From Fault Detection to Agency: A Framework for Industrial Autonomy

Doctoral Consortium

Dhiraj Neupane
Deakin University, Geelong (Australia)
d.neupane@deakin.edu.au

ABSTRACT

Developing autonomous agents for industrial maintenance requires bridging the gap between passive fault detection and active, intelligent decision-making. Current approaches mostly treat machinery fault detection (MFD) as a static classification task, ignoring the sequential nature of machine degradation and the structural complexity of multi-sensor environments. My research proposes a unified *perception-action framework* for autonomous industrial agents. To address the challenge of defining healthy behavior without fault labels, I formulate MFD as an offline inverse reinforcement learning problem. By employing *adversarial inverse reinforcement learning*, the agent recovers robust reward functions from expert demonstrations to detect anomalies. Furthermore, to address partial observability in multi-sensor settings, I propose a spatiotemporal state estimation module that leverages *structured relational modeling* and *selective state space architectures* to capture high-order sensor correlations and long-term degradation dynamics. Together, these components provide the normative reasoning and structured perception required for autonomous maintenance agents.

KEYWORDS

Industrial Agents; Anomaly Detection; Inverse Reinforcement Learning; State Space Models; Sensor Fusion

ACM Reference Format:
Dhiraj Neupane. 2026. From Fault Detection to Agency: A Framework for Industrial Autonomy: Doctoral Consortium. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/ORJT7684

1 INTRODUCTION

The transition to Industry 4.0 requires a paradigm shift in industrial cyber-physical systems, i.e., moving from automated monitoring tools to *autonomous maintenance agents*. While current deep learning approaches excel at pattern recognition, they remain fundamentally passive. They treat fault detection as a static classification task, relying on supervised labels that are inherently scarce in industrial settings [2]. Furthermore, they often struggle to model the physical topology of complex machinery, treating multi-sensor streams as flat feature vectors rather than interconnected components.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/ORJT7684

To achieve true autonomy, an industrial agent must possess two core capabilities: *normative reasoning*, the ability to intrinsically understand what constitutes "health" without constant supervision; and *structured perception*, the ability to estimate the state of complex, coupled components over time. These capabilities are analogous to a *conscience* (knowing what is normal) and a *brain* (perceiving the environment), respectively, though the technical contributions of this work are grounded in inverse reinforcement learning (IRL) and spatiotemporal modeling rather than cognitive analogy.

My thesis proposes a *perception-action framework* for industrial autonomy. I argue that by combining *adversarial inverse reinforcement learning (AIRL)* [5] with *structured state space models* for multi-sensor fusion, we can move beyond simple detection toward fault-aware agency. This framework enables an agent to learn robust reward functions from normal operating data (*defining health*) and capture high-order spatiotemporal correlations (*estimating state*), effectively bridging the gap between signal processing and autonomous agency. The primary focus is on fault detection (identifying the occurrence of faults), though the recovered reward functions also support fault diagnosis (identifying the cause) via an interpretable reward space.

2 RESEARCH QUESTIONS

To formalize this framework, my research addresses two fundamental challenges:

RQ1 (Normative Reasoning): *How can an agent autonomously define "healthy" behavior in the absence of fault labels?*
Standard anomaly detection relies on thresholding fragile reconstruction errors [1], which fluctuate wildly under varying operating conditions. I investigate how IRL can recover a robust *reward function* for health solely from expert (normal) demonstrations, a stable metric that generalizes across domain shifts better than raw signal error.

RQ2 (Structured Perception): *How can an agent perceive spatiotemporal dependencies in complex multi-sensor environments?*
Machinery sensors exhibit non-pairwise correlations (structural) and long-term degradation trends (temporal). I investigate how *structured relational models* (e.g., hypergraph representations) and *selective state space architectures* (e.g., Mamba [6]) can be combined to model these dependencies more efficiently than standard transformers [9] or pairwise graph networks [3].

3 PROGRESS TO DATE

My research follows a structured evolution from benchmarking standard unsupervised baselines to developing advanced agent-centric architectures.

3.1 Phase 1: Benchmarking Baselines

[Addressing the limitations of one-class classification (OCC)]
My initial research [7] conducted a comprehensive evaluation of unsupervised OCC methods, including isolation forests, one-class SVM, autoencoders, and variational autoencoders. These were tested on three major datasets: the Case Western Reserve University and Paderborn University bearing datasets, as well as the HUMS2023 gearbox dataset.

- **Finding:** While effective for simple, static faults, these methods failed to generalize across varying operating conditions. They treat data points as static vectors, lacking the normative reasoning required to distinguish between a *change in operation* (e.g., speed increase or load change) and a *degradation in health*.

3.2 Phase 2: Normative Learning via AIRL

[Addressing RQ1 - Completed]
To address the fragility of reconstruction errors, I developed an *AIRL framework*, presented concurrently as an extended abstract at this conference [8].

