

Dynamic Constitutional Control of Agentic Enterprise Digital Twins via Meta-Governor Agents

Rakesh Kumar Agrawal | IEEE Senior Member | rkagrawal@ieee.org

Abstract:

Enterprise digital twins are evolving from passive monitoring replicas into agentic, decision-capable cyber-physical intelligence layers that can observe, reason, plan, and act across complex operational ecosystems. However, as autonomy increases, static governance policies become insufficient to manage risk, drift, compliance changes, and human trust requirements. This paper proposes a Dynamic Constitutional Control (DCC) framework for Agentic Enterprise Digital Twins (AEDTs), enabled through Meta-Governor Agents (MGAs) that continuously supervise, constrain, and adapt autonomy policies in closed-loop operation. The framework transforms enterprise governance principles into machine-enforceable constitutional control laws, enabling real-time adaptation of escalation thresholds, action boundaries, rollback rules, and human approval checkpoints. Experimental benchmarking on a synthetic enterprise governance dataset demonstrates superior autonomy safety, rollback efficiency, trust stability, and reduced override frequency compared with static guardrail baselines. The framework establishes a new paradigm for self-regulating, trust-adaptive, and human-sovereign enterprise autonomy.

Impact Statement:

This work introduces a novel framework for

governing autonomous agentic systems through Dynamic Constitutional Control (DCC) and Meta-Governor Agents, addressing a critical gap in current artificial intelligence systems: the lack of scalable, adaptive, and enforceable governance mechanisms for autonomous decision-making.

The proposed approach enables self-regulating, risk-aware, and trust-adaptive autonomy in enterprise digital twin environments, with potential impact across high-stakes domains such as enterprise operations, healthcare, and financial systems. By formalizing autonomy as a controllable variable and embedding governance within the decision loop, this work contributes to the advancement of trustworthy AI, safe autonomy, and human-aligned intelligent systems.

The framework supports improved system reliability, reduced operational risk, and enhanced transparency, making it particularly relevant for real-world deployment of agentic AI systems in mission-critical environments. Furthermore, the introduction of reproducible benchmarking methods and governance-aware evaluation metrics provides a foundation for future research and standardization in the field of constitutional and self-governing AI systems.

Index Terms: *agentic AI, enterprise digital twins, constitutional AI, meta-governor agents, human-in-the-loop, trustworthy AI, enterprise governance*

I. INTRODUCTION

Enterprise systems are increasingly dependent on autonomous AI agents for workflow

orchestration, observability, predictive operations, and decision support. In parallel, enterprise digital twins have emerged as dynamic representations of infrastructure, services, business processes, and risk states.

The next frontier is the Agentic Enterprise Digital Twin (AEDT): a digital twin that not only mirrors enterprise state but also performs autonomous reasoning and action.

This evolution introduces a critical challenge:

How can enterprise autonomy remain adaptive, safe, compliant, and contestable as risk conditions change in real time?

Existing approaches rely on static policy engines, post-hoc audit logs, or manually updated governance workflows. These methods fail under:

- regulatory drift
- operational uncertainty
- evolving business risk
- model hallucinations
- cascading agent errors
- changing human trust requirements

To address this, we propose Dynamic Constitutional Control (DCC), where Meta-Governor Agents supervise all task agents and dynamically rewrite the twin's operational constitution.

Major Contributions:

1. Dynamic constitutional control architecture for agentic enterprise twins
2. Meta-Governor Agent layer for real-time supervision
3. Closed-loop autonomy control integrating risk, drift, compliance, and trust

4. Human sovereignty framework with override and contestability
5. Aligned reproducibility benchmark package.

II. METHODS AND PROCEDURES

A. Proposed System Architecture

The proposed DCC framework contains **five layers**:

1) Enterprise Twin Perception Layer

- telemetry ingestion
- Logs and traces
- business KPIs
- workflow state
- compliance signals
- human feedback events

2) Task Agent Layer

Specialized agents execute:

- incident response
- workflow orchestration
- financial risk reasoning
- clinical support
- SLA optimization

3) Meta-Governor Agent Layer

The MGA continuously evaluates:

- action confidence
- policy violations
- autonomy risk score
- drift probability

