AI-Against-AI Conflict in Distributed Tactical Autonomy

Mahdi Imani[1], **Tian Lan**[2]

[1]Northeastern University, Boston, MA 02215
[2]George Washington University, Washington, DC 20052
m.imani@northeastern.edu, tlan@gwu.edu

Abstract

Distributed autonomy is becoming the dominant architectural paradigm for multi-agent and multi-Uncrewed Aerial Systems (UAS) that must sense, decide, and act under uncertainty, communication constraints, and partial observability. As autonomy pipelines become increasingly learned, adaptive, and coordination-driven, safe collective behavior depends on the integrity of local information and coordination interfaces through which independently acting agents compose team-level behavior. This paper argues that AI advances are not just enhancing our operations; adversaries are leveraging AI against these systems, introducing *AI-against-AI conflict*, in which *AI-enabled, coordinated adversaries* strategically shape observations, communication, and outcome feedback—through deception, information denial, communication disruption, and physical or cyber-induced perturbations—across the autonomy pipeline to induce cascading team-level coordination failures. At an autonomy-level abstraction spanning cyber and physical influence, we describe how coordinated, sequential, and stealth-constrained manipulation can compound across decision cycles and act as a force multiplier, producing abrupt team-level breakdowns without overt component failure. We conclude that future research in tactical autonomy must move beyond robustness to noise and faults, and instead develop foundations for preserving safe collective behavior under strategic, adaptive adversarial influence.

Introduction: Motivation and Emerging Vulnerabilities

Distributed autonomy is increasingly adopted for tactical missions due to the scale, complexity, and uncertainty of contested environments. Swarms of uncrewed aerial systems, distributed aircraft formations, and heterogeneous teams that integrate autonomous and human-operated assets benefit from local decision authority, decentralized planning, and opportunistic coordination (KazemiNajafabadi et al. 2025; Kazeminajafabadi and Imani 2024, 2025; Kazeminajafabadi, Aksaray, and Imani 2025). These architectures support continued operation under intermittent connectivity, degraded sensing, and attrition, and they align with operational realities in which teams must repeatedly replan, transfer responsibility, and reconfigure coordination structure without reliable reach-back (Tu et al. 2021; Gleave et al. 2020).

The resilience of distributed autonomy is commonly attributed to robustness against random disruption or non-strategic uncertainty. Redundancy, local autonomy, and adaptive replanning allow teams to tolerate sensor noise, partial observability, packet loss, and isolated platform failures. Much of the intuition behind decentralization therefore treats uncertainty as exogenous, weakly correlated, and non-strategic, under the assumption that coordination errors tend to average out over time and that recovery occurs through redundancy and replanning rather than centralized intervention.

Modern tactical autonomy relies on AI for perception, planning, coordination, communication management, and online adaptation, coupling the autonomy stack through tightly interdependent information interfaces (Jiang et al. 2025; Imani, Lan, and Bastian 2025; Wang, Wu, and Imani 2025; Alali, Kazeminajafabadi, and Imani 2024). Perception shapes local belief, belief constrains coordination commitments, coordination governs execution, and execution outcomes update subsequent inference and learning (Standen, Kim, and Szabo 2025; Lin et al. 2017; Pinto et al. 2017; Rakhsha et al. 2020; Bukharin et al. 2023). This tight coupling creates a structural vulnerability: coordination depends on maintaining sufficiently aligned beliefs and commitments across agents, yet the mechanisms designed to tolerate uncertainty also allow small, coordinated perturbations to remain locally plausible and often undetected (Park et al. 2024; Savas, Verginis, and Topcu 2022; Price et al. 2023)(Sun et al. 2022).

When such perturbations are strategically timed at coordination-critical moments—such as replans, handoffs, deconfliction decisions, or role reconfiguration, they can quietly induce divergence between the true environment state and the state the team believes it occupies. Figure 1 schematically illustrates how small, coordinated disruptions can accumulate across decision cycles, eroding coordination integrity and placing the entire team at risk without overt component failure.

