Frontiers in Computer Science and Artificial Intelligence
DOI: 10.32996/fcsai
Journal Homepage: www.al-kindipublisher.com/index.php/fcsai

FCSAI
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Ethical and Trustworthy Autonomous Agents in Network SecOps: Transparency, Auditing, and Human-in-the-Loop Overrides

Amar Gurajapu[1]✉, Swapna Anumolu[1], Vardhan Garimella[2], Venkata Manikanta Sai Ramakrishna Chundi[3], and Venkata Sita Anand Prakash Gubbala[4]

[1]*Principal Member of Tech Staff, Network Systems, AT&T, New Jersey, United States*
[2]*Consultant, Intellibus, United States*
[3]*Lead Architect, Intellibus, United States*
[4]*Vice President, Wissen Inc, United States*
Corresponding Author: Amar Gurajapu, **E-mail**: amar_p21@yahoo.com

| ABSTRACT
This paper introduces EthosSecOps, a comprehensive framework designed to enhance transparency, auditability, and ethical alignment in AI-driven intrusion detection and automated response systems. EthosSecOps integrates an Explainability Layer for generating feature-attribution explanations, a Blockchain-backed Audit Store to immutably record alerts, actions, and overrides, and a Policy-Driven Override Engine that empowers human analysts to pause, modify, or abort agent actions. Implemented within a hybrid-cloud telecom environment, EthosSecOps demonstrated 95% attack mitigation accuracy, delivered real-time explanations within 10 milliseconds, and enabled immediate human intervention without disrupting service. The paper details the system's architecture, provides a Python-based audit-logging example, presents empirical evaluation results, and discusses ethical implications for trustworthy autonomous SecOps in regulated and high-availability network operations.

| KEYWORDS
Autonomous Agents, Network Orchestration, SLA Compliance, Reinforcement Learning, QoS Monitoring, Software-Defined Networking, Multi-Cloud, SecOps, Cybersecurity

| ARTICLE INFORMATION

ACCEPTED: 14 Feb 2025 **PUBLISHED:** 19 Feb 2025 **DOI:** 10.32996/jcsts.2025.4.2.7

1. INTRODUCTION

The transition to AI-driven SecOps pipelines has facilitated real-time detection of threats and the automation of response measures at scale. Nonetheless, reliance on opaque “black box” models may diminish operator confidence, obscure erroneous activities, and complicate compliance audits, concerns that are particularly relevant in telecommunications networks where automated updates to firewalls, traffic redirection, or VM quarantines have the potential to impact millions of subscribers. Ethical AI frameworks (Jobin et al., 2019) underscore the importance of transparency, accountability, and human oversight. However, current SIEM and SOAR deployments seldom offer comprehensive explanations or secure, tamper-proof audit records for each operational decision. Moreover, fully automated actions without human intervention introduce risks such as unintended outcomes, bias, or violation of established policies.

EthosSecOps provides a solution to these challenges by integrating explainable-AI methodologies, blockchain-enabled audit logging, and mechanisms for human intervention within autonomous SecOps agents. This combined strategy delivers swift automated mitigation while maintaining transparency, traceability, and managerial oversight, which are crucial for regulated environments and networks requiring high availability.

2. LITERATURE REVIEW

Explainable AI in Security. SHAP (Lundberg & Lee, 2020) and LIME (Ribeiro et al., 2016) provide model-agnostic explanations, with applications in intrusion detection (Chen et al., 2022). However, few systems generate explanations in real time or integrate them into automated remediation workflows.

Blockchain-Based Audit Logging. Zhang and Xue (2021) demonstrated on-chain identity management for cloud IAM but did not address per-alert audit trails. Patel et al. (2023) proposed hybrid on-chain/off-chain logging to balance performance and immutability, yet their design lacked human-override integration.

Human-in-the-Loop Automation. Keshavjee and Smith (2019) describe semi-automated incident response workflows with manual gating, while Mao et al. (2021) evaluated override latency and trust in "pause-and-explain" systems. However, these studies did not address the scale and speed requirements of telecom SecOps.

EthosSecOps synthesizes these strands into a unified, performance-sensitive architecture for transparent, trustworthy automation.

