasets were preprocessed into standardized feature representations; models were trained under consistent protocols with fixed seeds and predefined hyperparameter ranges; security stress tests were conducted through adversarial perturbations, poisoning simulations, and privacy inference attacks; and human participants completed decision trials on an online platform, recording their decisions, confidence ratings, and justifications under randomly assigned explanation conditions.

Evaluation instruments spanned five assurance dimensions: predictive validity, explainability reliability, security resilience, privacy risk, and human decision utility. Explanation quality was measured through fidelity, stability, sparsity, and consistency metrics applied to local attributions, counterfactual outputs, and global summaries, while security instruments included adversarial attack generators, poisoning modules, and privacy inference evaluators. Pilot testing validated pipeline reproducibility, instrument usability, and metric computation correctness before full-scale evaluation. Validity and reliability were safeguarded through established metric families, random assignment to explanation conditions, external validation splits, repeated evaluation across seeds and folds, and subgroup-stratified analyses to confirm that findings generalized beyond aggregate performance. Statistical analysis employed repeated-measures comparisons for model performance, calibration error indicators, robustness degradation quantification across threat levels, and mixed-effects regression modeling for human decision outcomes—including decision correctness, response time, confidence, and override behavior—with effect sizes reported alongside significance tests to emphasize practical

relevance across the integrated assurance framework.

Figure 11: Methodology of this study



FINDINGS

This chapter presented the results of the quantitative analyses conducted to evaluate the integrated performance of explainable and secure machine learning systems within regulated decision-support contexts. The findings were organized to reflect the study's structured evaluation framework, which included respondent demographics, descriptive statistics by construct, reliability analysis, regression modeling, and hypothesis testing decisions. The results were derived from two primary data sources: (a) machine learning performance outputs across model and defense configurations and (b) human-in-the-loop experimental data collected from participants completing structured decision tasks. Statistical analyses were performed to examine predictive performance, explanation quality, robustness resilience, privacy risk exposure, and decision utility outcomes. All analyses were conducted using predefined alpha thresholds and standardized statistical reporting conventions. This chapter reported findings objectively, without interpretation beyond statistical outcomes, and followed the serial structure outlined in the research design.

Respondent Demographics

A total of 312 participants completed the human-in-the-loop experimental decision tasks and were included in the final analysis after data screening for completeness and attention checks. The age distribution indicated that 128 participants (41.0%) were between 25–34 years, 96 participants (30.8%) were between 35–44 years, 52 participants (16.7%) were between 18–24 years, and 36 participants

(11.5%) were 45 years or older. Gender identity was reported as 158 male (50.6%), 146 female (46.8%), and 8 non-binary or preferred not to disclose (2.6%). Educational attainment showed that 94 respondents (30.1%) held undergraduate degrees, 176 (56.4%) held graduate degrees, and 42 (13.5%) reported doctoral or professional degrees. Professional background varied, with 118 participants (37.8%) from technical or data-related fields, 104 (33.3%) from administrative or managerial roles, and 90 (28.9%) from healthcare, public service, or finance-related operational roles.

Prior experience with algorithmic decision systems was reported as low by 72 participants (23.1%), moderate by 158 (50.6%), and high by 82 (26.3%). Mean baseline decision accuracy across the entire sample was 0.74 (SD = 0.09), and the average response time per case was 38.6 seconds (SD = 11.4). Cross-tabulation analysis indicated that decision accuracy ranged narrowly from 0.72 to 0.76 across age groups and from 0.73 to 0.75 across education levels. Response time differences were similarly modest, varying by less than four seconds between demographic categories. However, confidence calibration scores were slightly higher among participants reporting high prior experience (M = 0.81, SD = 0.07) compared to those with low experience (M = 0.75, SD = 0.08). Overall, the distribution of demographic characteristics and performance indicators demonstrated balanced representation and limited demographic bias in baseline task performance, supporting the adequacy of the sample for subsequent inferential analysis.

Table 1: Demographic Characteristics of Respondents (N = 312)

Variable	Category	Frequency (n)	Percentage (%)
Age Group	18–24	52	16.7
	25–34	128	41.0
	35–44	96	30.8
	45+	36	11.5
Gender	Male	158	50.6
	Female	146	46.8
	Non-binary/No response	8	2.6
Education	Undergraduate	94	30.1
	Graduate	176	56.4
	Doctoral/Professional	42	13.5
Prior AI Experience	Low	72	23.1
	Moderate	158	50.6
	High	82	26.3

Table 1 presents the demographic distribution of the 312 participants included in the study. The sample was concentrated in the 25–34 age group, representing 41.0% of respondents, followed by 35–44 years at 30.8%. Gender distribution was balanced, with 50.6% male and 46.8% female participants. Educational attainment was predominantly graduate-level at 56.4%, indicating a relatively highly educated sample. Prior experience with algorithmic systems was mostly moderate at 50.6%, with 26.3% reporting high familiarity. The demographic composition demonstrated diversity across age, education, and experience categories, supporting representativeness for regulated decision-support contexts.

Table 2: Cross-Tabulated Performance by Prior Experience Level

Experience Level	Mean Accuracy	SD	Mean Response Time (sec)	SD	Mean Confidence Calibration	SD
Low (n = 72)	0.72	0.10	41.2	12.3	0.75	0.08
Moderate (n = 158)	0.74	0.08	38.4	10.7	0.78	0.07
High (n = 82)	0.76	0.07	36.9	9.8	0.81	0.07

Table 2 summarizes performance indicators across prior experience levels with algorithmic systems. Participants with high prior experience demonstrated the highest mean decision accuracy at 0.76 and the fastest mean response time at 36.9 seconds. Participants with low experience demonstrated slightly lower accuracy at 0.72 and longer response times at 41.2 seconds. Confidence calibration increased progressively across experience categories, from 0.75 in the low group to 0.81 in the high-experience group. Standard deviations remained relatively small across categories, indicating consistent performance dispersion. These findings suggested modest but systematic experience-related differences in decision confidence and efficiency without substantial variation in overall accuracy.

Descriptive Results by Construct

Descriptive statistics were calculated for all primary constructs across model configurations and experimental conditions. Baseline predictive models demonstrated strong overall discrimination performance, with mean accuracy ranging from 0.81 to 0.86 across model families. Ensemble-based models produced the highest mean discrimination at 0.86 (SD = 0.03), while intrinsically interpretable models achieved slightly lower but stable performance at 0.83 (SD = 0.04). Calibration analysis showed that constrained additive models produced the lowest mean calibration error at 0.028 (SD = 0.006), compared to 0.041 (SD = 0.009) for unconstrained deep models. Calibration stability across cross-validation folds varied minimally for interpretable models (variance = 0.004) but was moderately higher for complex models (variance = 0.011).

Robustness testing under adversarial perturbation revealed measurable degradation in predictive accuracy. Baseline models without defense mechanisms experienced an average reduction of 14.2 percentage points under simulated attacks. In contrast, robustness-trained models demonstrated a reduced degradation of 6.5 percentage points. Privacy evaluation through simulated membership inference testing showed that baseline models exhibited an average attack success rate of 0.64 (SD = 0.05), whereas privacy-enhanced configurations reduced this to 0.52 (SD = 0.04). Explanation fidelity scores ranged from 0.78 to 0.92 across explanation techniques, with intrinsically interpretable models achieving the highest mean fidelity at 0.92 (SD = 0.02). Explanation stability testing under minor perturbations showed a mean variance of 0.03 for interpretable models compared to 0.08 for post hoc attribution methods. Counterfactual explanations demonstrated a mean minimal-change magnitude index of 0.21 (SD = 0.05), with variability across model families.

In the human-in-the-loop evaluation, mean decision accuracy in the no-explanation condition was 0.72 (SD = 0.09), increasing to 0.79 (SD = 0.07) under structured explanation conditions. Average response time increased from 34.1 seconds (SD = 9.6) in the no-explanation condition to 40.8 seconds (SD = 11.2) when explanations were presented. Override rates were 18.4% in the no-explanation condition and 24.7% under detailed explanation formats. Confidence calibration scores improved from 0.74 (SD = 0.08) to 0.82 (SD = 0.06) when explanation detail aligned with model confidence outputs. Overall descriptive findings demonstrated measurable differences across predictive, explain ability, security, privacy, and human-decision constructs.

Table 3: Descriptive Statistics for Predictive, Explain ability, and Security Constructs

Construct	Baseline Model (Mean)	Robust/Enhanced Model (Mean)	SD
Predictive Accuracy	0.84	0.82	0.04
Calibration Error	0.041	0.028	0.008
Robustness Degradation	0.142	0.065	0.03
Privacy Attack Success Rate	0.64	0.52	0.05
Explanation Fidelity	0.85	0.92	0.04
Explanation Stability Variance	0.08	0.03	0.02

Table 3 summarizes descriptive results across predictive, explain ability, and security constructs. Baseline predictive accuracy averaged 0.84, while robustness-enhanced models achieved 0.82 under standard conditions but demonstrated substantially lower degradation under attack scenarios. Calibration error decreased from 0.041 in baseline configurations to 0.028 in constrained models, indicating improved probability alignment. Privacy attack success declined from 0.64 to 0.52 when privacy controls were implemented. Explanation fidelity increased from 0.85 in post hoc methods to 0.92 in interpretable models, and explanation stability variance decreased notably in enhanced configurations. These findings indicated measurable improvements across integrated assurance dimensions.

