Poster: Beyond STRIDE: A Unified Multi-Attribute Framework for Trustworthy AI

Jun Hyeong Lee and Kwangsoo Cho
School of Cybersecurity
Korea University
Seoul, South Korea

Jieun Lee, Jae Hyuk Suk,
Yoon-chan Jhi and Jihoon Cho
Samsung SDS
Seoul, South Korea

Seungjoo Kim
School of Cybersecurity
Korea University
Seoul, South Korea

***Abstract*—Existing studies on trustworthy AI tend to adopt a narrow definition, often focus on only one or two attributes such as security or fairness, failing to comprehensively address the risks of AI-based services. In this work, we systematically analyze the literature to clearly define a set of non-overlapping trustworthy AI attributes and propose a design-phase framework that integrates eight domain-specific methods into a unified process to verify whether these attributes are operationally incorporated into AI system design. Validation shows 98% alignment with the EU AI Act's technical obligations and 92.2% inter-rater consistency among independent evaluators assessing deployed LLM-based services. We also demonstrate the practicality of the proposed approach through a national-level AI red team competition.**

1. Introduction and Motivation

As AI-based services operate with increasing autonomy, ensuring their trustworthiness has become a global priority [1], [2]. However, despite extensive discourse on AI trustworthiness, current approaches often fail to cover all relevant attributes. Specifically, traditional threat models like STRIDE support security-by-design but lack the scope to handle AI-specific issues like non-determinism, opacity, and data dependency. Consequently, a critical question remains: Are these existing methods truly adequate for managing risks in AI-driven environments? We structure our investigation around four Research Questions: **RQ1.** Are traditional IT threat modeling methods sufficient for AI-based services? **RQ2.** What multi-attribute framework should be employed for trustworthy AI design? **RQ3.** What additional requirements are identified through the proposed framework? **RQ4.** Are the derived requirements aligned with regulations (e.g., EU AI Act) and consistently interpretable?

2. Insufficiency of Prior Threat and Risk Analysis Approaches for AI-Based Services (RQ1)

Leading regulatory and technical frameworks, such as the NIST AI Risk Management Framework (AI RMF), identify up to eight attributes that trustworthy AI systems should possess (Table 1). Each of these attributes has a distinct definition; for example, accountability concerns the assignment of decision-making responsibility through RACI roles, whereas fairness relates to whether system outputs ensure equitable treatment across protected groups. However, existing requirements analysis and design methodologies address only a subset of these attributes. Our review of 40 prior studies reveals that even the most comprehensive works cover at most two of the eight attributes, with the vast majority focusing solely on security through STRIDE or its variants [6]. Moreover, even MAESTRO [9], a recently proposed framework for agentic AI, remains focused on security analysis and does not comprehensively cover all eight attributes. To address this gap, we propose a unified multi-attribute framework that integrates eight domain-specific methods within a design-phase process (RQ2–RQ4).

TABLE 1. TRUSTWORTHINESS ATTRIBUTES IN MAJOR FRAMEWORKS AND PRIOR WORK COVERAGE.

Framework	Sec	Saf	Rel	Tra	Acc	Fair	Hum	Priv
NIST AI RMF [1], Khlaaf (2023) [4]	✓	✓	✓	✓	✓	✓		✓
EU AI Act [2], ISO/IEC 22989 [3]	✓	✓	✓	✓	✓	✓	✓	✓
Japan AISI (2024) [5]	✓	✓		✓		✓	✓	✓
Max coverage of prior work	**2/8** (Mauri 2022 [6])							

Sec: Security; Saf: Safety; Rel: Reliability; Tra: Transparency; Acc: Accountability; Fair: Fairness; Hum: Human-centricity; Priv: Privacy.

