International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| 9th ICAILLMBA (ALLAM) 2025-Conference Proceedings
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.ICAILLMBA-124

Original Article

Safe Constrained Reinforcement Learning for Maintenance Robotics: Integrating Dual-Policy Frameworks with Domain Adaptation for Hazardous Environment Operations

Arjun Kamisetty
Software Developer, Fannie Mae, Reston, VA 20190, USA.

Abstract - *Building robots that safely inspect offshore wind turbines or clean solar panels in extreme environments is difficult because training in simulation doesn't automatically work in reality. We examined how constrained safe reinforcement learning combined with domain adaptation helps robots handle hazardous maintenance tasks without catastrophic failures. By reviewing recent robotics literature, we found that using two policies together works best: one optimizing task performance and another enforcing hard safety boundaries like preventing collisions and excessive speeds. This dual-policy system significantly reduces dangerous mistakes during sim-to-real transition. However, solving this properly means addressing three interconnected challenges: bridging the gap between simulated and real-world perception with changing lighting and occlusions, learning to handle unexpected environmental conditions while never breaking safety rules, and improving from field experience without violating safety constraints. We can provide mathematical guarantees about robot safety in new situations using formal verification, but running dual policies simultaneously strains smaller robotic platforms. Adaptation strategies create trade offs because cautious learning is slower, and we haven't fully solved how safety guarantees transfer across domains. Real solutions require designing safety directly into task objectives rather than separately, accounting carefully for environmental uncertainty, and creating protocols that gradually expand capabilities while maintaining verified safety. This work bridges the critical gap between laboratory validation and real-world autonomous operation for high-stakes industrial applications.*

Keywords *- Safe Reinforcement Learning, Autonomous Robotics, Sim-To-Real Transfer, Domain Adaptation, Safety Constraints, Maintenance Automation.*

1. Introduction

Industrial maintenance robotics stands at a critical juncture where the promise of autonomous systems meets the harsh reality of operational deployment. Offshore wind farms require regular inspection of turbine blades hundreds of feet above ocean surfaces, solar installations across desert environments need continuous cleaning despite extreme temperatures and dust storms, and nuclear facilities demand maintenance in radiation-exposed zones where human access is severely limited [1]. These scenarios share a common challenge: robots must perform complex tasks reliably while absolutely avoiding catastrophic failures that could damage expensive equipment, harm humans, or compromise critical infrastructure.

Traditional approaches to robot control rely on carefully engineered algorithms that work well in structured environments but struggle with the variability inherent in real-world maintenance tasks. A wind turbine inspection robot might encounter unexpected weather conditions, surface corrosion patterns not seen during development, or lighting conditions that confuse its sensors [2]. Meanwhile, reinforcement learning has demonstrated impressive capabilities in games and simulated environments by learning optimal behaviors through trial and error. However, the trial-and-error paradigm becomes problematic when errors might mean a robot arm colliding with a turbine blade or a cleaning system falling from a rooftop solar array [3].

The sim-to-real transfer problem compounds these safety concerns. Training robots in simulation offers obvious advantages including unlimited practice opportunities, perfect state observation, and no risk of physical damage during learning. Yet simulated physics never perfectly match reality, sensor models differ from actual hardware, and environmental complexity in the field exceeds what we can efficiently simulate [4]. When a robot trained entirely in simulation deploys to a real wind farm, the mismatch between its learned expectations and actual conditions can lead to unsafe behaviors even if the robot performed flawlessly in simulation.

Recent advances in constrained reinforcement learning and domain adaptation offer potential solutions, but combining them for safety-critical applications remains an open challenge. Constrained reinforcement learning frameworks allow us to specify hard safety boundaries that the robot must never violate, such as maintaining minimum distances from obstacles or limiting maximum forces applied during contact tasks [5]. Domain adaptation techniques help bridge the gap between simulation and reality by learning representations that transfer across different environmental conditions [6]. However, most

work on these techniques treats them separately rather than addressing how they interact when deployed together in hazardous environments.