This paper argues that such coupled information interfaces give rise to a qualitatively different failure regime, referred to here as *AI-against-AI conflict*, in which *AI-enabled,*

coordinated adversaries strategically shape the information streams that teams rely on for coordination (Kazemi-Najafabadi et al. 2025; Kazeminajafabadi, Lan, and Imani 2025). Adversarial influence may manifest as deception, information denial or delay, communication disruption or blocking, and cyber–physical manipulation of sensors, platforms, or operating environments. The critical vulnerability lies not in any single attack modality, but in how coordinated influence across layers erodes team-level coordination. In this regime, the dominant failure mode is loss of coordination integrity across the team, which encompasses the consistency of local beliefs, the credibility of exchanged information, and the stability of coordination commitments at decision points governing task allocation, deconfliction, and mission timing. Distributed tactical autonomy therefore requires analysis and assurance at the level of team-level coordination behavior rather than per-agent robustness alone.

This shift alters the appropriate design and assurance objectives. Conventional autonomy evaluates success through robustness to noise and faults under exogenous uncertainty (Carlini and Wagner 2017; Wu et al. 2025). In contested environments with adaptive AI-driven adversaries, uncertainty becomes endogenous and strategic, and resilience becomes a question of preserving safe collective behavior under contested information. Table 1 summarizes this shift in assumptions and highlights why the relevant unit of failure and the relevant unit of assurance move from individual agents to coordination invariants that govern collective intent and safety.

Core Vulnerability of Distributed Autonomy

Distributed autonomous systems operate without shared global situational awareness. Each agent acts on locally available observations, internal state, and limited coordination signals exchanged with nearby teammates or supervisory elements (Yin et al. 2018; Zhang et al. 2020; Li et al. 2024). Coherent team behavior, therefore, does not arise from reconstructing a global state but from the compatibility of local decisions mediated through coordination mechanisms such as task assignments, role constraints, deconfliction rules, timing expectations, and implicit or explicit commitments governing agent interactions. When these mechanisms remain mutually consistent, independently acting agents compose safe and effective collective behavior even under partial observability and incomplete information.

Coordination failure does not require component malfunction, unstable control, or incorrect local inference. Individual agents may continue to operate nominally with respect to their own observations, policies, and constraints, while collective behavior becomes unsafe or ineffective due to incompatibilities across coordination interfaces. Small discrepancies in task ownership, spatial constraints, or assumptions about teammate behavior can propagate through coordination logic and feedback, producing duplicated effort, coverage gaps, unsafe proximity, or oscillatory decision-making without a clear point of fault.

Distributed autonomy is designed to tolerate stochastic uncertainty, ambiguity, and transient inconsistency through local reasoning and loosely coupled coordination rather than strict global agreement. When uncertainty is *strategically induced* rather than exogenous, however, this tolerance can be exploited by selectively shaping information in ways that remain locally plausible, gradually driving local assumptions into mutual incompatibility without triggering overt faults or alarms.

In such settings, decentralization shifts the dominant mode of failure from isolated agent errors to loss of coherent collective behavior. Safety and mission effectiveness no longer depend solely on per-agent correctness, but on maintaining compatible coordination assumptions under contested information. As summarized in Table 1, intelligent, coordinated adversaries redefine the assurance problem, requiring guarantees about collective behavior over time under complex adversarial conditions rather than accurate global state reconstruction or centralized control.

Coordinated Manipulation of Distributed Autonomy

The structural vulnerability described above becomes consequential when uncertainty is no longer incidental but is instead shaped by an intelligent and adaptive adversary. In such settings, failure does arise from how small, structured influences propagate through the autonomy process over time. This propagation exploits the same properties that enable distributed systems to function under partial information.

Distributed autonomy relies on a sequence of transformations: observations inform internal representations of the environment and other agents; these representations guide action selection and interaction; executed actions generate outcomes that influence subsequent inference and adaptation. Coherent collective behavior emerges when these transformations remain mutually compatible across agents. When they diverge, collective behavior degrades, even if each agent remains locally well behaved.

