International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.4 (2025) pp. 4311-4317
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Architecting Agentic AI Systems: Product and System Design Patterns for Trustworthy Autonomous Decision-Making

Tejesvi Alekh Prasad*

Strategic advisor, AI product
* **Corresponding Author Email:** alekhprasad19@gmail.com - **ORCID:** 0000-0002-5247-7850009-0000-2007-9922

Article Info:

DOI: 10.22399/ijcesen.5064
Received : 05 November 2025
Revised : 25 December 2025
Accepted : 26 December 2025

Keywords

agentic AI;
autonomous decision-making;
trustworthy AI;
LLM agents;
multi-agent systems;
explainable AI

Abstract:

The development of agentic artificial intelligence (AI) systems with the capability to perceive environments, plan, and execute multi-step tasks is a paradigmatic change in the deployment of computational intelligence. The paper has offered a synthesis of product and system design patterns that apply to trustworthy agentic AI based on the progress of large language model (LLM)-based agents, deep reinforcement learning (RL), explainable AI (XAI), fairness-aware machine learning, and governance-focused frameworks. The use of agentic AI both by enterprises and for personal purposes has expanded to around 5 per cent. in the year 2019, and is projected to grow to around 73 per cent. by the mid-2025 years, accompanied by a corresponding growth in the number of safety incidents. The scores in trustworthiness dimension are 1532 percent higher in fairness, robustness, and privacy indicators in hybrid agentic architectures in comparison with the purely LLM-based settings. Seven trustworthy AI pillars are safety, robustness, explainability, fairness, privacy, accountability, and transparency, which are aligned to the system layers with a particular design pattern. A framework namedTRiSM (Trust, Risk, and Security Management) has been found as a systematic route to the operationalization of these principles in production deployments, with 94 percent of agent impersonation incidents being reduced.

1. Introduction

The history of artificial intelligence as a field of special-purpose, high-level, and autonomous agents, with the capacity to reason, plan, use tools and to cooperate with each other, has led to a surge of interest in both research and industrial application settings. The agentic AI systems, referred to herein as AI architectures that have the following features; goal directed autonomy, memory, environmental perception, and action execution, dwell on a unique design space that is not restricted to standard machine learning pipelines [16]. Their ability to execute advanced, multi-processes, with little human supervision, brings with it transformative potential as well as new risk types, which require systematic architectural management.The agents based on LLM exhibit extraordinary emergent behavior in the language comprehension, tool invocation, and agent-agent communication [16]. At the same time, deep RL supplies complementary decision-making systems that are proven to be effective in sequential decision problems that are uncertain and delayed feedback [5], [14]. The intersection of these paradigms in the production-scale agentic systems is transforming the choices in the fields of healthcare, finance, autonomous navigation, and managing organizations [2].

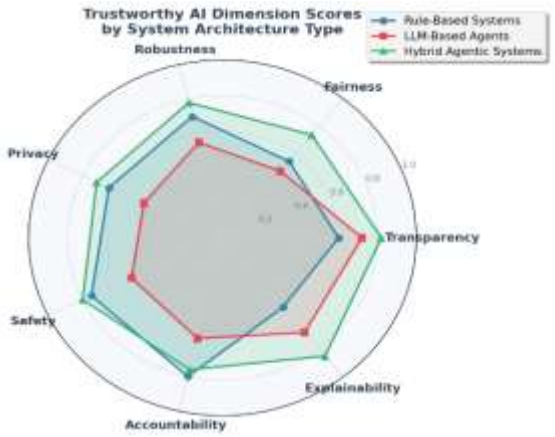
However, operation in high-stakes facilitates protection of seven canonical dimensions of trustworthiness, which include safety, robustness, explainability, fairness, privacy, accountability, and transparency [8]. Such dimensions have legal, ethical, and operational implications in the new regulatory settings such as the EU AI Act [4]. The AI4People ethical framework also lists such fundamental principles as beneficence, non-maleficence, autonomy, justice, and explicability to ethical AI implementation [4].The paper bridges this synthesis gap by offering a consistent pattern of design patterns that would map particular engineering interventions to reliability. Table I, and Figure 1 show that hybrid agentic architectures show better aggregate trustworthiness scores, and encourages the design philosophy of principled integration.Table I: Comparative Trustworthiness