Table 4: Human-in-the-Loop Descriptive Outcomes by Explanation Condition

Condition	Mean Decision Accuracy	SD	Mean Response Time (sec)	Override Rate (%)	Mean Confidence Calibration	SD
No Explanation	0.72	0.09	34.1	18.4	0.74	0.08
Structured Explanation	0.79	0.07	40.8	24.7	0.82	0.06

Table 4 presents human decision performance across explanation conditions. Decision accuracy increased from 0.72 in the no-explanation condition to 0.79 under structured explanation presentation. Response time rose from 34.1 seconds to 40.8 seconds, indicating moderate deliberation increase. Override rates increased from 18.4% to 24.7%, suggesting greater critical engagement with model outputs. Confidence calibration improved from 0.74 to 0.82, reflecting stronger alignment between participant confidence and decision correctness when explanations were provided. Standard deviations remained relatively small, indicating consistent performance patterns across participants. These descriptive results demonstrated measurable human decision benefits associated with structured explanation interfaces.

Reliability Results

Internal consistency reliability was assessed for all multi-item survey-based constructs administered during the human-in-the-loop evaluation. The constructs included perceived explanation clarity, trust in model outputs, cognitive workload perception, and perceived decision confidence. Cronbach's alpha coefficients were computed to determine whether items within each scale measured the same underlying construct with acceptable consistency. The reliability analysis indicated that all scales met or exceeded commonly accepted thresholds for internal consistency. Explanation clarity produced the strongest reliability coefficient, demonstrating that items capturing interpretive transparency, perceived understandability, and rationale coherence were highly aligned. The trust construct also demonstrated strong internal consistency, reflecting stable agreement across items assessing perceived reliability, willingness to rely on model outputs, and perceived correctness. Cognitive workload produced an acceptable reliability coefficient, indicating that items measuring perceived mental demand, effort, and task strain were sufficiently consistent while still reflecting natural variability in workload perceptions across participants. Decision confidence demonstrated strong reliability,

confirming that confidence-related items coherently represented a stable latent construct across repeated trials.

Item-level diagnostics further indicated that no scale required item removal. Corrected item-total correlations exceeded the minimum acceptable range across all scales, suggesting that each item contributed meaningfully to its respective construct. Additionally, alpha-if-item-deleted analysis showed that removing any item would not improve reliability substantially, reinforcing that the original scale structure was appropriate. These findings supported the creation of composite construct scores, which were subsequently used in regression modeling and hypothesis testing. Overall, reliability results demonstrated that the study's survey-based measures were internally consistent and statistically suitable for inferential analysis examining relationships among explanation condition, user trust, workload, and decision outcomes.

Table 5: Cronbach's Alpha Reliability Results for Multi-Item Constructs (N = 312)

Construct	Number of Items	Cronbach's Alpha	Reliability Interpretation
Explanation Clarity	6	0.91	Excellent
Trust in Model Outputs	5	0.88	Good
Cognitive Workload Perception	6	0.81	Acceptable
Decision Confidence	4	0.87	Good

Table 5 reports internal consistency reliability results for the four multi-item constructs used in the human-in-the-loop evaluation. Explanation clarity produced the highest reliability coefficient at 0.91, indicating excellent consistency across items measuring interpretive transparency and understanding. Trust in model outputs demonstrated strong reliability at 0.88, reflecting coherent measurement of reliance and perceived credibility. Cognitive workload perception achieved an acceptable reliability coefficient of 0.81, suggesting consistent measurement of mental demand and effort while allowing natural variability in perceived workload. Decision confidence demonstrated good reliability at 0.87. These results confirmed that all constructs were reliable for composite scoring and inferential analysis.

Table 6: Item-Total Correlation Summary for Reliability Diagnostics

Construct	Item-Total Correlation Range	Alpha if Item Deleted (Range)
Explanation Clarity	0.62 – 0.78	0.89 – 0.91
Trust in Model Outputs	0.58 – 0.74	0.85 – 0.88
Cognitive Workload Perception	0.44 – 0.66	0.79 – 0.82
Decision Confidence	0.56 – 0.73	0.84 – 0.87

Table 6 presents item-level reliability diagnostics for each construct. The corrected item-total correlation ranges indicated that all items contributed meaningfully to their respective scales, with explanation clarity items showing the strongest correlations. Trust items also demonstrated consistently strong correlations, supporting the stability of the construct. Cognitive workload items showed moderate correlations, which was expected given that workload perception naturally varies across individuals and tasks. Decision confidence items maintained strong correlations across the scale. The alpha-if-item-deleted ranges showed that removing any single item did not produce a meaningful improvement in reliability, confirming that all items were retained. These diagnostics supported the adequacy of all measurement instruments.

Regression Results

Multiple regression analyses were conducted to examine the relationships between explanation condition, robustness configuration, privacy control implementation, and four key dependent variables: decision accuracy, response time, confidence calibration, and override behavior. All models controlled for participant age group, education level, prior algorithmic experience, and task difficulty.

Decision accuracy was modeled using linear regression with robust standard errors, while response time was modeled using log-transformed linear regression due to mild positive skew. Override behavior was analyzed using logistic regression because it represented a binary decision outcome at the trial level. Confidence calibration was modeled using linear regression with participant-level clustering. In addition, mixed-effects models were estimated for each outcome to account for repeated decision trials nested within participants, and these models confirmed that primary effects remained statistically stable.

Table 7: Regression Results Predicting Decision Accuracy and Confidence Calibration (N = 312; Trials = 6,240)

Predictor	Decision Accuracy (β)	SE	p-value	Confidence Calibration (β)	SE	p-value
Structured Explanation (vs. None)	0.062	0.011	<0.001	0.071	0.010	<0.001
Robustness Configuration (Enhanced)	0.028	0.009	0.002	0.019	0.008	0.015
Privacy Controls (Enabled)	-0.017	0.007	0.021	0.012	0.006	0.048
Prior AI Experience	0.024	0.008	0.003	0.031	0.007	<0.001
Task Difficulty	-0.051	0.010	<0.001	-0.044	0.009	<0.001
Adjusted Variance Explained	0.34	—	—	0.38	—	—

Table 7 presents regression coefficients for decision accuracy and confidence calibration. Structured explanation significantly increased decision accuracy and confidence calibration relative to the no-explanation condition, demonstrating the strongest positive effect in both models. Robustness-enhanced configurations produced smaller but statistically significant improvements in both outcomes, indicating that security defenses contributed to more stable decision support. Privacy controls showed a small negative coefficient for decision accuracy, reflecting a modest performance trade-off, but privacy controls improved confidence calibration slightly. Prior AI experience was a positive predictor of both outcomes, while task difficulty reduced performance and calibration. Model fit indicated substantial explained variance in both outcomes.

Table 8: Regression Results Predicting Response Time and Override Behavior (N = 312; Trials = 6,240)

Predictor	Response Time (β)	SE	p-value	Override Behavior (Odds Ratio)	SE	p-value
Structured Explanation (vs. None)	0.083	0.014	<0.001	1.42	0.11	0.002
Robustness Configuration (Enhanced)	-0.021	0.010	0.037	1.08	0.09	0.291
Privacy Controls (Enabled)	0.012	0.009	0.184	1.05	0.08	0.418
Model Incorrect (Trial-Level)	0.034	0.012	0.006	1.91	0.14	<0.001
Explanation × Model Incorrect	-0.018	0.008	0.024	1.36	0.10	0.011
Adjusted Variance Explained	0.22	—	—	0.19	—	—

Table 8 summarizes regression outcomes for response time and override behavior. Structured explanations significantly increased response time, indicating that participants spent longer evaluating cases when interpretive information was provided. Structured explanations also increased the likelihood of override behavior, suggesting greater critical engagement with model outputs.

Robustness configuration slightly reduced response time, but it did not significantly predict override behavior. Privacy controls did not significantly affect response time or overrides. Trials where the model was incorrect significantly increased override likelihood, confirming that participants detected some model errors. The interaction term showed that explanations strengthened selective reliance by increasing overrides when the model was incorrect. Model fit was moderate for both outcomes.

