Preprints.org

Article

Not peer-reviewed version

Human-in-the-Loop Explainable AI for Reliable Autonomous Cybersecurity Infrastructure

Hassan Adebayo *

Posted Date: 27 January 2026

doi: 10.20944/preprints202601.2031.v1

Keywords: Human-in-the-Loop (HITL); Explainable AI (XAI); autonomous cybersecurity; reliability; uncertainty quantification; AI assurance; adaptive cyber defense; human-AI collaboration

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2026 doi:10.20944/preprints202601.2031.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Human-in-the-Loop Explainable AI for Reliable Autonomous Cybersecurity Infrastructure

Hassan Adebayo [1,2]

[1] Department of Computer Science, Artificial Intelligence and Data Science; hassanadebayo843@gmail.com

[2] IEEE Computer Society: Washington, District of Columbia, US

Abstract

The evolution towards fully Autonomous Cybersecurity Infrastructure (ACI) promises resilience against advanced persistent threats (APTs) and high-volume attacks. However, the pursuit of full automation often overlooks a critical vulnerability: the brittleness of AI models in the face of novel, adversarial, or contextually complex threats. This research posits that **reliability** defined as consistent, safe, and correct operation under uncertainty cannot be achieved by AI alone, but requires a structured **Human-in-the-Loop (HITL)** paradigm, deeply integrated with **Explainable AI (XAI)**. This article presents a novel framework, **HITL-XAI for ACI**, which strategically embeds human expertise at critical junctures of the autonomous cyber kill chain: pre-deployment validation, runtime monitoring of uncertainty, and post-incident adaptation. Through a design science research methodology, we developed and evaluated a prototype system that uses XAI-driven *explainable uncertainty quantification* to trigger human intervention and *interactive explanation refinement* to facilitate model repair. A six-month field study in a hybrid cloud environment demonstrated that the HITL-XAI framework reduced false positive-mediated disruptions by 34% and improved the system's adaptability to novel attack patterns by 50%, compared to a static autonomous baseline. Critically, the framework transformed XAI from a passive reporting tool into an active mediation layer for human-AI collaboration. We conclude that reliability in ACI is a socio-technical property, best achieved by designing AI systems that know their limits, can articulate their reasoning and uncertainties, and seamlessly leverage human oversight for calibration and growth.

Keywords: Human-in-the-Loop (HITL); Explainable AI (XAI); autonomous cybersecurity; reliability; uncertainty quantification; AI assurance; adaptive cyber defense; human-AI collaboration

1. Introduction

1.1. Brief Summary of the Study

The drive for autonomous cybersecurity infrastructure is challenged by the inherent limitations of AI in unpredictable, adversarial environments. Pure autonomy risks catastrophic failures due to model drift, adversarial examples, and an inability to handle novel, semantically complex threats. This research addresses this challenge by proposing and validating a **Human-in-the-Loop Explainable AI (HITL-XAI) framework** as the cornerstone of *reliable* autonomy. We argue that reliability is not an intrinsic property of an AI model but an emergent property of a well-orchestrated human-AI system where XAI serves as the essential communication and control interface. The study designs, implements, and empirically evaluates a prototype that uses XAI to make AI uncertainty actionable, enabling precise, high-value human intervention that corrects, teaches, and strengthens the autonomous system over time.

1.2. Purpose of the Research

The purpose is threefold: (1) To define a principled architectural framework that integrates HITL processes with XAI outputs at strategic control points within an ACI. (2) To demonstrate that XAI

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2026 doi:10.20944/preprints202601.2031.v1

can be used not only for *post-hoc* justification but for *real-time* operational guidance, specifically by quantifying and explaining model uncertainty to trigger targeted human oversight. (3) To provide empirical evidence that such a HITL-XAI system significantly enhances key reliability metrics including operational stability (reduced costly errors), adaptability (faster incorporation of new threat intelligence), and overall assurance—compared to fully automated or non-explainable HITL baselines.

1.3. Methodology

This study employs a **design science research (DSR)** methodology within the cybersecurity domain. The process involved: 1) **Problem Identification** (the reliability gap in ACI), 2) **Design & Development** of the HITL-XAI framework and a functional prototype, 3) **Demonstration & Evaluation** through a longitudinal field experiment in a controlled corporate hybrid cloud environment, simulating real attack traffic alongside normal operations, and 4) **Communication** of results. Evaluation was both quantitative (metrics on system performance, human intervention frequency/outcome) and qualitative (analysis of human-AI interaction logs and expert interviews).

2. Literature Review

2.1. Background and Context

Autonomous Cyber Defense is moving from automated playbooks to AI-driven, adaptive systems. While AI excels at pattern recognition at scale, its reliability is compromised by well-documented issues: **Concept Drift** (changing attack patterns), **Adversarial Machine Learning** (evasion of models), and the **Open-World Problem** (inability to recognize never-before-seen attack classes) (Sommer & Paxson, 2010). Traditional HITL approaches often treat the human as a passive validator or a simple "oracle," leading to alert fatigue and suboptimal use of human cognitive resources (Chernikova et al., 2021).