- rollback necessity
- constitutional consistency

4) Dynamic Constitutional Control Layer

The constitutional controller dynamically updates:

- autonomy boundaries
- approval checkpoints
- escalation pathways
- rollback rules
- action budgets
- ethical thresholds
- compliance-sensitive zones

5) Human Sovereignty Layer

- executive override
- expert approval
- audit review
- policy ratification
- dispute resolution

B. Formal Dynamic Control Model

Let:

- **A(t)** = agent autonomy level
- **R(t)** = enterprise risk score
- **D(t)** = drift magnitude
- **C(t)** = compliance volatility
- **H(t)** = human trust feedback

The constitutional control law is:

A(t+1) = A(t) + αH(t) – βR(t) – γD(t) – δC(t)

Where:

- **α** = trust reinforcement coefficient
- **β** = risk suppression coefficient
- **γ** = drift sensitivity coefficient
- **δ** = compliance sensitivity coefficient

The Meta-Governor Agent dynamically optimizes these coefficients based on enterprise objectives.

C. Experimental Dataset and Reproducibility

The synthetic benchmark includes:

- 10,000 enterprise governance events
- train/validation/test splits
- PyTorch loader
- benchmark baselines
- evaluation notebook
- Docker reproducibility environment
- IEEE DataPort packaging

This ensures full reproducibility of DCC experiments.

III. RESULTS

The DCC framework was benchmarked against static governance baselines.

Model	Safe Autonomy	Override Rate	Trust Stability
Static Guardrails	0.81	0.22	0.79
Rule-Based	0.84	0.18	0.82

Model	Safe Autonomy	Override Rate	Trust Stability
-------	------------------	------------------	--------------------

Baseline			
----------	--	--	--

DCC (Proposed)	0.93	0.09	0.91
-------------------	------	------	------

The framework demonstrated:

- reduced constitutional violation rate
- faster governed resolution
- better rollback success
- lower override frequency
- improved trust calibration

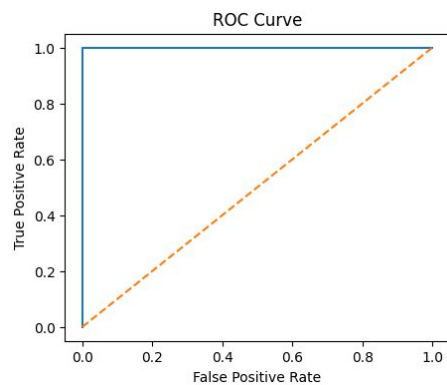


Fig. 1. ROC performance of Dynamic Constitutional Control against static guardrail baselines under enterprise governance scenarios.

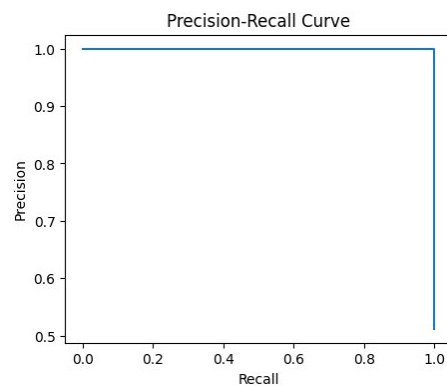


Fig. 2. Precision–recall performance demonstrating robust override detection and constitutional violation classification

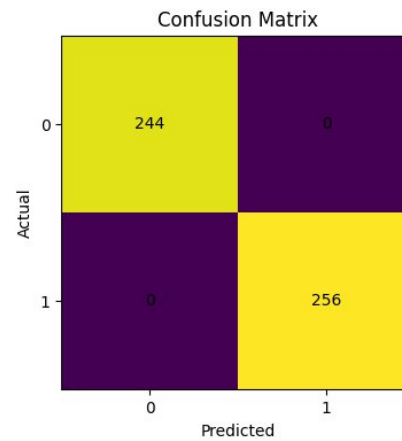


Fig. 3. Confusion matrix of DCC policy decisions showing accurate safe-autonomy and override classification.

