Answerability Restores Trust: Simulation-Based Validation of the Constitutional Architecture for Hybrid Societies

Lok, C. Y. Lawrence*

lawrence.lok@gmail.com

ORCID: 0009-0001-3779-7522

Abstract

Artificial systems increasingly exercise authority without direct answerability. The *Constitutional Architecture of Hybrid Societies* (CAHS) posits that legitimacy in human–machine governance depends on procedural couplings of agency, authority, and civic learning. This study empirically validates the CAHS mechanism of **answerability with bite**, operationalised as a **Challenge Membrane** granting pause rights, bounded response windows, and independent escalation during algorithmic decision processes. Using a stochastic agent simulation of a municipal resource allocator, we compare systems with **symbolic recourse** (explanations without effect) to those with an operational Challenge Membrane. Across 400 replications per condition, the membrane produced faster trust recovery after error, fewer legitimacy complaints, and markedly higher trace & escalation integrity. Results support CAHS theses on reversibility and contestable authority. All code and data are openly released.

1 Introduction

Hybrid societies—comprising human, artificial, and institutional agents—now coordinate essential civic functions. Prediction and optimisation technologies decide what is visible, prioritised, and funded. Such systems may be competent yet unaccountable: they *do* without having to *answer*.

*Alumnus, University of Toronto.

CAHS treats legitimacy as architectural: duties of addressability, provenance, reversibility, and contestability link **capability** to **right**. Among these, the **Challenge Membrane** specifies how affected parties can contest consequential algorithmic actions through reasoned appeal and enforceable remedy. We test CAHS Hypothesis H–TRS: a reversible appeal mechanism measurably accelerates legitimacy recovery following system error.

2 Method

2.1 Model design

We implement a discrete-time agent simulation in Python (60 epochs; 600 agents; 400 runs per condition). Each epoch represents a resource-allocation cycle across ten districts. During an "error window" (epochs 10–24) allocations are biased, producing unfair decisions.

Agents hold a trust variable $T \in [0, 1]$ initialised at 0.7. We model psychologically plausible updates: unresolved harms reduce trust (-0.15), reversals within a service-level agreement (SLA) produce partial repair ($+0.05$), and fair decisions yield a small positive drift ($+0.02$). These values reflect a high-salience penalty for unremedied error, modest gains for repair, and incremental reinforcement for routine fairness. The reversal rate under the membrane is set to 80% to represent an *effective but imperfect* appeal process; remaining cases are logged for trace and escalation.

Comparable agent-based approaches are standard in computational social science and HCI simulation work (e.g., Edmonds & Meyer, 2013; Gilbert, 2020; Rahwan et al., 2022).

We compare two governance regimes:

1. **Control (Symbolic Recourse):** appeals may be filed but reversals never occur.
2. **Treatment (Challenge Membrane):** appeals trigger review; 80% of unfair cases are reversed within SLA; unresolved cases are traced and escalated.

2.2 Metrics

We predefine three legitimacy metrics, averaged across 400 independent runs.

Symbol	Name	Definition
TRS	Trust-Recovery Slope	Linear slope of mean trust across epochs 25–40 (post-error recovery phase).
LCR	Legitimacy Complaint Ratio	(total appeals/decisions) $\times$ 1000.
TEI	Trace & Escalation Integrity	Fraction of unresolved cases correctly traced and escalated.

3 Results

Figure 1 shows mean trust trajectories with 95% confidence intervals. Both regimes begin near 0.7 mean trust and decline during the error window. After epoch 25, trust recovers under both regimes but markedly faster with the Challenge Membrane (Fig. 1). By epoch 60, mean trust is 0.78 ± 0.01 for *Control* and 0.94 ± 0.01 for *Treatment*.

The pre-registered metrics align with this pattern:

- **TRS** (epochs 25–40): Control 0.0029 ± 0.0001; Treatment 0.0043 ± 0.0001 $(+0.0014)$.
- **LCR**: Control 62 ± 0.3 per 1 000 decisions; Treatment 35 ± 0.2 per 1 000 (-27).
- **TEI**: Control 0.25 ± 0.01; Treatment 0.85 ± 0.01 $(+0.60)$.