- **Method:** Instead of manually defining a threshold, the agent treats normal operation data as *expert trajectories*. It learns a discriminator reward function $D(s, a)$ that assigns high rewards to healthy states and low rewards to anomalies.
- **Contribution:** This gives the agent an intrinsic definition of health. Evaluated on three run-to-failure datasets (IMS, XJTU-SY, and HUMS2023), the model successfully tracked *degradation progression*. Crucially, it demonstrated superior timeliness and robustness. Unlike Phase 1 baselines which suffered from premature false alarms (reacting to operational noise), AIRL accurately aligned with the ground truth onset of degradation, avoiding both hypersensitivity and missed detections.
- **Current Limitation:** This framework relies on a single sensor channel (vibration), yet complex industrial faults *propagate* across multiple physical domains (thermal, acoustic, electrical, and mechanical). Single-modality reliance limits the agent's ability to cross-verify anomalies, leaving it susceptible to sensor-specific noise.

4 PROPOSED METHODOLOGY

To resolve the limitation of single-modality perception and address RQ2, I am developing a spatiotemporal state estimation architecture. This module provides structured perception, fusing multi-physics sensor inputs into a coherent state representation that the normative module (*AIRL*) can evaluate (see Figure 1).

4.1 Structural Perception

Current multi-modal fusion methods often rely on simple feature concatenation, effectively treating the machine as a loose bag of sensors while ignoring its physical topology.

- **Approach:** I investigate structured relational models that explicitly encode multi-sensor topology. One candidate is *hypergraph* modeling [4], where a single hyperedge connects multiple heterogeneous nodes (e.g., vibration, current, and thermal sensors on the same drive shaft), capturing *one-to-many* physical couplings. I am also exploring *causal relational architectures* that model directed dependencies between sensor channels, enabling the agent to distinguish correlated sensor responses from genuine fault propagation pathways.

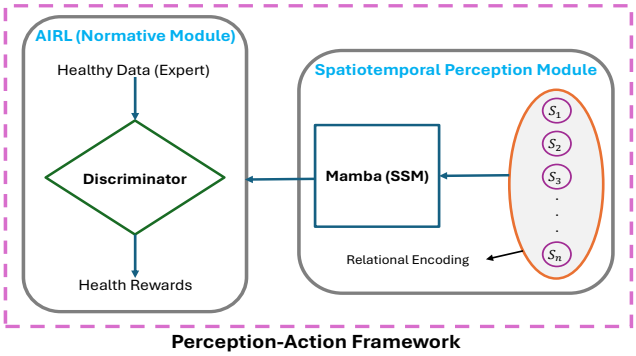

Figure 1: The Perception-Action Framework. The normative module (AIRL) learns health rewards from expert data (*Left*). The perception module estimates state from multi-physics sensors via structured relational and state space models (*Right*).

4.2 Temporal Perception

As observed in Phase 2, incipient faults evolve slowly over thousands of cycles. Capturing this *health trajectory* requires modeling extremely long sequences, yet transformers' quadratic complexity $O(L^2)$ is prohibitive for on-board industrial agents.

- **Approach:** I utilize selective state space models, specifically the *Mamba* architecture [6], which provides long-range reasoning with *linear complexity* $O(L)$. Ongoing work focuses on incorporating *causal structure* into state transitions, ensuring the temporal model respects directed degradation dynamics. This enables real-time tracking on edge hardware, satisfying the efficiency requirement of industrial autonomy.

4.3 Synthesis: The Perception-Action Loop

The final stage of my PhD focuses on integrating these modules into a unified architecture. The spatiotemporal perception module estimates the latent state trajectory $s_{1:t}$. The AIRL module then acts as the normative critic, computing the reward $R(s_{1:t})$ to assess the health of this trajectory. This closes the loop, creating an agent that perceives structure and evaluates health autonomously.

5 GOALS FOR THE DOCTORAL CONSORTIUM

I am in the final year of my candidature (completion expected October 2026). The AIRL-based normative module (Phase 2) is complete and published, and the spatiotemporal perception module (Phase 3) is under active development. I seek feedback from the AAMAS community on the positioning of this framework within the autonomous agents literature, and on design trade-offs in coupling the learned reward function with the state estimator. I also welcome career guidance as I transition to a postdoctoral role. The DC's peer interaction and mentorship would be valuable for refining both the scope of this integration and my broader research trajectory.

REFERENCES

[1] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. 2020. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 3395–3404.

[2] Zhiqiang Bao, Changfu Liu, Hui Yang, Jiayao Zhang, and Yuqi Li. 2026. From theory to industry: A survey of deep learning-enabled bearing fault diagnosis in complex environments. *Engineering Applications of Artificial Intelligence* 163 (2026), 113068.

[3] Gabriele Corso, Hannes Stark, Stefanie Jegelka, Tommi Jaakkola, and Regina Barzilay. 2024. Graph neural networks. *Nature Reviews Methods Primers* 4, 1 (2024), 17.

[4] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 3558–3565.

[5] Justin Fu, Katie Luo, and Sergey Levine. 2018. Learning Robust Rewards with Adversarial Inverse Reinforcement Learning. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=rkHywl-A-

[6] Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*.

[7] Dhiraj Neupane, Mohamed Reda Bouadjenek, Richard Dazeley, and Sunil Aryal. 2024. A Comparative Study of Semi-Supervised Anomaly Detection Methods for Machine Fault Detection. In *PHM Society European Conference*, Vol. 8. 10–10.

[8] Dhiraj Neupane, Richard Dazeley, Mohamed Reda Bouadjenek, and Sunil Aryal. 2026. Learning Rewards, Not Labels: Adversarial Inverse Reinforcement Learning for Machinery Fault Detection. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. https://doi.org/10.65109/AXYX4522

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).