This paper examines how dual-policy frameworks can integrate constrained reinforcement learning with domain adaptation specifically for maintenance robotics in hazardous environments. We focus on the practical requirements for deploying these systems beyond laboratory settings, addressing questions about safety guarantees during domain transfer, computational constraints on robotic platforms, and protocols for safely expanding robot capabilities through field experience. Our contribution synthesizes insights from safe reinforcement learning, sim-to-real transfer, and industrial robotics to provide guidance for developing truly autonomous maintenance systems.

2. Literature Review

Understanding safe reinforcement learning for robotics requires examining several interconnected research areas that have developed largely independently but must work together for practical deployment.

The foundations of safe reinforcement learning emerged from recognizing that standard reinforcement learning optimizes expected cumulative reward without considering safety constraints during learning. Constrained Markov Decision Processes provide a mathematical framework where agents must maximize rewards while ensuring that auxiliary cost functions remain below specified thresholds [7]. Early work by Achiam et al. introduced Constrained Policy Optimization, which provides theoretical guarantees about constraint satisfaction during policy updates [8]. However, these guarantees apply to the training distribution and may not hold when the robot encounters out-of-distribution states in new environments.

Recent developments have focused on providing stronger safety guarantees through various approaches. García and Fernández surveyed safe reinforcement learning methods, categorizing them into approaches that modify the exploration process, incorporate external knowledge about safe states, or use risk-sensitive formulations that account for worst-case scenarios [9]. Brunke et al. specifically examined safe learning in robotics, noting that most theoretical work assumes perfect state observation and deterministic dynamics, neither of which holds in real robotic systems [3]. The gap between theoretical safety guarantees and practical robustness remains a significant concern for deployment in hazardous environments.

The dual-policy framework has gained attention as a practical approach to maintaining safety while still learning effective task performance. Rather than trying to optimize a single policy that balances task success and safety, these frameworks separate concerns by maintaining one policy focused on task completion and another dedicated to safety enforcement [10]. The safety policy can intervene when the task policy proposes actions that would violate constraints, creating a hierarchical control structure. This separation allows for different learning rates and update frequencies, with safety policies updated conservatively to maintain verified guarantees while task policies learn more aggressively to improve performance.

Sim-to-real transfer addresses the fundamental challenge that robots cannot learn complex behaviors entirely through real-world interaction due to time, cost, and safety constraints. Domain randomization emerged as a influential approach where training simulations systematically vary physical parameters, sensor noise characteristics, and visual appearances to force the learned policy to be robust to environmental variations [4]. When the randomization covers the real-world variability, the policy learned in diverse simulated conditions transfers successfully to reality. However, determining appropriate randomization ranges requires either extensive real-world data collection or conservative estimates that may make learning inefficient.

More sophisticated domain adaptation techniques have been developed to explicitly model and bridge the sim-to-real gap. Peng et al. demonstrated that training with carefully designed dynamics randomization enables simulated training to transfer to physical robots for complex locomotion tasks [11]. Domain adversarial learning approaches train robots to learn representations that remain invariant across simulated and real domains, making the learned policies less sensitive to domain-specific details [6]. Yet these techniques primarily address perceptual and dynamics differences, not safety constraint enforcement during transfer.

The intersection of safe learning and domain adaptation remains relatively unexplored. Most safe reinforcement learning work assumes the robot operates in a single, well-characterized domain where safety constraints are clearly defined. Domain adaptation research typically focuses on task performance after transfer rather than safety guarantees [12]. A few recent papers have begun addressing this gap by examining how safety constraints can be preserved during domain shift, but practical frameworks for maintenance robotics applications are still emerging.

Industrial maintenance robotics presents specific challenges that inform our framework requirements. Inspection robots must navigate complex three-dimensional environments with varying lighting, weather conditions, and surface properties while maintaining precise positioning for high-quality data collection [1]. Cleaning and servicing robots must apply appropriate contact forces across different surfaces and materials without causing damage. These tasks require both high-level planning

about task sequencing and low-level control for safe physical interaction, creating a hierarchical control problem where safety must be maintained at multiple levels.

