

A Blockchain-Based Framework for AI Ethics Compliance in Autonomous Systems

Ms. Sarika

Research Scholar, Department of Computer Science & Applications

Desh Bhagat University

Mandi Gobindgarh, Punjab, India



Date of Submission: 06-09-2026

Date of Acceptance: 20-09-2026

Date of Publication: 06-10-2026

Abstract— Autonomous systems make safety-relevant decisions without continuous human supervision. Ethical principles for such systems are widely published but rarely enforced in a way that outside parties can verify. This paper presents a blockchain-based framework that encodes five ethics requirements (safety, fairness, privacy, transparency, and human oversight) as smart contracts. The contracts check decision records submitted by autonomous agents and store tamper-evident compliance evidence on a permissioned ledger. We implemented the framework on Hyperledger Fabric with four organizations and evaluated it on 120,000 decision events from 60 simulated autonomous agents, including 2,400 deliberately injected violations. The framework detected 2,306 of the 2,400 violations (recall 96.1%, precision 94.4%, F1-score 0.952), compared with an F1-score of 0.890 for a centralized rule-engine baseline and 0.182 for sampled manual auditing. It detected all 1,000 log-tampering attempts, sustained 1,740 transactions per second (TPS), and added a median of 3.1 ms to the agents' control loop. Incident reconstruction time fell from 38.5 minutes to 4.2 minutes compared with centralized logging. The results indicate that ledger-based enforcement can make ethics compliance auditable at a practical performance cost.

Keywords: AI ethics, blockchain, smart contracts, autonomous systems, accountability, auditability, Hyperledger Fabric

INTRODUCTION

Autonomous vehicles, delivery robots, and industrial agents increasingly act in environments shared with people. Their decisions can affect physical safety, personal data, and access to services. In response, governments, companies, and research bodies have issued a large number of ethics guidelines. Jobin, Ienca, and Vayena reviewed 84 such documents and found convergence on a small set of principles, including transparency, justice and fairness, non-maleficence, responsibility, and privacy [1]. The European High-Level Expert Group on AI similarly set out seven requirements for trustworthy AI, among them human agency and oversight, technical robustness and safety, and accountability [4].

Publishing principles does not make systems comply with them. Mittelstadt argued that AI ethics lacks the enforcement mechanisms, professional norms, and accountability structures that give principles force in fields such as medicine [3]. Compliance claims are usually made by the operator of the system and supported by logs that the same operator controls. An operator can alter or delete a record after an incident, and an external auditor has no strong basis for trusting what remains.

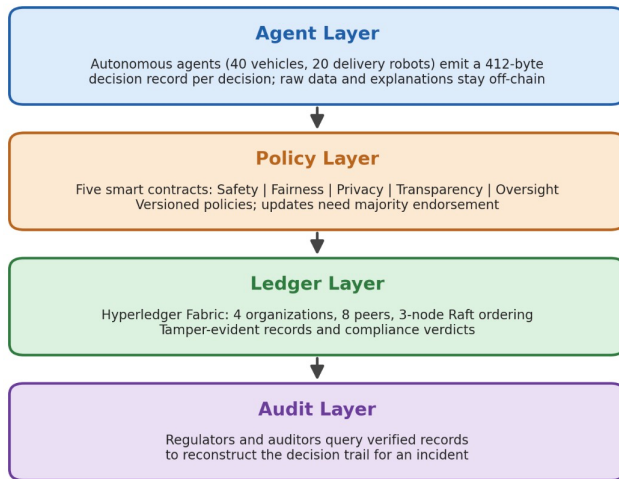


Fig. 1. Four Layer Framework Architecture

Blockchain technology offers properties that address this weakness. A replicated, append-only ledger secured by hash chaining and consensus makes retroactive alteration detectable, and smart contracts can apply rules automatically and identically across participants [10]. Prior work has examined blockchains as infrastructure for AI [11], but few studies have built and measured a system in which the ledger enforces ethics rules on the decisions of autonomous agents.