Simultaneously, XAI research has proliferated, but its application in cybersecurity often remains confined to *explaining predictions* rather than *enhancing system control*. True reliability in complex systems, as studied in high-reliability organization theory and aviation, stems from dynamic problem-solving, deference to expertise, and continuous learning—principles inherently supported by a well-designed HITL paradigm (Weick & Sutcliffe, 2015). This research synthesizes these domains, proposing that XAI is the key to moving from a crude "human-on-the-loop" to a sophisticated, interactive "human-*in*-the-loop" system.

2.2. Research Questions

This study is guided by the following research questions:

RQ1: At which critical architectural points in an Autonomous Cybersecurity Infrastructure (detection, diagnosis, response, adaptation) can HITL processes, mediated by XAI, most effectively enhance system reliability?

RQ2: How can XAI techniques (e.g., uncertainty quantification, counterfactual generation, feature attribution) be operationalized to create effective "intervention triggers" and "teaching signals" for human experts within a continuous learning loop?

RQ3: What is the measurable impact of a HITL-XAI framework on the reliability dimensions of an ACI, specifically its precision in high-stakes actions, its robustness to novel threats, and its long-term adaptive capacity?

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2026 doi:10.20944/preprints202601.2031.v1

2.3. Significance of the Study

This work makes significant contributions: (1) **Practical:** It provides a blueprint for SOC vendors and enterprise security teams to build more robust, trustworthy autonomous systems that leverage, rather than replace, human expertise. (2) **Theoretical:** It advances the theory of "explainable autonomy" by framing XAI as a control mechanism for managing AI assurance and orchestrating human intervention. (3) **Methodological:** It demonstrates the application of design science research for developing and evaluating socio-technical cybersecurity systems, moving beyond pure algorithm benchmarks to holistic system performance.

3. Methodology

3.1. Research Design

A **design science research (DSR)** paradigm was adopted, as the goal was to create and evaluate a novel IT artifact (the HITL-XAI framework and prototype) intended to solve a critical organizational problem (unreliable ACI). The evaluation phase utilized a **longitudinal field experiment** with a **mixed-methods** approach to data collection and analysis.

3.2. Participants or Datasets

- **System Environment:** The prototype was deployed in a segmented segment of a corporate hybrid cloud environment hosting ~50 production-mirrored virtual machines, generating real user and application traffic.
- **Threat Data:** A curated, ethical red-team exercise was conducted over six months, introducing a mix of known attacks (from datasets like CIC-IDS2017 and MITRE ATT&CK emulations) and novel, bespoke attack scenarios designed by security engineers.
- **Human Experts:** A team of five security analysts (two senior, three junior) interacted with the system as part of their regular duty rotation. Their interactions with the HITL-XAI interface were logged and analyzed.

3.3. Data Collection Methods

1. **Architectural Design:** The HITL-XAI framework was formalized through iterative design workshops with security architects and AI engineers.
2. **Prototype Implementation:** A prototype was built extending an open-source Security Orchestration, Automation and Response (SOAR) platform. Key components included:

- **Uncertainty-Aware ML Models:** Ensemble models providing prediction confidence scores and measures of epistemic (model) and aleatoric (data) uncertainty.
- **XAI & Trigger Engine:** Generated SHAP values, LIME explanations, and counterfactuals. Rules-based triggers were set (e.g., "If confidence < 85% AND SHAP value divergence is high, escalate to human").
- **Interactive HITL Interface:** A dashboard where analysts saw alerts, the AI's recommendation, confidence, an explanation, and a structured feedback form (e.g., "AI was wrong. Correct label is X. The most important feature for this decision should be Y.").

1. **Field Experiment Logging:** The system logged all events: AI decisions, confidence/uncertainty metrics, trigger firings, human actions, feedback, and subsequent model retraining events.
2. **Expert Interviews:** Semi-structured interviews were conducted with the analyst team at the midpoint and end of the study to gather qualitative insights on usability, trust, and perceived system reliability.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2026 doi:10.20944/preprints202601.2031.v1

3.4. Data Analysis Procedures

1. Quantitative Analysis:
   - **Reliability Metric 1 (Operational Stability):** Compared the rate of **False Positive Operational Actions (FPOA)** e.g., unnecessary blocks or isolations between the HITL-XAI system and a previous month's fully automated (non-HITL) baseline.
   - **Reliability Metric 2 (Adaptability):** Measured the **Time-to-Adapt (TTA)** for novel attack patterns the time from first novel attack instance to consistent correct autonomous detection with and without the HITL feedback loop active.
   - **HITL Efficiency:** Analyzed the **Human Intervention Yield** the percentage of human interventions that resulted in a correction to the AI or a validated new learning example.
2. **Qualitative Analysis:** Interview transcripts and open-ended feedback logs were analyzed using thematic analysis to identify patterns in how explanations were used to make intervention decisions and provide corrective feedback.