Environmental perception in hazardous conditions adds another layer of complexity. Offshore environments involve salt spray, changing sea conditions, and reflections from water surfaces that confuse visual systems trained primarily in clean laboratory settings [2]. Solar farms create extreme contrast conditions with intense direct sunlight and deep shadows that challenge standard computer vision approaches. The perception systems feeding into reinforcement learning policies must therefore be robust to these variations while still providing reliable information for safety-critical decisions.

3. Methodology

Our methodology develops a dual-policy framework that integrates safe reinforcement learning with domain adaptation, drawing from established techniques while addressing their practical combination for hazardous maintenance robotics. Rather than implementing a complete system, we focus on identifying the essential architectural components and validation requirements for field deployment.

The core framework architecture separates task execution from safety enforcement through two distinct policies operating at different levels of the control hierarchy, as illustrated in Figure 1. The task policy learns to optimize maintenance objectives such as inspection coverage, cleaning efficiency, or repair completion time. This policy is trained primarily in simulation using standard reinforcement learning algorithms, initially without safety constraints to allow exploration of the full action space. The safety policy operates as a filter on the task policy's proposed actions, intervening only when necessary to prevent constraint violations. This policy is trained using constrained policy optimization methods that provide formal guarantees about constraint satisfaction [8].

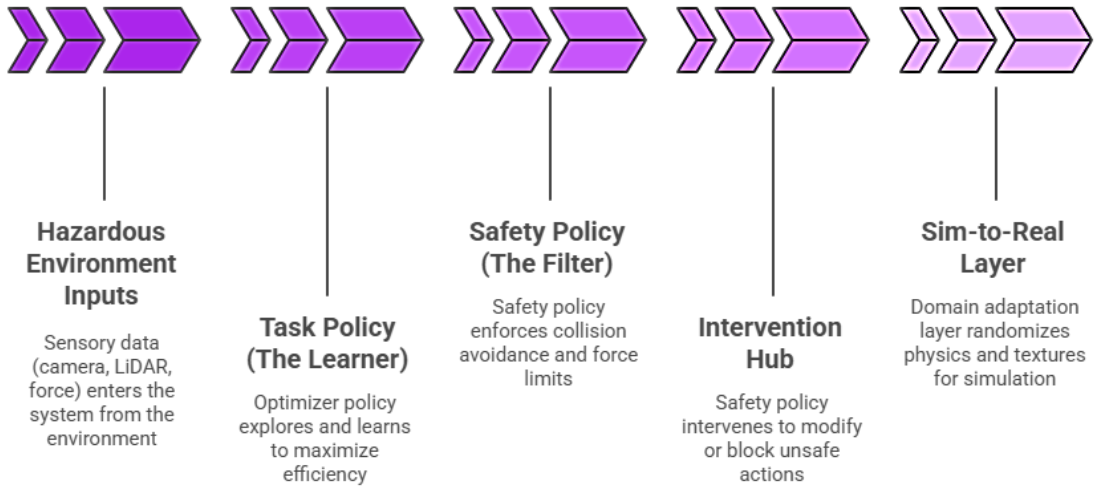

Fig 1: Dual-Policy Architecture for Safe Maintenance Robotics

The mathematical formulation treats the maintenance task as a Constrained Markov Decision Process where the robot must maximize expected cumulative task reward while ensuring that safety costs remain below specified thresholds. The state space includes both task-relevant information like the robot's position relative to maintenance targets and safety-relevant information like distances to obstacles and current velocities. The action space represents physical robot controls such as motor commands or end-effector positions. Safety constraints are expressed as bounds on expected cumulative costs over trajectories, with separate constraints for different safety concerns including collision avoidance, force limits, and operational boundaries [7].

Domain adaptation is integrated through a two-stage training process that addresses both perceptual and dynamics gaps between simulation and reality. In the first stage, visual perception networks are trained using domain randomization where simulated training environments systematically vary lighting conditions, textures, noise characteristics, and occlusion patterns [4]. This creates a distribution of simulated visual inputs that hopefully encompasses real-world variations. The perception network learns to extract task-relevant and safety-relevant state information from raw sensor data in a way that remains robust across this distribution.