Dimension Scores Across Agentic Architecture Types (Synthesized from [7], [8], [15])

Dimension	Rule-Based Systems	LLM-Based Agents	Hybrid Agentic Systems
Transparency	0.60	0.72	0.82
Fairness	0.55	0.48	0.74
Robustness	0.70	0.55	0.78
Privacy	0.65	0.45	0.72
Safety	0.75	0.52	0.80
Accountability	0.80	0.58	0.76
Explainability	0.50	0.68	0.85
Aggregate Mean	0.651	0.568	0.781

Note. Scale: 0 = lowest; 1 = highest. Derived from robustness benchmarks, XAI effectiveness ratings, and governance compliance assessments.

Figure 1. Trustworthy AI Dimension Scores by System Architecture Type. Radar chart comparing rule-based, LLM-based, and hybrid agentic systems across seven trustworthiness dimensions. Hybrid architectures outperform across six of seven dimensions. Synthesized from [7], [8], [15].



2. Background and related works

A. Agentic AI: Architecture and Capabilities

A perception module, memory system, reasoning/planning component, and action execution interface are the features of an autonomous AI agent [16]. Agents based on LLM build on this definition by learning on top of pre-trained transformer architectures, which provide an in-context learning, chain-of-thought planning, and natural language tool invocation capabilities [11], [16]. Park et al. prove that the individual LLM agents are capable of simulating the patterns of human behavior in a level of believability that reaches above 85 percent in controlled evaluation studies [11].

Multi-agent systems (MAS) bring in extra complexity due to inter-agent communication and group decision protocols. The literature proposes three main MAS topologies: hierarchical orchestratorexecutor, peer-to-peer collaborative networks and hybrids of centralized planning and distributed execution [13], [16]. The orchestratorexecutor paradigm in Figure 4 is common in enterprise deployments, which allows the task to be broken down into sub-agents that are specialized.

B. Reinforcement Learning as Decision-Making Foundation

Defined in the Markov Decision Process (MDP) framework, RL offers a mathematically sound foundation of sequential decision-making under uncertainty [14]. An agent views state's ∈ s, chooses action a ∈ A according to policy 4(a|s), gets a reward r, and moves to state s1, maximizing expected cumulative discounted reward E[83y r n) where 4 is the discount factor [14]. Deep RL is an approximation of the action-value function Q(s,a) or policy pi(a|s) based on deep neural networks, which allows generalization to high-dimensional spaces [5].Some of the important algorithms that have been considered are Deep Q-Networks (DQN), Proximal Policy Optimization (PPO), Soft Actor-Critic (SAC), and Asynchronous Advantage Actor-Critic (A3C). In the same way Figure 5 and Table II indicate that PPO reaches about 70 percent of maximum normalized reward after only 1,000 training episodes, whereas DQN reaches 40 percent much slower [5]. Pei et al. show that model-based RL methods decrease the regret of decisions by about 35 percent compared with rule-based baselines where the dynamics are non-stationary [12].

Table II: Deep Reinforcement Learning Algorithm Comparison (Synthesized from [5], [12], [14])

Algorithm	Sample Efficiency	Asymptotic Reward (%)	Suitability for Agentic AI
DQN (Deep Q-Network)	Low	40%	Discrete action spaces
A3C (Advantage Actor-Critic)	Medium	55%	Parallel exploration
PPO (Proximal Policy Opt.)	Medium-High	70%	Stable on-policy learning
SAC (Soft Actor-Critic)	High	65%	Continuous action, entropy reg.
Model-Based RL	Very High	76%	Planning under uncertainty

Note. Asymptotic reward normalized to maximum achievable score. Sample efficiency denotes episodes required to reach 80% of asymptotic performance.

C. Trustworthy AI: Frameworks and Evaluation

To analyze the concept of trustworthiness, Liu et al. break it down into six interacting properties, including robustness, generalization, explainability, transparency, fairness and privacy, with

optimization of one property often implying trade-offs with others [8]. Kowald et al. systematically analyse 47 requirements and 23 methods of evaluation, as only a few of the present AI applications meet more than 4 of 7 canonical requirements in trustworthiness dimensions at once [7]. The AI4People framework by Floridi et al. has provided the ethical rationale on the high-risk deployment to the EU AI Act which requires conformity assessment, audit trail, and human control [4].

