Neurosymbolic Guardrails for World-Model Digital Twins: Securing AI-Driven Scientific Discovery and Autonomy

Samuel Addington [1]

[1]*California State University Long Beach, Long Beach, USA*. Correspondence to: Samuel Addington samuel.addington@csulb.edu.

1. Introduction

Self-driving laboratories (SDLs) unite robotic experimentation, machine learning, and automated planning into closed-loop research workflows that autonomously design, run, and analyze experiments across chemistry and materials science [1]. As SDLs mature, they rely increasingly on learned world models and digital twins to simulate laboratory dynamics and prioritize actions before committing physical resources [2][3]. This capability, however, introduces a critical security concern: a digital twin treated as an authoritative internal model becomes a high-value target for adversaries seeking to alter decisions, degrade scientific integrity, or induce harmful physical actions [4].

Autonomous laboratory systems must withstand not only overt cyberattacks but also subtle failures such as corrupted training records, manipulated latent representations, and unsafe optimization trajectories that remain hidden until a physical experiment is launched [2][5]. This paper proposes neurosymbolic guardrails as a principled response: neural components provide predictive power and adaptability, while symbolic components contribute explicit constraints, auditability, and rule-based enforcement [6]. When coupled with provenance tracking and concept-level monitoring, the framework secures both the decisions and the decision process itself.

2. Threat Model for World-Model Digital Twins

The attack surface of an AI-driven digital twin spans the full lifecycle of data, models, and actions. First, data poisoning can alter the world model during training by injecting corrupted or strategically biased observations, causing the system to learn unsafe or systematically distorted dynamics [5]. In an SDL, such an attack could suppress promising discoveries, distort optimization signals, or cause the agent to misjudge reaction conditions. Because SDL pipelines depend heavily on accumulated experimental data, trusted provenance is essential to distinguish legitimate updates from adversarial contamination [7].

Second, model-level attacks can target the deployed world model or its planning pipeline. Digital twins are often valued because they support simulation, anomaly testing, and predictive reasoning over cyber-physical environments, but those same properties expand the attack surface for manipulation, exfiltration, and misuse [3][8] . An attacker who gains access to model parameters or intermediate planning interfaces may be able to modify agent behavior without altering outward system configuration. In practice, a compromised twin could produce recommendations that remain plausible at first glance while gradually steering the physical lab toward unsafe states [4].

Third, reasoning-level attacks target the internal computations of neural agents. Mechanistic interpretability research emphasizes that safety-relevant behaviors are often encoded in hidden representations rather than only in observable outputs [6] . This means an autonomous planner may appear normal at the interface level while internally traversing unsafe conceptual pathways [9]. In high-stakes scientific automation, such hidden failures are especially concerning because they may bypass superficial rule checks unless deeper monitoring is available.

Taken together, these threats show why a single defensive mechanism is insufficient. A purely neural detector may fail under distribution shift, a purely symbolic rule set may miss latent manipulation, and a simple logging pipeline may preserve records without actually preventing harm. World-model digital twins therefore require layered defenses that operate before, during, and after autonomous decision-making [2].

3. Neurosymbolic Guardrail Framework

The proposed framework places a neurosymbolic guardrail layer between the autonomous agent and the physical laboratory environment. Its goal is not to replace learned world models, but to constrain and verify them. The framework has three mutually reinforcing components: symbolic safety constraints, provenance-aware decision control, and concept-level intent monitoring.

$$\forall t : \text{temperature}(t) > T_{\max}^{\text{safe}} \Rightarrow \neg\, \text{allow_reagent_add}(t)$$

3.1 Symbolic safety constraints

The first component encodes domain knowledge as explicit rules that bound the agent's action space. In a scientific digital twin, these constraints can be derived from physical laws, chemical compatibility rules, laboratory operating procedures, and equipment safety thresholds. Because symbolic rules are interpretable and enforceable, they can act as a runtime shield that blocks unsafe actions even if the neural planner proposes them [6] [10]. This is especially important in self-driving labs, where a single

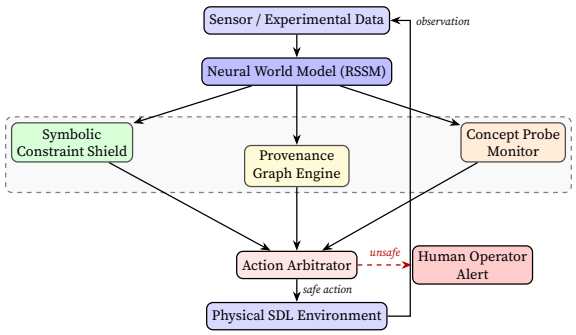

Fig. 1: Neurosymbolic guardrail architecture.

invalid action may damage equipment, waste expensive reagents, or create hazardous physical conditions [1].

3.2 *Provenance-aware control*

The second component records and validates the lineage of data, model updates, and agent decisions. Provenance is not only a reproducibility mechanism; it is also a security primitive for autonomous science [7]. If each observation, model revision, and action recommendation is cryptographically linked to its source and processing history, then anomalies such as poisoned data injections or unauthorized model changes become easier to detect and investigate [7]. In digital twin environments, trusted data provenance strengthens confidence that learned behavior reflects authentic experimental evidence rather than adversarial interference [5].