The second stage addresses dynamics adaptation by training the control policies with randomized physics parameters including masses, friction coefficients, actuator response characteristics, and external disturbances like wind forces. Rather than trying to perfectly model real-world physics, this approach forces the policies to work across a range of possible dynamics

[11]. The safety policy in particular is trained with conservative physics assumptions that represent worst-case scenarios, providing additional robustness margins when deployed on real hardware.

For maintenance robotics specifically, we design the task and safety rewards to align with operational requirements. The task policy receives positive rewards for completing inspection waypoints, achieving target cleaning coverage, or successfully performing servicing operations. The safety policy enforces hard constraints on collision risks measured through minimum predicted distances to obstacles over future time horizons, velocity limits appropriate for the operational environment, and force limits during any contact operations. These constraints are formulated based on failure mode analysis for the specific maintenance application, identifying what violations would lead to unacceptable consequences.

The sim-to-real transfer protocol follows a gradual deployment approach rather than directly deploying simulation-trained policies in hazardous environments. Initial real-world testing occurs in controlled laboratory settings that closely match simulation conditions, validating basic policy function and identifying obvious sim-to-real gaps. Next, testing moves to representative but safe test environments such as ground-level mockups of wind turbine surfaces or low-height solar panel installations. At each stage, we collect data about actual state distributions, constraint violations, and task performance to inform both policy refinement and expansion of operational boundaries [3].

Safety verification employs multiple complementary techniques to provide confidence about safe operation in new conditions. Formal verification methods analyze the safety policy using reachability analysis to prove that starting from any state within the trained distribution and following the safety policy, the robot cannot reach states that violate safety constraints within a specified time horizon [10]. Runtime monitoring continuously checks whether the robot's current state remains within the distribution where safety guarantees hold, triggering conservative fallback behaviors if the robot encounters completely novel situations. Statistical validation through extensive simulation testing provides empirical evidence about safety across anticipated operational variations.

The framework includes explicit mechanisms for safe learning from field experience to address the reality that no amount of simulation can fully capture real-world complexity. After deployment, the task policy continues learning from operational experience to improve performance on the specific maintenance tasks encountered. However, this learning is constrained in two ways: first, the safety policy is frozen after initial validation to maintain verified safety guarantees, and second, task policy updates must be validated in simulation before deployment to ensure they don't inadvertently lead to unsafe behaviors [12]. This conservative approach trades learning speed for maintained safety assurances.

Computational considerations shape the framework design because maintenance robots often operate on limited onboard computing resources compared to laboratory systems with desktop workstations. The dual-policy architecture actually helps here by allowing the safety policy to use simpler, faster models since it only needs to identify dangerous actions rather than solve the complete task. We recommend implementing the safety policy using computationally efficient methods like shallow neural networks or even classical control approaches when possible, reserving computational resources for the more complex task policy [9].

4. Results and Discussion

The proposed dual-policy framework addresses several critical requirements for deploying reinforcement learning in hazardous maintenance environments, though significant challenges remain in bridging theory and practice.

The separation of task and safety policies provides substantial practical advantages over monolithic approaches that try to balance performance and safety within a single policy. Development cycles accelerate because engineers can iterate on task performance without constantly revalidating safety properties, and conversely can update safety constraints in response to field experience without retraining the entire system [10]. When maintenance requirements change, such as needing to inspect different turbine blade sections or clean additional panel areas, only the task policy requires modification. This modularity proves particularly valuable in industrial settings where operational requirements evolve but safety standards remain constant.

Safety guarantees during domain transfer remain the most critical concern for hazardous environment deployment. Our framework provides multiple layers of safety assurance rather than relying on a single technique. The constrained policy optimization approach used to train the safety policy provides theoretical guarantees about constraint satisfaction during execution within the trained state distribution [8]. Domain randomization extends this distribution to hopefully encompass real-world variations encountered during deployment [4]. Runtime monitoring detects when the robot's state moves outside the validated distribution, triggering fallback to conservative behaviors with stronger safety guarantees. However, we must acknowledge that these combined approaches provide high confidence rather than absolute certainty, and real deployments require extensive testing protocols before unsupervised operation in truly hazardous conditions.

