

Mediagents: System of systems mediators inhibit unwanted emergence in artificial intelligence pipelines

Luke H. Douglas
Timothy J. Eveleigh

AI4SE & SE4AI Workshop 2026

School of Engineering & Applied Science



THE GEORGE
WASHINGTON
UNIVERSITY
WASHINGTON, DC

Abstract

Mediagents are a novel application of System of Systems (SoS) mediators that inhibit unwanted, emergent behavior in artificial intelligence (AI) pipelines

- Mediators promote collaboration between constituent systems in an SoS, which can result in emergent behavior
- Mediagents inhibit collaboration by monitoring for emergent behavior - examining interactions between agents or the system and its deployed context and intervening when necessary

Mediagents add an additional layer of safety to AI pipelines, beyond agent or pipeline output assessment

- Can catch out-of-policy system behaviors missed by other AI governance frameworks
- Applicable for both weak and strong emergence
- Examples from customer service and military target acquisition AI pipelines show improvement in performance via reductions in adverse outcomes

Terminology note:
“Mediagent” = “Mediator” + “Agent”

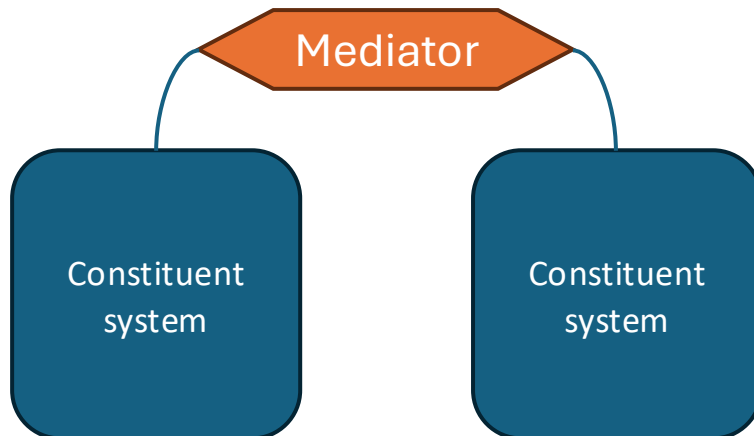
 Many AI governance frameworks exist; mediagents add additional, SE-driven layer

Framework	Description	Component governed	Violation monitored	Resolution method
Enforcement agent (Tamang & Bora, 2025)	Supervisory agent monitoring peers	Individual agent	Out-of-policy agent behavior	Reformation
Governance as a Service (Gaurav et al., 2025)	Enforcement layer governing agent outputs	Individual agent	Out-of-policy agent behavior	Escalation, Containment
MI9 (Wang et al., 2025)	Runtime governance framework	Individual agent, agent-agent	Out-of-policy agent behavior	Escalation, Containment
RV4JaCa (Engelmann et al. 2023)	Runtime verification of multi-agent systems	Agent-agent	Out-of-policy communication	Not specified
LOKA Protocol (Ranjan et al., 2025)	Structured orchestration layer for agent-agent communication	Agent-agent, system-system, system-environment	Out-of-consensus agent behavior, out-of-consensus system behavior	Escalation, Reformation
Constitutional reasoning module (Birari, 2025)	Evaluation of potential system actions against ethical principles	System-environment	Out-of-policy system behavior	Reformation
Mediagent (This work)	Supervisory interaction monitoring system	Agent-agent, system-environment	Out-of-policy system behavior	Escalation, Reformation, Containment

Sources: (Tamang and Bora 2025), (Gaurav et al. 2025), (Wang et al. 2025), (Engelmann et al. 2023), (Ranjan et al. 2025), (Birari 2025)

Terminology note:
Escalation = warn about the potential output
Reformation = change the output
Containment = stop the process

Mediators (can) create emergence...

