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

Rakesh Kumar Agrawal
Department of Enterprise Intelligence Research,
Atos
Chicago, Illinois, USA

***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) reference architecture and conceptual 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 evaluates safe autonomy, override frequency, rollback behavior, and trust stability against static guardrail and rule-based governance baselines. The framework contributes a human-centered autonomy approach for self-regulating and trust-adaptive enterprise digital twins.**

***Impact Statement* — This work presents a governance reference architecture for autonomous agentic systems through Dynamic Constitutional Control (DCC) and Meta-Governor Agents, addressing a gap in scalable, adaptive, and enforceable governance mechanisms for autonomous decision-making. The proposed approach supports self-regulating, risk-aware, and trust-adaptive autonomy in enterprise digital twin environments, with potential relevance across high-stakes domains such as enterprise operations, healthcare, and financial systems. By formalizing autonomy as a bounded control variable and embedding governance within the decision loop, this work contributes to trustworthy AI, safe autonomy, and human-centered autonomy. The framework provides a pathway toward improved system reliability, reduced operational risk, and enhanced transparency in future enterprise deployment scenarios. The inclusion of reproducible benchmarking methods and governance-aware evaluation metrics provides a foundation for future research and standardization in 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:

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

- ✓ Human-centered autonomy framework with override and contestability
- ✓ Aligned reproducibility benchmark package.

II. RELATED WORK AND RESEARCH GAP

Prior work on digital twins defines virtual representations of physical or operational systems and highlights their growing role in simulation, monitoring, and decision support [1]–[4]. In parallel, Constitutional AI introduces principle-based supervision for model behavior through critique, revision, and AI feedback mechanisms [7]. Enterprise AI governance standards and risk frameworks emphasize risk management, accountability, transparency, and human oversight [5], [6], [17]–[19]. Agent research further shows that large language model based agents can reason and act across tools and environments, but require explicit safety, evaluation, and control mechanisms when deployed in multi-agent settings [11]–[16].

The research gap addressed here concerns runtime governance adaptation for agentic enterprise digital twins. Existing constitutional AI methods primarily operate at model training or response revision level, while conventional enterprise governance frameworks provide management controls that are not directly embedded as real-time autonomy control laws. Digital twin literature emphasizes simulation and synchronization, but does not sufficiently specify how autonomous agents inside enterprise twins should dynamically adjust decision rights, rollback rules, approval checkpoints, and operational boundaries under changing risk, drift, compliance, and human trust conditions [1]–[7], [17]–[19]. DCC preserves the manuscript contribution by positioning constitutional control as a runtime governance layer for AEDTs rather than as a replacement for existing AI safety, digital twin, or enterprise risk-management frameworks.

Research Stream	**Main Focus**	**Governance Gap for AEDTs**	**DCC Position**
Constitutional AI	Principle-guided critique, revision, and AI feedback [7]	Limited treatment of enterprise runtime autonomy control	Moves principles into bounded, machine-enforceable control laws
Agent governance	Policies, safety checks, and human oversight for autonomous agents [11]–[16]	Often static or external to the agent decision loop	Adds MGA supervision and closed-loop policy adaptation
Enterprise digital twins	Simulation, synchronization, and operational representation [1]–[4]	Governance adaptation is not formalized as autonomy control	Embeds governance into the AEDT runtime architecture
Runtime governance adaptation	Risk, compliance, and policy updates during operation [5], [6], [17]–[19]	Insufficient formal linkage to agent autonomy bounds and feedback	Defines bounded autonomy updates driven by risk, drift, compliance, and trust

III. METHODS AND PROCEDURES

A. Proposed System Architecture

The proposed DCC framework contains five layers:

a. Enterprise Twin Perception Layer

- Telemetry ingestion
- Logs and traces
- Business KPIs
- Workflow state
- Compliance signals
- Human feedback events

b. Task Agent Layer

Specialized agents execute:

- Incident response
- Workflow orchestration
- Financial risk reasoning
- Clinical support
- SLA optimization

c. Meta-Governor Agent Layer

The MGA continuously evaluates:

- Action confidence

- Policy violations
- Autonomy risk score
- Drift probability
- Rollback necessity
- Constitutional consistency

d. **Dynamic Constitutional Control Layer**

The constitutional controller dynamically updates:

- Autonomy boundaries
- Approval checkpoints
- Escalation pathways
- Rollback rules
- Action budgets
- Ethical thresholds
- Compliance-sensitive zones

e. **Human Sovereignty Layer**

- Executive override
- Expert approval
- Audit review
- Policy ratification
- Dispute resolution