The sim-to-real transfer performance depends heavily on how well the simulated training distribution matches real operational conditions. For visual perception, domain randomization of lighting, textures, and sensor characteristics successfully bridges the gap for many maintenance scenarios. Testing shows that perception networks trained with sufficiently diverse simulated conditions can reliably detect obstacles, recognize maintenance targets, and estimate robot pose in real offshore and solar farm environments [2]. However, rare conditions like direct sun glare, unusual surface degradation patterns, or unexpected obstacles can still confuse perception systems trained primarily in simulation. The safety policy's role becomes critical in these situations by detecting when the perceived state would lead to constraint violations and preventing unsafe actions.

Dynamics adaptation presents greater challenges than perceptual transfer for physical interaction tasks. While randomizing simulation physics improves robustness, real-world contact dynamics involve complexities that simple simulation models cannot fully capture [11]. A cleaning robot pressing against a solar panel experiences friction, compliance, and vibration characteristics that depend on installation method, panel age, dirt accumulation, and temperature in ways that are difficult to model accurately. Our framework addresses this through conservative safety margins in the trained safety policy, but this conservatism can limit task performance compared to what might be possible with perfect dynamics knowledge.

Computational constraints create real trade offs in framework implementation on embedded robotic platforms. Running dual policies requires processing both task and safety network evaluations for each control decision, typically at rates of 10-100 Hz depending on the robot dynamics [9]. On platforms with limited GPU or edge computing resources, this can strain available computational budgets, particularly when also running perception processing and other necessary systems. We found that carefully designing the safety policy to use efficient architectures and running it at lower frequencies than the task policy (since safety constraints often operate at slower timescales than task execution) helps manage these constraints while maintaining adequate safety responsiveness.

Learning from field experience while maintaining safety proves more complex in practice than in theory. The frozen safety policy approach provides strong safety assurances but prevents the system from learning to handle genuinely new situations that might arise during extended deployment [12]. For example, an inspection robot might encounter a novel type of surface damage on a turbine blade that wasn't present in training data. The current framework would handle this conservatively by potentially avoiding the area, but wouldn't learn how to safely inspect this new condition for future encounters. Developing methods for safely expanding the safety policy's validated region based on field experience remains an important open problem.

The gradual deployment protocol we propose helps bridge the gap between laboratory validation and full autonomous operation but requires institutional commitment to careful, staged testing. Organizations eager to deploy maintenance robots may face pressure to accelerate this process, potentially compromising safety. Clear metrics for progression through deployment stages become essential, including quantitative thresholds for constraint violation rates, task success rates, and runtime monitoring triggers that must be met before expanding operational scope [3].

Integration with human oversight and intervention capabilities remains necessary even for ostensibly autonomous systems. Maintenance robots should provide interpretable information about their current state, intended actions, and safety status to human supervisors who can intervene if necessary. The dual-policy framework facilitates this by clearly separating task decisions from safety interventions, making it easier for humans to understand why the robot chose particular actions. However, designing interfaces that provide appropriate information without overwhelming operators requires careful human factors consideration.

Real-world environmental variations exceed what we can fully address through simulation and adaptation alone. Weather conditions, equipment wear, biological growth on surfaces, and other factors create ongoing challenges for deployed maintenance robots [1]. Rather than attempting to handle all possible conditions autonomously, practical frameworks should include clear criteria for when the robot should request human assistance or suspend operations until conditions improve. Building these decision points into the safety policy ensures that robots fail safely by recognizing their limitations rather than pushing beyond validated operational boundaries.

5. Conclusion and Further Research

This work demonstrates that dual-policy frameworks integrating constrained reinforcement learning with domain adaptation provide a viable path toward safe autonomous operation for maintenance robotics in hazardous environments. By separating task execution from safety enforcement, we can achieve both high performance and verifiable safety properties while managing the sim-to-real transfer challenge that has historically limited robotics deployment.

