International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Designing Trustworthy Agentic AI Systems for Autonomous Decision-Making: A Framework for Safety, Explainability, Security, and Accountability

Dr. Pankaj Kumar
Associate Professor in Computer Science, Government College for Women, Shahzadpur (Ambala), Haryana

Abstract - The rapid development of generative artificial intelligence has led to a new generation of systems known as agentic AI, which can perceive information, reason about objectives, plan tasks, use digital tools, and perform actions with limited human intervention. Unlike conventional AI applications that primarily generate predictions or content, agentic AI systems can operate through multi-step decision-making processes and interact with external environments. This increased autonomy creates significant opportunities in healthcare, education, cybersecurity, finance, public administration, and software engineering, but it also introduces new risks related to safety, security, transparency, privacy, and accountability. This paper proposes a framework for designing trustworthy agentic AI systems based on four central principles: safety, explainability, security, and accountability. The study examines the major challenges associated with autonomous decision-making and presents a conceptual framework integrating human oversight, risk assessment, explainable reasoning, continuous monitoring, secure tool use, auditability, and governance. Analytical tables are used to compare risks, mitigation strategies, trust dimensions, and evaluation criteria. The paper argues that trustworthy agentic AI should not be designed merely as a more capable form of automation; rather, it should be developed as a controlled socio-technical system in which autonomy is proportional to risk and human oversight remains meaningful. The proposed framework can serve as a foundation for future research and practical implementation of responsible autonomous AI systems.

Keywords: Agentic AI, Autonomous Decision-Making, Trustworthy AI, Explainable AI, AI Safety, AI Security, Accountability, Human Oversight

1. INTRODUCTION

Artificial intelligence is undergoing a significant transition from systems that primarily respond to user instructions toward systems capable of independently planning and executing multi-step tasks. These systems are increasingly described as **agentic AI** because they possess capabilities associated with perception, reasoning, planning, tool use, memory, and action. An agentic system may receive a broad objective, decompose it into smaller tasks, select appropriate tools, evaluate intermediate results, and continue working until the objective is achieved.

This development has considerable potential to improve productivity and decision-making. AI agents may assist researchers in analyzing large collections of information, support programmers in software development, help organizations automate complex workflows, and provide decision support in areas where rapid analysis is required. However, greater autonomy also increases the consequences of errors. A traditional chatbot that generates an incorrect answer may cause inconvenience, whereas an autonomous agent with access to financial systems, databases, or operational tools could potentially take an incorrect action with significant consequences.

The central challenge, therefore, is not simply how to make AI agents more intelligent, but how to make them **trustworthy**. Trustworthy agentic AI requires mechanisms that ensure that systems operate safely, explain important decisions, resist malicious manipulation, protect sensitive information, and remain accountable to identifiable human or institutional authorities [1][2].

The objective of this paper is to propose a conceptual framework for trustworthy agentic AI based on four fundamental dimensions: **safety, explainability, security, and accountability**. The paper also examines how human oversight, continuous monitoring, risk assessment, and governance can be integrated into autonomous decision-making systems.

2. UNDERSTANDING AGENTIC AI AND AUTONOMOUS DECISION-MAKING

Agentic AI can be understood as an AI system capable of pursuing a specified objective through a sequence of decisions and actions. While the exact definition varies across research and industry, an agentic system generally includes an AI model, memory, planning mechanisms, external tools, and an environment in which actions are performed.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

The autonomous decision-making process typically begins with a goal. The agent interprets the goal, gathers relevant information, creates a plan, performs actions, observes outcomes, and modifies the plan when necessary. This creates a feedback loop that distinguishes agentic systems from static predictive models [3].

Table 1. Major Characteristics of Agentic AI

Characteristic	Description	Trust Concern
Goal-directed behavior	The system works toward a defined objective	The objective may be misunderstood
Planning	The agent creates multi-step strategies	Incorrect plans may produce harmful outcomes
Tool use	The agent interacts with APIs, databases, and software	Unauthorized actions may occur
Memory	Previous information may influence future decisions	Incorrect or sensitive information may persist
Adaptation	The agent changes its approach based on feedback	Unexpected behavior may emerge
Autonomy	The system can act with limited human intervention	Accountability may become unclear

The table demonstrates why agentic AI creates a different risk profile from conventional AI. The combination of planning, memory, tool use, and autonomy allows an agent to perform useful complex tasks, but these same capabilities can amplify errors. Therefore, trustworthy design must address not only the accuracy of an AI model but also the complete decision-and-action cycle.

