Resolving Reward Misspecification in Deep Reinforcement Learning Agents Deployed Across Non-Stationary Industrial Control Environments

Pat Davis
Professor
Korea Advanced Institute of Science and Technology (KAIST)
291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

Adrian Rodriguez
Associate Professor
Technical University of Munich (TUM)
Arcisstraße 21, 80333 Munich, Bavaria, Germany

Avery Garcia
PhD
University of Waterloo
200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada

Skyler Evans
D.Sc
Nanyang Technological University (NTU)
50 Nanyang Avenue, Singapore 639798, Republic of Singapore

Abstract. Reward misspecification remains a critical failure mode for deep reinforcement learning (DRL) agents transitioning from simulated training pipelines to non-stationary industrial control environments, where dynamic process shifts invalidate handcrafted reward shaping assumptions. This paper proposes an Adaptive Reward Correction Network (ARCN), a meta-learning augmented framework that continuously re-calibrates reward signals via latent environment embeddings derived from variational autoencoders. ARCN integrates a distributional critic architecture with uncertainty-aware policy gradient updates, enabling robust agent behavior under covariate shift and partial observability. Empirical evaluations conducted on three benchmark industrial control testbeds—chemical reactor regulation, autonomous grid load balancing, and multi-axis robotic assembly—demonstrate a 31.4% reduction in cumulative constraint violations and a 22.7% improvement in long-horizon task completion relative to state-of-the-art baselines. Ablation studies confirm the necessity of each architectural component.

Keywords: deep reinforcement learning, reward misspecification, non-stationary environments, meta-learning reward correction, distributional reinforcement learning, variational autoencoder latent embedding, industrial process control, covariate shift robustness, uncertainty-aware policy optimization

Introduction

The deployment of autonomous decision-making agents in safety-critical industrial systems represents one of the most consequential frontiers of artificial intelligence engineering as of 2024–2026. Deep reinforcement learning has demonstrated remarkable proficiency in controlled simulation environments; however, the translation of trained policies to real-world industrial process control consistently exposes a fundamental vulnerability: reward misspecification. When the reward function—hand-engineered to encode domain-expert objectives—encounters the statistical non-stationarity inherent to live industrial processes, the agent's behavioral policy degrades rapidly, often producing constraint-violating or economically catastrophic decisions. This problem is not merely theoretical. Industrial automation deployments in chemical processing, smart grid management, and flexible robotic manufacturing routinely exhibit distributional shifts driven by equipment aging, feedstock variability, seasonal demand fluctuations, and unforeseen fault conditions. Under such shifts, a reward signal that accurately proxied the true objective during training may actively misdirect agent behavior during deployment, a phenomenon formally characterized as Goodhart's Law in the context of learned proxies. Existing mitigation strategies fall into three broad categories: reward learning from human feedback, constrained policy optimization, and domain randomization during training. While each addresses a subset of the misspecification problem, none provides a principled mechanism for online reward recalibration conditioned on latent representations of the current environmental regime. Reward learning methods require costly human annotation during deployment; constrained optimization methods assume reward structure validity within predefined constraint envelopes; and domain randomization, although effective for sim-to-real transfer in robotics, cannot anticipate the full distribution of real industrial disturbances without prohibitive computational overhead. This paper addresses the gap by proposing the Adaptive Reward Correction Network (ARCN), a unified meta-learning framework that treats the reward function as a dynamically adjustable module rather than a static design artifact. ARCN leverages variational autoencoder-based latent environment embeddings, extracted from short observation windows, to condition a reward correction head that modulates the agent's instantaneous reward signal in alignment with the inferred environmental regime. The corrected reward is then consumed by a distributional soft actor-critic backbone equipped with epistemic uncertainty quantification, ensuring that policy gradient updates remain conservative under high-ambiguity conditions. The remainder of this paper is organized as follows. Section 2 reviews related literature on reward shaping, meta-reinforcement learning, and non-stationary control. Section 3 formally defines the reward misspecification problem under Markov decision process non-stationarity. Section 4 details the ARCN architecture, training protocol, and theoretical convergence properties. Section 5 describes the three industrial benchmark environments and evaluation methodology. Section 6 presents quantitative results, ablation studies, and

qualitative policy analysis. Section 7 discusses practical deployment considerations, limitations, and directions for future work, followed by concluding remarks in Section 8.

References

1. Semeniuk, V. V. (2025). OPTIMIZATION OF LOCAL DEVELOPMENT PROCESS USING DOCKER PHP IMAGE THAT COMES WITH A FULL SET OF TOOLS OUT OF THE BOX–PERFORMANCE AND OPTIMIZATION EXTENSIONS. ІНФОРМАЦІЙНЕ ЗАБЕЗПЕЧЕННЯ БАГАТОІНДЕКСНОЇ ТРАНСПОРТНОЇ ЗАДАЧІ З НЕЧІТКИМИ ІНТЕРВАЛАМИ.
2. Semeniuk, V. V. (2025). OPTIMIZATION OF LOCAL DEVELOPMENT PROCESS USING DOCKER PHP IMAGE THAT COMES WITH A FULL SET OF TOOLS OUT OF THE BOX: DATABASE AND INTERNATIONALIZATION EXTENSIONS. ВЧЕНІ ЗАПИСКИ, 12025226.