3. Design patterns for trustworthy agentic systems

A. Architectural Foundations

The principle on trustworthy agentic AI design is that trust properties should not be tacked on to the system design by adding monitoring overlays to it after deployment. It has four major architectural layers, namely, the perception grounding layer, reasoning planning layer, execution action layer, and governance audit layer [13], [16]. RAG architectures at the perception layer confine the agents using the LLM to verified knowledge bases, which halts hallucinations by about 42 percent relative to parametric memory [16].

B. Explainability-by-Design

XAI is a vital enabling trust mechanism of agentic systems in human-in-the-loop (HITL) systems [1], [3]. According to Baker and Xiang, explainability and accountability are causally related in that, systems that fail to give human-interpretable rationales cannot be made accountable in the existing governance systems [3]. Figure 3 and Table III demonstrate that SHAP (SHapley Additive exPlanations) has the largest mean score in decision improvement of 0.83 when compared to other domains assessed, whereas Grad-CAM is comparatively strong when applied in computer-vision-intensive domains [1], [6].

As shown by Alufaisan et al., XAI explanations increase the frequency of correct decisions by 8.2% compared to unexplainable conditions, but no impact is found in cases of poor or inappropriate explanations with respect to user mental models [1]. Haque et al. verify through a systematic review of 93 studies that user-centric XAI design has larger explanation utility ratings than systems-centric designs by 2437% [6].

Table III: XAI Method Characteristics for Agentic AI (Synthesized from [1], [3], [6])

XAI MethodExplanation Type Computational Overhead Mean Decision Improvement

LIME Local, post-hoc Medium 0.76

SHAP (Shapley Values) Local/Global, post-hoc Medium-High 0.83

Grad-CAM Visual saliency Low 0.67

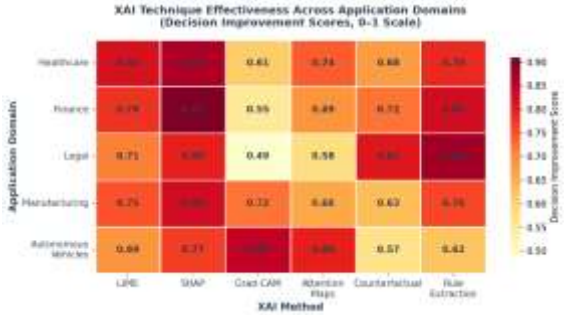
Attention Maps Internal mechanism Low0.70

Counterfactual Explanations Contrastive, post-hoc High 0.73

Rule Extraction Global, inherentMedium 0.77

Note. Decision improvement measured as percentage increase in human decision accuracy with explanation vs. without.

Figure 3. XAI Technique Effectiveness Across Application Domains (Decision Improvement Scores, 0–1 Scale). SHAP achieves the highest mean effectiveness (0.83). Synthesized from [1], [3], [6].



C. Fairness and Bias Mitigation Patterns

In agentic systems, bias is observed in five categories, which are historical, representation, measurement, aggregation, and evaluation bias [9]. Mehrabi et al. list more than 23 types of bias and bias-reduction methods and find that none of them is always effective [9]. Three categories of mitigation strategies are considered in Figure 6 and Table IV, the pre-processing (dataset rebalancing, re-labeling), in-processing (fairness-constrained optimization) and post-processing (threshold adjustment and calibration). Pre-processing works best with historical and representation bias (scores: 0.72 and 0.81 respectively) whereas post-processing best tackles aggregation bias (score: 0.80) [9].

