From Black Box to Glass Box: Implementing Explainable AI in Insurance Underwriting

Jalees Ahmad

jaleesahmad07@gmail.com

Abstract:
The insurance industry is currently navigating a profound structural transformation, moving away from historical, manual-intensive underwriting toward a digital-first paradigm powered by artificial intelligence and high-dimensional data analytics. While machine learning (ML) architecture specifically deep learning and ensemble methodologies—have demonstrated the ability to reduce operational costs by up to 50% and quote cycle times by 90%, their inherent opacity presents a formidable barrier to full-scale adoption. This white paper examines the transition from "black box" systems to "glass box" architectures through the systematic implementation of Explainable AI (XAI) frameworks. By synthesizing current research on local and global interpretability techniques, such as SHAP and LIME, this analysis evaluates the technological mechanisms required to satisfy rigorous global regulatory standards, including the EU AI Act and the NAIC Principles on AI. The study further explores the necessity of human-in-the-loop (HITL) governance to reconcile the tensions between predictive accuracy and ethical accountability. Ultimately, the findings suggest that the long-term viability of AI in insurance underwriting is predicated on achieving a "predict and prevent" model that prioritizes transparency as highly as accuracy, thereby fostering institutional trust and consumer confidence in the modern algorithmic economy.

Keywords: Explainable AI (XAI), Insurance Underwriting, SHAP, LIME, Regulatory Compliance, Human-in-the-Loop (HITL), Risk Assessment, Model Transparency, Algorithmic Fairness.

INTRODUCTION

Insurance underwriting serves as the fundamental mechanism for risk selection and pricing, historically relying on human expertise, deterministic rulesets, and simple linear statistical models. However, the contemporary landscape is defined by a data explosion—ranging from granular electronic health records (EHR) and real-time telematics to vast geospatial datasets—that traditional methods are ill-equipped to process with the necessary speed and precision. The integration of artificial intelligence (AI) has emerged as the essential solution to this "time tax," shifting underwriting from a reactive document-centric assessment to a proactive, data-driven prediction.

Despite the clear efficiency gains, the adoption of advanced machine learning models has introduced a significant interpretability crisis. Complex models, such as gradient-boosted decision trees (GBDT) and deep neural networks, often operate as "black boxes" where the logic behind a specific decision is hidden within thousands of non-linear parameters. This opacity creates substantial "examination risk," where insurers find it difficult to justify adverse underwriting decisions—such as the denial of a policy or a significant premium increase—to regulators, auditors, and consumers.

The transition to "glass box" machine learning represents a strategic pivot toward transparency, traceability, and accountability. Explainable AI (XAI) provides the mathematical and operational toolkit to deconstruct these complex models, offering insights into why a specific decision was reached. This shift is not merely a technical preference but a regulatory and ethical imperative. With the arrival of the EU AI Act and updated NAIC principles, the ability to provide "meaningful explanations" for AI-driven decisions is becoming a legal requirement for market access.

This paper provides an exhaustive exploration of the implementation of XAI in insurance underwriting. It analyzes technical methodologies, such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations), and their application across diverse insurance domains including health, life, and property and casualty (P&C). Furthermore, it investigates the role of human-in-the-loop governance in ensuring that automated systems remain anchored in ethical judgment and professional accountability. By moving from the darkness of the black box to the clarity of the glass box, the insurance industry can leverage the full potential of AI while maintaining the trust and transparency essential to its societal function.

The Evolution of Underwriting Architecture: From Reactive to Proactive

Traditional underwriting processes have historically been characterized by slow, manual, and document-heavy workflows. Underwriters frequently spend up to 40% of their time on low-value tasks, resulting in protracted cycle times that can stretch from weeks to months for complex cases. This inefficiency not only incurs high operational costs but also leads to "premium leakage," estimated at between 3% and 10%, and a general inability to respond to market dynamics in real time.