3. Proposed Unified Multi-Attribute Framework (RQ2)

To address RQ2, we established a four-stage filtering process (S1–S4), reviewing 51 literature: **▲S1** classified trustworthy AI attributes using two AI conceptual standards; **▲S2** surveyed seven AI-specific standards and directives; **▲S3** identified 21 methods from two domain-specific standards and two SoK/survey papers [7], [8]; and **▲S4** verified method's applicability using 38 papers. The results of this process are shown in Figure 1 and 2. Notably, as shown in Figure 2, MAESTRO was not selected as one of the eight analysis methods because, having been published in early 2025, it has not yet accumulated sufficient peer-reviewed case studies to meet our credible criterion (S4-2).

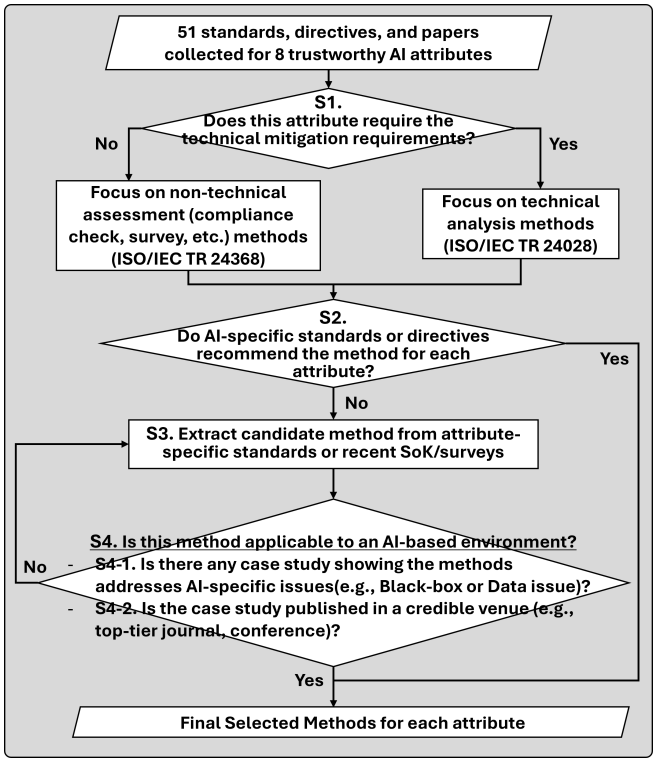

Figure 1. Four-stage filtering process (S1–S4).

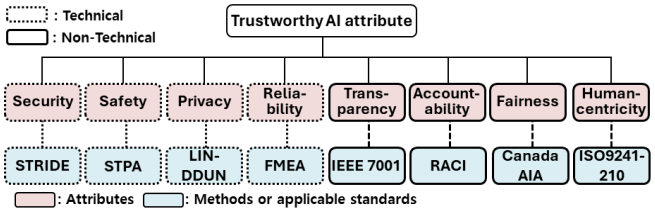

Figure 2. Eight analysis methods selected through Figure 1.

4. What STRIDE Misses: Coverage Gap Analysis (RQ3)

A STRIDE-only baseline produced 27 requirements, covering only 24% of the EU AI Act's technical obligations—with zero coverage in Transparency (Art. 13) and Human oversight (Art. 14). In contrast, our multi-attribute methodology derived 103 requirements, capturing risk classes that single-method approaches fail to identify. For example, STPA identified hazards in which AI agents cause harm through fully authorized actions—risks that remain undetected by threat-centric methods such as STRIDE.

5. Regulatory Alignment and Real-World Validation (RQ4)

- **Regulatory Alignment:** Mapping the 103 requirements to the EU AI Act achieved 98% coverage (98 of 100 technical obligations), compared to 24% with STRIDE alone. Detailed data is available online at https://github.com/HackProof/trustworthy_ai_by_design.
- **Empirical Evaluation:** External evaluators assessed three deployed models (GPT-4o, Gemini 2.5 flash, DeepSeek V3.1) via structured audits and prompt-based red teaming, achieving 92.2% inter-rater consistency.
- **Real-World Practice:** We placed second among 47 teams at the AI Red Teaming Challenge organized by the Korean Ministry of Food and Drug Safety. The attack techniques employed by our team in this competition—including role-play jailbreaking and encoding bypass—directly map to threat classes in our analysis.

