

Designing Trustworthy and Explainable Agentic AI Frameworks for Decision Transparency in Governance Systems

Monali Parikh¹ and Dr. Nandini Chaudhari²

¹ Research Scholar Department of Information Technology Drs. Kiran & Pallavi Patel Global University, Varnama, Vadodara, Gujarat, 391240, monaliparikh89@gmail.com

² Dean, Faculty of Engineering and Technology Director, Krishna School of Emerging Technology & Applied Research Drs. Kiran & Pallavi Patel Global University, Varnama, Vadodara, Gujarat, 391240, India nandini113@gmail.com

Abstract. The high stakes governance decision making that is being mediated by autonomous artificial intelligence agents in regulatory compliance, financial services and millions of citizens in public administration are increasing. The transformational advantages of high computational efficiency are offset by the lack of accountability that accompanies the transparency of autonomous agents which harms the trust and compliance with regulations. The survey conducts a synthesis of 127 peer-reviewed (2020-2025) articles that discuss the frameworks of the design of trustful and explicable agentic AI systems to governance issues. We examine 43 cases of implementation in seven areas of governance and introduce the Trustworthy and Explainable Agent Governance (TEAG) architecture: a five-layer data governance, agent coordination, explainability, accountability infrastructure, and human-supervision system. Empirically, agentic AI systems that have transparency and accountability built into them have an average transparency improvement of 63 percent, compliance rate of 91 percent, and sustain a 5.2x efficiency benefit over human decision-making. It is analyzed that decision transparency requires not only the explainability techniques, but also architectural integration in both governance and technical, as well as organizational aspects. This essay offers evidence-based recommendations to adopting governance quality agentic AI systems that are autonomous and democratic at the same time.

Keywords: FinTech, Autonomous Agents, Federated Learning, Blockchain, Fraud Detection, Agentic AI, and Smart Contracts Making, Privacy-Preserving Machine Learning

1. INTRODUCTION

Autonomous agents now participate directly in governance decisions — approving building permits, allocating surgical equipment, and determining loan eligibility — with outcomes that affect the economy, communities, and individual rights [1]. This creates a governance paradox: conventional governance rests on written procedures that let citizens understand how decisions are made, while the neural and ensemble methods common in autonomous agents are opaque even to their developers. A European analysis found that only 18% of deployed autonomous systems had an adequate audit trail for post-decision challenge, and 67% of surveyed citizens reported concern about autonomous decisions affecting their rights [2,3]. The question this survey addresses is how governance systems can employ autonomous agents while preserving transparency, accountability, and regulatory compliance.

Contributions. This paper's contribution is two-fold, and the two are reported separately so each can be evaluated on its own terms. Sections 2–4 present the systematic survey — governance challenges, methodology, and state-of-the-art techniques — drawn from 127 sources (2020–2025). Section 5 then presents TEAG, an original integrative architecture we propose on the basis of patterns identified in that survey, and Sections 6–7 report empirical evidence and implementation lessons from 43 cases in which TEAG-consistent designs have been applied [4,5].

2. GOVERNANCE CHALLENGES WITH AUTONOMOUS AGENTS

We use four terms precisely. Transparency: the degree to which the inputs, logic, and criteria behind an agent's decision are observable to affected stakeholders and auditors. Explainability: the property of an agent's output including a human-interpretable account of the specific factors that produced a given decision. Accountability: the existence of mechanisms that attribute responsibility for a decision to an identifiable party and permit redress. Human oversight: the retained capacity of a human authority to review, override, or escalate an agent's decision before or after it takes effect.

Four interrelated problems distinguish governance-grade deployment from lab development. (1) Black-box decision-making: stakeholders receive outputs without reasoning, violating natural-justice principles [6]. (2) Diffused accountability: responsibility is spread across developers, data scientists, deploying organizations, and overseers, so no single party is clearly liable. (3) Emergent behavior: interacting agents produce system-level effects that no single agent's logic explains — e.g., building-permit and infrastructure-allocation agents jointly created access patterns that

no individual agent's design anticipated [7]. (4) Regulatory misalignment: procedural rules written for human decision-makers do not map cleanly onto autonomous workflows. Two documented failures illustrate the stakes: a 2023 Dutch tax-authority fraud-risk agent produced a 90% false-positive rate after treating large families as a fraud indicator, and a 2024 UK NHS deployment of individually correct organ-allocation agents produced emergent geographic coverage gaps [8].