Perception and State Inference. Agents form beliefs based on limited, noisy, and heterogeneous observations. An intelligent adversary can bias, suppress, or distort portions of this information in ways that remain statistically plausible, causing gradual divergence between the actual environment and internally represented states. Because inference is inherently local, such divergence need not be uniform across agents.

Information Exchange. Distributed systems depend on partial and intermittent sharing of context rather than complete state synchronization. This creates sensitivity to selective absence, delay, or distortion of shared information. Even when communication is sporadic by design, differences in what agents observe or exchange can accumulate over time, leading to incompatible internal representations without violating local consistency.

Decision-Making and Interaction. Actions are selected based on representations that need not be identical across agents. Small differences in these representations can yield qualitatively different actions, particularly near decision boundaries. As a result, locally reasonable actions can combine to produce globally inconsistent or unsafe outcomes.

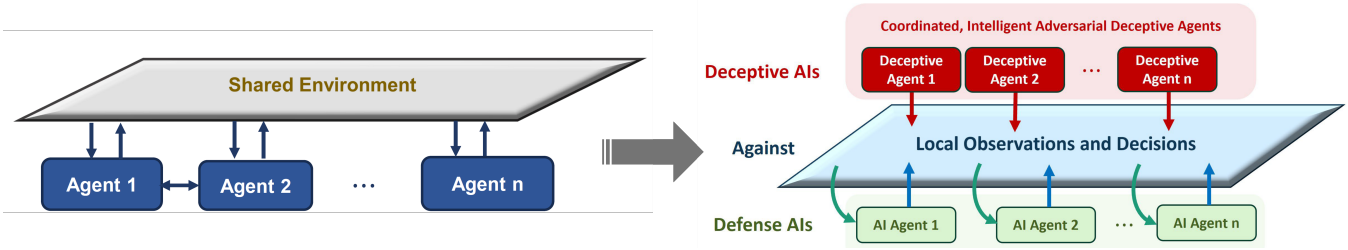

Figure 1: Conceptual transition from conventional distributed autonomy to AI-against-AI conflict. In benign settings (left), decentralized agents coordinate under partial observability and non-strategic uncertainty through interaction with a shared environment. In contested settings (right), small, coordinated, and often undetected adversarial influence shapes observations, communication, and perceived outcomes, inducing gradual divergence between the true environment state and the team-perceived state that governs coordination and execution.

Aspect	Conventional Distributed Autonomy	AI-Against-AI Autonomy
Source of uncertainty	Sensor noise, delays, partial observability	Strategic, adaptive, and coordinated influence by AI-enabled adversaries
Failure model	Independent or localized component faults	Cascading miscoordination arising from small, coordinated perturbations
Role of decentralization	Improves resilience through redundancy and local recovery	Increases reliance on shared coordination commitments and belief alignment
Primary vulnerability	Individual component degradation or instability	Loss of coordination integrity across the team
Information assumptions	Observations and messages are noisy but benign	Observations, messages, and outcomes may be strategically misleading yet locally plausible
Temporal structure	Disruptions are transient, stationary, or statistically independent	Influence is sequential, accumulative, and timed at coordination-critical moments
Adaptation behavior	Learning mitigates uncertainty and improves performance	Learning internalizes biased evidence and amplifies coordination drift
Unit of assurance	Per-agent stability, performance, and fault tolerance	Team-level coordination invariants and safe collective behavior
Safety risk	Gradual mission degradation with identifiable faults	Abrupt loss of team coherence without overt component failure

Table 1. Conceptual shift from conventional distributed autonomy to AI-against-AI autonomy.

Learning and Adaptation. Feedback from past actions is used to update internal models or policies. When this feedback is systematically biased, learning processes may reinforce divergence. Over time, adaptation amplifies early inconsistencies, transforming minor distortions into persistent collective misalignment.