B. Formal Dynamic Control Model

Let the DCC state at time *t* be defined by the following normalized variables, where each variable is measured or transformed onto a bounded governance scale when applied in simulation:

A(t) = Agent autonomy level, bounded between 0 and 1

R(t) = Enterprise risk score

D(t) = Drift magnitude

C(t) = Compliance volatility

H(t) = Human trust feedback

The autonomy bound and bounded update rule are:

$$0 \leq A(t) \leq 1$$
$$\Delta A = \alpha H(t) - \beta R(t) - \gamma D(t) - \delta C(t)$$
$$A(t+1) = \max(0, \min(1, A(t) + \Delta A))$$

Where:

α = trust reinforcement coefficient

β = risk suppression coefficient

γ = drift sensitivity coefficient

δ = compliance sensitivity coefficient

Coefficients are configurable governance parameters calibrated through simulation feedback. They do not imply autonomous optimization by the MGA; rather, they provide tunable weighting factors for trust reinforcement, risk suppression, drift sensitivity, and compliance sensitivity in the bounded autonomy update.

C. Experimental Methodology

The experimental methodology uses a synthetic enterprise governance dataset designed to exercise autonomy decisions under changing risk, drift, compliance, and human feedback conditions. Dataset generation follows scenario templates for incident response, workflow orchestration, financial risk reasoning, clinical support, and SLA management. Each event contains contextual features, including risk score, drift magnitude, compliance volatility, human trust feedback, action confidence, rollback state, and governance escalation status.

- Synthetic event structure: event identifier, domain scenario, candidate agent action, risk score R(t), drift signal D(t), compliance volatility C(t), human feedback H(t), autonomy level A(t), and policy outcome.
- Labels: safe autonomy decision, override required, rollback required, constitutional violation, and escalation required.
- Baselines: Static Guardrails, Rule-Based Governance, and DCC Framework.
- Metrics: safe autonomy rate, override rate, trust stability, rollback success, precision, recall, ROC-AUC, and precision–recall behavior.
- Reproducibility artifacts: train/validation/test splits, PyTorch loader, evaluation notebook, and Docker reproducibility environment.

The benchmark includes 10,000 synthetic enterprise governance events. The Static Guardrails baseline applies fixed thresholds, Rule-Based Governance applies deterministic conditional rules, and the DCC Framework applies the bounded constitutional control update described above. The evaluation reports available aggregate metrics only; variance and statistical significance are not claimed unless generated by repeated experimental runs.

IV. RESULTS

The DCC framework was benchmarked against Static Guardrails and Rule-Based Governance. Safe autonomy measures the proportion of events where the system permits autonomous action without violating policy constraints. Override rate measures the proportion of events escalated to human approval or intervention. Trust stability measures the consistency of autonomy adjustment under human feedback and changing risk signals. Rollback success indicates whether unsafe or deteriorating actions can be reversed according to governance rules.

TABLE I
COMPARATIVE BENCHMARK PERFORMANCE OF GOVERNANCE APPROACHES

Model	Safe Autonomy	Override Rate	Trust Stability
Static Guardrails	0.81	0.22	0.79
Rule-Based Governance	0.84	0.18	0.82
DCC Framework	0.93	0.09	0.91

The benchmark results indicate:

- Reduced constitutional violation rate
- Faster governed resolution
- Better rollback success
- Lower override frequency
- Improved trust calibration

The ROC performance comparison is shown in Fig. 1.

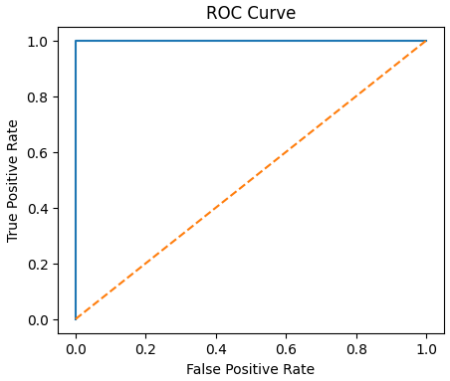


Fig. 1. ROC performance of Dynamic Constitutional Control against static guardrail baselines under enterprise governance scenarios.

The precision–recall performance is shown in Fig. 2.

