Conflict as Telemetry for Illegible AI: Governing LLM Agent Workflows (Extended Abstract)

Paul LaPosta

Independent Researcher
paul@theherongroupllc.com

Modern enterprises are moving from generative AI as a productivity layer to autonomous large language model (LLM) agents that execute multi-step workflows with limited human oversight. This shift amplifies an under-instrumented governance risk: organizational illegibility, where artifact polish and apparent operational smoothness decouple from understanding, accountability, and recoverability (LaPosta 2025b). When decisions cannot be reconstructed, governance becomes performative, auditability becomes brittle, and incident response degrades into blame allocation instead of safe reversal (NIST 2023, 2025; Sculley et al. 2015).

We propose a runnable field protocol that treats recurring conflict around mission-critical agent workflows as early-warning telemetry for illegibility. Conflict here means repeated escalations, reversals, exception requests, rework loops, and blame cycles. The method couples five illegibility fractures with six recurrent patterns under load. The fractures are Decision Fog, where teams cannot reconstruct the decision chain; Knowledge Drift, where operational expertise migrates into prompts, tools, or vendors; Synthetic Competence, where throughput is mistaken for capability; Ghost Apprenticeship, where learning becomes tool-shaped rather than judgment-shaped; and Promotion Blindness, where visible velocity outranks independent capability (LaPosta 2025b). The recurrent patterns are Cynic, Ghost, Hero-Martyr, Passive Burnout, Underminer, and False Compliance (LaPosta 2025a). Teams score fracture-pattern pairs, select the top two, then run the loop. The coupling is diagnostic rather than moralizing: conflict is treated as a signal that decision rights, knowledge distribution, and risk ownership have drifted out of alignment under automation pressure.

Field Procedure: Map-Probe-Trace-Teach (MPTT). Inputs: active conflict, agent workflow, ownership map. **Steps:** score fracture-pattern pairs and select the top two; map the end-to-end decision chain across humans, tools, and models; probe for gaps in explanation and ownership; trace lineage to specific owners and artifacts; teach by updating runbooks, tests, review gates, and escalation rules. **Outputs:** owner map, decision record, rollback test, gate update. **Stop-ship:** if time-to-explain is unchanged after two cycles, re-scope the workflow.

The protocol is designed for workflows where agents can trigger costly or irreversible actions, where ownership is shared across teams, or where oversight must be demonstrable to auditors, regulators, or customers. It is least effective when conflict stems primarily from resource scarcity, interpersonal dynamics unrelated to work structure, or external constraints no governance change can address. These boundary conditions make the approach falsifiable in practice. We define four operational outcomes: shorter time-to-explain and time-to-reverse during incidents, more complete decision records, lower recurrence of the same conflict around the same workflow, and fewer unowned risk transfers between teams.

The contribution is not a new model or agent architecture. It is a management instrument for intelligent transformation that raises reconstructability as agent autonomy scales. The core claim is operational: if organizations cannot explain how agent-mediated decisions were made, who could change them, and who carries the risk when they fail, then conflict should be read as telemetry that the system has become illegible.

References

LaPosta, P. 2025a. *Crafting Conflict Volume 1: Managing Saboteur Patterns in High-Performing Teams*. Leanpub.

LaPosta, P. 2025b. *The Illegibility Crisis: Instrumentation for AI-Era Leadership*. Leanpub.

National Institute of Standards and Technology (NIST). 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1.

National Institute of Standards and Technology (NIST). 2025. *Incident Response Recommendations and Considerations for Cybersecurity Risk Management*. NIST SP 800-61 Rev. 3.

Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.-F.; and Dennison, D. 2015. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems 28*.