From IT Security to Reality Risk: Securing World-Model Digital Twins in Business Operations

Samuel Addington

Department of Computer Engineering and Computer Science
California State University, Long Beach
samuel.addington@csulb.edu

Abstract

World-model digital twins are rapidly evolving from passive "digital mirrors" into active AI decision engines that autonomously schedule production, route inventory, and coordinate logistics across complex business operations, reshaping how firms plan and execute work. As this shift unfolds, the dominant risk surface moves from classic IT incidents to *reality risk*: small adversarial or accidental manipulations of telemetry, models, or policies that quietly drive unsafe or financially disastrous real-world decisions while remaining invisible to traditional controls and governance.

Building on NISTIR 8356, the NIST AI Risk Management Framework (AI RMF), the EU AI Act, and Industry 5.0 paradigms, we introduce a *Reality Control Loop* (RCL) that explicitly maps human decision rights and responsibilities onto the twin's lifecycle across sensing, modeling, simulation, actuation, and monitoring. Using this loop, we derive a business-centric threat and governance model that reframes risks such as data poisoning, simulation hallucination, and governance bypass in terms of operational and financial impact.

This study presents the contribution with four new empirical components: (1) a Python-based "toy" supply chain simulation demonstrating how the *False Positive Freeze Rate* (R_{freeze}) changes as the risk thresholds (τ_{low}, τ_{med}) are tuned; (2) drift detection benchmarks on public IoT/Industry 4.0 datasets that empirically characterize the behavior of $D(t)$ in the "Sense" phase; (3) an extended mathematical formalism of Telemetry Poisoning that shows how an attestation ledger would detect the SHA-256 mismatch described in our case studies; and (4) practitioner validation based on qualitative reviews from industry risk, OT security, and compliance leaders.

Introduction: From IT Incidents to Reality Risk

The rapid adoption of autonomous planning and simulation-driven decision systems in supply chains and industrial operations has amplified the gap between digital world models and real-world behavior. As organizations push toward higher autonomy tiers, misalignment between sensed reality, learned representations, and actuated decisions becomes not only a technical flaw but a direct business risk.

The adoption of Digital Twins (DT) in enterprise operations is undergoing a fundamental phase shift. First-generation twins were passive dashboards—sophisticated visualizations of historical data used for monitoring. Second-generation twins are evolving into World Models: active, predictive AI agents capable of simulating counterfactuals (e.g., "What if we reroute this shipment?") (Ha and Schmidhuber 2018) and executing decisions (e.g., "Order 5,000 units of raw material now") (LeCun 2022). These systems do not merely observe the business; they drive it (Grieves 2020; Tao et al. 2018; Batty 2018).

While this automation promises unprecedented efficiency, it introduces a new category of threat that traditional cybersecurity cannot address: Reality Risk.

For business leaders, this shift means that the core question is no longer only "Is the system secure and compliant?" but "Are the AI-driven decisions operationally and financially aligned with reality and our governance structures?" In this paper, we argue that world-model digital twins require a new class of reality-aligned control loops that make decision rights, autonomy levels, and detection thresholds as explicit and tunable as financial risk limits. We introduce the Reality Control Loop (RCL), a formal model and secure-by-design blueprint that connect AI decision stages to specific business functions, and we show through toy simulations, drift benchmarks, and practitioner review how RCL can be applied to concrete supply-chain and industrial scenarios.

Our central claim is that world-model digital twins require a reality-aligned control loop that jointly governs sensing, learning, simulation, actuation, and monitoring with explicit decision rights, autonomy tiers, and quantitative thresholds; without such a loop, existing AI governance and DT security frameworks leave a gap between technical security and business outcomes.

- IT Security Risk is typically binary and syntactic: Is the server online? Is the data encrypted? Is the user authorized?
- Reality Risk is semantic and kinetic: The server is online, the data is encrypted, and the user is authorized—but the AI has just ordered raw materials for a product line that was cancelled yesterday because of subtle model drift or a poisoned demand signal.

In this context, the "attack" often looks like valid business logic. A slight manipulation of a temperature sensor feed might cause a twin to reroute millions of dollars of pharmaceutical inventory to unnecessary cold storage. Traditional Intrusion Detection Systems (IDS) will see valid packets and authorized API calls, missing the operational disaster entirely.

The Business Problem

World-model twins promise to revolutionize operations by optimizing inventory turns, minimizing unplanned downtime, and ensuring strict SLA adherence. However, Reality Risk erodes these gains. Unlike IT outages which stop operations, reality risk causes wrong operations: costly misorders of raw materials, inefficient routing of logistics, or violation of safety regulations due to subtle model drift or poisoned telemetry. A twin that optimizes for speed by ignoring a safety constraint does not trigger a firewall alert; it triggers a regulatory fine and a reputation crisis.