6. Conclusion

Security-focused threat modeling alone is insufficient for AI-based services. By integrating eight domain-specific analysis methods into a unified design-phase process, we demonstrate that trustworthy AI by design can be systematically achieved. Our multi-attribute framework provides a rigorous and comprehensive foundation that addresses the structural limitations inherent in single-attribute approaches.

Acknowledgments

This work was supported by Samsung SDS. Jun Hyeong Lee and Kwangsoo Cho contributed equally to this work. Seungjoo Kim is the corresponding author of this paper (E-mail: skim71@korea.ac.kr).

References

[1] NIST, "Artificial intelligence risk management framework (AI RMF 1.0)," Tech. Rep., Jan. 2023.

[2] European Parliament and Council of the EU, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act)," *Official J. Eur. Union*, L, Jul. 2024.

[3] ISO/IEC, "ISO/IEC 22989: Information technology - Artificial intelligence concepts and terminology," 2022.

[4] H. Khlaaf, "Toward comprehensive risk assessments and assurance of AI-based systems," Trail of Bits, Tech. Rep., 2023.

[5] AI Safety Institute (AISI), Japan, "Guide to red teaming methodology on AI safety," Tech. Rep., 2024.

[6] L. Mauri and E. Damiani, "STRIDE-AI: An approach to identifying vulnerabilities of machine learning assets," in *Proc. IEEE Int. Conf. Cyber Security Resilience (CSR)*, 2021, pp. 147–154.

[7] H. Cho and S. Kim, "Threat modeling for the defense industry: Past, present, and future," *IEEE Access*, vol. 13, pp. 53276–53288, 2025.

[8] N. Azam, L. Michala, S. Ansari, and N. B. Truong, "Data privacy threat modelling for autonomous systems: A survey from the GDPR's perspective," *IEEE Transactions on Big Data*, vol. 9, no. 2, pp. 388–398, 2023.

[9] K. Huang, "Agentic AI threat modeling framework: MAESTRO," Cloud Security Alliance (CSA), Feb. 2025. [Online]. Available: https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro

Beyond STRIDE: A Unified Multi-Attribute Framework for Trustworthy AI

Jun Hyeong Lee[1], Kwangsoo Cho[1], Jieun Lee[2], Jae Hyuk Suk[2], Yoon-Chan Jhi[2], Jihoon Cho[2], Seungjoo Kim[1]
[1]School of Cybersecurity, Korea University, [2]Samsung SDS

SANE LAB 保安工學研究室

1 Introduction & Motivation

- **Global priority:** Ensuring trustworthiness as AI-based services operate with increasing autonomy
- **Gap in existing methods:** Traditional threat models (e.g., STRIDE) fail to handle AI-specific issues such as non-determinism, opacity, and data dependency

Research Questions(RQs):

- **RQ1.** Are traditional IT threat modeling methods sufficient for AI-based services?
- **RQ2.** What multi-attribute framework should be employed for trustworthy AI design?
- **RQ3.** What additional requirements are identified through the proposed framework?
- **RQ4.** Are the derived requirements aligned with regulatory obligations (e.g., EU AI Act) and consistently interpretable by external evaluators?

2 Insufficiency of Prior Threat and Risk Analysis Approaches for AI-Based Services RQ1

- **Up to 8 Attributes**: Leading frameworks (e.g., NIST AI RMF, EU AI Act) identify up to eight attributes that trustworthy AI systems should possess (Table 1).
- **Review of 40 Prior Studies:** Even the most comprehensive works cover at most 2/8 attributes, with the vast majority focusing solely on security via STRIDE or its variants.
- **Gap:** Even MAESTRO, a recent framework for agentic AI, remains security-focused and does not cover all eight attributes — motivating a unified multi-attribute framework

Table 1. Trustworthiness attributes in major frameworks and prior work coverage.