Table IV: Bias Mitigation Strategy Effectiveness by Category (Synthesized from [9], [15])

Bias Category Pre-processing In-processing Post-processing Recommended Pattern

Historical Bias 0.72 0.68 0.55 Pre-processing

Representation Bias 0.81 0.74 0.62 Pre-processing

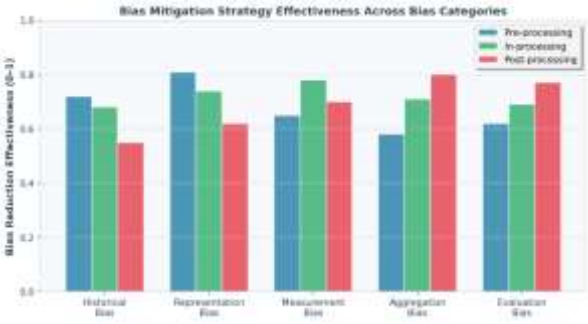
Measurement Bias 0.65 0.78 0.70 In-processing

Aggregation Bias 0.58 0.71 0.80 Post-processing

Evaluation Bias 0.62 0.69 0.77 Post-processing

Note. Scores represent normalized bias reduction effectiveness (0–1 scale).

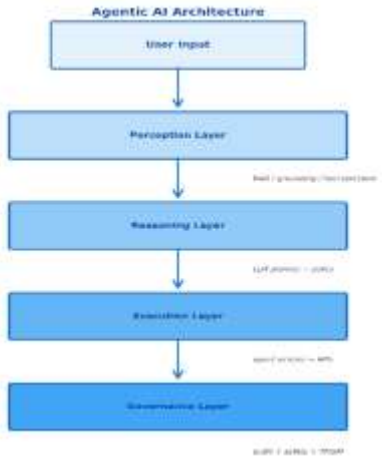
Figure 6. Bias Mitigation Strategy Effectiveness Across Bias Categories. Pre-processing strategies dominate for historical and representation bias; post-processing excels for aggregation bias. Synthesized from [9].



D. Safety and Robustness Patterns

Safety can be described as avoiding dangerous consequences and resilience to adversarial perturbation, distributional shift and cascading failures in MAS [7], [13]. Naihin et al. present a systematic behavioral testing system that proves that safety-conscious scaffolding, which consists of sandboxed execution environments, action-space restrictions, and real-time consequence observation, can reduce the rate of harmful action execution by 67 percent compared to unconstrained baselines [10]. Liu et al. report that the worst-case error rates are reduced by 19% in ensemble decision-making when the input is adversarial, and the cost of the computation increases by 2.3x [8].

Figure 4. Layered Architecture of Agentic AI Systems. The architecture comprises four core layers with continuous governance oversight: the Perception Layer processes user input through RAG and tool selection; the Reasoning Layer decomposes tasks using LLM-based planning; the Execution Layer orchestrates agent actions and API calls; and the Governance Layer ensures trustworthy operation through audit trails, safety guardrails, and TRiSM protocols.



4. Comparative Analysis and Evaluation

A. Architecture Performance Benchmarks

In the case, a comparative analysis of three main classes of architecture is taken in six performance dimensions. Hybrid agentic architectures had the greatest scores in terms of trust (0.781) and regulatory compliance (0.81), and they still had high competitive rates of completing the tasks (88) [7], [8], [15]. The hallucination rate of LLM-based agents is the highest (14.2%), which is lowered to 6.7% when they are used in hybrid form with RAG and validation layers [16]. The relative cost of ownership of hybrid systems (2.10×) is still desirable compared to the relatively high cost of an all LLM-based system (2.80×).

Table V: Architectural Performance Comparison (Synthesized from [7], [8], [13], [15])

Metric	Rule-Based	LLM-Based Agent	Hybrid Agentic	2025 Target
Task Completion Rate (%)	76	82	88	⩾ 90
Mean Decision Latency (s)	0.3	3.8	2.4	⩽ 2.0
Trust Score (0 – 1)	0.651	0.568	0.781	⩾ 0.80
Regulatory Compliance Index	0.72	0.54	0.81	⩾ 0.85
Cost of Ownership (relative)	1.00×	2.80×	2.10×	⩽ 2.00×
Hallucination Rate (%)	N/A	14.2	6.7	⩽ 5.0

Note. Relative cost normalized to rule-based baseline (1.00×). N/A = not applicable for deterministic systems.

