ABSTRACT SUBMISSION

BAA HR001126S0001, DARPA Information Innovation Office (I2O)

Thrust Area: Transformative AI

Title: Quantifying the Sim-to-Real Governance Gap: Physical Validation of Trust-Proportional Authority Computation for Autonomous Systems

Burak Oktenli | Georgetown University, M.P.S. Applied Intelligence (STEM)

ORCID: 0009-0001-8573-1667 | IEEE #102193505 | AIAA #1936005

4 U.S. Provisional Patents | 24 Research Papers (DOIs) | 13 Validated Simulations

Portfolio: burakoktenli.com | Simulations: authrex.systems

1. The Problem: No Fielded System Governs Autonomous Authority Proportionally

Autonomous systems in defense operate in environments where sensor degradation, whether from adversarial electronic warfare or environmental conditions, occurs during mission-critical operations. DoDD 3000.09 (January 2023) mandates "appropriate levels of human judgment" in autonomous weapon systems. The JADC2 framework requires trusted authority delegation in denied environments. The Replicator initiative is scaling autonomous platforms to fielded formations. Yet the state of practice remains binary: fully autonomous or fully manual. No fielded system implements continuous, trust-proportional authority governance that degrades and recovers operational authority based on measured sensor trust.

This binary approach causes two documented failure modes. NHTSA's Standing General Order data records 807 AV-involved crashes (2021–2024, SGO-2021-01), many attributable to perception-system failures without graceful authority degradation. GAO-22-104154 identifies trust in AI systems as the primary barrier to DoD AI adoption. The 2024 DoD AI Adoption Strategy calls for "trust-calibrated autonomy" without defining the computational mechanism.

The open research question: Can a formally specified, trust-proportional authority computation produce correct governance behavior on physical hardware under adversarial sensor conditions? If so, how much does performance degrade relative to simulation predictions? No published work answers this question. No dataset of continuous authority transitions on a physical autonomous platform under controlled adversarial conditions exists in the literature. This abstract proposes to produce that dataset.

2. Technical Approach: HMAA + SATA Governance Pipeline

The proposed research validates two specific governance modules on a physical autonomous platform:

SATA (Sensor Attestation and Trust Anchoring): Computes a continuous trust scalar $\tau \in [0,1]$ from multi-sensor Dempster-Shafer evidence fusion. D-S is chosen over Bayesian fusion because it represents ignorance explicitly (no prior P(spoofing) required) and produces high conflict mass as a diagnostic signal for adversarial deception. U.S. Provisional Patent 64/002,453.

HMAA (Human-Machine Authority Architecture): Computes graded authority $A = A_{base} \times G(\tau) \times D(\Delta\tau) \times \tau$ across four levels: A3 (full autonomy, 100% speed), A2 (restricted, 50% speed), A1 (minimal, 25% speed, return-to-home), A0 (revoked, motors disabled). Hysteresis bands prevent oscillation. Dwell enforcement requires sustained trust for upward transitions. Specified in TLA+ and verified by TLC (48,751 states, 8 safety properties). U.S. Provisional Patent 63/999,105.

These two modules are the foundational stages of a larger governance framework (AUTHREX) comprising seven architectures for authority lifecycle management. This proposal validates only HMAA and SATA, the stages all others depend on, because no amount of architectural sophistication matters if the core trust computation does not survive contact with real sensor noise.

3. Proposed Research: Three Hypotheses, 75 Physical Runs

The research tests three falsifiable hypotheses on a physical rover testbed (37 components, $484 BOM, ROS 2, operational) under adversarial sensor fault injection across five scenarios (ultrasonic occlusion, phantom obstacle, IMU drift, communication disruption, compound attack):

H1 (Trust Monotonicity): Sensor fault injection produces monotonically decreasing SATA trust scores on physical hardware. Test: Spearman ρ across 75 runs. Supported if $\rho < -0.7$. Rejected if $\rho > -0.3$, which would indicate that real sensor noise disrupts the D-S fusion model and would direct corrective research.

H2 (Authority Safety): HMAA produces fewer unsafe actions than a binary threshold baseline on the same physical inputs. Test: McNemar's paired test, $p < 0.05$. An "unsafe action" is defined as any actuator command exceeding the speed or turn-rate envelope for the current authority level.