Framework	Sec	Saf	Rel	Tra	Acc	Fair	Hum	Priv
NIST AI RMF [1], Khlaaf (2023) [4]	✓	✓	✓	✓	✓	✓		✓
EU AI Act [2], ISO/IEC 22989 [3]	✓	✓	✓	✓	✓	✓	✓	✓
Japan AISI (2024) [5]	✓	✓		✓		✓	✓	✓
Max coverage of prior work	**2/8** (Mauri 2022 [6])							

Sec: Security; Saf: Safety; Rel: Reliability; Tra: Transparency; Acc: Accountability; Fair: Fairness; Hum: Human-centricity; Priv: Privacy.

3 Proposed Unified Multi-Attribute Framework RQ2

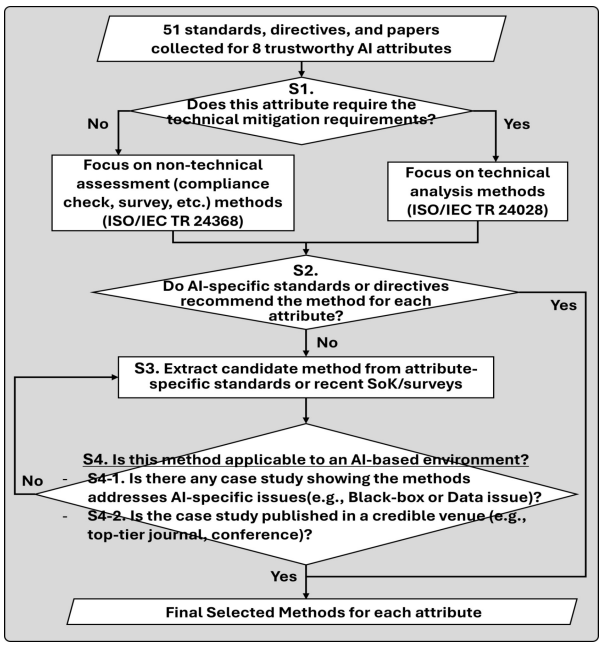

Figure 1. Four-stage filtering process (S1–S4).

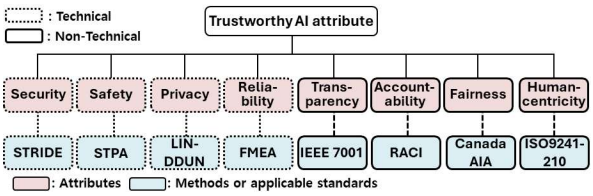

Figure 2. Eight analysis methods selected through Figure 1.

- Four-stage filtering process (S1–S4) reviewing 51 standards, directives, and papers to identify domain-specific analysis methods(Fig. 1)
- S1: Classify attributes → S2: Review AI standards/directives → S3: Review domain standards & SoK/surveys → S4: Check AI applicability and credibility → Result in Fig. 2
- Case Study: Deriving unified trustworthy AI requirements by applying Fig.3 to reference architecture of AI-based services → 103 unified requirements derived