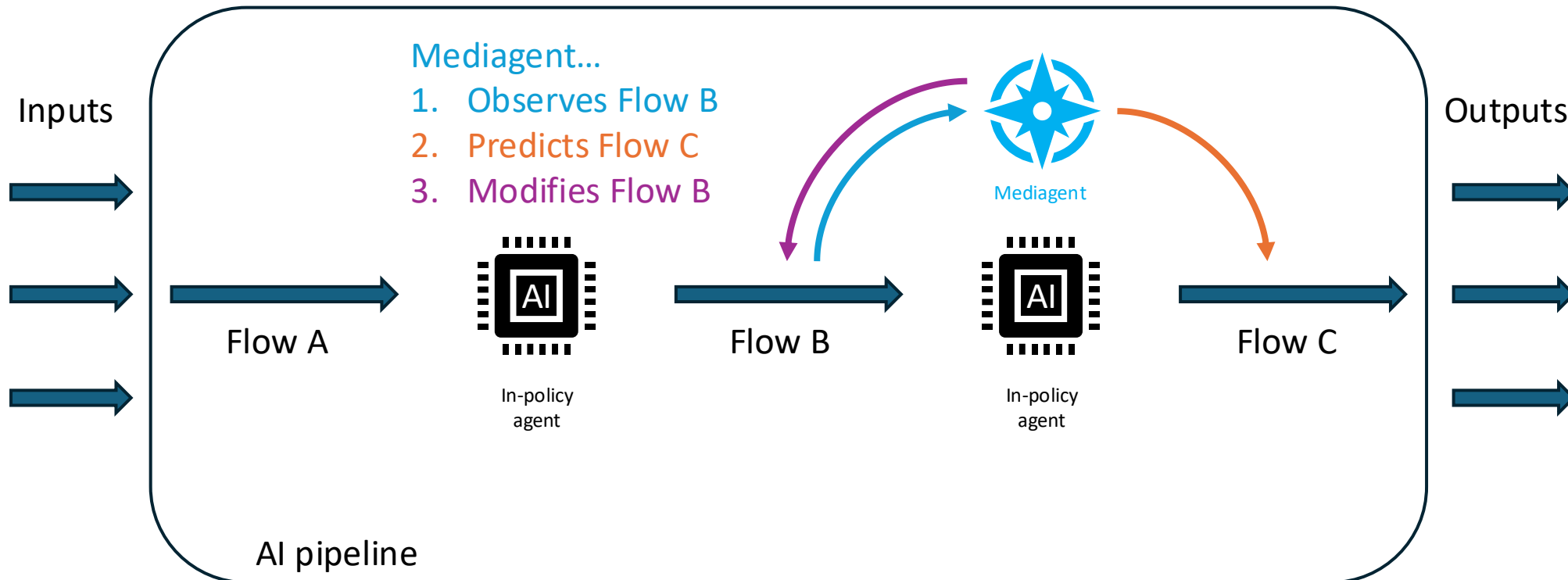


*“**Mediators** are elements that facilitate the collaboration between the constituents of the SoS or ecosystem, without having a role outside it.”*

Emergence Type	Description
Simple	Predictable behavior that arises from well-understood system elements and relationships
Weak	Behavior that cannot be analytically predicted from system elements or relationships alone, but require simulation or observation to explore
Strong	Unexpected behavior that is not anticipated during design or analysis

*“**Emergence** refers to system-level properties or behaviors that arise from these interactions and cannot be attributed to any single component.”*