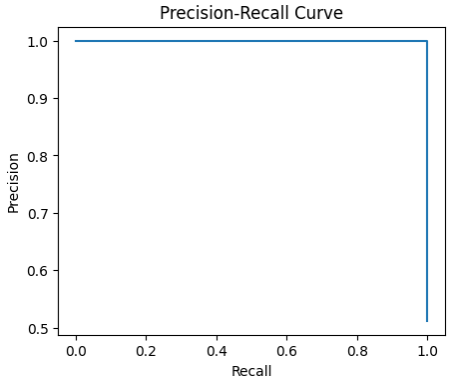
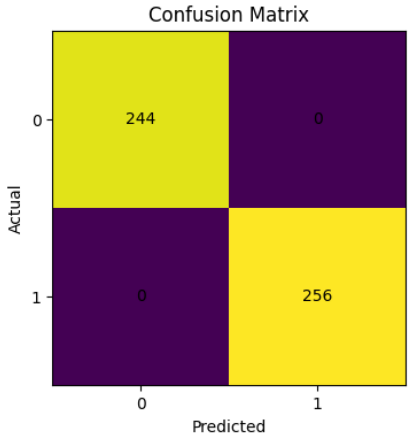
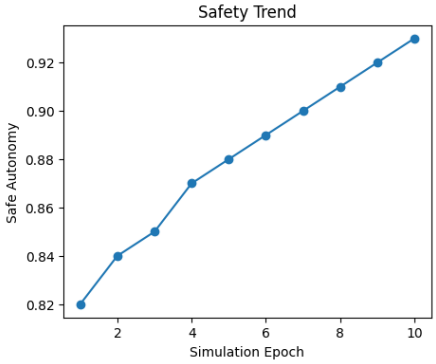


Fig. 2. Precision–recall performance demonstrating robust override detection and constitutional violation classification.



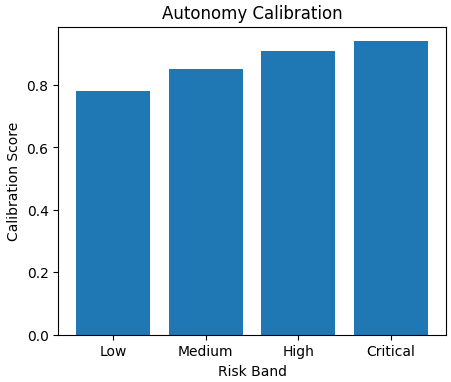
The DCC policy decision confusion matrix is shown in Fig. 3.

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



The evolution of safe autonomy utilization across simulation epochs is illustrated in Fig. 4.

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



The autonomy calibration behavior across different enterprise risk bands is presented in Fig. 5.

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

The visual evaluation suggests that Dynamic Constitutional Control improves safe autonomy utilization, reduces override dependency, and maintains trust calibration stability under evolving enterprise risk conditions in the synthetic benchmark. Because the current experiment reports aggregate benchmark outputs, variance and statistical significance are not asserted here.

Performance remained stable under:

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

V. DISCUSSION

The results indicate that DCC is most useful when autonomy must be continuously adjusted rather than governed by one-time policy thresholds. Unlike static guardrails, DCC treats autonomy as a bounded, feedback-sensitive control variable. This enables the MGA layer to reduce autonomy when risk, drift, or compliance volatility increases, and to cautiously restore autonomy when human trust feedback and operational conditions improve. The framework therefore complements existing AI risk-management standards by translating governance principles into runtime control behavior rather than treating governance as a post-hoc audit activity [5]–[7], [17]–[19].

For IJGIS readers, the relevant contribution is not a claim of a universal governance solution, but a conceptual framework and reference architecture for embedding adaptive governance into agentic enterprise digital twins. The approach remains compatible with enterprise digital twin literature, constitutional AI, agent governance, and formal risk-management standards while focusing on the unresolved problem of runtime governance adaptation.

VI. LIMITATIONS AND FUTURE RESEARCH

This study has limitations. The evaluation uses a synthetic enterprise governance dataset rather than production deployment data, and the reported results should therefore be interpreted as benchmark evidence for the proposed reference architecture rather than as field validation. The mathematical model defines bounded autonomy adaptation but does not yet include formal stability proofs, causal identification of trust feedback, or repeated-run statistical analysis. Future research should evaluate DCC in operational digital twin environments, compare additional agent-governance baselines, test domain-specific policy libraries, conduct sensitivity analysis for configurable coefficients, and examine human-centered autonomy under organizational, legal, and safety constraints.

