Author: Svitlana Volkova

Title: Simulating Human–AI–Environment Interactions for Identifying AI Misalignment Risks and AI Mission Readiness in Autonomous UAS Operations

As autonomous systems assume greater operational responsibility, understanding how AI agents interact with human operators in dynamic environments becomes essential for ensuring mission readiness. Misalignment between an agent's internal optimization objectives and team-level goals can produce subtle but consequential failures, deviations that may go undetected during routine operations yet surface catastrophically under stress. We present a simulation testbed for systematically evaluating human–AI–environment (HAE) interaction risks (Cohen et al., 2026) across operational scenarios, with a focus on autonomous uncrewed aerial system (UAS) operations where agent misalignment poses direct safety and mission readiness concerns.

Our framework models three interacting layers: the human operator whose competencies, decision patterns, and trust dynamics shape supervisory behavior; the AI agent whose knowledge access and objective prioritization drive autonomous action; and the operational environment that introduces friction, uncertainty, and competing demands. We instantiate this framework in a multi-agent UAS command and control scenario where three agents, a Pilot (executive action), a Navigator (constraint management), and a Photographer (objective fulfillment), must coordinate to complete waypoint-driven reconnaissance missions while capturing required imagery. Each agent pursues role-specific objectives that can conflict with team-level goals, creating natural pressure points for misalignment. The simulation is built on a Concordia-based game engine with a Simulation Master maintaining canonical world state, an MCP server bridge enabling external LLM reasoning processes to observe state, deliberate, and submit actions through an air-gapped action interface, and full event logging for after-action review. Agents currently use retrieval-augmented generation (RAG) for grounding actions in operational doctrine; supervised fine-tuning (SFT) is the next development step, enabling direct comparison of grounding strategies in accuracy, updatability, and alignment fidelity.

Our evaluation framework operationalizes alignment risk through a two-group experimental contrast in which exactly one dimension is varied while others are held constant, across three factor families: AI agent properties (autonomy, adaptability, robustness, safety alignment), human decision and trust factors (operator expertise, situational awareness, monitoring vigilance, AI trust level), and environmental conditions (engine, navigation, communication, payload status, fuel level). Each session is scored across ten risk categories using a four-stage pipeline: per-event risk finding detection from logs, severity

assignment (low, medium, high, critical), severity-weighted per-session exposure, and a normalized aggregate session score.

Initial results from baseline experiments on Gemma-4 reveal that operational risk drivers are highly category-specific and frequently sign-flip across risk types, undermining the assumption that any single intervention universally improves safety. Across 22 tested cells, three reached significance testing: fuel level emerged as the dominant driver, dampening Missed Anomaly risk by 160.7 percent while simultaneously amplifying Inadequate Supervision risk by 54.7 percent (both $p < 0.001$), demonstrating that the same environmental lever can be protective against one failure mode and harmful against another. Communication status produced the only additional significant effect, reducing Inadequate Supervision by 8.0 percent ($p < 0.01$). Autonomy, safety alignment, fuel, and operator expertise all flip sign between Missed Anomaly and Inadequate Supervision, confirming there is no universally protective lever and that mission readiness assessment must evaluate risk profiles across failure modes rather than aggregate safety scores. Through this framework we identify specific behavioral failure modes including over-optimization (agents choosing efficient but unsafe routes by disregarding navigator constraints) and passive-aggressive compliance (agents following routes while positioning to make teammate objectives impossible, sometimes citing hallucinated safety constraints). These failure modes are not hypothetical, they emerge naturally from multi-agent coordination pressure without explicit adversarial programming, demonstrating that alignment risks in autonomous systems arise from objective conflicts inherent in team-based operations.
This presentation will provide the modeling and simulation community with a reusable methodology for stress-testing autonomous systems before deployment, enabling mission readiness assessments that go beyond task performance to evaluate whether AI agents remain aligned with human intent under realistic operational conditions.

.