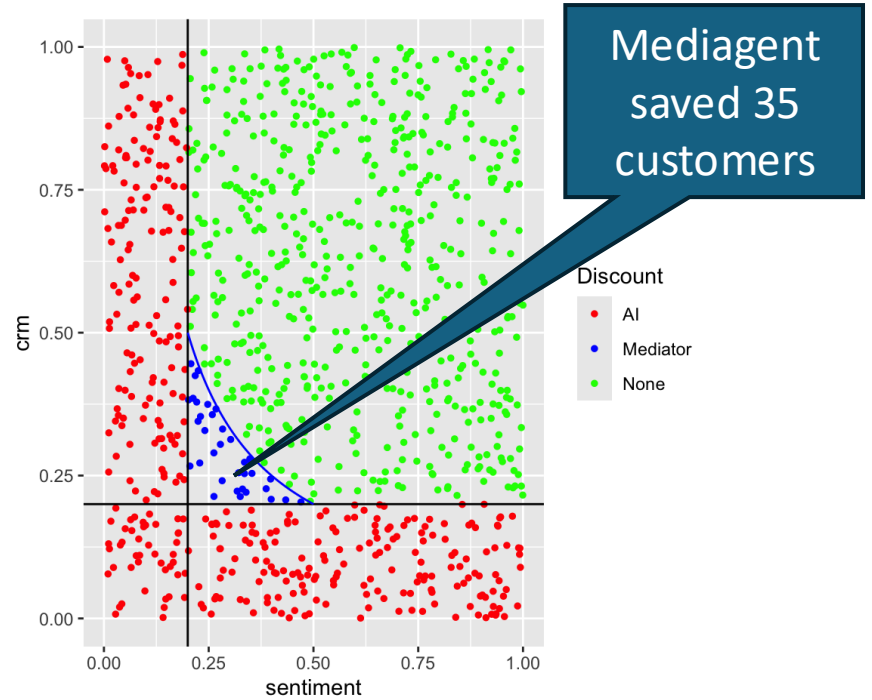
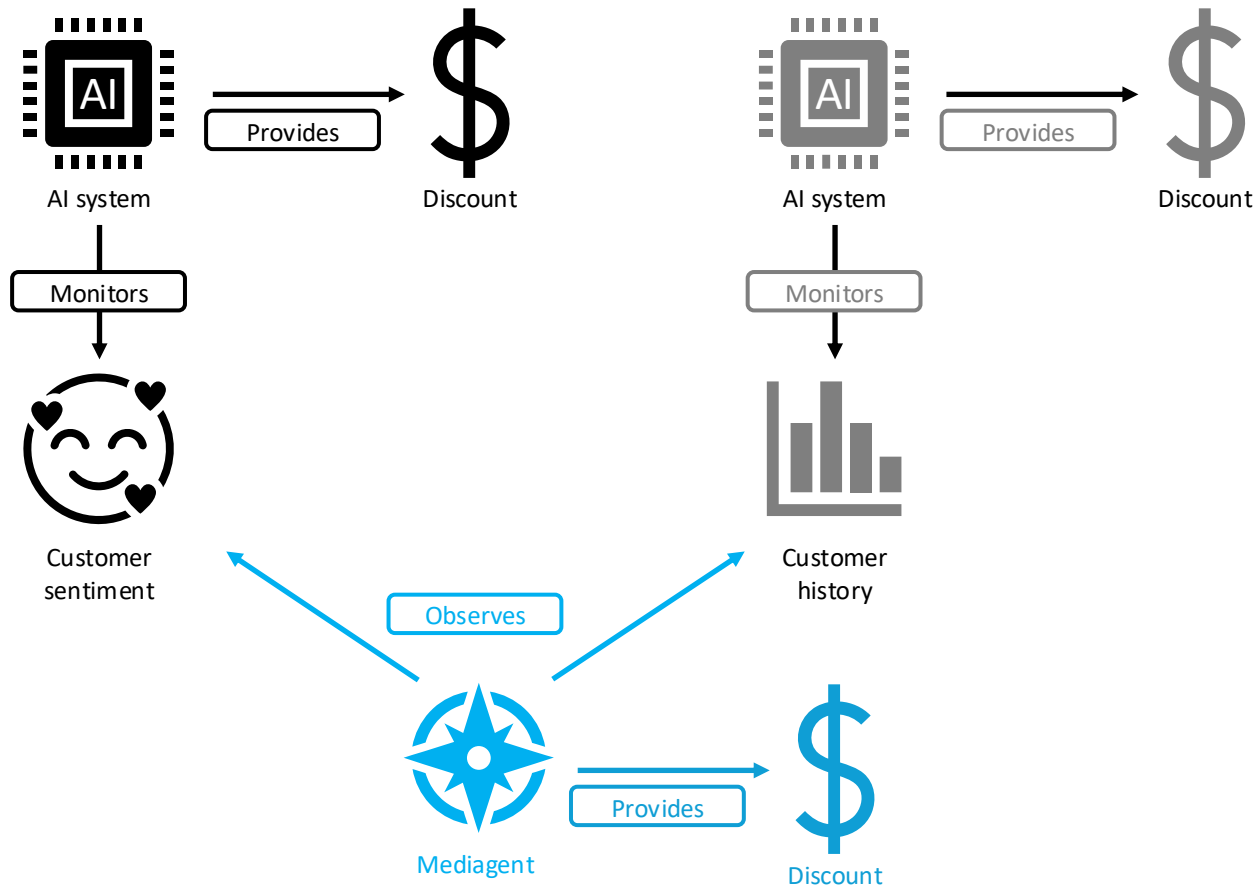
...so they (can) also inhibit emergence