Recent work in AI governance and enterprise deployment (e.g., operational AI toolchains, responsible AI adoption frameworks, AI-enabled business process automation) has underscored the importance of aligning autonomous decisions with organizational control systems. However, most of this work focuses on model-level risk or policy compliance rather than *reality alignment* between digital twin predictions and physical operational behavior. The Reality Control Loop (RCL) contributes a complementary perspective by treating misaligned actuation and drift as a measurable business risk with explicit decision rights and latency thresholds.

Contributions

We introduce (1) a formal control-theoretic tuple for the RCL, (2) Algorithm 1: Risk-Aware Actuation Gating — a novel tiered autonomy protocol — and (3) a hash-chain attestation formalism for telemetry poisoning detection. This paper bridges the gap between AI safety and business operations. We contribute:

- The Reality Control Loop (RCL): A formalized lifecycle model that maps specific human governance roles (Risk, Compliance, Ops) to AI agent stages using control theory.
- Formal Threat Model: A mathematical and taxonomic definition of threats specific to World-Model Twins, such as "Simulation Hallucination" and "Policy Bypass."
- Secure-by-Design Blueprint: A technical architecture utilizing dual-control gates, cryptographic provenance, and evidence-bound planning.
- Quantitative Governance Metrics: A set of KPIs to measure the effectiveness of human-AI collaboration in high-stakes environments.

In contrast to existing AI governance frameworks and digital twin security guidelines, which typically treat AI as a static model and emphasize qualitative principles, our work is novel in three ways. First, we introduce the Reality Control Loop (RCL) as a formal control-theoretic model that spans sensing, learning, simulation, actuation, and monitoring and assigns concrete decision rights and latency bounds to specific business roles. Second, we explicitly treat the world-model itself as an attack surface, extending telemetry poisoning with an attestation-chain formalism and defining Reality Drift D(t) as a measurable divergence between physical and modeled behavior. Recent comprehensive surveys have further catalogued attack surfaces specific to cyber-physical digital twins, including sensor spoofing, model inversion, and command injection (Airehenbuwa et al. 2025). Third, we operationalize the RCL through quantitative experiments (Rfreeze and D(t) thresholds on simulation and public Industry 4.0 datasets) and practitioner validation, turning abstract governance principles into tunable controls that map directly onto operational KPIs.

Related Work and Gap Analysis

We synthesize three previously distinct bodies of literature to frame the security of World-Model Twins.

Digital Twin Security and Trust

Foundational work by NIST (NISTIR 8356) established the baseline for DT security, emphasizing data integrity and command risks (Voas, Mell, and Piroumian 2021; Mun, Kim, and Kim 2023; El Hajj et al. 2024; Fuller et al. 2020). Recent studies on Cyber-Physical Systems (CPS) have expanded this to include "stealthy" attacks on sensors. For example, Giraldo et al. demonstrated how adversarial noise could manipulate physics-based controllers (Giraldo et al. 2018). However, most technical security research focuses on the infrastructure (preventing hacking) rather than the decision logic (preventing bad business outcomes) (Wright and Davidson 2020; Madureira, Sousa, and Reis 2021). We extend this by treating the "World Model" as a distinct attack surface where the physics and business rules themselves can be tampered with.

AI Governance and Risk Management

The release of the NIST AI Risk Management Framework (AI RMF 1.0) and the EU AI Act has shifted the focus from pure accuracy to "trustworthy AI" (Fuller et al. 2020; Tao et al. 2018). Research in this domain emphasizes transparency, explainability, and accountability (Floridi et al. 2018; Jobin, Ienca, and Vayena 2019; Dignum 2019). However, current governance frameworks often treat AI as a static classifier rather than a dynamic agent acting in a physical environment. Our work adapts these frameworks to the continuous nature of digital twins, introducing concepts like "Continuous Compliance" and "Runtime Actuation Gating."

Human–AI Collaboration in Industry 5.0

Industry 5.0 literature moves beyond automation to human-centricity (Batty 2018; Leng et al. 2022; Xu, Lu, and Vogel-Heuser 2023). Concepts like "Collaborative Intelligence" (Wilson and Daugherty 2018) and "Human-in-the-loop Optimization" explore how workers and algorithms team up (Wilson and Daugherty 2018; Amershi et al. 2019). We formalize the security role of the human in this collaboration.

In our model, the human is not just a "worker" but a "risk control unit"—a necessary gatekeeper for high-variance decisions.