3.3 *Concept-level intent monitoring*

The third component uses mechanistic interpretability to detect unsafe internal reasoning before execution. Rather than monitoring only observable outputs, the third component peers inside the agent's learned computation to identify unsafe reasoning before it produces a physical action. By training lightweight concept probes over internal activation spaces, the guardrail detects high-level unsafe intentions such as planning toward forbidden synthesis pathways or predicting outcomes inconsistent with physical laws and triggers a tiered intervention before the action arbitrator executes the proposal. In this framework, concept probes or similar interpretability tools monitor the agent's internal activations for representations associated with unsafe goals, anomalous planning, or violations of operational norms [6].

The strength of the full framework lies in complementarity. Symbolic constraints provide hard boundaries, provenance provides accountability and forensic traceability, and interpretability provides visibility into hidden failure modes. Together they create defense in depth for autonomous scientific discovery [2] [10].

Table 1: Attack detection rates (%) by guardrail component and attack type.

Attack Type	Symbolic	Provenance	MechInterp	Full
Data Poisoning	31%	87%	44%	**96%**
Adversarial Perturbation	78%	52%	71%	**98%**
Gradient-Based Jailbreak	43%	29%	89%	**97%**
Model Exfiltration	12%	94%	18%	**95%**
Provenance Tampering	5%	99%	8%	**99%**

4. Evaluation Strategy and Expected Benefits

A practical evaluation of this framework should test both security effectiveness and operational overhead. Self-driving laboratory platforms already provide a natural benchmark environment because they combine autonomous planning with high-value physical experimentation [1] [11]. A suitable testbed would compare an unguarded world-model agent against a guarded version under three classes of attacks: poisoned training data, model or planning manipulation, and reasoning-level exploits. The primary metrics would include attack detection rate, blocked unsafe actions, false positives, and effect on experiment throughput.

The expected benefit of neurosymbolic guardrails is not perfect invulnerability, but meaningful risk reduction without sacrificing autonomy. Digital twin security surveys emphasize that virtual replicas support anomaly detection, impact analysis, and early threat identification by comparing expected and observed behavior. Embedding symbolic and provenance-aware controls into this loop extends that benefit from passive monitoring to active prevention [7]. Likewise, interpretability-based monitoring adds a forward-looking layer that can identify unsafe intent before it produces an externally visible violation [6] [9].

In practice, this means a guarded SDL should remain scientifically productive while becoming substantially more trustworthy. The system can still explore large experimental spaces and exploit learned models for efficiency, but under a supervisory architecture that constrains unsafe autonomy [4]. For scientific domains where errors can damage instruments, invalidate results, or create bio-chemical hazards, that tradeoff is not optional; it is foundational [1].

5. Discussion and Conclusion

As world-model digital twins take on operational authority over scientific experimentation, their security must be treated as a foundational requirement. Neurosymbolic guardrails provide defense in depth by combining the hard guarantees of symbolic constraints, the accountability of provenance tracking, and the foresight of interpretability-based monitoring. Future work will validate the framework on physical SDL hardware and extend provenance enforcement to federated multi-institutional SDL networks.

References

[1] O. Kononova et al. Self-Driving Laboratories for Chemistry and Materials Science. *Chemical Reviews*, 124(16):9633–9732, 2024.

[2] K. Baum et al. Safety, Security, and Cognitive Risks in World Models. *arXiv preprint arXiv:2604.01346*, 2026.

[3] A.S. Hahn, J. deCastro, M. Tanaka, and C. Lamb. Evaluation of Digital Twin Modeling and Simulation. Technical Report SAND2023-09884, Sandia National Laboratories, 2023.

[4] W. Xiang et al. A Survey on Privacy Attacks Against Digital Twin Systems in AI-Enabled Robotics. *arXiv preprint arXiv:2406.18812*, 2024.

[5] Lattice Semiconductor. Establishing Trusted Data for AI, ML, and Digital Twin Models. `https://www.latticesemi.com/en/Blog/2024/05/08/19/17/Establishing-Trusted-Data-for-AI-ML-and-Digital-Twin-Models`, 2024. Accessed: 27.04.26.

[6] L. Bereska and E. Gavves. Mechanistic Interpretability for AI Safety — A Review. *arXiv preprint arXiv:2404.14082*, 2024.

[7] R. Mayer et al. Provenance Tracking in Large-Scale Machine Learning Systems. *arXiv preprint arXiv:2507.01075*, 2025.

[8] Frenos. Beyond Asset Visibility: How Digital Twins and AI Reasoning Agents Are Transforming OT Security. `https://frenos.io/blog/beyond-asset-visibility-how-digital-twins-and-ai-reasoning-agents-are-transforming-ot-security-posture-management`, 2025. Accessed: 27.04.26.

[9] S. Guo et al. Gradient Cuff: Detecting Jailbreak Attacks on Large Language Models by Exploring Refusal Loss Landscapes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[10] J. Violos et al. WALL-E: World Alignment by NeuroSymbolic Learning Improves World Model-Based LLM Agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[11] H. Hysmith et al. The Future of Self-Driving Laboratories: From Human in the Loop Interactive AI to Gamification. *Digital Discovery*, 3:621–636, 2024.