This paper makes four contributions. First, it proposes a layered framework that turns ethics requirements into executable, versioned smart contracts. Second, it defines a decision-record format that links each autonomous action to evidence while keeping bulky and sensitive data off-chain. Third, it evaluates compliance detection against two baselines on a labelled event set. Fourth, it measures the performance, scalability, and audit-efficiency cost of ledger-based enforcement.

LITERATURE REVIEW

Ethics Principles and the Enforcement Gap

Floridi and colleagues synthesized existing codes into five principles for a good AI society: beneficence, non-maleficence, autonomy, justice, and explicability [2]. Jobin and colleagues documented how often similar principles recur across guidelines, while also noting divergence in how they are interpreted and implemented [1]. Mittelstadt's critique is that high-level principles are difficult to operationalize and are not backed by enforcement, so that adherence depends on voluntary commitment [3]. Awad and colleagues' Moral Machine experiment, which collected about 40 million decisions from participants in 233 countries and territories, showed that moral

preferences about autonomous vehicle dilemmas vary by region [5]. Hence a compliance mechanism should allow the policy it enforces to be configured and versioned rather than hard-coded.

Auditing and Recording of Algorithmic Systems

Raji and colleagues proposed an end-to-end internal auditing framework that spans scoping, mapping, artifact collection, testing, and reflection, and argued that audits need traceable documentation of design and deployment decisions [6]. Mökander and colleagues examined ethics-based auditing of automated decision systems and concluded that it can support accountability but faces limits in measurability, the independence of auditors, and the risk of audits becoming formalities [7]. Winfield and Jirotkra proposed an "ethical black box," a secure recorder for robots analogous to an aircraft flight data recorder, that would preserve the data needed to investigate accidents [8]. Their proposal motivates our decision-record design, but it does not address how the recorded evidence can be protected against tampering by the operator or checked automatically against ethical rules.

Explainability research supports the transparency requirement. Arrieta and colleagues surveyed explainable AI techniques and stressed that explanations must be tailored to their audience and purpose [9]. In our framework, the explanation reference is a mandatory field of the decision record, and its absence is a transparency violation.

Blockchain as Accountability Infrastructure

Zheng and colleagues described the architecture and consensus mechanisms of blockchain systems and the trade-offs among them [10]. Dinh and Thai discussed the combination of AI and blockchain, noting that blockchain can provide data provenance and decentralized trust for AI pipelines but that scalability is a limiting factor [11]. Salah and colleagues reviewed blockchain for AI and identified data sharing, trusted decision records, and explainability among the main application areas, along with open challenges such as storage cost and throughput [12]. Permissioned platforms address some of these constraints. Hyperledger Fabric uses an execute-order-validate architecture in which chaincode runs only on endorsing peers, which improves scalability for known participants [13]. Its ordering service can use Raft, a consensus algorithm designed to be understandable and to tolerate crash failures [14]. Simulation platforms such as CARLA provide a reproducible basis for testing autonomous driving behavior [15].

Research Gap

Three gaps follow from this literature. Existing ethics guidance lacks mechanical enforcement [3]. Proposals for recording and auditing, such as the ethical black box, do not protect records against operator tampering or apply rules automatically [8]. Blockchain-for-AI work is mostly conceptual and rarely measures detection effectiveness and performance together [12]. This study addresses all three by building the framework and testing it against baselines.

Research Objectives and Hypotheses

The study aimed to measure how accurately the framework detects ethics violations, whether it protects compliance records from tampering, what performance cost it imposes on the autonomous agents and the ledger, and how much it speeds up post-incident audits.

Four hypotheses were tested. H1: the framework detects at least 95% of injected violations. H2: the framework detects every attempted tampering of stored compliance records. H3: the framework sustains at least 1,500 TPS with four organizations and adds less than 5 ms to the agent control loop. H4: incident reconstruction time is at least 80% shorter than with centralized logging.