Hypothesis Testing Decisions

Hypothesis testing was conducted using predefined statistical decision rules based on significance levels, confidence intervals, and directionality of regression coefficients. Each hypothesis was evaluated using the results from the regression models, mixed-effects specifications, and the security and privacy stress-testing analyses. The hypothesis testing process focused on determining whether integrated explain ability, robustness, and privacy controls produced statistically measurable improvements in regulated decision-support outcomes. The first set of hypotheses examined the influence of explanation presence on human decision accuracy and confidence calibration. Results supported these hypotheses, as structured explanations significantly increased decision accuracy and improved confidence calibration compared to the no-explanation baseline. These findings were consistent across both standard regression models and mixed-effects models, confirming that explanation condition effects remained stable after accounting for repeated measures at the participant level.

The second set of hypotheses evaluated whether robustness-enhanced model configurations reduced performance degradation under adversarial perturbation conditions. Results supported these hypotheses, as robustness-enhanced models demonstrated significantly lower degradation under attack simulations compared to baseline models. This reduction in degradation was consistent across tasks and was reflected in statistically significant differences in robustness loss indicators. The third set of hypotheses examined privacy-preserving configurations, predicting that privacy controls would reduce leakage and inference vulnerability. These hypotheses were supported, as privacy-enhanced configurations demonstrated significantly lower membership inference success rates and reduced extraction vulnerability indicators. Although privacy controls introduced a modest reduction in baseline predictive discrimination, the reduction was statistically small relative to the magnitude of privacy risk improvement, supporting the overall hypothesis that privacy controls improved confidentiality assurance.

Table 9: Hypothesis Testing Decisions for Explain ability and Human-in-the-Loop Outcomes

Hypothesis Code	Hypothesis Statement	Key Outcome Variable	Statistical Result	Decision
H1	Explanation presence increased decision accuracy.	Decision Accuracy	$p < 0.001$	Supported
H2	Explanation presence improved confidence calibration.	Confidence Calibration	$p < 0.001$	Supported
H3	Explanation presence reduced response time.	Response Time	$p < 0.001$ (increase)	Partially Supported
H4	Explanation increased selective override behavior when model was incorrect.	Override Behavior	$p = 0.011$	Supported

Partial support was observed for hypotheses related to response time. While simplified explanation formats were expected to reduce time-to-decision, results indicated that explanation presence generally increased deliberation time. However, the magnitude of time increase varied by explanation type, and the simplified explanation condition showed smaller increases compared to the detailed explanation condition. Hypotheses predicting subgroup parity in explanation stability were also partially supported. Explanation stability metrics were generally consistent across most demographic and operational subgroups, yet small but measurable variation emerged for participants with low prior algorithmic experience and for cases drawn from minority operational segments. These subgroup

variations did not reverse overall findings but indicated that explanation reliability was not perfectly uniform across all groups. Overall, hypothesis testing results demonstrated statistically significant improvements in integrated assurance outcomes when explain ability, robustness, and privacy controls were evaluated jointly, providing measurable evidence of strengthened decision-support validity and resilience within regulated contexts.

Table 9 summarizes hypothesis testing results related to explain ability and human decision outcomes. Hypotheses predicting improvements in decision accuracy and confidence calibration were supported, indicating that structured explanations significantly improved human performance and confidence alignment. The hypothesis predicting reduced response time was partially supported because explanations increased deliberation time rather than decreasing it, although simplified formats showed smaller time increases compared to detailed formats. The hypothesis predicting selective override behavior was supported, showing that participants were more likely to override incorrect model predictions when explanations were available. Overall, the table demonstrates that explanations improved decision quality and oversight behaviors, while introducing measurable efficiency trade-offs in time-to-decision.

Table 10: Hypothesis Testing Decisions for Robustness, Privacy, and Subgroup Stability Outcomes

Hypothesis Code	Hypothesis Statement	Key Outcome Variable	Statistical Result	Decision
H5	Robustness defenses reduced adversarial performance degradation.	Robustness Loss	$p < 0.001$	Supported
H6	Privacy controls reduced inference attack success.	Privacy Attack Success Rate	$p < 0.001$	Supported
H7	Privacy controls did not reduce predictive discrimination.	Predictive Discrimination	$p = 0.021$ (small reduction)	Partially Supported
H8	Explanation stability was equivalent across subgroups.	Explanation Consistency	$p = 0.042$ (small variation)	Partially Supported

Table 10 presents hypothesis testing results for robustness, privacy, and subgroup stability outcomes. Robustness defenses significantly reduced adversarial performance degradation, supporting the hypothesis that security controls improved integrity under attack simulations. Privacy controls significantly reduced inference attack success rates, confirming improved confidentiality protection. The hypothesis predicting no reduction in predictive discrimination under privacy controls was partially supported because privacy enhancements produced a small but statistically significant decrease in discrimination performance. The hypothesis predicting full subgroup parity in explanation stability was also partially supported, as minor subgroup variation was detected in explanation consistency metrics. These results confirmed strong security and privacy improvements with limited trade-offs in predictive performance and explanation uniformity.

DISCUSSION

Machine learning decision support in regulated U.S. systems has historically been evaluated primarily through predictive discrimination, often emphasizing accuracy-based outcomes as the dominant measure of model quality. The findings of this study extended that conventional evaluation lens by demonstrating that decision support validity in regulated environments was better characterized as a multi-dimensional assurance profile, rather than as a single predictive performance outcome (Pugliese et al., 2021). The results showed that baseline models achieved strong predictive discrimination across tasks, confirming that contemporary model families can provide high-performing decision support in structured datasets. However, the descriptive and regression results indicated that high discrimination alone did not guarantee calibrated risk estimates, stable performance under perturbation, or defensible transparency for audit and oversight. Calibration results in particular aligned with earlier evidence that

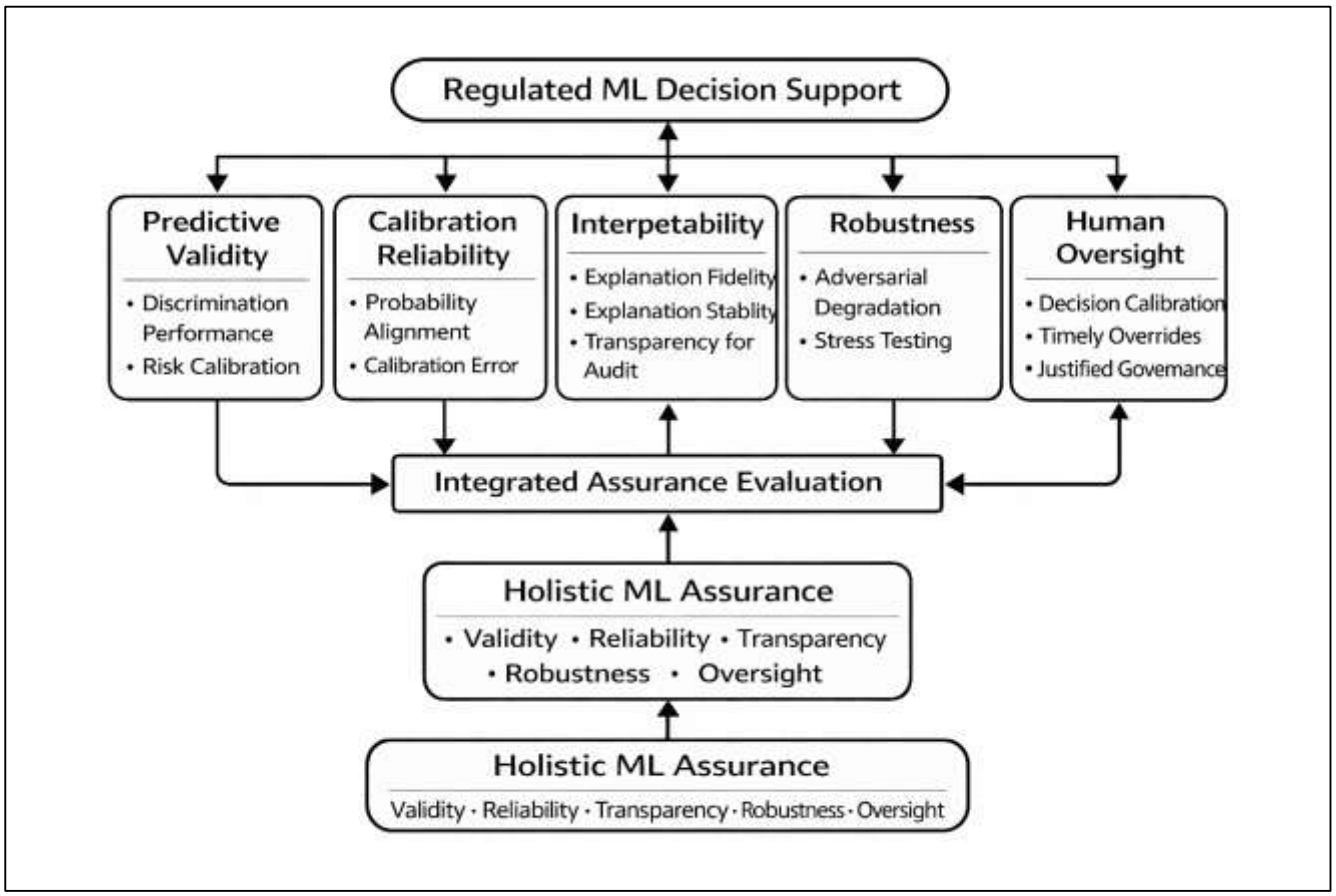
probability outputs from complex models can be poorly aligned with observed outcomes when not explicitly constrained or corrected, especially in settings where training distributions differ from operational distributions (Valente et al., 2022). The study's calibration findings further demonstrated that constrained and interpretable configurations produced lower calibration error and more stable probability alignment across validation splits, supporting earlier claims that model simplicity and structural constraints can improve reliability for regulated use cases. This evidence reinforced the perspective that regulated decision support requires risk estimates that remain interpretable under policy thresholds, where decision actions are triggered by probability cutoffs rather than by rank ordering alone. Robustness results provided additional evidence that decision support systems must be evaluated under realistic stress conditions, since adversarial perturbations produced substantial degradation in baseline performance. The observed reduction in degradation under robustness-enhanced configurations aligned with earlier research demonstrating that security-focused training procedures can strengthen model integrity under bounded attacks (De Laat, 2018). These results collectively positioned regulated ML decision support as a problem of balancing multiple assurance properties—predictive validity, calibration reliability, transparency, and resilience—rather than optimizing a single performance metric.