More formally: "Given an overall pipeline goal G , a pipeline function $f()$, a set of rules R , a mediagent correction function $C()$, and output from a process step O , the mediagent assesses if $f(O)$ will result in non-compliance with R , correcting O such that $f(C(O))$ complies with R while still satisfying G ."



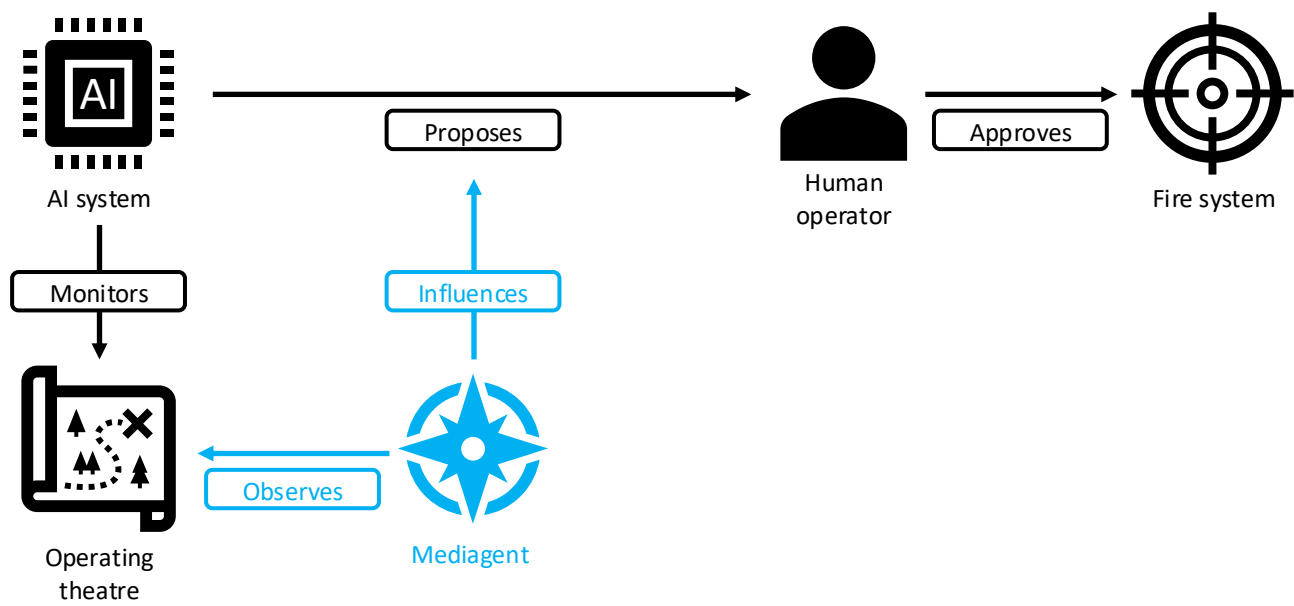
What it looks like: Automated customer service interactions



1000 simulated customer interactions, CRM & sentiment thresholds < 0.2, mediator threshold of CRM*sentiment < 0.1