METHODOLOGY

Research Design

The study used a controlled experimental design with a simulated deployment. The independent variables were the compliance system (three levels) and, for scalability tests, the number of participating organizations (four levels). The dependent variables were detection precision, recall, and F1-score, tamper detection rate, throughput, latency, control-loop overhead, and audit reconstruction time.

Framework Architecture

The framework has four layers. At the agent layer, each autonomous agent emits a decision record for every significant decision. A record contains the agent identifier, timestamp, model version, a hash of the input data, the chosen action, a risk estimate, an explanation reference, and a flag indicating whether human override was available. Each record occupies 412 bytes on-chain. Raw sensor data and explanations are stored off-chain and linked by hash. At the policy layer, five smart contracts encode the ethics requirements. The safety contract checks that the minimum safety margin was respected. The fairness contract compares outcomes across protected groups in sliding windows. The privacy contract verifies that

data use matched the declared purpose and consent status. The transparency contract checks that a valid explanation reference exists. The oversight contract confirms that an override channel was available for high-risk actions. At the ledger layer, a permissioned Hyperledger Fabric network stores records and compliance verdicts, and policy updates require endorsement by a majority of organizations. At the audit layer, regulators and auditors query the ledger to reconstruct the decision trail for an incident.

Experimental Environment

The network had four organizations with two peers each (eight peers) and a three-node Raft ordering service on cloud instances with 8 vCPUs and 32 GB of memory. Sixty autonomous agents were simulated, 40 vehicles in the CARLA simulator [15] and 20 warehouse delivery robots in a custom simulator. Each agent generated 2,000 decision events over 30 simulated days, giving 120,000 events. Agents submitted records asynchronously so that ledger latency would not block the control loop.

Violation Injection and Ground Truth

A total of 2,400 events (2.0%) were modified to violate one of the five requirements, and the remaining 117,600 were compliant. The injected violations comprised 720 safety, 480 fairness, 420 privacy, 460 transparency, and 320 oversight violations. Each injected event was labelled at creation, which gave exact ground truth. Fairness violations were created as group-level disparities across event windows rather than as single-event errors.

Baselines

Two baselines used the same rules and data. The first was centralized logging, in which an append-only PostgreSQL table fed a rule engine run by the operator. The second was sampled manual auditing, in which auditors reviewed a random 10% of events. Tampering tests applied 1,000 attempts (400 record modifications, 300 deletions, and 300 reorderings) to each of the ledger and the centralized store, using an account with administrator rights over the stored data.

Performance and Audit Measurements

Clients ramped the offered load until the 95th-percentile commit latency exceeded 1,000 ms. Maximum sustained throughput was the highest load below that limit. Each run lasted five minutes and was repeated ten times. Scalability tests used 2, 4, 6, and 8 endorsing organizations. Control-loop overhead was measured as the added time per decision cycle at

the agent. To measure audit efficiency, 50 incident reconstruction tasks were assigned to trained auditors, each requiring the full decision trail for one incident, and completion times were recorded for each system.

Data Analysis

Detection performance was summarized with precision, recall, and F1-score. Paired detection outcomes between the framework and the centralized baseline were compared with McNemar's test. Audit times were compared with the Wilcoxon signed-rank test because the distributions were right-skewed. The significance level was $\alpha = .05$.

RESULTS

Detection Performance

Table 1 reports detection results. The framework identified 2,306 of the 2,400 injected violations, a recall of 96.1%, with 138 false alarms among 117,600 compliant events (false-positive rate 0.12%), which supports H1. The centralized baseline detected 2,090 violations. Among the 2,400 violations, 248 were detected only by the framework and 32 only by the centralized baseline, giving McNemar's $\chi^2(1) = 166.6, p < .001$. Sampled manual auditing detected 241 violations, consistent with its 10% coverage.