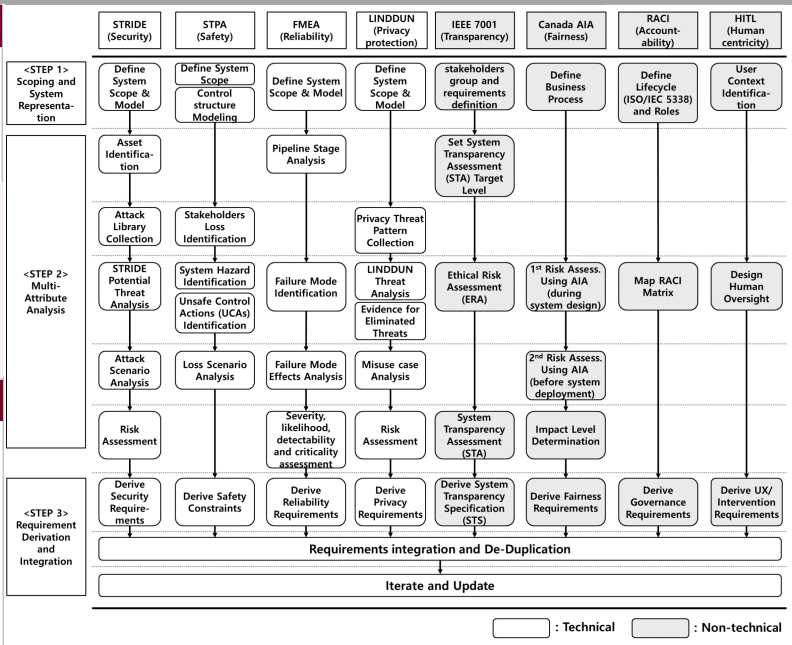

Figure 3. A Unified multi-attribute framework

4 What STRIDE Misses: Coverage Gap Analysis RQ3

- **STRIDE-only:** 27 reqs. — 24% EU AI Act coverage;
  ※ Zero coverage (0%) in Transparency (Art. 13) & Human Oversight (Art. 14).
- **Our Framework:** 103 requirements across all 8 attributes
  - e.g., STPA captured hazards from fully authorized AI actions — invisible in STRIDE

5 Regulatory Alignment and Real-World Validation RQ4

- **Regulatory Alignment:** 98% EU AI Act technical obligation coverage (vs. 24% STRIDE-only)
  * Detailed data is available on-line at https://github.com/HackProof/trustworthy_ai_by_design
- **Empirical Evaluation:** GPT-4o, Gemini 2.5 Flash, DeepSeek V3.1 — 92.2% inter-rater consistency
- **Real-World Validation:** 2nd / 47 teams at AI Red Teaming Challenge hosted by the MFDS (Ministry of Food and Drug Safety) — employing attack scenarios we identified

Table 2. EU AI Act coverage: STRIDE-only vs. multi-attribute.

EU AI Act Category	Units	STRIDE	Multi-Attr.
Art. 9 Risk management	18	4	18
Art. 10 Data governance	15	3	15
Art. 12 Logging & monitoring	8	4	8
Art. 13 Transparency	19	0	19
Art. 14 Human oversight	10	0	10
Art. 15 Technical robustness	20	12	19
Other obligations	10	1	9
Total	**100**	**24 (24%)**	**98 (98%)**

6 Conclusion

Security-focused threat modeling alone is insufficient for AI-based services. By integrating eight domain-specific methods, our framework demonstrates that trustworthy AI by design can be systematically achieved.

Acknowledgments

This work was supported by Samsung SDS. Jun Hyeong Lee and Kwangsoo Cho contributed equally to this work. Seungjoo Kim is the corresponding author of this paper (E-mail: skim71@korea.ac.kr).

References

[1] NIST, "Artificial intelligence risk management framework (AI RMF1.0)," NIST, Tech. Rep., 2023.
[2] European Parliament, "Regulation (EU) 2024/1689 on artificial intel-ligence (EU AI Act)," 2024.
[3] ISO/IEC, "ISO/IEC 22989: Information technology - Artificial intelli-gence concepts and terminology," 2022.
[4] H. Khlaaf, "Toward comprehensive risk assessments and assurance ofai-based systems," tech. rep., Trail of Bits, 2023.
[5] AI Safety Institute (AISI), Japan, "Guide to red teaming methodologyon ai safety," tech. rep., 2024.
[6] L. Mauri and E. Damiani, "STRIDE-AI: An approach to identifyingvulnerabilities of machine learning assets," in Proc. IEEE Int. Conf.Cyber Security Resilience (CSR), 2021, pp. 147–154.
[7] H. Cho and S. Kim, "Threat Modeling for the DefenseIndustry: Past, Present, and Future," IEEE Access, vol. 13, 2025,pp. 53276–53288.
[8] N. Azam, L. Michala, S. Ansari, and N. B. Truong, "Data privacythreat modelling for autonomous systems: A survey from the GDPR'sperspective," IEEE Transactions on Big Data, vol. 9, no. 2, pp. 388–398, 2023.
[9] K. Huang, "Agentic AI threat modeling framework: MAESTRO," Cloud Security Alliance (CSA), Feb. 2025. [Online].Available: https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro

If you have any feedback on our research, please feel free to share it in the survey!