The integration of AI transforms this architecture into a multi-layered, proactive system. The modern AI-driven underwriting engine typically combines several key technologies into a unified framework:

Architecture Layer	Technology Employed	Operational Impact
Data Ingestion	Intelligent Document Processing (IDP)	Automates the extraction of data from unstructured forms and EHRs.
Business Logic	Deterministic Rules Engines	Ensures compliance with base-level underwriting guidelines.
Risk Scoring	Probabilistic ML (XGBoost, Random Forest)	Identifies non-linear patterns and interactions for precise risk stratification.
Decision Refinement	Agentic AI (Multi-LLM Orchestration)	Provides reasoning and context for complex, non-standard risks.

This shift significantly reduces the "time tax" associated with traditional methods. Research indicates that AI-enabled systems can reduce quote times by approximately 90% and operational costs by roughly 50%. In specific sectors such as health insurance, AI systems have demonstrated an accurate improvement of 30-50% over manual underwriting by utilizing large volumes of health data, including lifestyle indicators and genetic history, to provide more granular risk assessments.

However, as these models move beyond simple linear regression, they capture sophisticated relationships that are difficult for human experts to audit. For example, a deep learning model may find that a specific combination of geographic, behavioral, and medical variables indicates a high mortality risk, but without XAI, it cannot explain *which* factor was the primary driver. This lack of transparency leads to the "black box" problem, where the institution gains accuracy at the cost of explainability, potentially violating the consumer's "right to explanation" under frameworks like the GDPR.

Transparency Imperative: Regulatory and Ethical Drivers

The necessity of moving to a "glass box" model is driven by a convergence of regulatory pressure and the need to maintain consumer trust. Insurance is a socially critical product, and the rationale behind its availability and pricing must be defensible.

Global Regulatory Dynamics

Regulators across the globe are increasingly focused on the governance of high-risk AI systems. The European Union's AI Act is the most comprehensive example, identifying AI systems used for risk assessment and pricing in life and health insurance as "high-risk". Under Article 13 of the Act, these systems must be designed for sufficient transparency, allowing deployers to interpret outputs and providing them with clear "instructions for use" that detail the system's capabilities and limitations.

In the United States, the National Association of Insurance Commissioners (NAIC) has established principles emphasizing that AI systems must be fair, transparent, and accountable. NAIC Model Law #670 explicitly defines "adverse underwriting decisions"—such as declination or termination—and mandates that insurers provide consumers with specific reasons for such actions. Furthermore, the NAIC focuses on "unfair discrimination," requiring insurers to ensure that predictive models do not use variables that serve as proxies for protected classes, such as race or gender.

Regulatory Framework	Core Requirement	Impact on Underwriting
EU AI Act	Transparency and traceability for high-risk systems.	Mandates technical documentation and human oversight for pricing models.
NAIC Principles	Fairness and actuarial soundness.	Requires rational explanations for correlations between data and risk.
GDPR (Article 22)	Right to human intervention and explanation.	Grants consumers the right to contest automated underwriting decisions.
UK FCA Consumer Duty	Delivery of good outcomes for customers.	Requires that AI does not embed or amplify biases leading to consumer harm.

Building Stakeholder Trust

Beyond compliance, transparency is a strategic asset. Research by FICO indicates that 65% of consumers would trust AI more if the decision-making process was understandable. In high-stakes environments like credit scoring and health insurance, "black box" models can alienate customers, particularly if they perceive a decision as arbitrary or biased. Explainable AI allows insurers to bridge this communication gap, providing "360-degree explanations" that serve different stakeholders: global explanations for regulators, local feature-based explanations for underwriters, and instance-based reasoning for consumers.

Methodologies for Glass Box Implementation: XAI Techniques

To transform an opaque model into a transparent one, researchers have developed various XAI techniques that interrogate the model's logic. These are generally classified into intrinsic (ante-hoc) and post-hoc methods.

Intrinsic vs. Post-hoc Interpretability

Intrinsic models are "glass box" by design. These includes decision trees, linear regression, and rule-based systems where the architecture is simple enough to be followed by a human. However, these models often lack the predictive power of more complex architectures. To overcome this, the industry frequently uses post-hoc methods, which are applied to a "black box" model after training to extract explanations.