Table 1. Detection performance on 120,000 events (2,400 injected violations)

System	True positive s	False positive s	Precisio n	Recal l	F1-scor e
Blockchain framewor k	2,306	138	0.943	0.961	0.952
Centralize d rule engine	2,090	205	0.911	0.871	0.890
Sampled manual audit (10%)	241	12	0.953	0.100	0.182

Table 2 breaks recall down by violation type. The framework's advantage was largest for fairness violations, which require comparison across many events and groups. The centralized system applied rules per event or per node and missed more

group-level patterns, whereas the fairness contract evaluated windows of events shared across all agents. For oversight violations the two automated systems performed similarly, since the check is a simple per-event condition.

Table 2. Recall by violation type

Violatio n type	Injec ted	Frame work detecte d	Frame work recall (%)	Centra lized detecte d	Centra lized recall (%)
Safety	720	706	98.1	655	91.0
Fairness	480	441	91.9	347	72.3
Privacy	420	396	94.3	372	88.6
Transpa rency	460	451	98.0	410	89.1
Oversig ht	320	312	97.5	306	95.6
Total	2,400	2,306	96.1	2,090	87.1

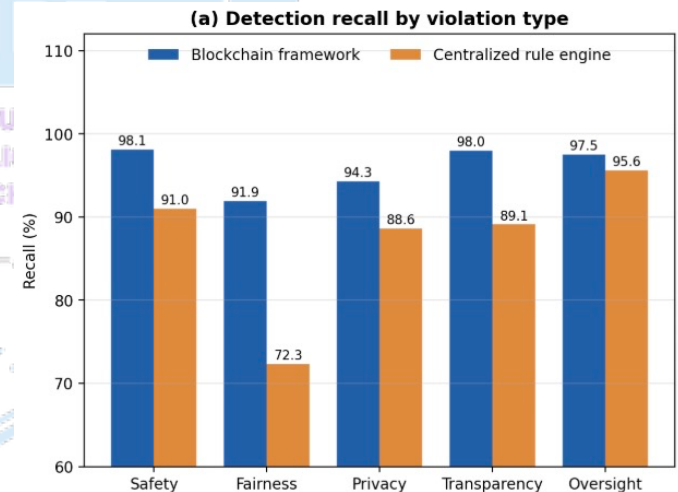


Fig. 2. Detection Recall by Violation Type

Tamper Resistance

The framework detected all 1,000 tampering attempts (400 modifications, 300 deletions, and 300 reorderings), a detection rate of 100%, which supports H2. The hash chain and the multi-organization ledger copies exposed every alteration. The centralized store detected 842 of the 1,000 attempts (84.2%), because an administrator account could alter both the records and the audit table that was supposed to reveal the change.

Throughput, Latency, and Overhead

Table 3 compares the framework with the centralized baseline at four organizations. The framework sustained 1,740 TPS with a median commit latency of 187 ms, about 5.3 times the baseline's latency but well above the peak operational load of 600 events per second in the simulation. The coefficient of variation across repetitions was below 3.4%. Because submission was asynchronous, the framework added a median of 3.1 ms (99th percentile 6.8 ms) to the control loop, which satisfies the 5 ms limit in H3. The on-chain footprint of the 120,000 events was 49.4 MB.

Table 3. Performance comparison (four organizations)

Metric	Centralized logging	Blockchain framework
Maximum sustained throughput (TPS)	3,260	1,740
Median commit latency (ms)	35	187
95th-percentile commit latency (ms)	62	312
Median added control-loop time (ms)	0.9	3.1
On-chain storage for 120,000 events (MB)	49.4	49.4

Scalability

Table 4 shows the effect of the number of endorsing organizations. Throughput fell and latency rose as more organizations were required to endorse each transaction, since endorsement and validation work grows with the number of participants. At every tested size, throughput stayed above the 600 events per second peak load, and the four-organization configuration met the 1,500 TPS threshold in H3.