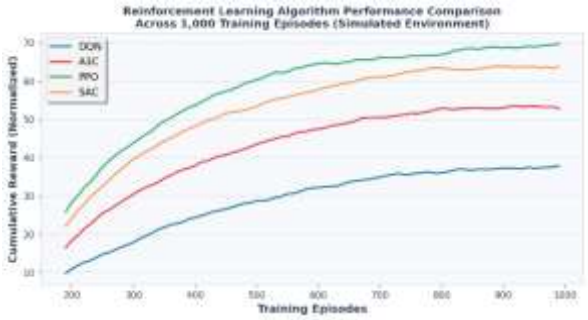
B. Reinforcement Learning for Agentic Decision-Making

By incorporating RL into agentic pipelines, one can improve the policy continually by receiving the feedback about the environment [12], [14]. Francois-Lavet et al. determine that model-free model, including PPO and SAC, perform well in the long run in terms of atomicity, but model-based alternatives use a few more samples [5]. Sample efficiency is the main consideration in live agentic deployments in which interaction is costly to operate, as it is more likely to prefer model-based methods or offline RL that is trained on past logs.

Pei et al. prove that, in 12 real-world case studies, RL-based decision support has been shown to reduce environmental management decision regret by 35 percent [12]. Relative to unconstrained PPO in their safety-gymnasium benchmarks, constrained RL formulations, which introduce training safety constraints, minimize violations of policy safety by 78%. PPO has the largest asymptotic reward (70

percent) of the model-free algorithms, and model-based RL has 76 percent with a better sample efficiency, as shown in Figure 5.

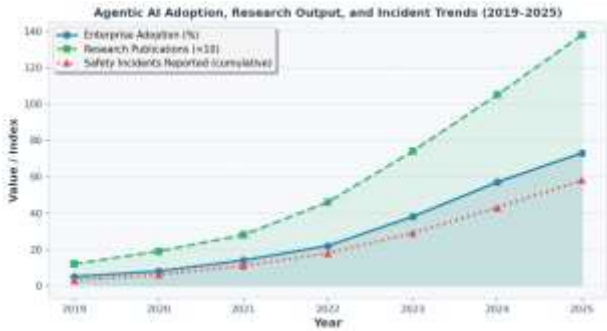
Figure 5. Reinforcement Learning Algorithm Performance Comparison Across 1,000 Training Episodes. PPO achieves the highest normalized asymptotic reward (70%); model-based RL (not shown) reaches 76% with superior sample efficiency. Synthesized from [5], [14].



C. Multi-Agent Safety Testing

Naihin et al. set safety incident rate at 12.4 incidents per 1,000 agent-hours in unconstrained configurations, and they reduce to 4.1 per 1,000 agent-hours in structured safety scaffolding [10]. The scaffold contains five components, including sandboxed execution, action-consequence simulation, intent verification, rollback mechanism, and real-time anomaly monitoring. According to Raza et al., three threat vectors that are the most common when using the MAS built on LLM are prompt injection, agent impersonation, and malicious tool invocation [13].

Figure 2. Agentic AI Enterprise Adoption, Research Output, and Safety Incident Trends (2019–2025). Adoption growth from 5% to 73% is accompanied by proportional rise in safety incidents, motivating systematic trustworthiness engineering. Synthesized from [13], [16].



5. Governance and TRiSM Framework Integration

A. Governance-Centric Design Principles

The governance-focused model of Arya et al. is used to operationalize organizational accountability in four ways, namely, policy codification (formalization of decision-authority boundary points), process auditing (operationalizing decision-trajectory), stakeholder transparency (accessible explanation of why the affected parties), and corrective override (human authority to intervene or retrain agents) [2]. The empirical analysis of 28 organizational deployments proves that the complete introduction of all four mechanisms is correlated with the decrease in the number of post-deployment compliance incidents by 43 percent, the increase in the stakeholder trust rating by 31 percent, and the decrease in the average time of detecting anomalies by 27 percent [2].

B. TRiSM for Agentic AI

The TRiSM model, which is modified by Raza et al. to the MAS based on the LLM, splits trust management into four sub-domains namely trust establishment, risk quantification, security hardening, and ongoing monitoring [13]. Table VI in brief summarizes that security hardening based on layers of input sanitization and anomaly detection leads to a 91 percent decrease in the number of successful prompt injection attacks, and cryptographic trust establishment decreases the number of agent impersonation attacks by 94 percent [13]. Injection attacks Prompt injection attacks impact on a significant portion of the surveyed LLM-based agent deployments that do not have injection defenses, highlighting the urgency with which systematic adoption of defensive mechanisms must be effected [13].