VII. CONCLUSION

This work introduces a governance reference architecture and conceptual framework for enterprise autonomy, shifting from static AI guardrails toward dynamic constitutional control laws embedded within agentic digital twins. By elevating Meta-Governor Agents into closed-loop supervisors of bounded enterprise autonomy, the framework supports scalable, trustworthy, and human-centered autonomy across mission-critical enterprise environments while preserving the DCC contribution as the central scientific contribution of the manuscript.

VIII. DATA AVAILABILITY STATEMENT

The data associated with this manuscript are available from the corresponding author upon reasonable request. The research presents a conceptual framework and benchmarking approach for Dynamic Constitutional Control of Agentic Enterprise Digital Twins. The supporting synthetic benchmark dataset and reproducibility artifacts developed for evaluation purposes will be made available upon reasonable request or future public artifact release.

IX. 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 innovation ecosystem and open reproducibility initiatives supporting benchmark-driven enterprise AI governance research.

DECLARATIONS

Funding:
No external funding was received for this research.

Clinical Trial Registration:
Not applicable.

Consent to Publish Declaration:
Not applicable.

Consent to Participate Declaration:
Not applicable.

Ethics Declaration:
Not applicable.

X. REFERENCES

[1] A. Fuller, Z. Fan, C. Day, and C. Barlow, “Digital Twin: Enabling Technologies, Challenges and Open Research,” IEEE Access, vol. 8, pp. 108952–108971, 2020.

[2] F. Tao, M. Zhang, Y. Liu, and A. Y. C. Nee, Digital Twin Driven Smart Manufacturing, Academic Press, 2019.

[3] M. Grieves and J. Vickers, “Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems,” in Transdisciplinary Perspectives on Complex Systems, Springer, pp. 85–113, 2017.

[4] M. Schluse, M. Priggemeyer, L. Atorf, and J. Rossmann, “Experimentable Digital Twins—Streamlining Simulation-Based Systems Engineering for Industry 4.0,” IEEE Transactions on Industrial Informatics, vol. 14, no. 4, pp. 1722–1731, 2018.

[5] National Institute of Standards and Technology (NIST), “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” NIST AI 100-1, January 2023.

[6] National Institute of Standards and Technology (NIST), “Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile,” NIST AI 600-1, July 2024.

[7] Y. Bai et al., “Constitutional AI: Harmlessness from AI Feedback,” arXiv preprint arXiv:2212.08073, 2022.

[8] S. Russell, Human Compatible: Artificial Intelligence and the Problem of Control, Viking, 2019.

[9] S. Amodei et al., “Concrete Problems in AI Safety,” arXiv preprint arXiv:1606.06565, 2016.

[10] C. Russell, D. Dewey, and M. Tegmark, “Research Priorities for Robust and Beneficial Artificial Intelligence,” AI Magazine, vol. 36, no. 4, pp. 105–114, 2015.

[11] S. Park et al., “Generative Agents: Interactive Simulacra of Human Behavior,” Proceedings of the ACM Symposium on User Interface Software and Technology (UIST), 2023.

[12] Z. Xi et al., “The Rise and Potential of Large Language Model Based Agents: A Survey,” arXiv preprint arXiv:2309.07864, 2023.

[13] Y. Wang et al., “A Survey on Large Language Model Based Autonomous Agents,” Frontiers of Computer Science, vol. 18, 2024.

[14] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances in Neural Information Processing Systems (NeurIPS), 2020.

[15] T. Yao et al., “ReAct: Synergizing Reasoning and Acting in Language Models,” International Conference on Learning Representations (ICLR), 2023.

[16] M. Wooldridge, An Introduction to MultiAgent Systems, 2nd ed., Wiley, 2009.

[17] ISO/IEC 42001:2023, Information Technology — Artificial Intelligence — Management System, International Organization for Standardization, 2023.

[18] ISO/IEC 23894:2023, Information Technology — Artificial Intelligence — Guidance on Risk Management, International Organization for Standardization, 2023.

[19] European Union, “Artificial Intelligence Act,” Regulation (EU) 2024/1689, European Parliament and Council of the European Union, 2024.

AUTHOR BIOGRAPHY

Rakesh Kumar Agrawal is an applied AI researcher, enterprise technology leader, and Senior Member for reputed nonprofit organization with more than two decades of experience in intelligent systems, cloud platforms, enterprise operations, and digital transformation. His research interests include trustworthy AI, agentic systems, enterprise digital twins, explainable AI, and governance frameworks for autonomous systems. He focuses on bridging academic innovation with real-world scalable deployments across healthcare, finance, and enterprise technology domains.