Local Interpretable Model-agnostic Explanations (LIME)

LIME is a widely adopted technique for generating local explanations. It works by perturbing the input data around a specific prediction and observing how the output changes. It then fits a simpler, surrogate model

(like a linear regression) to these perturbations to identify the features that were most influential for that specific decision. While useful for providing immediate reasoning for a single case, LIME can be unstable and less robust than game-theoretic approaches.

Shapley Additive Explanations (SHAP)

SHAP is the current state-of-the-art in XAI for insurance and finance. It is based on cooperative game theory, assigning a "Shapley value" to each feature that represents its fair contribution to the final prediction. The mathematical foundation of SHAP ensures consistency and accuracy in variable attribution.

The model prediction $f(x^*)$ is decomposed as follows:

$$f(x^*) = V_0 + \sum_{\{j=1\}}^{p} V_j(f, x^*)$$

where V_0 is the average model response and $V_j(f, x^*)$ is the contribution of the j^{th} variable for that specific instance.

In the context of actuarial science, standard SHAP's additive nature can conflict with the multiplicative risk relationships typically found in insurance pricing. To address this, "Multiplicative SHAP" has been developed, transforming predictions into logarithmic space:

$$\ln(\text{Premium}) = \ln(\text{Baseline}) + \ln\sum(\phi_j)$$

This allows for the exponentiation of contributions back into traditional rating factors, where a factor $\phi > 1$ represents a premium increase and $\phi < 1$ represents a discount. This framework bridges the gap between sophisticated ML and traditional Generalized Linear Models (GLMs).

Explainable Boosting Machines (EBMs)

EBMs represent a specialized class of "glass box" models that utilize gradient boosting and automatic interaction detection to achieve the accuracy of Random Forests while remaining fully interpretable. EBMs are additive models where each feature's contribution can be visualized on a graph, allowing domain experts to edit the model's logic directly if it discovers a spurious correlation.

XAI Technique	Type	Primary Benefit	Technical Basis
Decision Trees	Intrinsic	Natural visual structure; easy to follow.	Rule-based branching.
LIME	Post-hoc	Model-agnostic; good for individual cases.	Local surrogate approximation.
SHAP	Post-hoc	Mathematically rigorous; consistent.	Cooperative Game Theory.
Multi-SHAP	Post-hoc	Aligns with actuarial rating factors.	Logarithmic space additive decomposition.
EBM	Intrinsic	Combines accuracy with full editability.	Generalized Additive Models (GAM).

Domain-Specific Applications: Case Studies in Underwriting

The application of XAI varies significantly across different insurance domains, each presenting unique data challenges and ethical considerations.

Health Insurance and Disease Prediction

In health insurance, XAI is used to stratify risk by predicting medical conditions like diabetes, heart disease, and anemia. Hybrid frameworks combining ensemble models (Random Forest, XGBoost) with SHAP and LIME have shown that transparency can be achieved without compromising diagnostic accuracy. For example, in a study of 2.9 million claims, the "Deep Claim" framework demonstrated that a white-box algorithm (AdaBoost) could yield an AUC of 0.83, helping hospitals and insurers identify potentially denied claims before they are filed. This proactive approach reduces operational friction and improves the patient's financial experience.

Life Insurance and Mortality Risk

Life insurance underwriting is increasingly utilizing unstructured data, such as doctor's notes and social media, via Natural Language Processing (NLP). XAI methodologies, including rule extraction and natural language generation, are being integrated into these systems to provide comprehensive understandings of mortality risk. This transformation allows insurers to move from fixed risk assessment to modern, dynamic risk understanding tools that can process real-time wearable data to encourage healthier lifestyles.

Property, Casualty, and Corporate Bonds

In P&C insurance, XAI is critical for maintaining the "actuarial justification" required for price changes. By using permutation feature importance and partial dependence plots (PDPs), actuaries can understand the "shape" of a model's logic—ensuring, for instance, that risk increases linearly with a variable like vehicle power or property age as expected. In corporate bond pricing, "glass box" ML has been used to uncover non-linear interactions related to macroeconomic uncertainty, achieving state-of-the-art performance while remaining fully interpretable for portfolio managers.