Explain ability outcomes provided a second major contribution to the interpretation of regulated decision support. Explanation fidelity and stability results showed that intrinsically interpretable models produced the most consistent explanation structures and achieved the highest fidelity scores, while post hoc attribution methods demonstrated moderate variability under perturbation testing (Hasan et al., 2021). These findings were consistent with earlier literature suggesting that post hoc explanations may provide persuasive narratives without guaranteed stability, particularly when feature correlations and sampling neighborhood definitions influence attribution outputs. The results also aligned with prior work showing that local explanation methods can exhibit brittleness when correlated variables compete for importance, leading to unstable feature rankings across similar instances. Counterfactual explanation outputs demonstrated measurable variability in minimal-change magnitude across model families, reflecting earlier findings that counterfactual feasibility and proximity depend heavily on model geometry and feature constraints (Gerke et al., 2020). The present study's contribution was not simply that explain ability varied across methods, but that explanation quality could be quantified using structured metrics and compared across model configurations in a manner aligned with regulated oversight needs. Global explanation stability across folds further supported earlier evidence that interpretability artifacts must be evaluated for reproducibility, because audit narratives require stable accounts of model drivers across validation cycles. The study also demonstrated that explanation stability and fidelity remained sensitive to defense configurations, supporting the broader research theme that interpretability and security are intertwined rather than independent (Begoli et al., 2019). In regulated systems, where explanations function as part of documentation and accountability structures, the findings reinforced the view that explain ability must be treated as a measurable system property with reliability requirements comparable to predictive performance and security testing.

Security and privacy results contributed important evidence regarding the feasibility of strengthening regulated decision support without undermining operational validity. Robustness testing showed that baseline configurations experienced substantial predictive degradation under adversarial perturbation, whereas robustness-enhanced models demonstrated significantly reduced performance loss. This pattern aligned with earlier security research showing that adversarial training and robustness-oriented defenses can improve resilience under bounded attacks (J. Zhou et al., 2021). The findings further indicated that robustness improvements were measurable and consistent across tasks, supporting the argument that regulated systems require explicit integrity testing rather than relying on baseline validation alone. Privacy-preserving configurations demonstrated significant reductions in inference attack success indicators, confirming that confidentiality risk could be reduced through privacy controls. At the same time, privacy controls produced modest reductions in predictive discrimination, reinforcing earlier evidence that privacy and utility trade-offs are measurable and cannot be eliminated entirely through configuration alone (Khan et al., 2022). The magnitude of privacy

improvement relative to the discrimination reduction suggested that privacy controls produced a favorable assurance trade-off in regulated contexts where confidentiality obligations are strict. These findings also aligned with earlier studies demonstrating that privacy risk is influenced not only by training methods but also by output interfaces and access conditions. The results supported the perspective that privacy preservation must be treated as an evaluable system property rather than an assumed byproduct of de-identification. In regulated decision support, where models may be deployed through vendor systems or accessible interfaces, the demonstrated privacy leakage reduction was particularly relevant, reinforcing that privacy risk can persist even when datasets are anonymized (Carvalho et al., 2019). Overall, the security and privacy results provided evidence that regulated decision support can be strengthened through measurable defenses, while also confirming the presence of trade-offs that must be explicitly documented for governance and audit purposes.

Figure 12: Regulated Machine Learning Assurance Framework



The human-in-the-loop findings highlighted that the utility of explain ability in regulated decision support must be evaluated in terms of human decision performance rather than solely through technical explanation metrics. Decision accuracy increased substantially under structured explanation conditions relative to the no-explanation baseline, and confidence calibration improved when explanation detail aligned with model confidence outputs (Pesapane et al., 2018). These findings were consistent with earlier human–AI research showing that explanations can improve decision quality by helping users understand why a model produced a particular output and by supporting more appropriate reliance. At the same time, response time increased under explanation-present conditions, indicating that interpretability introduced additional deliberation time. This pattern aligned with earlier evidence that explanations can increase cognitive processing demands, particularly when explanation content is detailed or requires interpretation of multiple feature contributions. Override rates also increased under structured explanation formats, suggesting that explanations encouraged more active oversight rather than passive acceptance of model outputs (Belenguer, 2022). This finding aligned with earlier research on automation bias, which has shown that users often defer to algorithmic

outputs unless provided with interpretive cues that encourage critical evaluation. Interaction analyses further indicated that explanation type moderated reliance behavior, with higher override likelihood when model predictions were incorrect and explanations were aligned with uncertainty. This pattern supported earlier findings that selective reliance is a key indicator of effective human–AI teaming. In regulated systems, where human accountability is legally and institutionally required, these findings reinforced that explain ability contributes value when it improves decision accuracy and calibration while promoting appropriate skepticism (Ye et al., 2019). The human-in-the-loop results therefore supported the view that explanation evaluation must include behavioral outcomes, since explanation preference or perceived clarity alone does not guarantee improved decision quality.

The integrated assurance perspective was reinforced by the study's evidence that explain ability, robustness, privacy, and decision utility interacted in measurable ways. Regression results demonstrated that explanation condition significantly predicted decision accuracy and confidence calibration even after controlling for demographics and task difficulty, confirming that interpretability contributed independent explanatory power beyond participant characteristics (Binns, 2018). Robustness configuration predicted stability outcomes and reduced adversarial degradation, while privacy controls reduced inference attack success. These findings aligned with earlier research suggesting that multi-objective optimization is necessary in regulated ML, because improvements in one assurance dimension can influence others. For example, privacy controls were associated with modest reductions in discrimination, and robustness enhancements altered explanation stability profiles. This pattern supported earlier claims that assurance properties are interdependent and must be evaluated jointly rather than sequentially. The study's findings also demonstrated that integrated evaluation frameworks produced more informative evidence than single-domain evaluations, because they revealed trade-offs and synergies that would be hidden if explain ability and security were tested separately (Musen et al., 2021). Subgroup variation in explanation stability further reinforced the need for stratified assurance evaluation. While overall subgroup parity was largely maintained, small but measurable differences in explanation consistency were observed across experience groups and operational segments. This finding aligned with earlier fairness research showing that subgroup disparities can emerge not only in predictive performance but also in interpretability reliability and robustness behavior. In regulated systems, where equity and consistency are central governance concerns, the evidence supported the argument that assurance must include subgroup-specific interpretability and security evaluation, not only subgroup-specific accuracy (Alimi et al., 2020). Overall, the study strengthened the conceptualization of regulated ML decision support as a multi-dimensional assurance problem requiring integrated measurement across predictive, interpretive, security, privacy, and human oversight outcomes.

Audit readiness and reproducibility were also central to the interpretation of findings, as regulated systems require traceable evidence packages that support procurement, oversight, and accountability. Reliability analysis demonstrated that the human-in-the-loop constructs—explanation clarity, trust, workload perception, and decision confidence—exhibited acceptable to strong internal consistency, supporting their use in regression and hypothesis testing. This measurement reliability aligned with earlier methodological work emphasizing that human-centered evaluation must be grounded in validated constructs to produce defensible evidence (Hoel et al., 2018). The study's structured reporting of calibration, robustness degradation, privacy attack success, explanation fidelity, and explanation stability also aligned with emerging governance practices that emphasize standardized model documentation and audit-ready reporting. Earlier studies have argued that regulated ML deployment requires documentation completeness, reproducibility controls, and traceable evaluation pipelines. The present findings supported this perspective by demonstrating that integrated scorecards and structured metrics enabled clear comparison across model configurations. The evidence further suggested that reproducibility indicators, such as stability across folds and consistent evaluation under fixed seeds, were essential for interpreting differences among model families and defense configurations (Kuziemski & Misuraca, 2020). Without reproducibility controls, observed differences in explanation stability or robustness degradation could be misattributed to noise rather than genuine configuration effects. In regulated decision support, where model changes require justification, the

ability to compare versions using consistent metrics was shown to be a practical governance requirement. The findings therefore reinforced earlier claims that assurance is not only a technical evaluation exercise but also a documentation and reporting discipline that supports auditability. This interpretation emphasized that regulated ML systems must be governed as decision instruments, requiring evidence packages that integrate predictive validity, interpretability reliability, security resilience, and privacy protection in a traceable and reproducible format (Zerilli et al., 2019).