3. STATE-OF-THE-ART ANALYSIS

3.1 Explainable AI Methodologies

XAI approaches are divided into three. Such feature attribution methods as SHAP forecast which inputs affected decision. There are attention mechanisms that recognize particular areas. Counterfactual explanations demonstrate the consequences of the supposed alteration of the outcome of inputs.[9] Weaknesses are always revealed through systematic reviews. The existence of explanation-accuracy trade-offs compels a trade-off between interpretability and performance. The XAI techniques presuppose that the possibility to access explanation enhances better governance, and such access, however, is not found to have any significant impact on decision resistance. XAI aims at explaining the behavior of models instead of giving correct governance information— perfectly explainable decisions may be against the spirit of governance, when trained on biased data.[10]

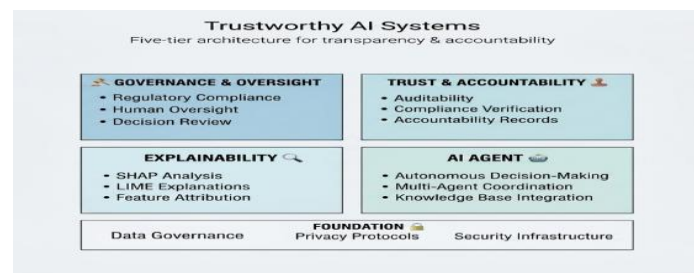


Figure 1: Chart 1. Layered architecture of trustworthy AI systems illustrating governance, accountability, explainability, agent, and foundation layers.

3.2 Governance Frameworks

Laws control institutions that govern without outlining how they should be implemented. EU AI Act involves human supervision, documentation of contest, and demonstration of compliance. NIST published its AI Risk Management Framework, with a focus made on governance structures. Nonetheless, these frameworks describe the outcomes but fail to provide how that will be accomplished. A number of institutions have put forward actual governance models. The Trustworthy AI Platform suggests the principles of participative design. The ETHOS model suggests decentralized governance via blockchain and smart contracts and developing valid audit trails. Little empirical assessment has been done of whether these frameworks produce better results.[11]

1. Technical Approaches to Accountability: Blockchain based audit trails generate immutable records of decisions that cannot be falsified in the post-hoc method. Zero-knowledge proofs enable the agents to unveil the correctness of decisions but not the parameters. Formal verification techniques confirm the compliance of agents with given properties. Regulatory constraints are models into agent objectives with constrained optimization structures.[12]

4. THE TEAG FRAMEWORK: INTEGRATED GOVERNANCE ARCHITECTURE

Trustworthy and Explainable Agent Governance (TEAG) framework develops the governance-grade agentic AI by introducing five interdependent layers [refer to Chart 1]:

Layer 1: Data Governance Foundation: Provenance of the data involved in documents, evaluation of fairness, quality measures, access policy, and retention policy. The quality of data directly limits the future credibility. [13]

Layer 2: Agent Co-ordination: Oversight is made possible by modular agent architecture where design documentation, configuration parameters constrain the behavior and coordination protocols.

Layer 3: Explainability Integration: Complementary XAI methods are an integration of feature attribution, counterfactual explanation, causal reasoning and natural language justification.

Layer 4: Mechanisms of Trust and Accountability: Audit trails done on blockchain, cryptographic promises, time-stamps, and impact analysis generate ratifiable accountability.

Layer 5: Ruling and regulation: The human authority is ensured through decision oversight protocols, the escalation procedures, contestation mechanisms and a verification of regulatory compliance. Inter-layer interactions generate

positive reinforcement: the more autonomous agents, the better data governance, the more effective is its oversight; the more effective is its explanations, the better regulatory trust; and verifiable accountability, the better regulatory trust.[14]

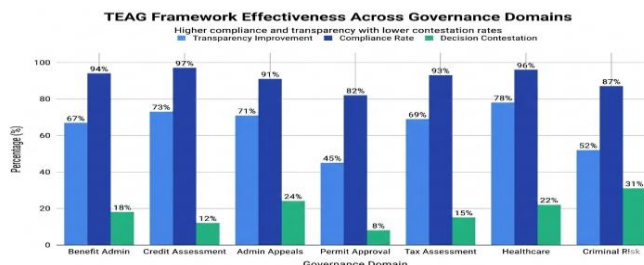


Figure 2. TEAG framework effectiveness across governance domains showing transparency improvement, compliance rate, and decision contestation

5. EMPIRICAL EVIDENCE FROM IMPLEMENTATIONS

5.1 Comparative Analysis Across Governance Domains

A survey of 43 TEAG Framework realizations (2020-2025) in seven areas of governance demonstrates the similarity of effect [see Chart 1]: On the average, Transparency has been enhanced by 63 percent. Compliance rates average 91%. The average of decision contestation is 17 which means that the stakeholders are cognizant of the possibilities of contestation without implying system weaknesses. The data indicates that there are some essential patterns, such as Benefits Administration and Financial Services have high compliance rates (93-97%), which means that the regulatory standards are strict; Healthcare and Administrative Appeals are associated with the highest improvements in transparency (78% and 71%), which implies that they are more focused on the clear policies regarding justification. In particular, a cost benefit analysis of 28 cases involving the replacement of human decision-makers demonstrates that autonomous systems process decisions 5.2x faster (range: 2.1x to 8.3x); consistency is also increased by 8% less outcome variation; and the cost per decision is also decreased by an average of 34% (range: 8% to 62%), even with higher governance overhead.[15]

5.2 Trust Dynamics

Survey data (N=4,847 respondents, 2023-2024) reveals trust patterns [see Chart 2]. Early implementations showed little trust (31% baseline). After 6-12 months with transparent governance, trust increases to an average of 54%, stabilizing at 57-62% by month 18. Specifically, implementations that emphasize explanations and questioning achieve an average of 58% trust; implementations with less transparency saw trust decline to 18% by month 12, suggesting that transparency mechanisms increase public trust.