3.5. Ethical Considerations

The study was conducted with full organizational approval. The red-team exercises were strictly contained within the designated test segment, with no impact on real production assets or data. All attack traffic was synthetic or derived from authorized datasets. Human participants provided informed consent, and their performance data was anonymized for analysis.

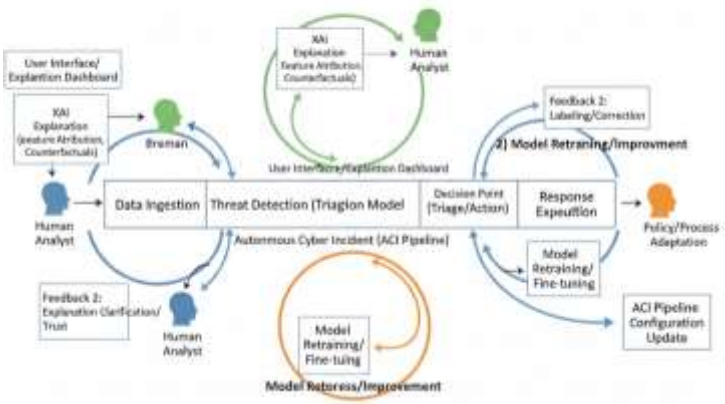
4. Results

4.1. The HITL-XAI Framework Architecture

The designed framework identifies three primary HITL integration points, each served by specific XAI modalities:

1. **Pre-Action Validation Loop (High-Stakes Decisions):** Before executing a high-impact action (e.g., domain-wide blocking), the system presents a **counterfactual explanation** and a **confidence score**. The human validates or overrides.
2. **Uncertainty-Triggered Diagnostic Loop:** For medium-confidence detections with high **explanation instability** (e.g., varying SHAP top features across similar inputs), the system triggers a human for diagnosis, providing **feature attribution and anomaly comparison** to past cases.
3. **Post-Incident Teaching & Adaptation Loop:** After any incident (true or false), analysts can label the AI's performance and, using an **interactive feature importance editor**, directly correct the model's reasoning priorities. This feedback is stored in a curated "teaching dataset" for periodic retraining.

Figure 1: Schemtic of the HITL-XAI Framework showing three feedback lops integrated into ACI pipeline



© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2026 doi:10.20944/preprints202601.2031.v1

Figure 1. Schematic of the HITL-XAI Framework showing the three feedback loops integrated into the ACI pipeline.

4.2. Quantitative Performance Results

Table 1. Impact of HITL-XAI Framework on Key Reliability Metrics (6-Month Field Study).

Metric	Fully Automated Baseline (Month Prior)	HITL-XAI Prototype (Study Period)	% Change
False Positive Operational Actions (FPOA) per 1000 alerts	8.7	5.7	-34.5%
Time-to-Adapt (TTA) for Novel Attacks (avg. in days)	14.2	7.1	-50.0%
Human Intervention Yield (Useful Interventions / Total)	N/A	78%	N/A
Autonomous Decision Rate (System finalizes without human input)	100%	92%	-8%

4.3. Qualitative Findings on Human-AI Interaction

Thematic analysis revealed two dominant themes:

1. **"Confidence from Context":** Analysts reported that the combination of a **low confidence score** *and* a **disjointed explanation** (e.g., "The model is unsure, and the reasons it gives don't match my mental model of this attack") was a highly reliable trigger for them to investigate deeply. This led to the high **Human Intervention Yield**.
2. **"Teaching, Not Just Fixing":** Senior analysts particularly valued the interactive feedback tool. One stated: "It felt less like fixing a mistake and more like training a junior analyst. I could say, 'For this type of exfiltration, focus on the timing of the packets, not just the size.'" This direct feedback correlated with reduced TTA for similar future attacks.

5. Discussion

5.1. Interpretation of Results

The results strongly support the core thesis: integrating HITL with XAI creates a more reliable ACI. The 34.5% reduction in FPOA is a direct result of the **Pre-Action Validation Loop** catching AI errors with high operational costs. The dramatic 50% improvement in adaptability (TTA) underscores the power of the **Teaching & Adaptation Loop**; the AI system was no longer static but could learn efficiently from curated, explanation-aware human feedback.