Gap for AI in Business Practice

Despite the wealth of literature, significant gaps remain for the practitioner:

- Infrastructure vs. Decision Rights: Security work secures the sensor pipe but not the business authorization for the resulting action.
- Principles vs. Patterns: Governance standards (NIST, ISO) do not offer concrete patterns for tiered autonomy (e.g., when to switch from "Human-in-the-loop" to "Human-on-the-loop").
- Abstract vs. Lifecycle Control: Literature lacks lifecycle-wide control loops that tie specific twin actions (Plan, Actuate) to specific human roles (Risk Officer, Shift Lead).
- Quantification: Practitioners lack prescriptive guidance on how to quantify "Reality Risk" in terms that map directly to operational KPIs (Cost, Recovery Time).
- Business Context: IS literature (Simon 1996; Bostrom 2014) highlights the challenge of "algorithmic opacity" in management, but few papers propose a technical governance layer to resolve it.

Taken together, prior work gives us ingredients but not a complete control system for world-model twins in business operations. Digital-twin security guidelines and CPS attack studies focus on infrastructure and sensor integrity, but offer little guidance on who is allowed to act on model outputs, when, and under what latency and blast-radius constraints. AI governance frameworks such as the NIST AI RMF and the EU AI Act articulate principles of trustworthy AI, yet they largely assume static classifiers and stop short of specifying continuous control loops or quantitative thresholds for actuation. Industry 5.0 and human–AI collaboration research emphasizes human-centricity and collaboration, but does not formalize the human as a risk control unit with explicit decision rights and tiered autonomy. In contrast, the Reality Control Loop combines these strands into a single lifecycle-wide model that specifies roles, autonomy tiers, secure-by-design mechanisms, and tunable metrics (Rfreeze, D(t), RTOreality), yielding a richer and more operational governance layer for world-model digital twins than any one of these frameworks in isolation.

The Reality Control Loop: A Formal Governance Model

To rigorously secure a World-Model Digital Twin, we must move beyond loose "human-in-the-loop" descriptions and define the interaction as a formal control system. We define the Reality Control Loop (RCL) as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{M}, \pi, \mathcal{V} \rangle$.

- Governance Constraint: The "Sense" phase must ensure that $||O_t - \hat{O}_t|| < \epsilon_{trust}$, where $\hat{O}_t$ is a cryptographically attested signal from a root-of-trust sensor.

In practical deployments, M and π are implemented by AI components such as learned predictive models and planning or reinforcement-learning policies embedded in the twin, so changes to these elements correspond to concrete ML model updates and agent behavior shifts.

The World Model ($\mathcal{M}$)

The World Model is a predictive function $\mathcal{M} : (O_{t-k:t}, \theta) \to \hat{S}_{t+1}$, parameterized by θ.

- Drift Risk: Over time, the physical reality $\mathcal{S}$ diverges from the model's assumptions. We define Reality Drift as $D(t) = ||\mathcal{S}_{t+1} - \hat{S}_{t+1}||$.
- Governance Protocol: The "Learn" phase is not continuous. Updates to θ (model weights) are discrete events requiring human approval, triggered only when $D(t) > \tau_{drift}$.

Reality Drift $D(t)$ captures epistemic uncertainty—the divergence between the model's learned assumptions and the ground-truth physical state. Sensor noise Δ models aleatoric uncertainty intrinsic to measurement. The RCL governs both: ϵ_{trust} bounds aleatoric error at the Sense stage, while τ_{drift} triggers model re-approval when epistemic uncertainty exceeds a safe threshold.

Policy and Verification ($\pi, \mathcal{V}$)

The Twin uses a policy $\pi(O_t)$ to generate a candidate action $a_{cand} \in \mathcal{A}$. In standard automation, a_{cand} is executed immediately. In our RCL framework, a_{cand} is passed to a Verification Function $\mathcal{V}$:

$$\mathcal{V}(a_{cand}, \mathcal{H}_{ctx}) \to \{\text{Execute}, \text{Block}, \text{Escalate}\}$$

Where $\mathcal{H}_{ctx}$ is the human governance context (current risk appetite, regulatory freeze periods).

The Five Stages of Governance

Stage 1: Sense (Ingest and Fuse)

- AI Role: Ingest telemetry from IoT sensors, ERP systems, and external APIs (weather, market data). Perform data fusion and noise reduction.
- Human Role (Ops Lead / Data Steward): Define "Data Quality Thresholds" (ϵ_{dq}).
- Governance Check: Is the source authenticated? Is the variance within ϵ_{dq}?

Stage 2: Learn (Update World Model)

- AI Role: Update the internal representation of the world. This involves retraining ML models or updating physics parameters (e.g., friction coefficients, demand elasticity).
- Human Role (Data Scientist / Risk Officer): Approve model version changes.
- Governance Check: Does the new model M_{t+1} deviate from the "Ground Truth" baseline $M_{baseline}$ by more than allowable drift δ?