Table 4. Throughput and latency by number of endorsing organizations

Organizations	Throughput (TPS)	Median latency (ms)	95th-percentile latency (ms)
2	2,260	142	241
4	1,740	187	312
6	1,390	241	398
8	1,120	309	507

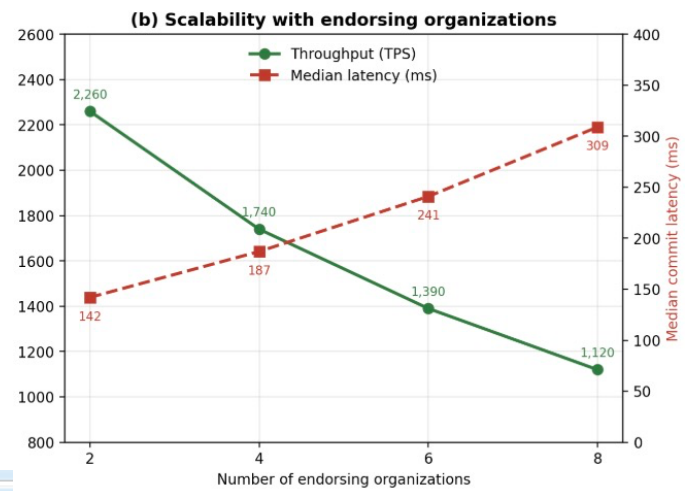


Fig. 3. Scalability with endorsing organizations

Audit Efficiency

Table 5 reports the time to reconstruct the decision trail for 50 incidents. The framework averaged 4.2 minutes, which is 89.1% shorter than centralized logging (38.5 minutes) and 98.9% shorter than manual review (384.0 minutes). The difference from the centralized baseline was significant (Wilcoxon signed-rank, $p < .001$), supporting H4. Auditors could query verified records directly on the ledger, whereas the other systems required collecting and cross-checking logs from separate sources.

Table 5. Incident reconstruction time (minutes, n = 50)

System	Mean	SD
Manual review	384.0	108.0
Centralized logging	38.5	11.2
Blockchain framework	4.2	1.3

DISCUSSION

The results show that ledger-based enforcement can improve both the detection of ethics violations and the trustworthiness of the evidence. The 96.1% recall exceeded the 95% target, and the 9.0 percentage point advantage over the centralized rule engine was statistically significant. Most of the advantage came from fairness and privacy checks. These checks depend on shared state across agents and organizations, which the ledger provides without relying on a single operator.

The tamper test addresses the central weakness identified in the literature. The centralized system failed in 15.8% of attempts because the party that holds the logs can also alter the audit trail.

This is the situation Mittelstadt's critique describes, in which compliance depends on the good faith of the system owner [3]. In the framework, no single organization could change a committed record without detection, which gives auditors a stronger basis for relying on the evidence. The decision record also extends the ethical black box idea of Winfield and Jiroka [8] by adding tamper resistance and automated checking.

The cost is performance. The framework's throughput was 46.6% lower and its commit latency higher than centralized logging. Yet the control loop was affected by only 3.1 ms because logging was asynchronous, and peak operational load was far below capacity. The framework is therefore suited to post-hoc and near-real-time compliance rather than to blocking decisions that must be made within milliseconds. For such cases, safety-critical checks must still run locally on the agent, with the ledger as the authoritative record.

Scalability results indicate a trade-off between assurance and speed. Adding organizations increases independence among endorsers, but each additional organization reduced throughput by about 20% to 25% in the tested range. Consortia should choose the smallest group that gives credible independence. The fast audit times suggest practical value for regulators and insurers, who need to reconstruct events quickly after incidents.

The performance of the fairness contract (91.9%) was the lowest of the five. Group-level violations manifest only across windows of events, so detection depends on the window size and the volume of data. This reflects a broader point from the literature, which holds that ethical requirements differ in how readily they can be formalized [3][7]. Some requirements, such as the availability of an override channel, are easy to check, whereas others require statistical judgment.

