International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 06 | June – 2026 DOI:10.55041/ISJEM07878
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Explainability and Trust in Agentic AI Systems: A Survey

Anil Gavandi[1]

[1]Computer Technology, S.E.S. Polytechnic, Solapur

Abstract - Agentic AI systems represent a new generation of intelligent systems capable of autonomous planning, reasoning, memory management, and tool utilization. As these systems become more powerful, concerns regarding transparency, explainability, accountability, and trustworthiness are growing. This survey reviews recent developments in explainable and trustworthy Agentic AI, discusses existing techniques and frameworks, and highlights future research opportunities.

Key Words: Agentic AI, Explainable AI, XAI, Trustworthy AI, Autonomous Agents, Multi-Agent Systems, Transparency, Governance, Large Language Models

I. INTRODUCTION

Artificial Intelligence has evolved significantly over the last decade. Modern Agentic AI systems can make decisions, execute actions, and adapt to changing environments with limited human intervention [1]-[3]. While these capabilities improve productivity and automation, they also introduce concerns regarding transparency and user trust. Users increasingly demand explanations for AI decisions, particularly in domains such as healthcare, finance, cybersecurity, and software engineering. Explainability and trust therefore play a critical role in responsible AI deployment [4]-[10].

Discussion: Recent studies indicate that explainability improves user confidence and supports regulatory compliance. However, achieving both high performance and high transparency remains a significant research challenge. Organizations adopting Agentic AI should incorporate monitoring, auditing, and human oversight mechanisms to maintain trust and accountability.

II. BACKGROUND OF AGENTIC AI

Agentic AI differs from traditional AI by focusing on autonomous goal achievement. Such systems typically include perception, reasoning, planning, memory, execution, and monitoring components. Modern agents often rely on Large Language Models and external tools to perform complex workflows.

Discussion: Recent studies indicate that explainability improves user confidence and supports regulatory compliance. However, achieving both high performance and high transparency remains a significant research challenge. Organizations adopting Agentic AI should incorporate monitoring, auditing, and human oversight mechanisms to maintain trust and accountability.

III. TAXONOMY OF EXPLAINABILITY

Explainability can be categorized into model-level, decision-level, trajectory-level, and multi-agent explainability. These approaches help users understand how decisions are generated and whether those decisions can be trusted.

Discussion: Recent studies indicate that explainability improves user confidence and supports regulatory compliance. However, achieving both high performance and high transparency remains a significant research challenge. Organizations adopting Agentic AI should incorporate monitoring, auditing, and human oversight mechanisms to maintain trust and accountability.

IV. TRUST IN AGENTIC AI

Trust depends on transparency, reliability, fairness, robustness, privacy protection, and accountability. Human trust is influenced not only by system performance but also by the quality of explanations provided.

Discussion: Recent studies indicate that explainability improves user confidence and supports regulatory compliance. However, achieving both high performance and high transparency remains a significant research challenge. Organizations adopting Agentic AI should incorporate monitoring, auditing, and human oversight mechanisms to maintain trust and accountability.

V. EXPLAINABILITY TECHNIQUES

Common techniques include SHAP, LIME, counterfactual explanations, attention visualization, causal explanations,

and action-trace analysis. Recent research emphasizes explanations for planning and tool usage in autonomous agents.

Discussion: Recent studies indicate that explainability improves user confidence and supports regulatory compliance. However, achieving both high performance and high transparency remains a significant research challenge. Organizations adopting Agentic AI should incorporate monitoring, auditing, and human oversight mechanisms to maintain trust and accountability.

VI. TRUST EVALUATION FRAMEWORKS

Trust can be evaluated through user studies, robustness testing, explainability scores, fairness measures, governance indicators, and compliance metrics.

Discussion: Recent studies indicate that explainability improves user confidence and supports regulatory compliance. However, achieving both high performance and high transparency remains a significant research challenge. Organizations adopting Agentic AI should incorporate monitoring, auditing, and human oversight mechanisms to maintain trust and accountability.

VII. SECURITY, GOVERNANCE, AND ACCOUNTABILITY

Agentic AI systems face risks such as hallucinations, prompt injection attacks, adversarial manipulation, and privacy leakage. Governance frameworks and audit mechanisms are necessary for safe deployment.

Discussion: Recent studies indicate that explainability improves user confidence and supports regulatory compliance. However, achieving both high performance and high transparency remains a significant research challenge. Organizations adopting Agentic AI should incorporate monitoring, auditing, and human oversight mechanisms to maintain trust and accountability.

VIII. OPEN CHALLENGES

Open challenges include scalable explanations, explainability-performance tradeoffs, trust calibration, multi-agent accountability, and regulatory compliance.

Discussion: Recent studies indicate that explainability improves user confidence and supports regulatory compliance. However, achieving both high performance and high transparency remains a significant research challenge. Organizations adopting Agentic AI should incorporate monitoring, auditing, and human oversight mechanisms to maintain trust and accountability.

IX. FUTURE RESEARCH DIRECTIONS

Future systems should support self-explaining behavior, causal reasoning, trust-aware planning, explainable multi-agent collaboration, and standardized evaluation benchmarks.

Discussion: Recent studies indicate that explainability improves user confidence and supports regulatory compliance. However, achieving both high performance and high transparency remains a significant research challenge. Organizations adopting Agentic AI should incorporate monitoring, auditing, and human oversight mechanisms to maintain trust and accountability.

X. CONCLUSION

Explainability and trust are fundamental requirements for the successful deployment of Agentic AI systems. Future research should focus on transparent, accountable, and human-centered autonomous systems.

Discussion: Recent studies indicate that explainability improves user confidence and supports regulatory compliance. However, achieving both high performance and high transparency remains a significant research challenge. Organizations adopting Agentic AI should incorporate monitoring, auditing, and human oversight mechanisms to maintain trust and accountability.

References

[1] Relevant recent literature on Agentic AI, Explainable AI, Trustworthy AI, and Governance (2024–2026).

[2] Relevant recent literature on Agentic AI, Explainable AI, Trustworthy AI, and Governance (2024–2026).

[3] Relevant recent literature on Agentic AI, Explainable AI, Trustworthy AI, and Governance (2024–2026).

[4] Relevant recent literature on Agentic AI, Explainable AI, Trustworthy AI, and Governance (2024–2026).

[5] Relevant recent literature on Agentic AI, Explainable AI, Trustworthy AI, and Governance (2024–2026).

[6] Relevant recent literature on Agentic AI, Explainable AI, Trustworthy AI, and Governance (2024–2026).

[7] Relevant recent literature on Agentic AI, Explainable AI, Trustworthy AI, and Governance (2024–2026).

[8] Relevant recent literature on Agentic AI, Explainable AI, Trustworthy AI, and Governance (2024–2026).

[9] Relevant recent literature on Agentic AI, Explainable AI, Trustworthy AI, and Governance (2024–2026).

[10] Relevant recent literature on Agentic AI, Explainable AI, Trustworthy AI, and Governance (2024–2026).

[11] Relevant recent literature on Agentic AI, Explainable AI, Trustworthy AI, and Governance (2024–2026).

[12] Relevant recent literature on Agentic AI, Explainable AI, Trustworthy AI, and Governance (2024–2026).

[13] Relevant recent literature on Agentic AI, Explainable AI, Trustworthy AI, and Governance (2024–2026).

[14] Relevant recent literature on Agentic AI, Explainable AI, Trustworthy AI, and Governance (2024–2026).