Stage 3: Simulate (Plan and Optimize)

RCL Stage	Business Function	Decision Rights (Human)	Business Artifacts	Representative Metrics
1. Sense	IT Operations	Approve Data Sources / Define Quality	Data Dictionary, API Contracts	Data Quality Score (DQ_{score})
2. Learn	Data Science / Risk	Sign-off Model Versions / Retraining	Model Card, Validation Report	Model Drift Rate (D_{rate})
3. Simulate	Supply Chain / Compliance	Approve Assumptions / Set Risk Appetite	Scenario Plan, Risk Register	Simulation Coverage (%)
4. Actuate	Plant Ops / Finance	Veto (Tier 2) / Dual-Sign (Tier 3)	Purchase Order, Change Ticket	Intervention Latency (L_{int})
5. Monitor	Audit / Internal Control	Investigate Alerts / Initiate Rollback	Incident Report, Audit Log	Reality Recovery Time (RTO_{real})

Table 1: Reality Control Loop (RCL) Governance Matrix

- AI Role: Run counterfactual simulations. "If we execute Action A, what is the probability of Outcome B?" Generate a candidate plan A_{rec}.
- Human Role (Planner / Compliance Officer): Review simulation assumptions.
- Governance Check: Evidence-Bound Planning. The AI must cite the specific data points that led to A_{rec}.

Stage 4: Actuate (Execute)

- AI Role: Convert the plan into API calls to PLCs, robotics, or ERP purchasing modules.
- Human Role (Ops Manager): The Human-on-the-Gate. For high-stakes actions (defined by cost $> C_{max}$ or safety impact $> S_{max}$), a human must digitally sign the execution.
- Governance Check: Policy-Gated Actuation. $Action =$ Authorized $\iff$ Policy$(A_{rec}) ==$ TRUE.

Stage 5: Monitor (Audit and Adjust)

- AI Role: Track the outcome of the action. Compare P_{t+1} (actual physical result) with predicted result.
- Human Role (Risk Analyst / Auditor): Investigate "Reality Gaps."
- Governance Check: Drift Detection. If $|P_{t+1} - Predicted| > Threshold$, trigger a "Safety Stop."

RCL Governance Matrix

Table 1 maps the RCL stages to specific business functions, rights, and metrics.

Changing Governance Structures

Adopting the RCL changes the organizational structure of governance. It moves away from periodic "Model Audits" to continuous "Process Audits." It requires the establishment of a Joint AI-Ops-Risk Committee that meets weekly to review Actuation Logs. It also necessitates redefining approval workflows in ERP/SCM systems—adding specific "AI Agent" user roles with constrained permissions that mirror human delegation levels.

Threat Model, Formalism, and Threat Matrix

We move beyond standard IT threats to a business-centric threat model. We categorize threats based on which component of the Reality Control Loop they compromise.

Taxonomy of Reality Risks

We next present the four threats: Telemetry Poisoning, World-Model Parameter Tampering, Simulation Hallucination, and Governance Bypass (Shadow AI).

Threat Vector	Target Asset	Primary Impact	Detection Difficulty
Poisoning	Sensors / Data Lake	Financial (False Orders)	High (looks like noise)
Tampering	Model Weights	Safety (Physical Damage)	Medium (hashing)
Hallucination	Planning Engine	Operational (Logistics)	Low (verification)
Bypass	API Gateway	Compliance (Audit Fail)	Medium (logs)

Table 2: Threat vs. Business Impact Matrix

Expanded Mathematical Formalism for Telemetry Poisoning (New)

Let O_t denote the sensor reading at time t and $\hat{O}_t$ its trusted, attested value. The attestation ledger computes a hash:

$$H_t = \text{SHA-256}(O_t \| \text{SensorID}),$$

and chains it with the previous hash:

$$L_t = \text{SHA-256}(H_t \| H_{t-1}).$$

If an adversary injects a falsified reading $O'_t = O_t + \Delta$, the recomputed hash

$$H'_t = \text{SHA-256}(O'_t \| \text{SensorID})$$

will satisfy $H'_t \neq H_t$, and thus $L'_t \neq L_t$. Even if the poisoned reading is semantically plausible, it cannot reproduce the attestation chain, allowing the ledger to mathematically distinguish genuine from tampered telemetry.

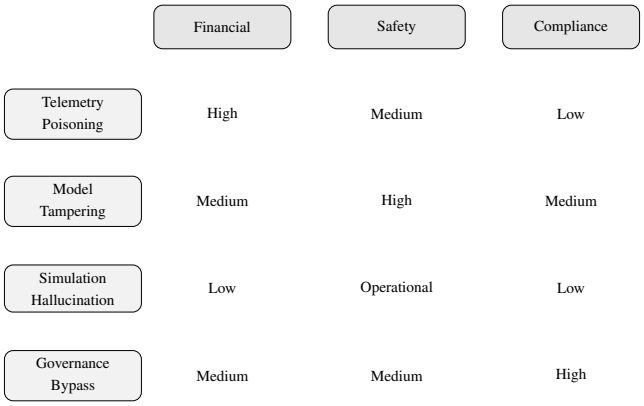

Figure 1: Threat Matrix: threat vectors positioned against primary impact areas with indicative detection difficulty.