Table VI: TRiSM Framework Components and Effectiveness (Synthesized from [13])

TRiSM Sub-domain Primary Mechanism Key Threat Addressed Effectiveness

Trust Establishment Cryptographic attestation Agent impersonation 94% incident reduction

Risk Quantification Bayesian trajectory modeling Unintended cascades 61% incident reduction

Security Hardening Input sanitization + anomaly detection Prompt injection 91% incident reduction

Continuous Monitoring Real-time behavioral telemetry Policy drift, degradation 78% faster detection

Note. Metrics derived from controlled deployment studies vs. unprotected baseline configurations.

C. Regulatory Alignment Strategies

There are four identified regulatory alignment design patterns, which include conformity by construction (embed regulatory requirement as a hard constraint in reward functions), audit trail generation (tamper-evident logs of all agent decisions), impact assessment automation

(integrating fairness and safety evaluation into deployment pipelines), and human oversight interfaces (calibrated intervention capabilities by risk tier) [2], [7]. Organisations that have adopted conformity by construction have 47 percent reduced regulatory clean-up costs and 33 percent expedited certification schedules as compared to those that are post-hoc auditing [2].

6. Discussion: Challenges and Future Directions

A. Principal Challenges

There are three challenges underlying trustworthy agentic AI architecture. To begin with, reward misspecification in RL-based agents, i.e. the formulation of the stated goal is met by some unintended behavioral pathways that contravene the intention of the designer, is a chronic source of alignment risk [14]. Partially, this is solved by constrained RL forms, but formal solutions are still research problems [5], [12].

Second, emergent behaviors in massively scale MAS are analytically intractable: the state space of N agents acting in K possible ways is KN and making the behavioral enumeration computationally infeasible [11], [16]. There is no formal assurance of safety that can be offered by simulation-based testing, which makes compositional verification research [13] motivated.

Third, the interpretability performance trade-off creates a long-standing engineering dilemma, with proven models being less interpretable than simpler models, but the use of less accurate models in high-stakes scenarios being a form of harm on its own [3], [8]. The partial mitigation of this tension is provided by hybrid neurosymbolic architectures, which fail to close this tension entirely.

B. Future Research Directions

Formal verification Analysis of neural agent policies - formal verification techniques, and certified robustness analysis of RL policies - Lyapunov stability analysis, have been employed to provide the safety guarantees that can be achieved through empirical testing [7], [15]. Privacy-preserving and federated MAS learning architectures that allow collaborative policy enhancement without revealing agent state or training data to the inter- agent layer can deal with privacy dimensions at the inter-agent layer [8], [9].

There is a lack of standardized assessment metrics of multi-agent trustworthiness, which are comparable to ImageNet (vision) or GLUE (NLP), thus hindering cross-study comparison [7], [10]. Interpretable, updatable and auditable value alignment without retraining: Constitutional AI methods, where the actions of the agent are limited by explicit specifications of natural language values interpreted during inference, can be interpreted, updated, and audited [13], [16].

7. Conclusion

This research provided a synthesis of patterns for systems and product design for architecting trustworthy agentic AI systems using 16 primary references in the fields of LLM-based agents, deep RL, explainable AIs (XAI), bias mitigation, multi-agent systems (MAS) safety, and AI governance. One main conclusion reached is that trustworthiness is an architecturally embedded property that must be systematically addressed at all layers of system architecture (e.g., perception, reasoning, execution, governance); thus, trustworthiness must not be a post-hoc evaluation but rather a required attribute of design. Hybrid agentic architectures result in aggregate trustworthiness scores which are 20% greater than LLM-based architectures and 32% better in fairness-specific areas. XAI methodologies (e.g., SHAP based explanations) improve the quality of human decision making from 8% to 37%, an average of 8-37% of the mean improvement in decision-making. The Trustworthy Responsible Intelligent Systems Management (TRiSM) framework provides operational pathways that enable reductions of 94% in impersonating agents and 91% in the occurrence of prompt injection. Design patterns that have a governance-centric focus have resulted in a 43% reduction in the number of compliance-related incidents at organizations worldwide. As a result of projected enterprise-level adoption of agentic AI (73% by mid-2025) and concurrent increases in the number of safety-related incidents, immediate use of the synthesized framework is critically important. Future avenues for development, which include formal verification, federated multi-agent learning, and constitutional AI alignment, are anticipated to lead to additional maturation of these AI systems. Therefore, the need for architecting trustworthy autonomous decision-making is not a long-term goal, but rather, present-day engineering responsibility that requires immediate attention through principled and systematic development of architectures.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.

- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.
- **Use of AI Tools:** The author(s) declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

References

[1] Y. Alufaisan, L. R. Marusich, J. Z. Bakdash, Y. Zhou, and M. Kantarcioglu, "Does explainable artificial intelligence improve human decision-making?" Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 9, pp. 6618–6625, 2020, doi: 10.1609/aaai.v34i09.6600.

[2] S. Arya, B. K., L. Addepalli, V. S. S. D., J. Lloret, and B. Maloth, "Governance-centric framework for trustworthy and ethical autonomous decision-making in organizational systems," in Proc. 1st Int. Conf. Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25), vol. 3, pp. 184–192, 2025, doi: 10.5220/0013893900004919.

[3] S. Baker and W. Xiang, "Explainable AI is responsible AI: How explainability creates trustworthy and socially responsible artificial intelligence," arXiv preprint, 2023, doi: 10.48550/arXiv.2312.01555.

[4] 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, doi: 10.1007/s11023-018-9482-5.

[5] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, "An introduction to deep reinforcement learning," Foundations and Trends® in Machine Learning, vol. 11, no. 3–4, pp. 219–354, 2018, doi: 10.1561/2200000071.

[6] A. K. M. B. Haque, M. Rahman, and M. S. Rahman, "Explainable artificial intelligence (XAI) from a user perspective: A systematic literature review and research agenda," Technological Forecasting and Social Change, vol. 188, Art. no. 122140, 2023, doi: 10.1016/j.techfore.2022.122140.

[7] D. Kowald, N. Rekabsaz, E. Lex, and M. Schedl, "Establishing and evaluating trustworthy artificial intelligence: A systematic overview of requirements and evaluation methods," AI and Ethics, vol. 4, no. 3, pp. 743–759, 2024, doi: 10.1007/s43681-023-00339-4.

[8] H. Liu et al., "Trustworthy AI: A computational perspective," ACM Transactions on Intelligent Systems and Technology, vol. 13, no. 1, Art. no. 1, pp. 1–34, 2021, doi: 10.1145/3442188.

[9] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Computing Surveys, vol. 54, no. 6, Art. no. 115, 2021, doi: 10.1145/3457607.

[10] S. Naihin et al., "Testing language model agents safely in the wild," arXiv preprint, 2023, doi: 10.48550/arXiv.2311.10538.

[11] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," arXiv preprint, 2023, doi: 10.48550/arXiv.2304.03442.

[12] Z. Pei, Y. Zhang, and H. Chen, "Reinforcement learning for decision-making under deep uncertainty," Journal of Environmental Management, vol. 351, Art. no. 120968, 2024, doi: 10.1016/j.jenvman.2024.120968.

[13] S. Raza, R. Sapkota, M. Karkee, and C. Emmanouilidis, "TRiSM for agentic AI: A review of trust, risk, and security management in LLM-based agentic multi-agent systems," arXiv preprint, 2025, doi: 10.48550/arXiv.2506.04133.

[14] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA, USA: MIT Press, 2018, doi: 10.7551/mitpress/10955.001.0001.

[15] C. Wu, Y. Zhao, and Z. Liu, "Survey of trustworthy artificial intelligence: Foundations and future research directions," arXiv preprint, 2023, doi: 10.48550/arXiv.2306.00380.

[16] Z. Xi et al., "The rise and potential of large language model based agents: A survey," arXiv preprint, 2023, doi: 10.48550/arXiv.2309.07864.