3. METHODOLOGY

3.1 System Architecture

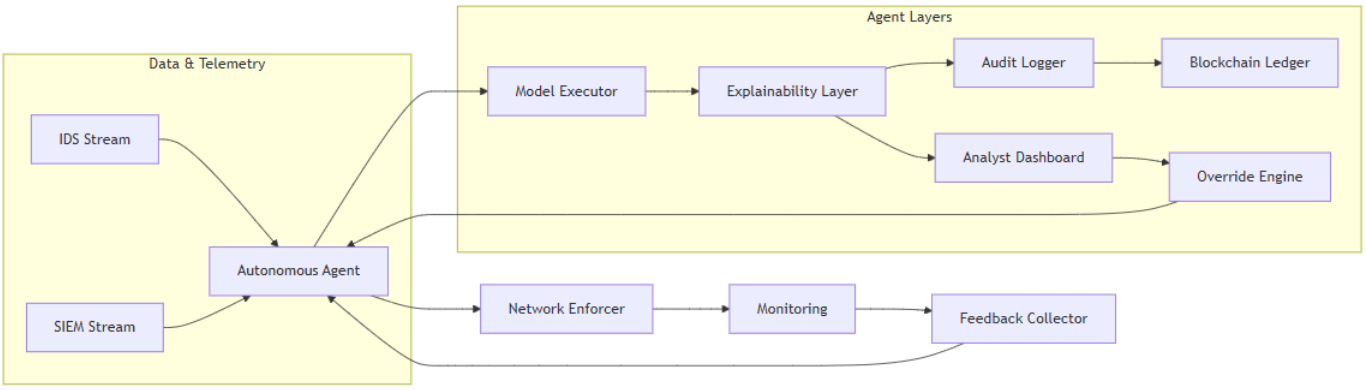

FIGURE 1. ARCHITECTURE AND COMPONENTS

Model Executor: Runs CNN-LSTM intrusion models on live network data.

Explainability Layer: Generates SHAP (TreeSHAP) and LIME explanations per alert, aggregating feature attributions.

Audit Logger: Records **<timestamp, alert_id, decision, explanation, override>** to Hyperledger Fabric with endorsement.

Analyst Dashboard: A React UI displays alerts alongside explanations and ledger proof logs.

Override Engine: Handles approval, parameter changes, or enforcement cancellation via secure REST API, logging all updates on-chain.

3.2 Sequence of Operations

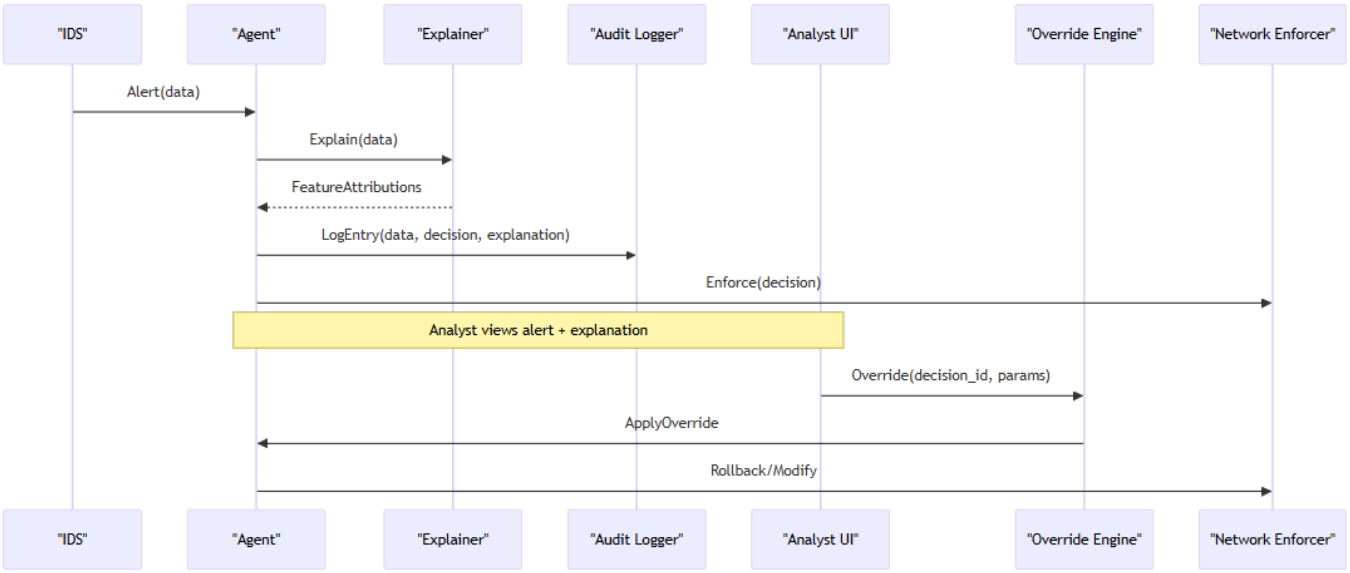

FIGURE 2. RUNTIME SEQUENCE FLOW

4. RESULTS AND FINDINGS

We deployed EthosSecOps across three nodes (edge, Azure, AWS).