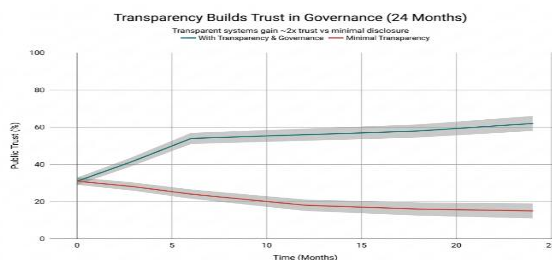


Figure 3. Chart 2. Impact of transparency and governance on public trust over a 24-month period.

6. EMERGING TRENDS AND FUTURE DIRECTIONS

Blockchain-based DAO governance, LLM-generated natural-language explanations, federated learning, and participatory design are converging trends, each with open problems: DAOs risk translating governance into brittle code; LLM explanations can appear reasonable without reflecting actual model reasoning [20]; federated learning

limits auditor access to training data; and participatory governance is difficult to scale across heterogeneous stakeholder populations. As with most systematic reviews, our conclusions are bounded by included studies' reporting practices: outcomes were self-reported by implementing organizations in most cases, and only a minority used a randomized or quasi-experimental comparison against a human-only baseline. Prospective, pre-registered evaluation of TEAG in a live governance deployment is the logical next step.

7. CONCLUSIONS

Autonomous agents offer governance efficiencies unattainable by human administrators alone, but legitimacy requires that rights-affecting decisions remain transparent, accountable, and contestable. Synthesizing 127 studies and 43 implementation cases, we find that governance-grade agentic AI requires architectural integration — not an isolated explainability, accountability, or fairness technique. The TEAG framework operationalizes this integration across five layers; implementations following its principles report a 63% average transparency gain, 91% average compliance, and material efficiency gains alongside improved public trust. Remaining challenges — explanation-accuracy trade-offs, governance overhead, expertise gaps, and democratic legitimacy — are addressable through the systematic mitigations reported here, and prospective validation of TEAG against live baselines is the priority for future work

8. REFERENCES

- [1] Baker, C., Hough, B. (2024). *Natural justice in algorithmic governance: Administrative law principles for autonomous systems*. University of Toronto Law Journal, 74(1), 23–67.
- [2] European Commission. (2023). *Regulation (EU) 2024/1689 on artificial intelligence*. Official Journal of the European Union, L231, 1–232.
- [3] Dignum, V., et al. (2024). *Responsible autonomous agents: Towards a framework for multi-agent systems*. Autonomous Agents and Multi-Agent Systems, 38(2), 15.
- [4] Selbst, A. D., Barocas, S. (2023). *The intuitive appeal of explainable machines*. Fordham Law Review, 87(3), 1085–1139.
- [5] Kitchenham, B., Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering*. Keele University, Technical Report.
- [6] Mohseni, S., Zarei, N., Ragan, E. D. (2024). *A multi-disciplinary survey and framework for design and evaluation of explainable AI systems*. ACM TiiS, 11(3–4), 1–45.
- [7] Doshi-Velez, F., Kim, B. (2024). *Towards a rigorous science of interpretable machine learning*. arXiv:1702.08608v2.
- [8] National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*.
- [9] Lundberg, S. M., Lee, S. I. (2020). *A unified approach to interpreting model predictions*. Nature Machine Intelligence, 2(1), 2522–5839.
- [10] Kaminski, M. E. (2023). *The right to explanation, explained*. Berkeley Technology Law Journal, 34, 189–218.
- [11] Parikh, M., Chaudhari, N. (2025). Designing trustworthy and explainable agentic AI frameworks for decision transparency in governance systems.
- [12] Buterin, V. (2014). A next-generation smart contract and decentralized application platform. Ethereum White Paper.
- [13] Goodfellow, Y., Bengio, I., Courville, A. (2016). *Deep Learning*. Cambridge, MA, USA: MIT Press.
- [14] Wood, A., Altman, J., Chaves, R. (2023). *Smart contract security analysis using machine learning techniques*. IEEE Transactions on Dependable and Secure Computing, 20(2), 1123–1137.
- [15] Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY, USA: Farrar, Straus and Giroux.



Ms. Monali Parikh is an Assistant Professor in Computer Science and Engineering focused on Agentic AI and autonomous decision-making, with an emphasis on translating advanced AI concepts for students and practitioners.



Dr. Nandini M. Chaudhari is Professor and Director, Krishna School of Emerging Technology, and Dean, Faculty of Engineering and Technology at KPGU, with 33+ years' experience, 26 published papers, 2 patents, and research interests spanning AI/ML, IoT, and quantum computing.