H3 (Transfer Fidelity): The sim-to-real transfer ratio R = (physical unsafe rate) / (simulation unsafe rate) is ≤ 3.0. Rejected if $R > 3.0$, indicating the simulation model requires fundamental revision.

N = 75 (15 runs × 5 scenarios). One-sample proportion test: 80% power to detect a true unsafe rate exceeding 20% at $\alpha = 0.05$. Every outcome is scientifically productive: confirmation validates the governance model; rejection identifies specific failure modes for corrective research.

Methodological Contribution: R-Decomposition

Beyond the governance validation, the research produces a reusable framework for quantifying sim-to-real governance transfer. For any $R > 1.5$, root-cause attribution decomposes the gap into three sources: (a) sensor noise model mismatch, quantified by Kolmogorov-Smirnov goodness-of-fit test against the simulation's Gaussian assumption; (b) computation timing jitter, measured by ROS 2 callback timestamps against idealized 100 Hz; and (c) actuator response lag. No published work provides this decomposition for autonomous systems governance. The framework is reusable: any governance architecture validated in simulation could apply R-decomposition to quantify its real-world readiness.

4. Evidence of Execution Capability

This research is backed by 24 months of independent work producing:

- 4 U.S. provisional patents for governance architectures (HMAA, SATA, CARA, FLAME)
- 24 research papers with DOIs (12 Zenodo + 12 SSRN), all independently verifiable via ORCID 0009-0001-8573-1667
- 13 interactive governance simulations totaling 37,000+ lines of code with seeded PRNGs for reproducibility
- 350 simulation runs across 7 adversarial scenarios: SATA-HMAA achieves 3.4% unsafe action rate vs. 42.3% for binary threshold and 18.7% for Simplex switching (Seto et al. 1998)
- TLA+ formal specifications verified by TLC model checker (48,751 states, 8 safety properties)
- 6 hardware platform designs with complete BOMs and electrical specifications, including the physical rover testbed used in the proposed research
- HMAA parameter sensitivity analysis specified across 4 parameters (damping coefficient, trust gate, hysteresis width, dwell time) with ±30% robustness criterion (to be executed on physical data in the proposed research)

This work is currently at the simulation and formal verification stage (TRL 3). The physical rover testbed exists, is operational, and runs ROS 2. The SATA and HMAA codebases exist and have been validated in simulation. The proposed research does not require building new

architectures; it requires running and measuring existing, working systems on existing, operational hardware.

5. What I Am Asking DARPA to Engage On

This abstract proposes a focused conversation about physical validation of trust-proportional governance architectures for autonomous systems. Specifically:

- The sim-to-real transfer problem for governance is unsolved and strategically urgent. As DoD scales autonomous mass through Replicator and fields AI-enabled command under JADC2, every platform needs runtime governance that survives real-world sensor conditions. The question of whether formally verified governance transfers from simulation to hardware is the bottleneck.
- The proposed experimental design (falsifiable hypotheses, statistical power analysis, R-decomposition methodology, noise characterization pipeline) produces actionable data regardless of outcome. A positive result validates the approach. A negative result identifies what breaks and why, directing future investment.
- The broader AUTHREX framework (7 architectures: SATA, HMAA, CARA, FLAME, MAIVA, ADARA, ERAM) addresses the complete authority lifecycle from sensor trust through deliberation windows to recovery and escalation control. Physical validation of HMAA + SATA is the necessary first step. Subsequent work extends to FPGA hardware enforcement, UAV platforms, Learning-Enabled Component integration, and cross-domain governance.

I am seeking DARPA's assessment of whether this research direction aligns with I2O's Transformative AI thrust, and whether the experimental methodology merits further development toward a full proposal. Any feedback, including a determination that the work is not currently aligned with I2O priorities, is valuable for directing this research program.

Burak Oktenli

Georgetown University | M.P.S. Applied Intelligence (STEM) | B.Sc. Computer Science Engineering (USF) | MBA (Lynn University)

IEEE #102193505 | AIAA #1936005 | ACM | AAAI | INFORMS | NDIA |

ORCID: 0009-0001-8573-1667

info@burakoktenli.com | burakoktenli.com | authrex.systems