3. THE NEED FOR TRUSTWORTHY AGENTIC AI

Trust is essential when AI systems influence decisions that affect individuals, organizations, or society. The National Institute of Standards and Technology (NIST) identifies characteristics such as validity, reliability, safety, security, resilience, accountability, transparency, explainability, privacy, and fairness as important elements of trustworthy AI [1].

For agentic AI, these principles become even more important because the system may independently determine *what to do next*. An autonomous agent may therefore introduce risks at multiple stages: interpreting the goal, selecting information, reasoning about alternatives, selecting tools, executing actions, and evaluating outcomes.

The European Union's AI Act also emphasizes a risk-based approach to artificial intelligence regulation, with stronger obligations applying to higher-risk AI applications [4]. This reinforces the importance of aligning the level of autonomy with the potential consequences of failure.

Table 2. Key Trust Dimensions for Agentic AI

Trust Dimension	Core Question	Example Evaluation Metric
Safety	Can the agent avoid harmful actions?	Unsafe action rate
Explainability	Can users understand important decisions?	Explanation completeness
Security	Can the system resist attacks and misuse?	Attack success rate
Accountability	Can responsibility be assigned?	Audit trace completeness
Reliability	Does the system perform consistently?	Task success rate
Privacy	Does the system protect sensitive data?	Privacy violation rate
Fairness	Does the system avoid unjustified bias?	Disparity indicators
Human Control	Can humans intervene effectively?	Intervention success rate

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Trustworthiness is multidimensional. High accuracy alone does not make an AI agent trustworthy. For example, an agent may be highly accurate but still unsafe if it can execute unauthorized actions. Similarly, a secure system may not be trustworthy if its decisions cannot be explained or audited. The dimensions in the table should therefore be evaluated together rather than independently.

4. PROPOSED FRAMEWORK FOR TRUSTWORTHY AGENTIC AI

This paper proposes a four-layer framework consisting of **Safety, Explainability, Security, and Accountability (SESA)**. These four dimensions are supported by two cross-cutting mechanisms: **human oversight** and **continuous monitoring**.

The framework follows the principle that the greater the potential impact of an AI agent's actions, the stronger the controls surrounding autonomy should be. Low-risk tasks may be fully automated, while high-risk tasks should require human approval.

Figure/Conceptual Model

User Goal → Risk Assessment → Agent Planning → Explainable Reasoning → Secure Tool Selection → Human Approval (if required) → Action → Monitoring → Audit → Feedback

The proposed architecture treats an AI agent as part of a broader socio-technical system rather than as an isolated AI model.

4.1. Safety Layer

Safety refers to the ability of an agentic AI system to operate without causing unacceptable harm. Safety must be considered throughout the entire agent lifecycle, including goal interpretation, planning, tool selection, execution, and post-action evaluation.

One important mechanism is **risk-sensitive autonomy**. The system should classify actions according to their potential impact. For example, generating a draft email may be low risk, while transferring money or modifying medical records may be high risk.

Table 3. Risk-Based Autonomy Model

Risk Level	Example Activity	Recommended Autonomy	Human Control
Low	Summarizing documents	High	Periodic review
Moderate	Preparing business reports	Medium-High	Human review before publication
High	Changing critical database records	Low	Mandatory approval
Very High	Medical or legal high-impact decisions	Very Low	Human expert decision required

The table proposes that autonomy should be proportional to risk. A trustworthy system should not provide identical levels of independence for every task. The use of human approval gates is particularly important for decisions that are irreversible, financially significant, legally sensitive, or potentially harmful to individuals.

Safety mechanisms should include action constraints, permission boundaries, sandbox environments, emergency shutdown mechanisms, and predefined policies. Agents should also be designed to recognize uncertainty and stop when they cannot confidently complete a task.

4.2. Explainability Layer

Explainability is the ability to provide understandable information about how and why an AI system reached a decision. This is particularly challenging for agentic AI because a decision may result from multiple intermediate steps rather than a single model output.

For trustworthy agentic systems, explainability should operate at three levels:

1. **Decision explanation** – Why was a particular decision made?
2. **Process explanation** – What steps were followed?
3. **Action explanation** – Why was a particular external action performed?

