UDC 004.415:005.334

TRUST-ADJUSTED RISK ASSESSMENT IN A MULTI-AGENT DECISION SUPPORT SYSTEM FOR SOFTWARE DEVELOPMENT

Mazepa A.S., Zadrykin S.O.
e-mail: andrii.mazepa@nure.ua, serhii.zadrykin@nure.ua
Scientific Supervisor – Cand. Sc. (Tech.), Ass. Prof. Brynza N.O.
Kharkiv National University of Radio Electronics,
Department of Applied Mathematics
Kharkiv, Ukraine

У роботі узагальнено сучасні підходи до оцінювання ризиків у багатоагентних системах підтримки прийняття рішень у межах парадигми TRiSM (Trust, Risk and Security Management). Показано, що сучасні агентні архітектури штучного інтелекту, які поєднують автономних агентів і центральний планувальник, формують багаторівневе поширення ризиків та каскадну динаміку відмов, що обумовлює обмеженість класичних моделей оцінювання ризиків, заснованих лише на ймовірності та впливі. Запропоновано підхід до оцінювання ризиків із урахуванням довіри як розширення традиційної моделі шляхом введення динамічного коефіцієнта довіри, який враховує відповідність політикам, результати валідації та моніторинг під час виконання. Встановлено узгодженість запропонованого підходу з принципами TRiSM та можливість його інтеграції з механізмами управління і валідації. Показано, що застосування підходу сприяє підвищенню системної стійкості та зниженню ефекту підсилення ризиків у багатоагентних СППР із централізованим планувальником.

In paper [4], a dynamic multi-agent decision support system (DSS) for risk management was considered, along with the implementation of an independent governance module incorporating a validation loop. The proposed architecture was based on the orchestration of functional agents responsible for log analysis, risk assessment, regulatory compliance monitoring, and audit procedures. The further development of agentic AI, particularly LLM-oriented multi-agent systems, necessitates the systematic incorporation of trust into the risk assessment process.

As noted in recent TRiSM research [1], “risk in agentic AI systems is distributed across interacting components rather than localized within a single model,” and “planner-driven coordination may amplify localized errors into systemic failures.” When an AI agent represents a composition of independent agents and a planner, risk acquires a multi-level structure: it emerges at the level of individual agents, at the planner level, and at the level of their coordination. Under such conditions, the classical risk assessment model becomes insufficient:

$$R_i = P_i \cdot I_i,$$

where P_i – the probability of occurrence of the event and I_i – the magnitude of its impact does not account for the structural reliability of the decision source.

The TRiSM concept states that "trust metrics must be continuously evaluated and enforced at runtime", and governance should operate as a "cross-layer control mechanism spanning planning, coordination, execution and monitoring." This implies that trust is a dynamic parameter that directly influences the interpretation of agent-generated analytical outcomes. Within the generalization of contemporary trust-weighted risk assessment approaches, a trust-adjusted risk model is considered in the context of a multi-agent DSS:

$$R_{i^*} = P_i \cdot I_i \cdot T_i(t),$$

where $T_i(t) \in (0,\ 1]$ is the dynamic trust coefficient assigned to an agent or to the planner at time t. The value of $T_i(t)$ reflects policy compliance, validation loop outcomes, as well as the history of anomalies recorded by the governance layer. This formulation is not introduced as a fundamentally new equation, but rather as a logical generalization of reliability-weighted and trust-weighted models adapted to a multi-agent DSS architecture.

The trust coefficient may be updated recursively according to:

$$T_i(t+1) = T_i(t) - \alpha V_i + \beta S_i,$$

where V_i – denotes the number of policy violations or detected anomalies, S_i – yrepresents successful validation outcomes, and α, β – are system sensitivity parameters to negative and positive events, respectively. Consequently, trust becomes a function of the agent's behavioral history and the effectiveness of its decisions.

The integration of the trust-adjusted risk assessment approach is implemented within the TRiSM control loop, which represents a continuous and iterative governance cycle including risk identification, multi-agent assessment, runtime governance, execution of the validation loop, decision acceptance or escalation to human-in-the-loop intervention when the level of trust is insufficient, as well as audit and compensation mechanisms [1, 4]. This structure corresponds to iterative risk management models [4] in which identification, assessment, mitigation, monitoring, and reporting form a cyclic process that continuously refines both risk evaluation and response strategies. Within this cycle, the governance layer operates [2, 4] as an independent supervisory component that enforces policy compliance, ensures traceability, and provides explainability of agent-generated decisions without modifying the internal logic of individual agents. The validation loop functions as a quality control mechanism by verifying agent outputs against reference datasets, validation rules, or simulation-based checks, while insufficient confidence triggers retry strategies or alternative analytical paths, and in high-uncertainty scenarios activates human-in-the-loop intervention to maintain reliability and auditability. The inclusion of compensation mechanisms, such as saga-based rollback strategies, ensures consistency in distributed workflows and enables controlled recovery from partial failures. In this context, the TRiSM layer acts as a cross-layer control mechanism spanning planning, coordination, execution, and monitoring, ensuring controlled interaction between independent

agents and the central planner and preventing cascading amplification of localized errors into systemic failures.

Trust-adjusted risk assessment in a multi-agent decision support system can therefore be interpreted as a generalization of contemporary TRiSM and Governance-as-a-Service paradigms, where trust is treated as a dynamic parameter that is continuously evaluated and directly influences both decision acceptance and the interpretation of risk. Unlike traditional models that assume static reliability, this approach incorporates behavioral history, validation outcomes, and policy compliance into the risk formulation, thereby reflecting the structural reliability of distributed decision sources. The proposed model enhances systemic robustness by stabilizing decision-making processes under uncertainty, reduces variability in outcomes by filtering low-trust agent contributions, and limits the propagation of structural risks across interconnected components in planner-based agentic architectures. Empirical evidence from related domains indicates that orchestrated and governed multi-agent systems improve adaptability, explainability, and compliance in dynamic environments, particularly in cybersecurity, healthcare, and infrastructure management. These results support the integration of trust-aware mechanisms into risk assessment frameworks for complex systems. Further research is aimed at formalizing stability criteria for multi-agent orchestration, including the analysis of risk propagation dynamics and inter-agent dependencies, as well as experimentally validating the impact of the dynamic trust coefficient on the system's integral risk and overall resilience under varying operational conditions.

References:

1. Raza S., Sapkota R., Karkee M., Emmanouilidis C. TRiSM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems. arXiv:2506.04133, 2025. URL: https://arxiv.org/abs/2506.04133.

2. Gaurav S., Heikkonen J., Chaudhary J. Governance-as-a-Service: A Multi-Agent Framework for AI System Compliance and Policy Enforcement. arXiv:2508.18765, 2025. URL: https://arxiv.org/abs/2508.18765.

3. Gosmar D., Dahl D.A. Sentinel Agents for Secure and Trustworthy Agentic AI in Multi-Agent Systems. arXiv:2509.14956, 2025. URL: https://arxiv.org/abs/2509.14956.

4. Mazepa A.S. Governance Module and Validation Loop in a Multi-Agent Decision Support System for Risk Management in IT Projects // Комп'ютерне моделювання та оптимізація складних систем (КМОСС-2025): матеріали IX Міжнародної науково-технічної конференції, 5-7 листопада 2025 р. Дніпро, 2025. С. 164-166. URL: https://udhtu.edu.ua/wp-content/uploads/2025/11/zbirnyk-kmoss-2025_compressed.pdf.