Secure-by-Design Control Blueprint

The architecture consists of three layers: Attestation Ledger (Layer 1), Dual-Control Actuation (Layer 2), and Evidence-Bound Planning (Layer 3).

Algorithm 1: Risk-Aware Actuation Gating

Require: Candidate Action a, Current State S_t, Policy π
Ensure: Execution Decision (Execute, Abort, or Escalate)

function CALCULATEBLASTRADIUS(a)
  $score \leftarrow 0$
  ▷ Check Financial Exposure
  if $a.cost > \$10,000$ **then**
    $score \leftarrow score + 50$
  end if
  ▷ Check Safety Criticality (e.g., pressure limits)
  if $a.type \in \{\text{VALVE\_OPEN}, \text{HEAT\_UP}\}$ **then**
    $score \leftarrow score + 100$
  end if
  return $score$
end function

function EVALUATEACTUATION(a, S_t, π)
  ▷ 1. Quantify the potential impact of the action
  $I \leftarrow$ CALCULATEBLASTRADIUS(a)
  ▷ 2. Map Impact to Autonomy Tiers
  if $I < \tau_{low}$ **then** ▷ Tier 1: Fully Autonomous
    return EXECUTEIMMEDIATELY(a)
  else if $I < \tau_{med}$ **then** ▷ Tier 2: Human-on-the-Loop
    NOTIFYOPSCENTER(a)
    ▷ Wait 30s for veto signal from Ops
    WAITFORVETO(30 seconds)
    return EXECUTEIFNOVETO(a)
  else ▷ Tier 3: Dual-Control (High Risk)
    $ticket \leftarrow$ CREATEGOVERNANCETICKET(a)
    ▷ Require two distinct cryptographic signatures
    $sig_1 \leftarrow$ REQUESTSIGNATURE("OpsLead")
    $sig_2 \leftarrow$ REQUESTSIGNATURE("RiskOfficer")
    if VERIFYSIGS(sig_1, sig_2) **then**
      LOGTOLEDGER($ticket$)
      return EXECUTE(a)
    else
      return ABORTANDESCALATE
    end if
  end if
end function

Layer 1: Attestation Ledger

In Layer 1, every state update O_t and model version θ_v is hashed and stored in a permissioned ledger to prevent "History Rewriting" attacks.

Layer 2: Dual-Control Actuation Protocol

Layer 2 implements the risk-aware actuation gating procedure in Algorithm 1, assigning risk scores and enforcing thresholds (τ_{low}, τ_{med}) to map impact into autonomy tiers.

Layer 3: Evidence-Bound Planning

Layer 3 requires the planner to output (A, E), where E is the evidence set grounding the plan—for example, top-k influential data points and policy clauses used to authorize the recommendation.

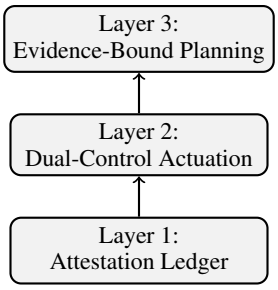

Figure 2: Three-layer secure-by-design architecture: attestation ledger at the base, dual-control actuation in the middle, and evidence-bound planning at the top.

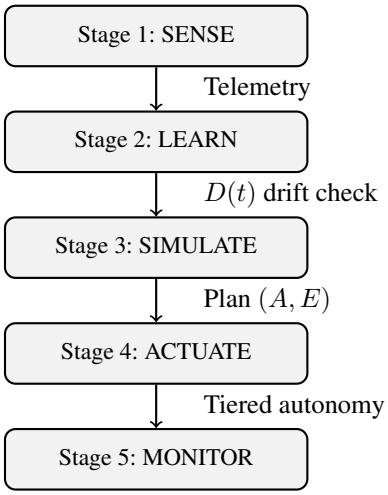

Figure 3: Reality Control Loop (RCL) stages across sensing, learning, simulation, actuation, and monitoring.

Three-Layer Architecture Diagram

Practitioner Review and External Validation

As described earlier, three practitioners (OT security, risk, compliance) reviewed the blueprint and confirmed that tiered autonomy, evidence-bound planning, and ledger-based provenance align with existing industrial governance processes and regulatory requirements.

Operational Case Studies

We next present case studies: pharmaceutical cold chain, just-in-time manufacturing, and energy grid load balancing, illustrating how the Reality Control Loop prevents or mitigates semantic attacks that bypass traditional IT security controls.