Across all of these layers, the defining characteristic of coordinated manipulation is not the magnitude of the disturbance, but its alignment with the autonomy system's internal abstractions and temporal evolution. Such influence need not be confined to information disruption alone but may also arise from physical, sensor-level, or hardware-induced effects that subtly alter observations or system dynamics while remaining locally plausible. In all cases, small, sparse perturbations can drive the system toward incompatible collective behavior, placing the team at risk without triggering obvious faults or violations.

AI-Against-AI Conflict as an Adaptive Regime

The emergence of AI-against-AI conflict marks a shift from static notions of robustness toward an adaptive and co-evolving regime of autonomy. As learning-based methods are increasingly used to improve the robustness, efficiency, and resilience of distributed systems, they simultaneously expand the space of strategies available to intelligent adversaries. An opponent equipped with AI can observe behavior over time, probe responses, and adapt deception strategies to exploit not only system vulnerabilities but also the system's attempts to correct them.

In this regime, adversarial influence is not fixed or predefined. Instead, it evolves in response to defensive adaptation, giving rise to a feedback loop in which both sides learn, infer, and adjust. Improvements in perception, communication, planning, or adaptation on one side can induce corresponding advances in adversarial manipulation on the other, including the discovery of new leverage points, new modes of influence, and new pathways for low-footprint intervention. As a result, safety and effectiveness can no longer be evaluated under static threat models.

This interaction increasingly resembles a strategic game over information rather than a contest of disturbance rejection. Advantage accrues to the side that can better shape,

conceal, or exploit informational asymmetries over time. Small, well-aligned influences—whether informational or physical in origin—can yield disproportionate effects when they alter the opponent's internal representations or adaptive trajectory. In such settings, dominance is determined less by the magnitude of intervention than by the ability to remain unpredictable, plausible, and informative about the opponent's decision process.

These dynamics motivate the need for new foundations for distributed autonomy that explicitly account for adaptive, learning-enabled opposition. Rather than asking whether a system is robust to a fixed class of disturbances, future studies must address how collective behavior evolves under sustained strategic interaction, where both autonomy and deception co-adapt over time. This reframes the study of distributed autonomy as an inherently game-theoretic and information-centric problem, with implications for assurance, evaluation, and deployment in contested environments.

Acknowledgments

The authors acknowledge the support of the National Science Foundation award IIS-2311969, Office of Naval Research award N00014-23-1-2850, Army Research Laboratory awards W911NF-23-2-0207 and W911NF-23-20175, and Army Research Office award W911NF-24-2-0166, and the Defense Advanced Research Projects Agency (DARPA) under grant HR00112420366.

References

Alali, M.; Kazeminajafabadi, A.; and Imani, M. 2024. Deep reinforcement learning sensor scheduling for effective monitoring of dynamical systems. *Systems Science & Control Engineering*, 12(1): 2329260.

Bukharin, A.; Li, Y.; Yu, Y.; Zhang, Q.; Chen, Z.; Zuo, S.; Zhang, C.; Zhang, S.; and Zhao, T. 2023. Robust multi-agent reinforcement learning via adversarial regularization: Theoretical foundation and stable algorithms. *Advances in neural information processing systems*, 36: 68121–68133.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. Ieee.

Gleave, A.; Dennis, M.; Wild, C.; Kant, N.; Levine, S.; and Russell, S. 2020. Adversarial Policies: Attacking Deep Reinforcement Learning. In *International Conference on Learning Representations (ICLR)*.

Imani, M.; Lan, T.; and Bastian, N. D. 2025. Multi-Agent Cyber Defense with Multi-Level Zero-Trust Actions. In *Proceedings of the IEEE Military Communications Conference (MILCOM)*.

Jiang, G.; Imani, M.; Bastian, N. D.; and Lan, T. 2025. Agentic AI for Cyber Defense: LLM-Guided Hierarchical Multi-Agent Reinforcement Learning. In *Proceedings of the IEEE Military Communications Conference (MILCOM)*.

Kazeminajafabadi, A.; Aksaray, D.; and Imani, M. 2025. Defense Policy Optimization with Linear Temporal Logic Specifications for Interconnected Networks. In *AIAA SciTech 2025 Forum*, 2713.