What it looks like: Target acquisition under domain shift



1000 simulations, potential targets are uniform random decimal between 0 and 1 (>0.8 considered a 'true' target). Gaussian noise added to each random decimal (mean = 0, sd = 0.1). AI assesses each noisy value, passing to human for review if >0.8 . Mediagent passes to human for review if >0.7 due to observed domain shift. Human reviews up to 100 targets, approving as a hit when base value above >0.8

	Base SoS	w/ mediagent
Targets monitored	501 (471, 533)	336 (319, 354)
Correct proposals	80 (78, 83)	64 (61, 67)
% correct	16%	19%
Median (Q1, Q3)		

Mediagent reduced throughput ~33%; improved % of correct proposals by 3%



Wrap up

Current status: Mediagent shows positive impact conceptually and in simple examples, not yet rigorously proven, and is not actively deployed

What's next:

- Deciding on level of evidence necessary to support idea as a plausible intervention
- Exploring impact of modeling adaptation into mediagent behavior

Open questions:

- How much application/evidence is necessary to prove the concept?
- Are there other AI pipelines/domains worth investigating as additional examples?

Contact: Luke Douglas (luke.douglas@gwu.edu)

Thank you!

School of Engineering & Applied Science



THE GEORGE
WASHINGTON
UNIVERSITY
WASHINGTON, DC

References

- Axelsson J. Mediators in Systems-of-Systems and Ecosystems: A Systematic Literature Review and Conceptualization. In: Proceedings of the 12th ACM/IEEE International Workshop on Software Engineering for Systems-of-Systems and Software Ecosystems. ACM; 2024:21-28. doi:10.1145/3643655.3643880
- Birari, Tanvi. 2025. "Constitutional AI for Autonomous Systems: Building Ethical Constraints into Decision-Making Agents." *International Journal Of Scientific Research In Engineering & Technology*, December 24, 45–54. <https://doi.org/10.59256/ijcreat.20250506008>.
- Engelmann, Debora C., Angelo Ferrando, Alison R. Panisson, Davide Ancona, Rafael H. Bordini, and Viviana Mascardi. 2023. "RV4JaCa—Towards Runtime Verification of Multi-Agent Systems and Robotic Applications." *Robotics* 12 (2): 49. <https://doi.org/10.3390/robotics12020049>.
- Gaurav, Suyash, Jukka Heikkonen, and Jatin Chaudhary. 2025. "Governance-as-a-Service: A Multi-Agent Framework for AI System Compliance and Policy Enforcement." arXiv:2508.18765. Preprint, arXiv, August 27. <https://doi.org/10.48550/arXiv.2508.18765>.
- Ranjan, Rajesh, Shailja Gupta, and Surya Narayan Singh. 2025. "LOKA Protocol: A Decentralized Framework for Trustworthy and Ethical AI Agent Ecosystems." arXiv:2504.10915. Preprint, arXiv, April 22. <https://doi.org/10.48550/arXiv.2504.10915>.
- Smith G, Sillitto H, Sheard S, Adcock R. Emergence and Complexity. In: The Guide to the Systems Engineering Body of Knowledge (SEBoK). v. 2.14. N. Hutchison (Editor in Chief). The Trustees of the Stevens Institute of Technology; 2026. Accessed August 31, 2026. https://sebokwiki.org/wiki/Emergence_and_Complexity
- Tamang, Sagar, and Dibya Jyoti Bora. 2025. "Enforcement Agents: Enhancing Accountability and Resilience in Multi-Agent AI Frameworks." arXiv:2504.04070. Preprint, arXiv, April 5. <https://doi.org/10.48550/arXiv.2504.04070>.
- Wang, Charles L., Trisha Singhal, Ameya Kelkar, and Jason Tuo. 2025. "MI9: An Integrated Runtime Governance Framework for Agentic AI." arXiv:2508.03858. Preprint, arXiv, November 18. <https://doi.org/10.48550/arXiv.2508.03858>.