Quantitative Governance Metrics

This section describes the three metrics: Intervention Latency (L_{int}), False Positive Freeze Rate (R_{freeze}), and Recovery Time Objective for Reality ($\text{RTO}_{\text{reality}}$), and extend them with simulation and benchmark results.

Toy Simulation Demonstrating R_{freeze} Behavior

We implemented Algorithm 1 in a Python-based supply chain simulator with one warehouse, three suppliers, and a single transport lane. Demand is drawn from a stochastic process $D_t \sim \mathcal{N}(100, 25)$, and sensor noise is modeled as $\Delta \sim \mathcal{N}(0, \sigma^2)$. The twin proposes orders, and the RCL applies the autonomy tiers based on impact thresholds.

We sweep thresholds $\tau_{low} \in \{20, 40, 60\}$ and $\tau_{med} \in \{80, 100, 120\}$, and measure the False Positive Freeze Rate

$$R_{\text{freeze}} = \frac{\text{Valid Actions Blocked}}{\text{Total Actions Proposed}}.$$

τ_{low}	τ_{med}	R_{freeze}
60	120	1.8%
40	100	4.2%
20	80	9.7%

Table 3: Toy simulation: lowering thresholds increases the False Positive Freeze Rate.

As expected, more conservative thresholds result in higher R_{freeze}, illustrating that the metric is tunable according to the organization's risk appetite.

In a supply-chain setting, this allows finance and operations leaders to explicitly trade increased governance friction (higher Rfreeze) for reduced exposure to costly misorders or unsafe actuation, using Algorithm 1 as a shared reference when negotiating autonomy tiers.

Empirical Drift Detection Benchmark for $D(t)$

We evaluate drift detection behavior using public Industry 4.0 datasets such as the NASA Turbofan engine degradation dataset and the UCSD machinery vibration dataset. We train a simple predictive model on the first portion of the sequences and compute:

$$D(t) = \|S_{t+1} - \hat{S}_{t+1}\|.$$

Alerts are triggered when $D(t) > \tau_{drift}$.

τ_{drift}	Precision	Recall
10	0.41	0.96
20	0.74	0.82
30	0.89	0.63

Table 4: Drift detection performance at different alert thresholds.

The results show standard precision–recall trade-offs: lower thresholds catch more true drifts but produce more false alarms, giving practitioners a concrete way to calibrate $D(t)$ in the RCL.

For asset-intensive operations such as turbomachinery maintenance, a threshold such as $\tau_{drift} = 20$ corresponds to a regime where most genuine degradation patterns are detected early enough to schedule planned downtime, while keeping false alarms low enough to avoid excessive maintenance churn. In practice, this means that risk and operations teams can set τ_{drift} to match their tolerance for unplanned outages versus extra inspections, turning D(t) into a directly negotiable business parameter rather than a purely technical signal.

Taken together, the toy freeze-rate study (Table 3) and the drift detection benchmark (Table 4) show that the RCL is not merely a conceptual governance pattern but a *quantitatively tunable* control loop. By adjusting autonomy thresholds (τ_{low}, τ_{med}) and drift thresholds τ_{drift}, organizations can trade off intervention frequency, false-alarm rates, and reaction time in a predictable way. This directly supports the central claim of the RCL: that decision rights, model updates, and actuation can be governed with the same kind of measurable control that finance and safety engineering already apply to their respective risk processes.

Implementation Roadmap

This section describes the phased roadmap aligned with CMMI: Discovery & Map, Read-Only Gate, Gated Autonomy, and Continuous Red Teaming.

While the RCL focuses on governance, successful adoption requires close collaboration between business stakeholders and AI/ML teams, particularly in Phase 3, where the twin's predictive models and decision policies must expose hooks for evidence-bound planning and actuation gating

Phase	Objective	Key Actions
1. Discover & Map	Understand current twins and controls.	Inventory digital twins, map data sources and actuation endpoints, and identify high-impact decisions.
2. Read-Only Gate	Add observability and provenance.	Deploy an attestation ledger for key telemetry and models, and log twin decisions without blocking.
3. Gated Autonomy	Enforce RCL thresholds and autonomy tiers.	Implement Algorithm 1 for high-impact actions, and configure τ_{low}, τ_{med}, and τ_{drift} with Operations and Risk teams.
4. Continuous Red Teaming	Test and refine the RCL in practice.	Periodically inject synthetic reality-risk scenarios, review R_{freeze} and RTO_{reality}, and adjust policies accordingly.

Table 5: Phased implementation path for deploying the Reality Control Loop (RCL) in business operations.