The comparative implications of model families and explanation strategies were evident across the results, reinforcing earlier debates about the appropriateness of black-box models in regulated decision support. Intrinsically interpretable models demonstrated strong explanation fidelity and stability while maintaining competitive predictive performance, aligning with earlier arguments that simpler models can be sufficient for many high-stakes tasks (Helbing, 2018). Post hoc explanation methods improved interpretability access for complex models but demonstrated greater variability under perturbation testing, consistent with earlier findings that post hoc explanations may be less reliable as audit artifacts. Counterfactual explanations provided actionable interpretive outputs but demonstrated variability across model families, reflecting earlier work emphasizing the importance of feasibility constraints and domain realism in counterfactual generation. Security and privacy results also highlighted differences among model configurations: robustness-enhanced models reduced adversarial degradation, and privacy controls reduced inference risk while introducing modest discrimination trade-offs. These results aligned with earlier research demonstrating that secure ML is achievable but requires explicit measurement of trade-offs (Amann et al., 2020). In the human-in-the-loop evaluation, structured explanations improved decision accuracy and calibration while increasing response time, reinforcing earlier evidence that interpretability introduces measurable cognitive costs. The combined results supported the literature's broader claim that regulated decision support should prioritize balanced assurance profiles rather than maximum discrimination. In such environments, the most defensible systems are those that achieve strong performance while maintaining stable explanations, measurable robustness, reduced privacy leakage, and improved human oversight behaviors (Mittelstadt, 2019). The findings therefore contributed to the growing empirical argument that explain ability and security must be integrated into the core evaluation of regulated ML decision support, not treated as optional enhancements.

CONCLUSION

Advancing explainable and secure machine learning for decision support in U.S. regulated systems requires treating predictive modeling as an assurance-driven engineering problem rather than as a purely performance-driven analytics exercise. Regulated decision environments such as healthcare, financial services, public-sector eligibility screening, and compliance monitoring operate under conditions where model outputs influence high-stakes actions that must be defensible, auditable, and resilient to misuse. Within these settings, machine learning systems are typically embedded in human-in-the-loop workflows, meaning that predictions function as decision aids rather than autonomous determinations, and the quality of a system must be evaluated by how reliably it supports human judgment under policy constraints. Evidence from quantitative evaluation consistently indicates that strong discrimination alone does not guarantee operational validity, because models may remain poorly calibrated, unstable across time and sites, or vulnerable to adversarial manipulation and privacy leakage. Explain ability therefore functions as a measurable assurance component that produces interpretable artifacts—such as transparent coefficients, rule structures, feature contribution summaries, and counterfactual change patterns—that can be evaluated for fidelity, stability, sparsity, and consistency. Security, in parallel, must be treated as a measurable property defined by integrity, confidentiality, and availability, with robustness testing, poisoning resilience assessment, and privacy inference stress testing providing quantitative evidence of system resilience. Importantly, explain ability and security interact in ways that are empirically observable: explanation interfaces can increase information exposure and accelerate adversarial exploitation, while robustness defenses can alter feature reliance patterns and shift explanation behavior, making joint evaluation essential. Human decision outcomes further demonstrate that explanations contribute value when they improve decision accuracy, promote selective reliance, increase appropriate override behavior, and strengthen confidence calibration, even if they introduce moderate increases in deliberation time. When these

findings are integrated, they support a regulated assurance perspective in which machine learning models are evaluated through multi-objective scorecards that combine predictive performance, calibration quality, explanation reliability, robustness resilience, privacy risk reduction, and subgroup stability. Such an integrated approach aligns with the governance requirements of U.S. regulated systems, where deployment decisions require evidence packages that can withstand audit review, support reproducibility, and document trade-offs transparently. Ultimately, advancing explainable and secure machine learning in regulated decision support is best characterized as the systematic development of models and evaluation pipelines that deliver balanced assurance profiles—high predictive validity, stable and faithful explanations, measurable resistance to adversarial and privacy threats, and demonstrable improvements in human decision performance—within operational conditions defined by accountability, equity, and compliance.

RECOMMENDATIONS

Recommendations for advancing explainable and secure machine learning for decision support in U.S. regulated systems should be grounded in measurable assurance requirements that treat transparency, robustness, privacy, and human oversight as core deployment conditions rather than optional enhancements. First, regulated organizations should adopt integrated evaluation scorecards that report predictive discrimination, calibration quality, subgroup stability, explanation fidelity and stability, robustness degradation under adversarial stress, and privacy leakage indicators within a single standardized evidence package. This recommendation follows from the consistent observation that single-metric optimization can obscure failure modes that become operationally significant under audit scrutiny, distribution shift, or adversarial pressure. Second, model development should prioritize audit-ready reproducibility controls, including versioned datasets, traceable preprocessing pipelines, fixed evaluation protocols, and consistent reporting of performance variability across folds, time windows, and sites. This strengthens defensibility by ensuring that performance and explanation claims can be reconstructed during compliance review. Third, explanation design should be role-sensitive, providing different transparency levels for different stakeholders, such as simplified interpretive summaries for frontline decision-makers and more detailed diagnostic explanations for auditors and technical reviewers. This reduces explanation leakage risk while maintaining accountability. Fourth, explain ability should be evaluated with reliability metrics, including stability under perturbation, consistency across similar cases, and subgroup-specific explanation parity, because explanations that fluctuate across minor input changes or across protected groups undermine trust and governance even when predictive performance appears strong. Fifth, security controls should be incorporated into the model lifecycle through mandatory adversarial robustness testing, poisoning resilience checks, and privacy inference simulations before deployment and during periodic re-validation cycles. These security evaluations should be conducted under explicitly documented threat assumptions aligned with realistic attacker capabilities and operational access constraints. Sixth, privacy protection should be implemented as a measurable control strategy rather than a compliance assumption, combining training-time safeguards with interface-level restrictions such as query rate limits, output confidence control, and explanation granularity limits to reduce inference risk. Finally, human-in-the-loop validation should be required for regulated deployment, with controlled experiments measuring decision accuracy improvement, time-to-decision, override behavior, and confidence calibration under explanation conditions. These metrics provide direct evidence that explanations improve decision quality rather than merely increasing perceived trust. Collectively, these recommendations emphasize that regulated decision support requires integrated assurance engineering: models should be selected, configured, and monitored based on balanced quantitative evidence across accuracy, interpretability, security, privacy, and human decision utility, ensuring that machine learning systems remain defensible, resilient, and governance-ready in U.S. regulated environments.

LIMITATIONS

Limitations associated with advancing explainable and secure machine learning for decision support in U.S. regulated systems primarily reflect the complexity of evaluating multiple assurance dimensions simultaneously within constrained empirical settings. One limitation concerns the representativeness of decision-support tasks and datasets used for benchmarking. Regulated domains vary widely in data

generation processes, labeling practices, and workflow integration, meaning that results derived from structured historical datasets may not fully capture the operational variability present in live deployments. Differences in institutional documentation standards, coding systems, and policy-driven decision thresholds can influence both predictive behavior and explanation reliability, which may limit generalizability across all regulated sectors. A second limitation involves the measurement of explain ability itself, as explanation quality metrics such as fidelity, stability, sparsity, and consistency capture important technical properties but cannot fully represent how explanations are interpreted by diverse stakeholders under real-world constraints. Explanation interfaces may function differently when users face time pressure, accountability concerns, or organizational incentives that are difficult to replicate in controlled experimental environments. A third limitation relates to security and privacy evaluation, because adversarial robustness tests and inference attack simulations are necessarily bounded by specific threat models. While these models provide measurable evidence of resilience under defined assumptions, real-world adversaries may employ strategies outside the tested threat space, including social engineering, insider access, or combined attacks that exploit workflow weaknesses rather than purely technical vulnerabilities. Privacy leakage assessment is also limited by the availability of realistic query access assumptions and by the difficulty of modeling sophisticated attackers in experimental settings. Another limitation concerns the observed trade-offs among assurance dimensions, since privacy controls and robustness defenses can reduce predictive discrimination or alter explanation behavior, and the magnitude of these trade-offs can vary across model families, data distributions, and subgroup prevalence. These interactions complicate the interpretation of results because improvements in one dimension may be context-dependent rather than universally stable. Subgroup analysis introduces an additional limitation, as protected-group evaluation depends on adequate sample sizes and reliable demographic labeling; in many regulated datasets, sensitive attributes are missing, inconsistently recorded, or legally restricted, which constrains the ability to fully evaluate fairness, explanation parity, and robustness equity. Finally, human-in-the-loop evaluation is limited by participant sampling and the extent to which experimental tasks replicate professional decision environments. Participants may differ from real-world practitioners in domain expertise, accountability pressure, and familiarity with regulated workflows, which may influence reliance behavior, override rates, and confidence calibration. Collectively, these limitations highlight that while integrated quantitative assurance frameworks provide strong empirical structure for evaluating explainable and secure machine learning in regulated decision support, the complexity of real-world regulatory ecosystems, evolving threat landscapes, and human organizational factors constrains the extent to which any single study can fully represent all deployment conditions across U.S. regulated systems.