Human-in-the-Loop (HITL): The Governance Paradigm

A common misconception is that AI explainability is intended only for external regulators. In practice, the primary beneficiary is the human underwriter. The "human-in-the-loop" model ensures that AI handles the "heavy lifting of data extraction" while humans apply judgment to nuanced or high-stakes cases.

Workflow and Quality Assurance

Effective HITL systems are designed to validate AI outputs *before* they impact the customer. For instance, in ACORD form automation, AI extracts data points, but an expert reviewer validates and normalizes that data to ensure 100% accuracy. This collaborative model can achieve a 5x improvement in processing speed.

Stage of HITL Workflow	AI Action	Human Intervention
Submission	Ingest data from emails/S3 buckets.	Monitors ingestion for missing documents.
Extraction	Identifies key data points in documents.	Validates accuracy of critical fields (e.g., coverage limits).
Reconciliation	Cross-references conflicting data points.	Resolves inconsistencies using domain expertise.
Decisioning	Generates risk score and recommendations.	Overrides anomalies and apply subjective judgment.

Continuous Learning and Accountability

Each human intervention serves as a training signal for AI. By correcting errors or validating complex edge cases, underwriters contribute to an iterative process where the model becomes smarter over time. Furthermore, HITL is essential for accountability. Under governance frameworks like SM&CR, a senior

manager must be responsible for AI-driven decisions; XAI provides the documentation and reasoning necessary for that manager to fulfill their legal obligations.

Ethical Challenges and Future Outlook

The transition to glass box AI does not eliminate all risks. New challenges are emerging as models become more sophisticated.

Proxy Discrimination and Fairness

One of the most significant concerns is "proxy discrimination," where a model avoids using a protected characteristic (like race) but utilizes another variable that is highly correlated with it (like zip code or purchasing history). XAI tools like SHAP are now being used alongside fairness metrics—such as demographic parity and equalized odds—to detect and mitigate these biases. By highlighting the demographic effects in model outcomes, XAI allows compliance officers to question and correct discriminatory patterns.

The Move to "Predict and Prevent"

Looking toward 2030, the insurance industry is expected to fully transition from a "detect and repair" model to a "predict and prevent" paradigm. This will be driven by the proliferation of connected devices and real-time risk assessments. In the future, XAI will be the cornerstone of customer service, providing personalized recommendations that help policyholders reduce their own risk, thereby lowering premiums and claims costs for the entire system.

Conclusion

The implementation of Explainable AI is the defining challenge for the next generation of insurance underwriting. While the efficiency gains of "black box" machine learning are undeniable, the operational and regulatory risks of opacity are too high to ignore. By adopting "glass box" architectures—leveraging SHAP, LIME, and EBMs—insurers can reconcile the tension between accuracy and interpretability.

The integration of XAI enables a robust governance framework that prioritizes actuarial soundness, ethical fairness, and regulatory compliance. Moreover, by keeping the human in the loop, insurers can ensure that AI augments professional expertise rather than replacing it. Ultimately, transparency is not a hurdle to innovation but its facilitator; only through clear, explainable, and accountable systems can the insurance industry maintain the trust necessary to navigate the complex risk landscape of the twenty-first century. As regulators mandate more rigorous oversight and consumers demand greater clarity, the transition from the black box to the glass box is not merely an option, it is a strategic imperative for the future of underwriting.

REFERENCES:

1. Vel, D. V. T., & Durgaraju, S. (2024). Explainable AI (XAI) for Health Insurance Underwriting. *International Journal of Communication Networks and Information Security*, 15(1), 285–306.
2. Gummadi, H. S. B. (2025). Artificial Intelligence in Life Insurance Underwriting: A Risk Assessment and Ethical Implications. *International Journal of Management Technology*, 12(1), 36–49.
3. Kumar, R. (2024). AI-Driven Transformation in Insurance Underwriting: Architecture, Efficiency, and Regulatory Alignment. *SSRN Electronic Journal*.
4. Gummadi, H. S. B. (2025). Explainable AI-Enhanced Underwriting Automation for Personalized Insurance Policy Recommendations. *European Journal of Computer Science and Information Technology*, 13(19), 24–40.
5. Kuo, K., & Lupton, D. (2023). Towards Explainability of Machine Learning Models in Insurance Pricing. *Variance Journal*, 16(1).
6. Gabelaia et al. (2024). Technology Acceptance Model in Risk-Sensitive Environments. *Journal of Insurance Regulation*.
7. National Association of Insurance Commissioners (NAIC). (2024). Insurance Information and Privacy Protection Model Act (Model #670).
8. MirrorWeb. (2025). Glass Box vs. Black Box AI: Why Compliance Teams Need Explainable Systems.

9. Shapley, L. S. (1953). A Value for n-person Games. *Contributions to the Theory of Games*, 2(28), 307–317.
10. Huang, H., & Huang, Z. (2023). SHAP for Transparency in Machine Learning Models. *Journal of Risk Management*.
11. European Parliament and Council. (2024). Artificial Intelligence Act (Regulation (EU) 2024/1689).
12. NAIC. (2020). Principles on Artificial Intelligence.
13. Wilson, C. A. (2025). Explainable AI in Finance: Addressing the Needs of Diverse Stakeholders. *CFA Institute Research Foundation*.
14. Bell, S., Kakhbod, A., Lettau, M., & Nazemi, A. (2024). Glass Box Machine Learning and Corporate Bond Returns. *NBER Working Paper 33320*.
15. Lee et al. (2021). Improvements in Health Insurance Processing using AI. *Health Systems Journal*.
16. Charpentier, A. (2024). Ethical Concerns in Opaque AI Risk Assessment. *Insurance Research Journal*.
17. GDPR Article 22. Automated individual decision-making, including profiling.
18. Valdrighi et al. (2025). Explainability in Credit Scoring: Addressing Discriminatory Patterns. *Journal of Financial Data Science*.
19. EU AI Act Article 13. Transparency and provision of information to deployers.
20. Annex III of the EU AI Act. High-risk AI systems.
21. Tu, J., & Wu, Z. (2025). Inherently interpretable machine learning for credit scoring. *European Journal of Operational Research*, 322(2), 647–664.
22. FICO. (2018). Consumer Trust in AI Report.
23. Bartlet et al. (2022). Socioeconomic Factors in Credit Scoring. *Review of Financial Studies*.
24. Fisher, A., Rudin, C., & Dominici, F. (2018). Model Class Reliance: Variable Importance for Any Machine Learning Model.
25. Actuaries Institute. (2025). Multiplicative SHAP: Bridging the Gap between Machine Learning and Actuarial Science.
26. Ortmann, K. M. (2016). Multiplicative Cooperative Games.
27. Microsoft Research. (2024). InterpretML: Explainable Boosting Machines (EBM).
28. Vel, D. V. T., & Durgaraju, S. (2024). Hybrid ML-XAI framework for predicting diseases. *International Journal of Communication Networks and Information Security*.
29. Deep Claim Framework. (2025). Assessment of healthcare claims rejection risk using machine learning.
30. AdaBoost for Claims Denials. (2025). *Journal of Healthcare Management*.
31. Dynamic Risk Assessment in Life Insurance. (2025). *International Journal of Management Technology*.
32. McKinsey & Company. (2022). AI-driven Transformation in Insurance.
33. IntellectAI. (2024). Why Human-in-the-Loop is Essential in Insurance AI.
34. Inube Solutions. (2025). Collaborative Human-AI Models in Underwriting.
35. ACORD Form Automation and Human Validation. (2025). *IntellectAI White Paper*.
36. Starmind Research. (2024). Continuous Model Training via Human Feedback.
37. UK Financial Conduct Authority (FCA). (2025). Senior Managers and Certification Regime (SM&CR) for AI.
38. ResearchGate. (2025). Financial Explainable AI for Credit Card Scoring: Fairness Challenges and Case Study.
39. PMC Scoping Review. (2025). Insurance 2030: The Impact of AI on the Future of Insurance.