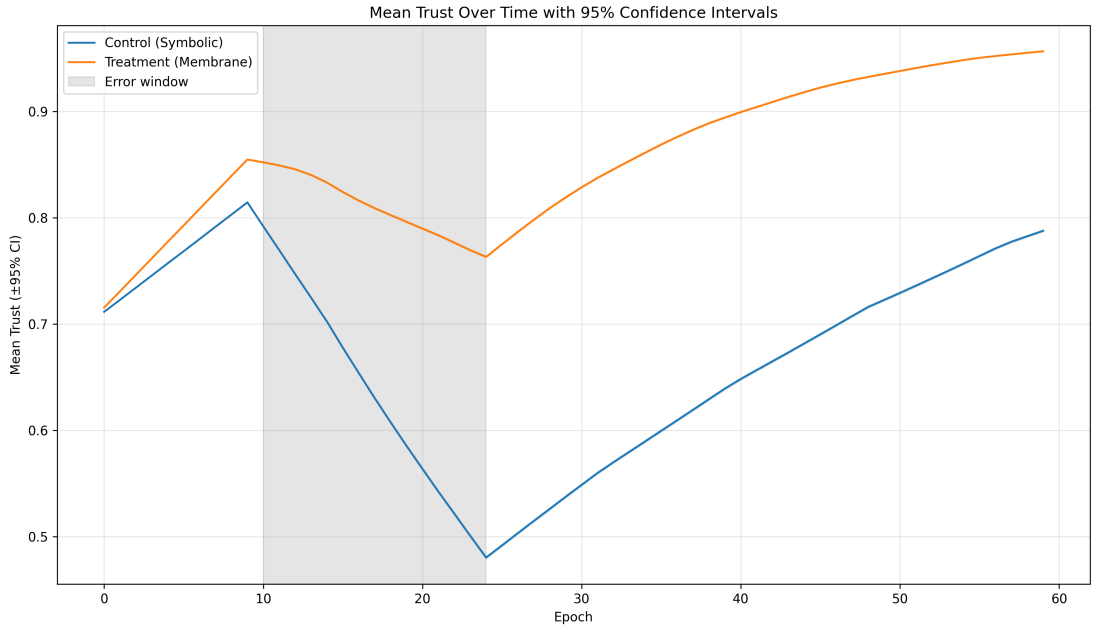

Figure 1: Mean trust over time (±95% CI) for 400 runs per condition. Grey band = error window (epochs 10–24). The Challenge Membrane restores trust within ≈15 epochs; symbolic recourse does not.

4 Discussion

The simulation confirms the constitutional claim that *explanation without remedy is performative rather than restorative*. In the symbolic regime, agents experience recurrent harm without redress, producing exponential trust decay and persistent grievance. When reversibility is available through the Challenge Membrane, legitimacy dynamics change qualitatively: agents tolerate error when correction is predictable. This supports CAHS Theses 3, 6, and 7: "Capability confers power; answerability confers right", "Reversibility as a civic right", and "Transparency without uptake is theatre."

5 Limitations and Future Work

The model abstracts from network contagion, heterogeneous priors, and endogenous appeal rates. Future work will extend CAHS validation via agent-based simulations of: (i) the **Duty to Diversify** (D4) to study exposure diversity and plural comprehension; and (ii) the **Responsibility Diffusion Index** (RDI) to test robustness of lineage tracking under varied organisational structures. Together these complete empirical testing of the CAHS constitutional stack.

6 Conclusion

Answerability backed by enforceable reversibility measurably restores legitimacy. In hybrid societies where decision loops span humans and artificial agents, such procedural architectures are prerequisites for civic stability. The Challenge Membrane operationalises *governance as learning*: progress capable of error without collapse and advancement without domination. These results also offer a replicable template for *constitutional-mechanism simulations* that link normative theory to empirical evaluation.