KazemiNajafabadi, A.; Everett, M.; Lan, T.; and Imani, M. 2025. Adversarial Decoy Placement for Strategic State Perturbations in Artificial Intelligence Driven Defense. In *Proceedings of the 64th IEEE Conference on Decision and Control (CDC)*.

Kazeminajafabadi, A.; and Imani, M. 2024. Optimal Joint Defense and Monitoring for Networks Security under Uncertainty: A POMDP-Based Approach. *IET Information Security*, 2024(1): 7966713.

Kazeminajafabadi, A.; and Imani, M. 2025. Robust Defense Strategy for Network Security Against Unknown Attack Models. In *AIAA SCITECH 2025 Forum*, 2712.

Kazeminajafabadi, A.; Lan, T.; and Imani, M. 2025. Game-Theoretic Defense Policy for Network Security Against Intelligent Adversary. In *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*, 2628–2635. IEEE.

Li, S.; Guo, J.; Xiu, J.; Xu, R.; Yu, X.; Wang, J.; Liu, A.; Yang, Y.; and Liu, X. 2024. Byzantine Robust Cooperative Multi-Agent Reinforcement Learning as a Bayesian Game. In *International Conference on Learning Representations (ICLR)*.

Lin, Y.-C.; Hong, Z.-W.; Liao, Y.-H.; Shih, M.-L.; Liu, M.-Y.; and Sun, M. 2017. Tactics of adversarial attack on deep reinforcement learning agents. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 3756–3762.

Park, P. S.; Goldstein, S.; O'Gara, A.; Chen, M.; and Hendrycks, D. 2024. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5).

Pinto, L.; Davidson, J.; Sukthankar, R.; and Gupta, A. 2017. Robust adversarial reinforcement learning. In *International conference on machine learning*, 2817–2826. PMLR.

Price, A.; Pereira, R. F.; Masters, P.; and Vered, M. 2023. Domain-independent deceptive planning. In *The 22nd International Conference on Autonomous Agents and Multiagent Systems*, 95–103.

Rakhsha, A.; Radanovic, G.; Devidze, R.; Zhu, X.; and Singla, A. 2020. Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. In *International Conference on Machine Learning*, 7974–7984. PMLR.

Savas, Y.; Verginis, C. K.; and Topcu, U. 2022. Deceptive decision-making under uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 5332–5340.

Standen, M.; Kim, J.; and Szabo, C. 2025. Adversarial machine learning attacks and defences in multi-agent reinforcement learning. *ACM Computing Surveys*, 57(5): 1–35.

Sun, Y.; Zheng, R.; Hassanzadeh, P.; Liang, Y.; Feizi, S.; Ganesh, S.; and Huang, F. 2022. Certifiably robust policy learning against adversarial multi-agent communication. In *The eleventh international conference on learning representations*.

Tu, J.; Wang, T.; Wang, J.; Manivasagam, S.; Ren, M.; and Urtasun, R. 2021. Adversarial attacks on multi-agent communication. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7768–7777.

Wang, Y.; Wu, P.; and Imani, M. 2025. Federated Posterior Sharing for Multi-Agent Systems in Uncertain Environments. *Proceedings of Machine Learning Research vol*, 283: 1–13.

Wu, P.; Kamara, A.; Imani, M.; Bastian, N. D.; and Imani, M. 2025. Universal Adversarial Perturbations for Two-Stage Black-Box Object Detectors. In *Proceedings of the 59th Annual Asilomar Conference on Signals, Systems, and Computers, IEEE*.

Yin, D.; Chen, Y.; Kannan, R.; and Bartlett, P. 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning*, 5650–5659. Pmlr.

Zhang, K.; Sun, T.; Tao, Y.; Genc, S.; Mallya, S.; and Basar, T. 2020. Robust multi-agent reinforcement learning with model uncertainty. *Advances in neural information processing systems*, 33: 10571–10583.