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Explainability does not necessarily require revealing every internal computational process. Instead, systems should provide useful information about objectives, evidence, constraints, uncertainty, and actions.

Table 4. Explainability Requirements

Explanation Type	Information Provided	Intended User
Goal explanation	What objective is being pursued?	General user
Decision explanation	Why was the decision selected?	User/manager
Evidence explanation	What information influenced the decision?	Expert
Action explanation	Why was an external action taken?	Auditor/operator
Uncertainty explanation	How confident is the system?	Decision-maker
Alternative explanation	What other options were considered?	Expert/auditor

Different stakeholders require different types of explanations. A general user may only need a concise reason for a recommendation, while an auditor may require a detailed record of evidence and actions. Therefore, explainability should be adaptive and role-based.

4.3. Security Layer

Agentic AI creates new cybersecurity challenges because agents may have access to external tools, applications, databases, and communication channels. A compromised or manipulated agent could potentially misuse these capabilities.

Important threats include prompt injection, malicious tool instructions, unauthorized access, data leakage, identity compromise, and excessive permissions. OWASP has identified several security risks associated with large language model applications, many of which are relevant to agentic systems [5].

A trustworthy agent should follow the principle of **least privilege**, meaning that it receives only the permissions necessary to perform a task.

Table 5. Security Risks and Mitigation Measures

Security Risk	Potential Impact	Recommended Control
Prompt injection	Manipulation of agent behavior	Input validation and isolation
Excessive permissions	Unauthorized actions	Least-privilege access
Data leakage	Exposure of confidential information	Data classification and access controls
Malicious tools	Harmful external actions	Tool verification
Credential theft	System compromise	Secure credential management
Supply-chain attacks	Compromised AI components	Dependency monitoring
Uncontrolled autonomy	Large-scale harmful actions	Action limits and approval gates

Security controls must be applied not only to the AI model but also to the surrounding ecosystem. An agent may be technically secure but still dangerous if it has unrestricted access to sensitive systems. Therefore, identity management, permission controls, secure APIs, logging, and continuous monitoring are essential components of trustworthy agentic AI.

4.4 Accountability Layer

Accountability addresses the question of who is responsible when an AI agent makes an incorrect or harmful decision. Autonomous systems can create a responsibility gap if organizations treat the AI system itself as the responsible actor.

The proposed framework assigns accountability across multiple stakeholders, including system developers, deploying organizations, system operators, and human decision-makers. Clear responsibility should be established before an agent is deployed.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Every significant action should generate an **audit trail** containing information about the goal, relevant inputs, decisions, tools used, actions taken, human approvals, and final outcomes.

Table 6. Accountability Framework

Stakeholder	Primary Responsibility
Developers	Design safe and reliable systems
AI system provider	Maintain system performance and security
Deploying organization	Establish appropriate policies
System operator	Monitor operation and respond to incidents
Human decision-maker	Review high-risk decisions
Auditor	Independently evaluate compliance and performance

Accountability should be distributed according to roles rather than assigned vaguely to "the AI." This creates a clear governance structure. An audit trail also enables organizations to reconstruct events after an incident and identify where a failure occurred.

5. PROPOSED EVALUATION FRAMEWORK AND ANALYTICAL DATA

To assess trustworthy agentic AI, organizations require measurable indicators. The following analytical model proposes a hypothetical evaluation of five AI agent prototypes. The values are illustrative and demonstrate how a trust score could be calculated.

Table 8. Illustrative Trustworthiness Evaluation of AI Agents

Agent	Safety Score (%)	Explainability (%)	Security (%)	Accountability (%)	Overall Trust Score (%)
Agent A	91	84	89	86	87.5
Agent B	86	92	83	90	87.8
Agent C	95	88	94	93	92.5
Agent D	80	78	90	82	82.5
Agent E	88	90	87	91	89.0

The data are illustrative rather than empirical measurements from a specific experiment. They demonstrate that an agent should not be evaluated using a single performance indicator. Agent C, for example, achieves the highest overall score because it performs consistently across all four trust dimensions. The table also shows that an agent with excellent security but weak explainability may still have limitations in practical deployment.

For research implementation, each dimension could be assigned a weighted score based on application context. A healthcare agent, for example, may assign greater weight to safety and accountability, while a cybersecurity agent may assign greater weight to security and reliability.