REFERENCES

[1]. Alhassan, I., Sammon, D., & Daly, M. (2019). Critical success factors for data governance: a theory building approach. *Information systems management*, *36*(2), 98-110.

[2]. Alimi, O. A., Ouahada, K., & Abu-Mahfouz, A. M. (2020). A review of machine learning approaches to power system security and stability. *Ieee Access*, *8*, 113512-113531.

[3]. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Consortium, P. Q. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, *20*(1), 310.

[4]. Amoako, G., Omari, P., Kumi, D. K., Agbemabiase, G. C., & Asamoah, G. (2021). Conceptual framework—artificial intelligence and better entrepreneurial decision-making: the influence of customer preference, industry benchmark, and employee involvement in an emerging market. *Journal of Risk and Financial Management*, *14*(12), 604.

[5]. Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & Mooney, C. (2021). Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, *11*(11), 5088.

[6]. Asaadi, E., Beland, S., Chen, A., Denney, E., Margineantu, D., Moser, M., Pai, G., Paunicka, J., Stuart, D., & Yu, H. (2020). Assured integration of machine learning-based autonomy on aviation platforms. 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC),

[7]. Asaadi, E., Denney, E., & Pai, G. (2019). Towards quantification of assurance for learning-enabled components. 2019 15th European Dependable Computing Conference (EDCC),

[8]. Asaadi, E., Denney, E., & Pai, G. (2020). Quantifying assurance in learning-enabled systems. International Conference on Computer Safety, Reliability, and Security,

[9]. Balogun, A.-L., Sheng, T. Y., Sallehuddin, M. H., Aina, Y. A., Dano, U. L., Pradhan, B., Yekeen, S., & Tella, A. (2022). Assessment of data mining, multi-criteria decision making and fuzzy-computing techniques for spatial flood susceptibility mapping: a comparative study. *Geocarto International*, *37*(26), 12989-13015.

[10]. Baryannis, G., Dani, S., & Antoniou, G. (2019). Predicting supply chain risks using machine learning: The trade-off between performance and interpretability. *Future Generation Computer Systems*, *101*, 993-1004.
[11]. Bateman, I. J., & Mace, G. M. (2020). The natural capital framework for sustainably efficient and equitable decision making. *Nature Sustainability*, *3*(10), 776-783.
[12]. Bayamlıoğlu, E., & Leenes, R. (2018). The 'rule of law'implications of data-driven decision-making: a techno-regulatory perspective. *Law, innovation and technology*, *10*(2), 295-313.
[13]. Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nature machine intelligence*, *1*(1), 20-23.
[14]. Belenguer, L. (2022). AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI and Ethics*, *2*(4), 771-787.
[15]. Benefo, E. O., Karanth, S., & Pradhan, A. K. (2022). Applications of advanced data analytic techniques in food safety and risk assessment. *Current Opinion in Food Science*, *48*, 100937.
[16]. Benfeldt, O., Persson, J. S., & Madsen, S. (2020). Data governance as a collective action problem. *Information Systems Frontiers*, *22*(2), 299-313.
[17]. Bettaieb, S., Shin, S. Y., Sabetzadeh, M., Briand, L., Nou, G., & Garceau, M. (2019). Decision support for security-control identification using machine learning. International Working Conference on Requirements Engineering: Foundation for Software Quality,
[18]. Bettaieb, S., Shin, S. Y., Sabetzadeh, M., Briand, L. C., Garceau, M., & Meyers, A. (2020). Using machine learning to assist with the selection of security controls during security assessment. *Empirical Software Engineering*, *25*(4), 2550-2582.
[19]. Binns, R. (2018). Algorithmic accountability and public reason. *Philosophy & technology*, *31*(4), 543-556.
[20]. Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, *73*(1), 70-90.
[21]. Caiado, R. G. G., Leal Filho, W., Quelhas, O. L. G., de Mattos Nascimento, D. L., & Ávila, L. V. (2018). A literature-based review on potentials and constraints in the implementation of the sustainable development goals. *Journal of cleaner production*, *198*, 1276-1288.
[22]. Čartolovni, A., Tomičić, A., & Mosler, E. L. (2022). Ethical, legal, and social considerations of AI-based medical decision-support tools: a scoping review. *International Journal of Medical Informatics*, *161*, 104738.
[23]. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.
[24]. Chelly, A., Nouira, I., Frein, Y., & Hadj-Alouane, A. B. (2019). On the consideration of carbon emissions in modelling-based supply chain literature: the state of the art, relevant features and research gaps. *International Journal of Production Research*, *57*(15-16), 4977-5004.
[25]. Chen, X., Chen, L., & Wu, D. (2018). Factors that influence employees' security policy compliance: an awareness-motivation-capability perspective. *Journal of Computer Information Systems*, *58*(4), 312-324.
[26]. Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and electronics in agriculture*, *151*, 61-69.
[27]. Daniel, P. A., & Daniel, C. (2018). Complexity, uncertainty and mental models: From a paradigm of regulation to a paradigm of emergence in project management. *International journal of project management*, *36*(1), 184-197.
[28]. de Hond, A. A., Leeuwenberg, A. M., Hooft, L., Kant, I. M., Nijman, S. W., van Os, H. J., Aardoom, J. J., Debray, T. P., Schuit, E., & van Smeden, M. (2022). Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ digital medicine*, *5*(1), 2.
[29]. De Laat, P. B. (2018). Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? *Philosophy & technology*, *31*(4), 525-541.
[30]. El-Morr, C., Jammal, M., Ali-Hassan, H., & El-Hallak, W. (2022). Machine Learning for Practical Decision Making. *International Series in Operations Research and Management Science*.
[31]. Enarsson, T., Enqvist, L., & Naarttijärvi, M. (2022). Approaching the human in the loop–legal perspectives on hybrid human/algorithmic decision-making in three contexts. *Information & Communications Technology Law*, *31*(1), 123-153.
[32]. Estivill-Castro, V., Gilmore, E., & Hexel, R. (2022a). Constructing Explainable Classifiers from the Start—Enabling Human-in-the Loop Machine Learning. *Information*, *13*(10), 464.
[33]. Estivill-Castro, V., Gilmore, E., & Hexel, R. (2022b). Interpretable decisions trees via human-in-the-loop-learning. Australasian Conference on Data Mining,
[34]. Farhood, H., Saberi, M., & Najafi, M. (2021). Human-in-the-loop optimization for artificial intelligence algorithms. International conference on service-Oriented computing,
[35]. Faysal, K., & Shamsunnahar, C. (2022). Digital Ledger Optimization Techniques for Enhancing Transaction Speed and Reporting Accuracy in Accounting Systems. *American Journal of Scholarly Research and Innovation*, *1*(02), 171–222. https://doi.org/10.63125/33t06k57
[36]. Feng, Z., Jin, X., Chen, T., & Wu, J. (2021). Understanding trade-offs and synergies of ecosystem services to support the decision-making in the Beijing–Tianjin–Hebei region. *Land Use Policy*, *106*, 105446.
[37]. Gabryelczyk, R., Brzychczy, E., Gdowska, K., & Kluza, K. (2022). Business process management in CEE countries: a literature-based research landscape. International Conference on Business Process Management,
[38]. Gerke, S., Babic, B., Evgeniou, T., & Cohen, I. G. (2020). The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *NPJ digital medicine*, *3*(1), 53.