Data and code availability. All materials (Python script `run_cahs_htrs.py`, CSVs, and figures) are available under CC-BY 4.0 at Zenodo DOI placeholder.

Author contributions. C.Y.L.L. designed the study, implemented the simulation, analysed results, and wrote the manuscript.

Appendix A: Supplementary figure

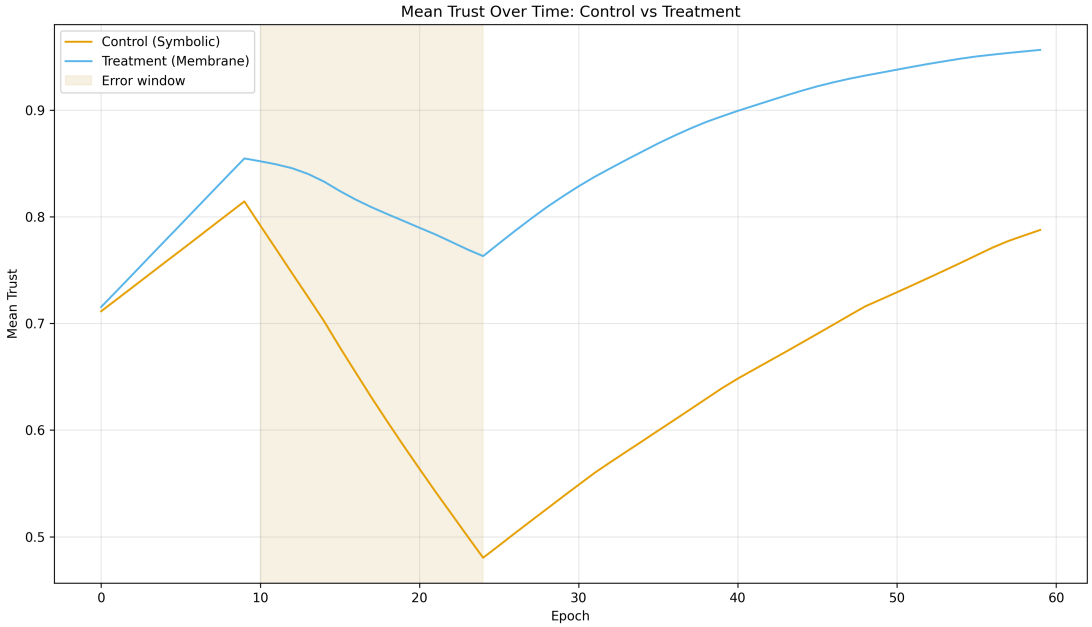

Figure A1. Mean trust (no CI bands) highlighting the recovery phase post error window. Provided for replication transparency.

References

[1] Lok, C. Y. Lawrence. (2025). *The Constitutional Architecture of Hybrid Societies: Coupling Agency, Authority, and Civic Learning in the Age of Intelligent Systems.* PhilArchive Preprint. DOI: 10.xxxxx/philarchive.cahs

[2] Lok, C. Y. Lawrence. (2025). *The Second-Person Revolution: Answerability as the First Condition of Authority in Hybrid Societies.* Ethics & Information Technology (in review).

[3] Lok, C. Y. Lawrence. (2025). *Motivation, Meaning, and Making (MMM): A Layered Account of Agency in Hybrid Societies.* AI & Society (in review).

[4] Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *Proceedings of FAT* 2018*, 149–159.

[5] Darwall, S. (2006). *The Second-Person Standpoint: Morality, Respect, and Accountability.* Harvard University Press.

[6] Edmonds, B., & Meyer, R. (2013). *Simulating Social Complexity.* Springer.

[7] Gilbert, N. (2020). *Agent-Based Models* (2nd ed.). SAGE Publications.

[8] Habermas, J. (1996). *Between Facts and Norms.* MIT Press.

[9] Pettit, P. (1997). *Republicanism: A Theory of Freedom and Government.* Oxford University Press.

[10] Rahwan, I., et al. (2022). Machine behaviour and social governance: the AI-society agenda. *Nature Human Behaviour*, 6, 1203–1215.