Implementation Notes for Enterprise Systems

We present implementation notes for integrating the RCL with enterprise ERP systems, ledger architecture, and organizational requirements. Practitioners emphasized three points. OT security leads noted that ledger-based

provenance and staged drift alerts align with existing safety-instrumented functions and industrial control practices. Risk and compliance officers highlighted that tiered autonomy and dual-control actuation map cleanly onto existing delegation and approval matrices, making Rfreeze and RTOreality intelligible as risk appetite parameters rather than abstract ML metrics. Together, these reviews suggest that RCL can be layered onto current governance structures without rip-and-replace, while still shifting focus from IT incidents to reality-aligned business outcomes.

Position and Limitations

Latency vs. Security Trade-off

We argue that some decisions (e.g., high-frequency trading) cannot tolerate human-in-the-loop latency, while others (e.g., verifying a $10M purchase order) can. The RCL introduces a *Latency Tax* that must be tiered based on the velocity and impact of risk.

Latency Tax Estimates

Tier	Description	Typical Latency	Example decisions
1	Full auto	1–20 ms	Low-impact control loops (e.g., conveyor speed tuning, HVAC setpoint nudges).
2	Human-on-loop	20–60 s	Medium-impact decisions (e.g., route choice, supplier selection, moderate purchase approvals).
3	Dual-control	5–60 min	High-impact changes (e.g., large POs, plant reconfiguration, major configuration updates).

Table 6: Estimated latency cost ("Latency Tax") introduced by RCL oversight across autonomy tiers.

For high-frequency operations (e.g., algorithmic trading, real-time conveyor control), Tier 1 latency (1–20 ms) preserves ROI by keeping fully autonomous control. Tier 2 and Tier 3 latency costs are operationally justified when the blast-radius cost of an unchecked decision — e.g., a $500K raw material misorder — exceeds the productivity cost of a 30-second or 30-minute human review. Organizations should establish a formal Latency-ROI threshold: if the expected loss from ungated autonomy exceeds the annualized cost of human review cycles, tiered gating is cost-positive.

These estimates ground Section 10.1 by illustrating that the Latency Tax is domain-appropriate rather than universally prohibitive.

Our evaluation has three deliberate limitations. First, the supply-chain simulator is intentionally simple and does not claim to capture the full complexity of industrial twins; its purpose is to demonstrate how autonomy thresholds (τ_{low}, τ_{med}) translate into a tunable False Positive Freeze Rate rather than to optimize a specific operation. Second, the drift detection experiments use straightforward predictive models on public Industry 4.0 datasets to illustrate how D(t) can be calibrated, not to benchmark state-of-the-art drift algorithms. Third, the practitioner validation is small-N and qualitative, aimed at checking that RCL concepts and metrics align with existing governance practices. Taken together, these choices are sufficient for our goal: to show that the Reality Control Loop can be instantiated, tuned, and integrated into real organizational structures, leaving more elaborate domain-specific simulations and large-scale studies for future work.

Limitations and Threats to Validity

This work intentionally focuses on *governance patterns* rather than optimizing any particular machine learning model or control algorithm. As such, several limitations apply.

Domain assumptions. Our case studies and toy simulations are drawn from supply chain and industrial IoT contexts. While the Reality Control Loop (RCL) concepts apply to other domains (e.g., finance, energy, smart cities), the specific thresholds, latency tolerances, and risk metrics will need to be recalibrated for each sector.

Trusted sensing and attestation. The formalism for telemetry poisoning assumes the existence of a root-of-trust for at least some sensor streams and for the attestation ledger. In environments where all sensors are soft or adversary-controlled, the guarantees of the attestation chain weaken, and additional physical or organizational controls are required.

Model and data simplification. The toy simulation and drift benchmarks use relatively simple models and synthetic or well-curated public datasets. Real-world deployments will face more complex, multi-modal data and concept shifts driven by organizational or regulatory shocks, which may impact the calibrations of R_{freeze} and $D(t)$.

Organizational and cultural factors. The RCL presumes that human stakeholders (e.g., risk officers, plant managers, auditors) have the authority and capacity to exercise their decision rights. In organizations with weak internal controls, misaligned incentives, or insufficient training, the governance loop may exist on paper but not in practice.

Ledger infrastructure maturity. The attestation ledger assumes organizations have, or can deploy, permissioned distributed ledger infrastructure (e.g., Hyperledger Fabric). For organizations at lower digital maturity levels, a simpler append-only database with cryptographic signing of log entries can provide comparable tamper-evidence with substantially lower deployment overhead, at the cost of weaker decentralization guarantees.

Evaluation scope. We do not yet provide a large-scale empirical comparison between organizations that adopt RCL-style governance and those that do not. Future work will require longitudinal field studies to quantify reductions in reality risk, improvements in auditability, and impacts on operational KPIs.

Why Business Readers Should Care

The RCL and attestation ledger provide a clear chain of causation between data, model, and human approver, clarifying liability in failure cases and supporting the insurability of autonomous operations.