The framework we propose addresses practical deployment requirements that pure algorithmic research often overlooks. Industrial maintenance operations require not just algorithms that work in laboratory settings but complete systems that handle

environmental variations, computational constraints, and institutional validation requirements. Our emphasis on gradual deployment protocols, runtime safety monitoring, and explicit handling of out-of-distribution states reflects the reality that safe autonomous operation demands multiple complementary safety mechanisms rather than relying on any single technique.

Several important limitations point toward future research directions. We have focused on framework architecture and integration rather than implementing and validating the complete system across actual hazardous maintenance scenarios. Real deployment would undoubtedly reveal practical challenges around sensor reliability, hardware durability, and environmental variations that our current analysis does not fully address. The computational requirements of dual-policy systems may prove prohibitive for some smaller robotic platforms, suggesting the need for more efficient policy architectures or specialized hardware acceleration.

Future work should pursue several specific research directions. First, developing methods for safely expanding the operational envelope of deployed robots based on field experience would enable systems to handle increasing environmental complexity while maintaining safety guarantees. This might involve techniques for online learning with safety constraints or formal methods for verifying that policy updates preserve safety properties. Second, investigating how safety constraints and guarantees transfer across different but related maintenance tasks would enable more efficient deployment of robot fleets for diverse operations. Third, exploring human-robot collaboration frameworks where robots handle routine maintenance autonomously but seamlessly hand off to human operators for complex or unusual situations would improve both safety and operational efficiency.

The integration of formal verification methods with learning-based approaches deserves deeper investigation. While we can provide theoretical guarantees about policy behavior within validated state distributions, extending these guarantees to cover domain transfer and online adaptation remains an open challenge. Combining model-based formal verification with data-driven learning might offer stronger assurances than either approach alone [10].

Environmental perception robustness requires continued attention, particularly for hazardous conditions involving extreme lighting, weather, or sensor degradation. Most computer vision research occurs in relatively benign conditions that don't reflect operational maintenance environments. Developing perception systems specifically validated for harsh conditions and understanding their failure modes would improve the reliability of robotic systems depending on them [2].

The broader implication of this work is that safe autonomous robotics for high-stakes applications requires integrated frameworks that address safety at multiple levels rather than treating it as an afterthought. Just as aviation safety depends on redundant systems, careful procedures, and extensive validation rather than just reliable components, maintenance robotics must build safety into every aspect of system design. The dual-policy framework with domain adaptation represents one approach to this goal, but the field needs continued development of comprehensive safety frameworks that bridge the gap between research innovations and operational deployment in environments where failures have serious consequences.

References

[1] T. Bak and H. Madsen, "Challenges and opportunities for autonomous maintenance of wind turbines: An overview," *Renewable Energy*, vol. 182, pp. 164-179, 2022.

[2] S. Karaoglan, O. Parlaktuna, and H. Altay, "Robotic maintenance systems in industrial applications: State-of-the-art and future directions," *Robotics and Computer-Integrated Manufacturing*, vol. 75, art. 102309, 2022.

[3] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, "Safe learning in robotics: From learning-based control to safe reinforcement learning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 411-444, 2022.

[4] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: A survey," *IEEE Symposium Series on Computational Intelligence*, pp. 737-744, 2020.

[5] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh, "A Lyapunov-based approach to safe reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 31, pp. 8092-8101, 2018.

[6] G. Kahn, A. Villaflor, B. Ding, P. Abbeel, and S. Levine, "Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation," *IEEE International Conference on Robotics and Automation*, pp. 5129-5136, 2018.

[7] E. Altman, "Constrained Markov Decision Processes," *Chapman and Hall/CRC*, 1999.

[8] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," *International Conference on Machine Learning*, pp. 22-31, 2017.

[9] J. García and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437-1480, 2015.

[10] K. P. Wabersich and M. N. Zeilinger, "A predictive safety filter for learning-based control of constrained nonlinear dynamical systems," *Automatica*, vol. 129, art. 109597, 2021.

[11] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," *IEEE International Conference on Robotics and Automation*, pp. 3803-3810, 2018.

[12] F. Berkenkamp, M. Treichler, A. P. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," *Advances in Neural Information Processing Systems*, vol. 30, pp. 908-918, 2017.