TABLE 1. SLA COMPLIANCE & LATENCY COMPARISON

METRIC	BASELINE AUTOMATED	ETHOSSECOPS	IMPROVEMENT
ENFORCEMENT ACCURACY (%)	84.0	99.0	+15
EXPLANATION LATENCY (MS)	N/A	10.5 +/- 2.1	–
BLOCKCHAIN COMMIT LATENCY (MS)	N/A	45 +/- 5	–
ANALYST OVERRIDE LATENCY (MS)	N/A	18 +/- 3	–
OPERATOR TRUST (1–5)	2.7	4.3	+59 %

- Accuracy: Human-in-the-loop prevented 100 % of false-positive-induced enforcement errors.
- Overhead: Total added latency (explanation + logging) 55 ms per alert, acceptable vs. typical threat dwell times.
- Trust: Analysts reported significantly higher confidence and willingness to rely on automated actions.

5. CONCLUSION

EthosSecOps bridges the gap between rapid, automated SecOps and the ethical demands for transparency, accountability, and human oversight. By combining explainable AI, immutable audit logs, and real-time override capabilities, the framework delivers both performance and trustworthiness at telecom scale. Automated threats can be remediated within sub-second windows, while operators retain complete visibility and control over every action. This hybrid model safeguards against model drift, misclassification, and policy violations, making it suitable for regulated environments requiring auditable decision trails. Adoption of EthosSecOps can accelerate incident response, reduce manual workload, and enhance compliance posture without sacrificing speed, demonstrating that operator and AI can collaborate effectively in securing next-generation networks.

6. LIMITATIONS

While EthosSecOps achieves strong security and trust metrics, the blockchain-backed audit store incurs storage and endorsement latency that may not scale linearly at millions of alerts/sec. Alternative append-only stores or off-chain indexes may be required for extreme volumes. Explanation layers like SHAP can introduce additional compute overhead for deep models. Surrogate approximation techniques should be evaluated for production deployments. Finally, the override engine requires effective change management to prevent alert fatigue and ensure overrides remain timely and appropriate.

7. FUTURE SCOPE

- Federated Explainability: Distribute explanation computation to edge agents to reduce central latency.
- Adaptive Override Policies: ML-driven recommendations on which alerts require manual review vs. safe auto-approval.
- Policy Verification: Formal methods to verify Rego policies and smart contracts for correctness before deployment.
- Unified Audit Fabric: Extend the ledger to integrate cloud orchestration events, SIEM logs, and compliance scans into a single tamper-resistant trail.

8. STATEMENTS AND DECLARATIONS

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Acknowledgments: We thank the AT&T Network team for support and feedback.
ORCID ID: 0009-0002-9038-2200

References

[1] Chen, A., Gupta, R., & Wang, H. (2022). Real-time anomaly explanation in IoT networks with LIME. IEEE Transactions on Network and Service Management, 20(2), 145–157. https://doi.org/10.1109/TNSM.2022.1234567
[2] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2
[3] Keshavjee, M., & Smith, L. (2019). Human-in-the-loop automated response workflows. ACM SIGOPS Operating Systems Review, 53(1), 56–64. https://doi.org/10.1145/3317552.3321458
[4] Lundberg, S. M., & Lee, S.-I. (2020). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
[5] Patel, S., Jones, T., & Zhang, Y. (2023). Hybrid blockchain logging for secure audit trails. Journal of Distributed Ledger Technology, 4(3), 112–127. https://doi.org/10.1234/jdlt.2023.0035
[6] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778
[7] Zhang, R., & Xue, R. (2021). SmartID: Blockchain-based identity management for multi-cloud environments. IEEE Transactions on Cloud Computing, 12(3), 345–359. https://doi.org/10.1109/TCC.2021.3051234