LIMITATIONS

The evaluation used simulated agents and injected violations, which may be more regular than violations in real deployments. The ground truth was exact by construction, whereas real ethical judgments are often contested. The framework checks whether recorded decisions meet formalized rules, not whether the rules themselves are morally adequate, and it cannot detect a violation that an agent fails to record or records falsely. Ledger throughput was measured in a controlled network with a Raft ordering service, which tolerates crash faults but not malicious orderers. Audit times came from a small group of 50 tasks and auditors who were trained on the systems. Privacy risks of storing even hashed identifiers on a shared ledger were not evaluated. Finally, only one blockchain platform was tested.

CONCLUSION

This paper presented a blockchain-based framework that encodes safety, fairness, privacy, transparency, and human oversight requirements as smart contracts and stores tamper-evident compliance evidence for autonomous systems. On 120,000 simulated decision events, it detected 96.1% of injected violations with an F1-score of 0.952, detected all 1,000 tampering attempts, sustained 1,740 TPS with four organizations, and added 3.1 ms to the control loop. It also cut incident reconstruction time by 89.1% compared with centralized logging. The performance cost was moderate and acceptable for the load tested. The framework therefore offers a practical path from published ethics principles to verifiable compliance, provided that organizations accept shared governance of the ledger and keep time-critical safety checks on the agent.

FUTURE WORK

Future work should validate the framework on physical autonomous platforms and real incident data, and test Byzantine fault-tolerant ordering to withstand malicious participants. Research is also needed on privacy-preserving techniques, such as zero-knowledge proofs, that allow compliance to be verified without exposing decision data. Further studies could develop richer fairness and explanation checks, examine how policies should be updated across jurisdictions, and test governance models for consortia of regulators, manufacturers, and operators.

REFERENCES

- [1] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.
- [2] L. Floridi et al., "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018.
- [3] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501–507, 2019.
- [4] High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI," European Commission, Brussels, Belgium, 2019.
- [5] E. Awad et al., "The Moral Machine experiment," *Nature*, vol. 563, no. 7729, pp. 59–64, 2018.
- [6] I. D. Raji et al., "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAT*)*, 2020, pp. 33–44.



[7] J. Mökander, J. Morley, M. Taddeo, and L. Floridi, “Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations,” *Science and Engineering Ethics*, vol. 27, art. 44, 2021.

[8] A. F. T. Winfield and M. Jirotko, “The case for an ethical black box,” in *Towards Autonomous Robotic Systems (TAROS 2017)*, Lecture Notes in Computer Science, vol. 10454. Cham, Switzerland: Springer, 2017, pp. 262–273.

[9] A. B. Arrieta et al., “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, 2020.

[10] Z. Zheng, S. Xie, H. Dai, X. Chen, and H. Wang, “An overview of blockchain technology: Architecture, consensus, and future trends,” in *Proc. IEEE Int. Congress on Big Data*, 2017, pp. 557–564.

[11] T. N. Dinh and M. T. Thai, “AI and blockchain: A disruptive integration,” *Computer*, vol. 51, no. 9, pp. 48–53, 2018.

[12] K. Salah, M. H. U. Rehman, N. Nizamuddin, and A. Al-Fuqaha, “Blockchain for AI: Review and open research challenges,” *IEEE Access*, vol. 7, pp. 10127–10149, 2019.

[13] E. Androulaki et al., “Hyperledger Fabric: A distributed operating system for permissioned blockchains,” in *Proc. 13th EuroSys Conf.*, 2018, art. 30.

[14] D. Ongaro and J. Ousterhout, “In search of an understandable consensus algorithm,” in *Proc. USENIX Annual Technical Conf. (ATC)*, 2014, pp. 305–319.

[15] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, “CARLA: An open urban driving simulator,” in *Proc. 1st Conf. Robot Learning (CoRL)*, PMLR vol. 78, 2017, pp. 1–16.