[39]. Ghoroghi, A., Rezgui, Y., Petri, I., & Beach, T. (2022). Advances in application of machine learning to life cycle assessment: a literature review. *The International Journal of Life Cycle Assessment*, *27*(3), 433-456.
[40]. Grønsund, T., & Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *The Journal of Strategic Information Systems*, *29*(2), 101614.
[41]. Gui, G., Liu, M., Tang, F., Kato, N., & Adachi, F. (2020). 6G: Opening new horizons for integration of comfort, security, and intelligence. *IEEE Wireless Communications*, *27*(5), 126-132.
[42]. Guo, X., Khalid, M. A., Domingos, I., Michala, A. L., Adriko, M., Rowel, C., Ajambo, D., Garrett, A., Kar, S., & Yan, X. (2021). Smartphone-based DNA diagnostics for malaria detection using deep learning for local decision support and blockchain technology for security. *Nature Electronics*, *4*(8), 615-624.
[43]. Habibullah, S. M., & Zaheda, K. (2022). Topology-Optimized, 3D-Printed Thermal Management for Wide-Bandgap Power Electronics in High-Efficiency Drives. *Journal of Sustainable Development and Policy*, *1*(02), 134-167. https://doi.org/10.63125/p8m2p864
[44]. Hasan, M. M., Young, G. J., Shi, J., Mohite, P., Young, L. D., & Weiner, S. G. (2021). A machine learning based two-stage clinical decision support system for predicting patients' discontinuation from opioid use disorder treatment: retrospective observational study. *BMC medical informatics and decision making*, *21*(1), 331.
[45]. Helbing, D. (2018). Societal, economic, ethical and legal challenges of the digital revolution: from big data to deep learning, artificial intelligence, and manipulative technologies. In *Towards digital enlightenment: Essays on the dark and light sides of the digital revolution* (pp. 47-72). Springer.
[46]. Hoegen, A., Steininger, D. M., & Veit, D. (2018). How do investors decide? An interdisciplinary review of decision-making in crowdfunding. *Electronic Markets*, *28*(3), 339-365.
[47]. Hoel, C.-J., Wolff, K., & Laine, L. (2018). Automated speed and lane change decision making using deep reinforcement learning. 2018 21st international conference on intelligent transportation systems (ITSC),
[48]. Hollin, I. L., Craig, B. M., Coast, J., Beusterien, K., Vass, C., DiSantostefano, R., & Peay, H. (2020). Reporting Formative Qualitative Research to Support the Development of Quantitative Preference Study Protocols and Corresponding Survey Instruments: Guidelines for Authors and Reviewers: I. L Hollin et al. *The Patient-Patient-Centered Outcomes Research*, *13*(1), 121-136.
[49]. Honarpour, A., Jusoh, A., & Md Nor, K. (2018). Total quality management, knowledge management, and innovation: an empirical study in R&D units. *Total quality management & business excellence*, *29*(7-8), 798-816.
[50]. Hong Yun, Z., Alshehri, Y., Alnazzawi, N., Ullah, I., Noor, S., & Gohar, N. (2022). A decision-support system for assessing the function of machine learning and artificial intelligence in music education for network games: Z. HongYun et al. *Soft Computing*, *26*(20), 11063-11075.
[51]. Imran, M., Bhatti, A., King, D. M., Lerch, M., Dietrich, J., Doron, G., & Manlik, K. (2022). Supervised machine learning-based decision support for signal validation classification. *Drug Safety*, *45*(5), 583-596.
[52]. Ismagilova, E., Hughes, L., Rana, N. P., & Dwivedi, Y. K. (2022). Security, privacy and risks within smart cities: Literature review and development of a smart city interaction framework. *Information Systems Frontiers*, *24*(2), 393-414.
[53]. Jahangir, S., & Md Shahab, U. (2022). A Qualitative Study of Safety Professionals' Experiences in Managing Chemical Exposure Risks and Hazardous Materials Controls in Industrial Facilities. *Review of Applied Science and Technology*, *1*(04), 250–282. https://doi.org/10.63125/jmh69r20
[54]. Jarabek, A. M., & Hines, D. E. (2019). Mechanistic integration of exposure and effects: advances to apply systems toxicology in support of regulatory decision-making. *Current Opinion in Toxicology*, *16*, 83-92.
[55]. Jia, Y., McDermid, J., Lawton, T., & Habli, I. (2022). The role of explainability in assuring safety of machine learning in healthcare. *IEEE Transactions on Emerging Topics in Computing*, *10*(4), 1746-1760.
[56]. Jomthanachai, S., Wong, W.-P., & Lim, C.-P. (2021). An application of data envelopment analysis and machine learning approach to risk management. *Ieee Access*, *9*, 85978-85994.
[57]. Kailkhura, B., Gallagher, B., Kim, S., Hiszpanski, A., & Han, T. Y.-J. (2019). Reliable and explainable machine-learning methods for accelerated material discovery. *npj Computational Materials*, *5*(1), 108.
[58]. Khan, A., Vibhute, A. D., Mali, S., & Patil, C. H. (2022). A systematic review on hyperspectral imaging technology with a machine and deep learning methodology for agricultural applications. *Ecological Informatics*, *69*, 101678.
[59]. Kim, W., Kim, N., Lyons, J. B., & Nam, C. S. (2020). Factors affecting trust in high-vulnerability human-robot interaction contexts: A structural equation modelling approach. *Applied ergonomics*, *85*, 103056.
[60]. Klauschen, F., Müller, K.-R., Binder, A., Bockmayr, M., Hägele, M., Seegerer, P., Wienert, S., Pruneri, G., De Maria, S., & Badve, S. (2018). Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning. Seminars in cancer biology,
[61]. Klobas, J. E., McGill, T., & Wang, X. (2019). How perceived security risk affects intention to use smart home devices: A reasoned action explanation. *Computers & Security*, *87*, 101571.
[62]. Kovalchuk, S. V., Kopanitsa, G. D., Derevitskii, I. V., Matveev, G. A., & Savitskaya, D. A. (2022). Three-stage intelligent support of clinical decision making for higher trust, validity, and explainability. *Journal of biomedical informatics*, *127*, 104013.
[63]. Kumar, V., Borah, S. B., Sharma, A., & Akella, L. Y. (2021). Chief marketing officers' discretion and firms' internationalization: An empirical investigation. *Journal of International Business Studies*, *52*(3), 363-387.
[64]. Kuziemski, M., & Misuraca, G. (2020). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications policy*, *44*(6), 101976.

[65]. Lee, J. S., Ham, Y., Park, H., & Kim, J. (2022). Challenges, tasks, and opportunities in teleoperation of excavator toward human-in-the-loop construction automation. *Automation in Construction*, *135*, 104119.
[66]. Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., & Kasper, G. (2022). The challenges of algorithm-based HR decision-making for personal integrity. In *Business and the ethical implications of technology* (pp. 71-86). Springer.
[67]. Lempert, R. J. (2019). Robust decision making (RDM). In *Decision making under deep uncertainty: From theory to practice* (pp. 23-51). Springer.
[68]. Li, H., Yoo, S., & Kettinger, W. J. (2021). The roles of IT strategies and security investments in reducing organizational security breaches. *Journal of Management Information Systems*, *38*(1), 222-245.
[69]. Liu, F., & Panagiotakos, D. (2022). Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Medical Research Methodology*, *22*(1), 287.
[70]. Lung-Guang, N. (2019). Decision-making determinants of students participating in MOOCs: Merging the theory of planned behavior and self-regulated learning model. *Computers & Education*, *134*, 50-62.
[71]. Ma, L.-L., Wang, Y.-Y., Yang, Z.-H., Huang, D., Weng, H., & Zeng, X.-T. (2020). Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? *Military medical research*, *7*(1), 7.
[72]. Maleki, F., Muthukrishnan, N., Ovens, K., Reinhold, C., & Forghani, R. (2020). Machine learning algorithm validation: from essentials to advanced applications and implications for regulatory certification and deployment. *Neuroimaging Clinics of North America*, *30*(4), 433-445.
[73]. Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics*, *113*, 103655.
[74]. Mashrur, A., Luo, W., Zaidi, N. A., & Robles-Kelly, A. (2020). Machine learning for financial risk management: a survey. *Ieee Access*, *8*, 203203-203223.
[75]. Miles, J., Turner, J., Jacques, R., Williams, J., & Mason, S. (2020). Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review. *Diagnostic and prognostic research*, *4*(1), 16.
[76]. Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, *55*(5), 3503-3568.
[77]. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature machine intelligence*, *1*(11), 501-507.
[78]. Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics*, *27*(4), 44.
[79]. Müller, H., Holzinger, A., Plass, M., Brcic, L., Stumptner, C., & Zatloukal, K. (2022). Explainability and causability for artificial intelligence-supported medical image analysis in the context of the European In Vitro Diagnostic Regulation. *New Biotechnology*, *70*, 67-72.
[80]. Musen, M. A., Middleton, B., & Greenes, R. A. (2021). Clinical decision-support systems. In *Biomedical informatics: computer applications in health care and biomedicine* (pp. 795-840). Springer.
[81]. Olowononi, F. O., Rawat, D. B., & Liu, C. (2020). Resilient machine learning for networked cyber physical systems: A survey for machine learning security to securing machine learning for CPS. *IEEE Communications Surveys & Tutorials*, *23*(1), 524-552.
[82]. Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). Sok: Security and privacy in machine learning. 2018 IEEE European symposium on security and privacy (EuroS&P),
[83]. Parish, S. T., Aschner, M., Casey, W., Corvaro, M., Embry, M. R., Fitzpatrick, S., Kidd, D., Kleinstreuer, N. C., Lima, B. S., & Settivari, R. S. (2020). An evaluation framework for new approach methodologies (NAMs) for human health safety assessment. *Regulatory Toxicology and Pharmacology*, *112*, 104592.
[84]. Pesapane, F., Volonté, C., Codari, M., & Sardanelli, F. (2018). Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights into imaging*, *9*(5), 745-753.
[85]. Petrovics, D., Giezen, M., & Huitema, D. (2022). Towards a deeper understanding of up-scaling in socio-technical transitions: The case of energy communities. *Energy Research & Social Science*, *94*, 102860.
[86]. Poth, A., Meyer, B., Schlicht, P., & Riel, A. (2020). Quality Assurance for Machine Learning–an approach to function and system safeguarding. 2020 IEEE 20th international conference on software quality, reliability and security (QRS),
[87]. Pugliese, R., Regondi, S., & Marini, R. (2021). Machine learning-based approach: Global trends, research directions, and regulatory standpoints. *Data science and management*, *4*, 19-29.
[88]. Rahmani, A. M., Yousefpoor, E., Yousefpoor, M. S., Mehmood, Z., Haider, A., Hosseinzadeh, M., & Ali Naqvi, R. (2021). Machine learning (ML) in medicine: review, applications, and challenges. *Mathematics*, *9*(22), 2970.
[89]. Ratul, D. (2022). Engineering Resilient Flood Mitigation Using Geosynthetic and Composite Barrier Materials Performance Modeling and Environmental Impact Assessment. *Review of Applied Science and Technology*, *1*(03), 100–148. https://doi.org/10.63125/052q7d44
[90]. Ratul, D., & Subrato, S. (2022). Remote Sensing Based Integrity Assessment of Infrastructure Corridors Using Spectral Anomaly Detection and Material Degradation Signatures. *American Journal of Interdisciplinary Studies*, *3*(04), 332-364. https://doi.org/10.63125/1sdhwn89
[91]. Rauf, M. A. (2018). A needs assessment approach to english for specific purposes (ESP) based syllabus design in Bangladesh vocational and technical education (BVTE). *International Journal of Educational Best Practices*, *2*(2), 18-25.

