Securing the AI Supply Chain: A Runtime Interceptor for Safe AI Execution

Shannie Chekani

5th KTH Workshop on the Software Supply Chain 2026

AI agents can reason freely — but cannot act without validation

1. FROM CHAT TO ACTION

- AI → autonomous agents
- Risks:
  - Distributed reasoning
  - Agentic drift
  - Action hallucination

- Unsafe actions reach production

- Solution:
Runtime interception layer

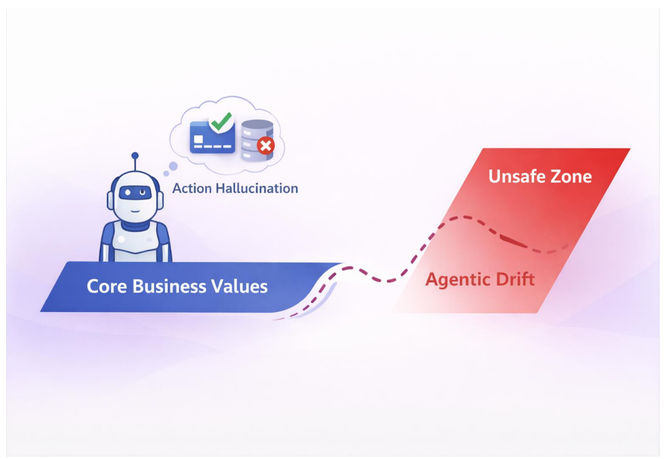

Figure 1: Agentic drift causing execution of unsafe or non-existent actions.

2. ARCHITECTURE

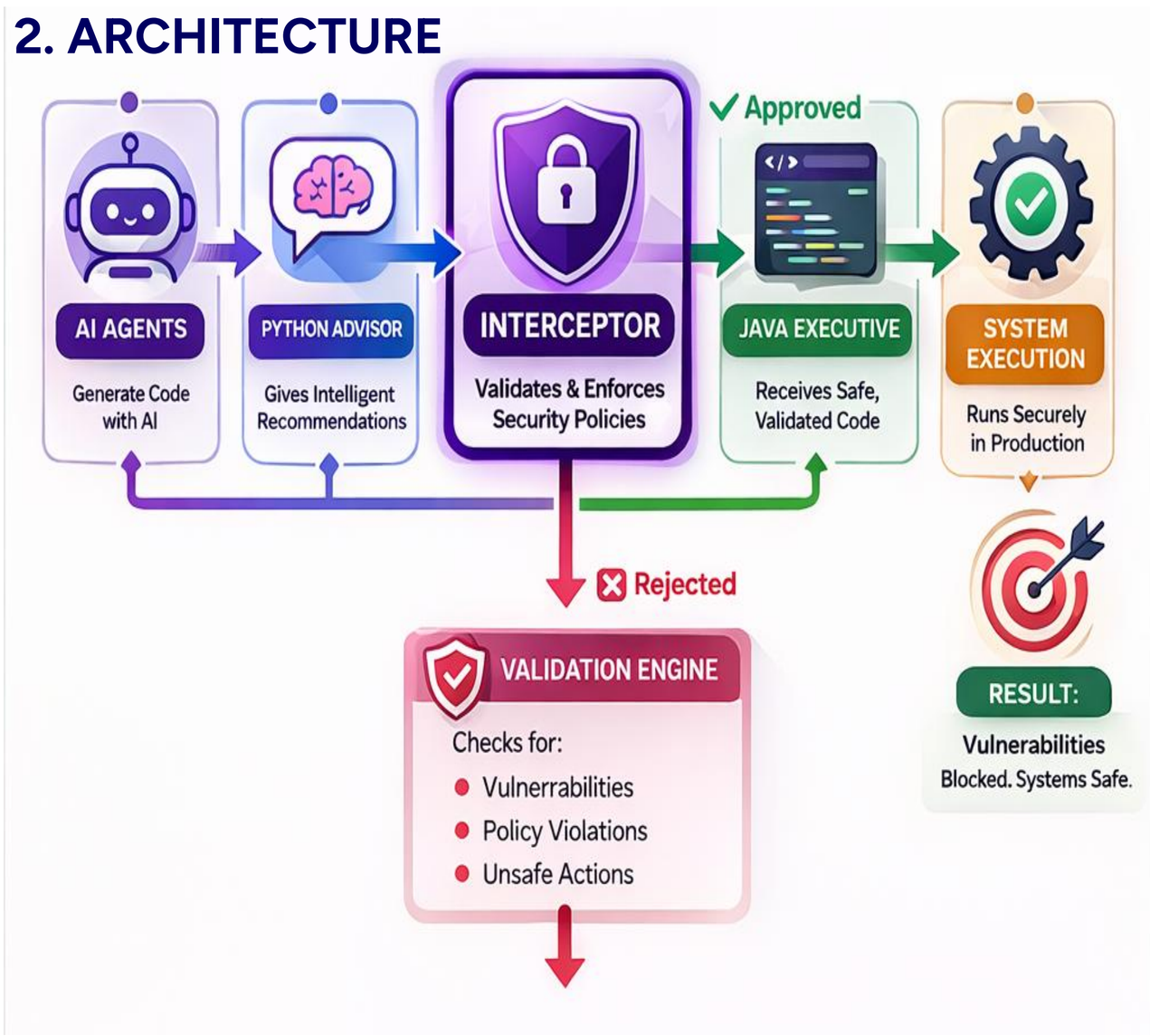

Figure 2: Runtime interception layer validating agent actions before execution.

3. SYSTEM IMPACT

- Safety guarantees
- Regulatory alignment
- Scalable automation

[1] Amodei, D., et al. 'Concrete Problems in AI Safety', arXiv *preprint*, 2016.

[2] Balliu, M., et al. 'CHAINS_ Consistent Hardening and Analysis of Software Chains (CHAINS)', *KTH Royal Institute of Technology Research*.

[3] European Parliament, 'Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations', *Official Journal of the European Union*.
Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[4] Huang, L., et al. 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', *ACM Transactions on Information Systems*, vol 43, pp. 1 - 55

[5] Pydantic Team. 'PydanticAI: Model-agnostic agent framework', Available: Pydantic AI | Pydantic Docs

[6] Wu, Q., et al. 'AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation', *Microsoft Research,* 2023.