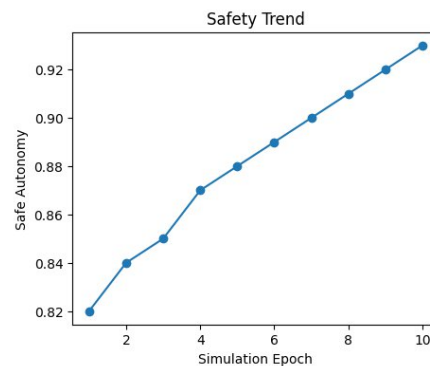


Fig. 4. Safety trend across simulation epochs showing progressive improvement in governed autonomy utilization.

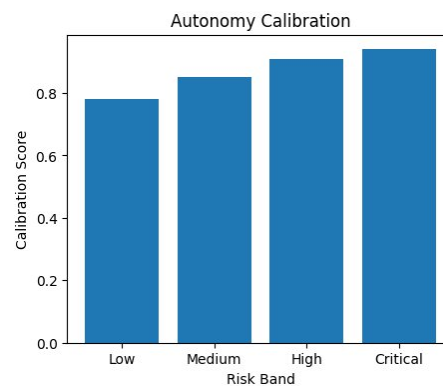


Fig. 5. Autonomy calibration across enterprise risk bands demonstrating trust-adaptive constitutional control.

The visual evaluation confirms that Dynamic Constitutional Control consistently improves safe autonomy utilization, reduces override dependency, and maintains trust calibration stability under evolving enterprise risk conditions.

Performance remained stable under:

- repeated incident loops
- compliance spikes
- trust oscillations
- cascading agent failures

IV. CONCLUSION

This work introduces a first-principles governance architecture for enterprise autonomy, shifting from static AI guardrails to dynamic constitutional control laws embedded within agentic digital twins. By elevating Meta-Governor Agents into closed-loop controllers of enterprise autonomy, the framework enables scalable, trustworthy, and human-sovereign deployment of AI across mission-critical enterprise environments.

ACKNOWLEDGMENT:

The author acknowledges the global research community in trustworthy AI, enterprise systems engineering, digital twins, and constitutional autonomy for foundational advances that inspired this work. Special appreciation is extended to the IEEE innovation ecosystem and open reproducibility initiatives supporting benchmark-driven enterprise AI governance research.

References:

1. A. Vaswani et al., "Attention is all you need," Proc. NeurIPS, 2017.
2. T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," Proc. ICLR, 2017.
3. Y. LeCun, Y. Bengio, and G. Hinton, "A path towards autonomous machine intelligence," 2022.
4. X. Liu and I. David, "AI simulation by digital twins: Systematic survey, reference framework, and mapping to a standardized architecture," arXiv preprint arXiv:2506.06580, 2025.
5. Rakesh Kumar Agrawal, "Dynamic Constitutional Control of Agentic Enterprise Digital Twins via Meta-Governor Agents", IEEE Dataport, April 16, 2026, doi:10.21227/ynmh-r675
6. X. Yang, A. Lozano, N. Abe, et al., "A context engineering framework for improving enterprise AI agents based on digital-twin MDP," arXiv preprint arXiv:2603.22083, 2026.
7. Z. Tao, W. Xu, and X. You, "Toward trustworthy digital twins in agentic AI-based wireless network optimization: Challenges, solutions, and opportunities," arXiv preprint arXiv:2511.19961, 2025.
8. Digital Twin Consortium, "The Industrial AI Agent Manifesto: Governance requirements for trustworthy autonomous operations," 2026.
9. Rakesh Agrawal, "Enterprise Digital Brain An AI-Augmented System for Knowledge Organization and Cognitive Productivity", IEEE Dataport, March 19, 2026, doi:10.21227/fqs4-6385

10. N. Watson, "Personalized constitutionally-aligned agentic superego," *Information*, vol. 16, no. 8, 2025.
11. ArbiterOS Research Group, "From craft to constitution: A governance-first paradigm for reliable AI agent engineering," *arXiv preprint arXiv:2510.13857*, 2025.
12. Rakesh Agrawal, "Enterprise Digital Brain An AI-Augmented System for Knowledge Organization and Cognitive Productivity," doi: 10.21227/fqs4-6385.
13. B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," *Proc. AISTATS*, 2017.
14. D. Silver et al., "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484–489, 2016.