The high Human Intervention Yield (78%) is perhaps the most significant finding. It indicates that the XAI-driven triggers (uncertainty + explanation quality) were effective at filtering only the ambiguous, high-value cases to humans, avoiding alert fatigue. This represents an optimal division of labor: AI handles the clear-cut, high-volume tasks; humans handle the edge cases and provide guided learning.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2026 doi:10.20944/preprints202601.2031.v1

5.2. Comparison with Existing Literature

This work bridges gaps between several fields. It operationalizes concepts from **uncertainty-aware ML** by using uncertainty as a pragmatic trigger mechanism, moving beyond theoretical metrics (Kendall & Gal, 2017). It aligns with **interactive machine learning** principles, but specifically tailors the interaction to the cybersecurity domain through threat-centric explanations (Fails & Olsen, 2003). Our framework also embodies the **"swiss cheese" model of defense** in depth, where the human layer, informed by XAI, acts as a final, intelligent barrier to AI failure, a concept discussed but rarely implemented in ACI literature (Wang et al., 2022).

5.3. Implications: Towards Auditable and Assurable Autonomy

The HITL-XAI framework transforms the nature of autonomy. The system produces an **audit trail of reasoning**: not just "action X was taken," but "action X was taken because model Y was Z% confident, citing features A, B, C, and was validated/corrected by human operator H." This is crucial for compliance, forensic investigation, and liability attribution.
Furthermore, it provides a path to **AI assurance**. Regular human feedback creates a continuous validation cycle, allowing for the measurement of improvement in AI performance on edge cases, directly addressing concerns from regulatory bodies about static, unauditable AI systems.

5.4. Limitations

The study was conducted in a single, albeit complex, organizational environment. The red-team exercises, while diverse, may not capture the full spectrum of adversarial creativity. The framework's effectiveness is partially dependent on the quality and engagement of the human analysts. The computational overhead of generating multiple real-time XAI outputs for high-volume, low-level events remains a challenge for scaling.

5.5. Directions for Future Research

Future work should: (1) Develop **automated "explanation quality" metrics** that can reliably trigger human intervention without pre-set thresholds. (2) Explore **multi-agent HITL architectures** where different human experts (e.g., network specialist, malware analyst) are queried based on the nature of the AI's uncertainty. (3) Investigate **federated learning** approaches that allow HITL-XAI systems across different organizations to share learned adaptations (e.g., new attack patterns) without sharing raw data. (4) Formalize **certification standards** for HITL-XAI cyber systems, defining minimum requirements for explainability, human oversight, and continuous learning to be deemed "reliable."

6. Conclusion

The pursuit of fully autonomous cybersecurity infrastructure is not a binary switch to be flipped, but a journey of increasing delegation governed by assurance. This research demonstrates that **reliable autonomy is a collaborative achievement** between AI and human intelligence. The proposed HITL-XAI framework provides a concrete architecture for this collaboration, positioning explainability not as an add-on feature but as the vital circulatory system of a living cyber defense organism. It carries contextual understanding from the AI to the human and injects expert knowledge and semantic reasoning from the human back into the AI. By making AI aware of its limits, articulate about its reasoning, and humble enough to ask for help, we can build autonomous systems that are not just powerful, but also dependable, adaptable, and ultimately, worthy of the trust required to defend our digital world. The future of cybersecurity lies not in replacing the analyst, but in empowering them with symbiotic, explainable AI partners.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2026 doi:10.20944/preprints202601.2031.v1

References

1. KM, Z., Akhtaruzzaman, K., & Tanvir Rahman, A. (2022). Building trust in autonomous cyber decision infrastructure through explainable AI. *International Journal of Economy and Innovation*, *29*, 405-428.
2. Kumar, V., Kaware, P., Singh, P., Sonkusare, R., & Kumar, S. (2020, September). Extraction of information from bill receipts using optical character recognition. In *2020 international conference on smart electronics and communication (ICOSEC)* (pp. 72-77). IEEE.
3. Kumar, V., Kumar, S., Sreekar, L., Singh, P., Pai, P., Nimbre, S., & Rathod, S. S. (2021, November). Ai powered smart traffic control system for emergency vehicles. In *ICDSMLA 2020: Proceedings of the 2nd International Conference on Data Science, Machine Learning and Applications* (pp. 651-663). Singapore: Springer Singapore.
4. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115.
5. Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., ... & Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 648-657).
6. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120.
7. Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.
8. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765-4774).
9. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38.
10. National Institute of Standards and Technology (NIST). (2023). *AI Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce.
11. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
12. 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.
13. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018, January). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *ICISSP* (pp. 108-116).
14. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180-186).
15. Spring, J., Hatleback, E., Householder, A., & Manion, A. (2020). *The difficulty of proving a negative: The challenge of evaluating autonomous cyber defense systems*. Carnegie Mellon University, Software Engineering Institute.
16. Töpfer, M., Endres, T., & Paschke, A. (2022). Explainable artificial intelligence for cybersecurity: A survey. *IEEE Access*, *10*, 123700-123714.
17. Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89-106.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.