6. CHALLENGES IN IMPLEMENTING TRUSTWORTHY AGENTIC AI

Despite significant progress, several challenges remain. First, autonomous agents may exhibit unpredictable behaviour when operating in complex environments. The combination of probabilistic AI models and external tools makes complete prediction of behaviour difficult.

Second, explainability becomes more complex as agents perform longer sequences of actions. A simple explanation may not adequately capture why an agent changed its plan after receiving new information.

Third, security threats evolve rapidly. Attackers may attempt to manipulate agents through carefully designed inputs or compromised external resources.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Fourth, accountability remains a major governance challenge. Organizations need clear policies defining who is responsible for an agent's decisions and how incidents should be investigated.

Table 9. Major Challenges and Research Priorities

Challenge	Research Priority
Unpredictable autonomous behaviour	Formal verification and behavioural testing
Complex decision chains	Structured decision logging
Prompt and tool manipulation	Agent-specific security architecture
Responsibility gaps	Clear governance frameworks
Human over-reliance	Human factors and user training
Excessive autonomy	Dynamic risk-based permission systems
Difficult auditing	Standardized audit protocols

The table indicates that trustworthy agentic AI requires interdisciplinary research. Technical improvements alone are insufficient. Advances in cybersecurity, human-computer interaction, law, ethics, governance, and organizational policy must complement AI engineering.

7. FUTURE SCOPE

Future research should investigate methods for creating agents that can recognize uncertainty, explain their decisions, and safely reduce their level of autonomy when risks increase. Dynamic autonomy is a particularly promising direction in which an agent continuously evaluates the risk associated with its current task and adjusts its permissions accordingly.

Another important research area is the development of standardized benchmarks for evaluating agentic AI. Existing AI evaluation methods often focus on accuracy or task completion, but autonomous systems require additional metrics for safety, robustness, security, explainability, and accountability.

Research should also examine multi-agent environments in which several AI agents collaborate. In such systems, responsibility and coordination become more complex because decisions may emerge from interactions among multiple autonomous components.

8. CONCLUSION

Agentic AI represents an important evolution in artificial intelligence because it enables systems to move beyond generating responses toward planning and performing actions. This increased autonomy has the potential to transform many sectors, but it also creates new risks. An autonomous AI agent that can access tools, modify data, or influence important decisions must be designed with stronger safeguards than a conventional AI application.

This paper proposed a **SESA framework—Safety, Explainability, Security, and Accountability**—for designing trustworthy agentic AI systems. The framework emphasizes risk-based autonomy, human oversight, secure tool access, transparent decision processes, continuous monitoring, and comprehensive audit trails.

The central argument of this study is that trustworthiness should be treated as a fundamental architectural requirement rather than an additional feature added after development. AI agents should be given only the level of autonomy appropriate to their risk, and high-impact decisions should remain subject to meaningful human control. By integrating technical safeguards with governance and accountability mechanisms, organizations can move toward AI systems that are not only capable and autonomous but also reliable, transparent, secure, and socially responsible.

The future of agentic AI will therefore depend not only on how intelligently these systems can act, but also on how effectively humans can understand, supervise, control, and hold them accountable.

REFERENCES

[1] National Institute of Standards and Technology (NIST), Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, 2023.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

[2] National Institute of Standards and Technology (NIST), Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1, 2024.
[3] Wang, L., Ma, C., Feng, X., et al., "A Survey on Large Language Model based Autonomous Agents," Frontiers of Computer Science, vol. 18, 2024.
[4] European Union, Regulation (EU) 2024/1689: Artificial Intelligence Act, Official Journal of the European Union, 2024.
[5] OWASP Foundation, OWASP Top 10 for Large Language Model Applications, 2025.
[6] UNESCO, Recommendation on the Ethics of Artificial Intelligence, United Nations Educational, Scientific and Cultural Organization, Paris, 2021.
[7] OECD, OECD Principles on Artificial Intelligence, Organisation for Economic Co-operation and Development, updated 2024.
[8] Stanford Institute for Human-Centered Artificial Intelligence, AI Index Report 2025, Stanford University, 2025.
[9] Bommasani, R., Hudson, D. A., Adeli, E., et al., "On the Opportunities and Risks of Foundation Models," arXiv preprint arXiv:2108.07258, 2021.
[10] Ji, Z., Lee, N., Frieske, R., et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, 2023.