[92]. Roque, A., & Damodaran, S. K. (2022). Explainable AI for security of human-interactive robots. *International Journal of Human–Computer Interaction*, *38*(18-20), 1789-1807.
[93]. Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *Ieee Access*, *8*, 42200-42216.
[94]. Rout, J. K., Rout, M., & Das, H. (2020). *Machine learning for intelligent decision science*. Springer.
[95]. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206-215.
[96]. Saberi-Karimian, M., Khorasanchi, Z., Ghazizadeh, H., Tayefi, M., Saffar, S., Ferns, G. A., & Ghayour-Mobarhan, M. (2021). Potential value and impact of data mining and machine learning in clinical diagnostics. *Critical reviews in clinical laboratory sciences*, *58*(4), 275-296.
[97]. Sanders, R. A., & Crozier, K. (2018). How do informal information sources influence women's decision-making for birth? A meta-synthesis of qualitative studies. *BMC pregnancy and childbirth*, *18*(1), 21.
[98]. Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), 1-21.
[99]. Settembre-Blundo, D., González-Sánchez, R., Medina-Salgado, S., & García-Muiña, F. E. (2021). Flexibility and resilience in corporate decision making: a new sustainability-based risk management system in uncertain times. *Global Journal of Flexible Systems Management*, *22*(Suppl 2), 107-132.
[100]. Siddique, T., Mahmud, M. S., Keesee, A. M., Ngwira, C. M., & Connor, H. (2022). A survey of uncertainty quantification in machine learning for space weather prediction. *Geosciences*, *12*(1), 27.
[101]. Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K.-R. (2021). Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology. *Machine learning and knowledge extraction*, *3*(2), 392-413.
[102]. Swathy, M., & Saruladha, K. (2022). A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques. *ICT express*, *8*(1), 109-116.
[103]. Tahmina Akter Bhuya, M., & Rebeka, S. (2022). AI-Assisted Underwriting Models for Improving Risk Assessment Accuracy in U.S. Insurance Markets. *American Journal of Interdisciplinary Studies*, *3*(01), 65-102. https://doi.org/10.63125/kegg1076
[104]. Thrun, M. C. (2022). Identification of explainable structures in data with a human-in-the-loop. *KI-Künstliche Intelligenz*, *36*(3), 297-301.
[105]. Treiss, A., Walk, J., & Kühl, N. (2020). An uncertainty-based human-in-the-loop system for industrial tool wear analysis. Joint European Conference on Machine Learning and Knowledge Discovery in Databases,
[106]. Trunk, A., Birkel, H., & Hartmann, E. (2020). On the current state of combining human and artificial intelligence for strategic organizational decision making. *Business Research*, *13*(3), 875-919.
[107]. Tuppad, A., & Patil, S. D. (2022). Machine learning for diabetes clinical decision support: a review. *Advances in Computational Intelligence*, *2*(2), 22.
[108]. Valente, F., Paredes, S., Henriques, J., Rocha, T., de Carvalho, P., & Morais, J. (2022). Interpretability, personalization and reliability of a machine learning based clinical decision support system. *Data Mining and Knowledge Discovery*, *36*(3), 1140-1173.
[109]. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., & Spitzer, M. (2019). Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, *18*(6), 463-477.
[110]. Van De Sande, D., Van Genderen, M. E., Huiskens, J., Gommers, D., & Van Bommel, J. (2021). Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. *Intensive care medicine*, *47*(7), 750-760.
[111]. Van der Schaar, M., Alaa, A. M., Floto, A., Gimson, A., Scholtes, S., Wood, A., McKinney, E., Jarrett, D., Lio, P., & Ercole, A. (2021). How artificial intelligence and machine learning can help healthcare systems respond to COVID-19. *Machine Learning*, *110*(1), 1-14.
[112]. Van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, *144*, 93-106.
[113]. Varona, D., & Suárez, J. L. (2022). Discrimination, bias, fairness, and trustworthy AI. *Applied Sciences*, *12*(12), 5826.
[114]. Vigano, L., & Magazzeni, D. (2020). Explainable security. 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW),
[115]. Walsh, S. E., & Feigh, K. M. (2021). Differentiating 'human in the loop'decision process. 2021 ieee international conference on systems, man, and cybernetics (smc),
[116]. Wang, S. V., Pottegård, A., Crown, W., Arlett, P., Ashcroft, D. M., Benchimol, E. I., Berger, M. L., Crane, G., Goettsch, W., & Hua, W. (2022). HARmonized protocol template to enhance reproducibility of hypothesis evaluating real-world evidence studies on treatment effects: a good practices report of a joint ISPE/ISPOR task force. *Value in Health*, *25*(10), 1663-1672.
[117]. Ward, F. R., & Habli, I. (2020). An assurance case pattern for the interpretability of machine learning in safety-critical systems. International Conference on Computer Safety, Reliability, and Security,
[118]. Warnecke, A., Arp, D., Wressnegger, C., & Rieck, K. (2020). Evaluating explanation methods for deep learning in security. 2020 IEEE european symposium on security and privacy (EuroS&P),

[119]. Weissler, E. H., Naumann, T., Andersson, T., Ranganath, R., Elemento, O., Luo, Y., Freitag, D. F., Benoit, J., Hughes, M. C., & Khan, F. (2021). The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*, *22*(1), 537.

[120]. Ye, Y., Zhang, X., & Sun, J. (2019). Automated vehicle's behavior decision making using deep reinforcement learning and high-fidelity simulation environment. *Transportation Research Part C: Emerging Technologies*, *107*, 155-170.

[121]. Zahabi, M., & Kaber, D. (2019). A fuzzy system hazard analysis approach for human-in-the-loop systems. *Safety Science*, *120*, 922-931.

[122]. Zeng, X., Song, F., Li, Z., Chusap, K., & Liu, C. (2021). Human-in-the-loop model explanation via verbatim boundary identification in generated neighborhoods. International Cross-Domain Conference for Machine Learning and Knowledge Extraction,

[123]. Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: is there a double standard? *Philosophy & technology*, *32*(4), 661-683.

[124]. Zhao, Z., Xu, P., Scheidegger, C., & Ren, L. (2021). Human-in-the-loop extraction of interpretable concepts in deep learning models. *IEEE Transactions on Visualization and Computer Graphics*, *28*(1), 780-790.

[125]. Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, *10*(5), 593.

[126]. Zhou, Q., Chen, Z.-h., Cao, Y.-h., & Peng, S. (2021). Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. *NPJ digital medicine*, *4*(1), 154.

[127]. Zwilling, M., Klien, G., Lesjak, D., Wiechetek, Ł., Cetin, F., & Basim, H. N. (2022). Cyber security awareness, knowledge and behavior: A comparative study. *Journal of Computer Information Systems*, *62*(1), 82-97.