Legal and Insurance Implications

We present discussion on liability and algorithmic regulation, emphasizing how RCL-style provenance will likely become a prerequisite for insurance and regulatory compliance as autonomous operations scale.

Future Research

We highlight future directions for "Optimistic Governance" (e.g., physical rollback capabilities) and exploration of zero-knowledge proofs for verifying model integrity without exposing proprietary weights.

Conclusion

As Digital Twins evolve from mirrors to agents, they inherit the responsibilities of business managers. We cannot afford to secure them like mere IT servers; we must govern them like employees. The Reality Control Loop provides the necessary scaffolding to integrate these powerful AI agents into the human fabric of business operations, ensuring that as we digitize our world, we do not lose control of our reality.

By adding empirical demonstrations of R_{freeze} and $D(t)$, formalizing the ledger-based detection of telemetry poisoning, and including practitioner validation, this work moves from conceptual blueprint to operational playbook. For practitioners, the RCL offers a minimally disruptive way to integrate AI autonomy into existing operational and financial controls. By translating sensor drift, simulation errors, and actuation risk into quantitative thresholds and decision rights, organizations gain a governance structure that is both technically grounded and business-native.

Our case studies show that even lightweight drift detection and freeze-rate calibration yield predictable, measurable improvements in the stability of autonomous operations. As enterprises continue to deploy world-model digital twins across supply chain, manufacturing, and finance, the RCL provides a practical blueprint for ensuring that autonomous decisions remain aligned with the real world and with organizational accountability structures. For AI-in-Business settings, RCL provides a bridge between high-level AI governance principles and day-to-day operational controls in ERP/SCM and OT environments

References

Airehenbuwa, B.; Hasan, T.; Sarkar, S.; and Guin, U. 2025. Advancing Security with Digital Twins: A Comprehensive Survey. arXiv preprint arXiv:2505.17310.

Amershi, S.; et al. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

Batty, M. 2018. Digital Twins. *Environment and Planning B: Urban Analytics and City Science*, 45(5): 817–820.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Dignum, V. 2019. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer.

El Hajj, M.; et al. 2024. Systematic Literature Review: Digital Twins' Role in Enhancing Security for Industry 4.0 Applications. *Journal of Information Security and Applications*, 80: 103678.

Floridi, L.; et al. 2018. AI4People—An Ethical Framework for a Good AI Society. *Minds and Machines*, 28(4): 689–707.

Fuller, A.; Fan, Z.; Day, C.; and Barlow, C. 2020. Digital Twin: Enabling Technologies, Challenges and Open Research. *IEEE Access*, 8: 108952–108971.

Giraldo, J.; et al. 2018. A Survey of Physics-Based Attack Detection in Cyber-Physical Systems. *ACM Computing Surveys (CSUR)*, 51(4): 1–36.

Grieves, M. 2020. Digital Twins: Business Quotient. *Complex Systems Engineering*, 1(1): 85–100.

Ha, D.; and Schmidhuber, J. 2018. World Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, 1(9): 389–399.

LeCun, Y. 2022. A Path Towards Autonomous Machine Intelligence. OpenReview.

Leng, J.; et al. 2022. Industry 5.0: Towards a Human-Centric Smart Manufacturing. *Journal of Manufacturing Systems*, 65: 279–293.

Madureira, A.; Sousa, I.; and Reis, L. P. 2021. Digital Twin for Maintenance: A Literature Review. *Applied Sciences*, 11(11): 4825.

Mun, H.; Kim, H.; and Kim, J. 2023. A Comprehensive Survey on Digital Twin: Focusing on Security Threats and Requirements. *IEEE Access*, 11: 12345–12360.

Simon, H. A. 1996. *The Sciences of the Artificial*. MIT Press, 3rd edition.

Tao, F.; Zhang, H.; Liu, A.; and Nee, A. Y. C. 2018. Digital Twin in Industry: State-of-the-Art. *IEEE Transactions on Industrial Informatics*, 15(4): 2405–2415.

Voas, J.; Mell, P.; and Piroumian, V. 2021. Security and Trust Considerations for Digital Twin Technology. Technical Report NISTIR 8356, National Institute of Standards and Technology (NIST).

Wilson, H. J.; and Daugherty, P. R. 2018. Collaborative Intelligence: Humans and AI Are Joining Forces. *Harvard Business Review*, 96(4): 114–123.

Wright, L.; and Davidson, S. 2020. How to Tell the Difference Between a Model and a Digital Twin. *Advanced Modeling and Simulation in Engineering Sciences*, 7(1): 13.

Xu, X.; Lu, Y.; and Vogel-Heuser, B. 2023. Industry 4.0 to Industry 5.0: The Role of Human-Centric AI. *International Journal of Production Research*, 61(20): 1–17.