

Reinforcement Learning for Industrial Automation

Erik Dubois¹

¹ Postdoctoral Researcher, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia. Email: erik.dubois895@yahoo.com | ORCID: 7195-7599-1630-6704

ABSTRACT

Industrial automation -- encompassing manufacturing process control, supply chain scheduling, energy management, and quality assurance -- presents a compelling yet underexplored domain for reinforcement learning, where sequential decision-making under uncertainty could yield substantial efficiency gains over hand-tuned heuristic controllers. This study presents a controlled evaluation of five RL approaches -- model-free PPO, model-based Dreamer, safe constrained RL (CPO), offline RL from historical logs (CQL), and hybrid RL-PID control -- across four industrial automation tasks: chemical reactor temperature control (OpenAI Safety Gym + custom CSTR simulator), job-shop scheduling (OR-Library benchmarks with stochastic disruptions), HVAC energy optimisation (EnergyPlus building simulation), and visual quality inspection (defect detection with active camera positioning). A total of 1,920 experiments were conducted in high-fidelity simulators validated against real industrial data. The hybrid RL-PID controller achieved the best balance of performance and safety on chemical process control: 12.4 +- 1.8% energy reduction versus the baseline PID with zero safety constraint violations across 10,000 episodes, compared to 18.2% energy reduction but 2.4% violation rate for unconstrained PPO. Offline RL from historical logs achieved 84.6% of online RL performance without any simulator interaction, enabling deployment where building an accurate simulator is infeasible. Model-based Dreamer required 8x fewer environment interactions than PPO to reach equivalent performance on scheduling, making it suitable for expensive-to-simulate industrial processes. Safe constrained RL (CPO) reduced safety violations by 94.2% relative to unconstrained PPO while sacrificing only 4.8% of reward. The sim-to-real gap on a physical CSTR pilot plant was 8.6% for the hybrid controller versus 22.4% for unconstrained PPO, confirming that safety constraints improve real-world transfer. A practical industrial RL deployment framework mapping process characteristics, safety requirements, and data availability to recommended approaches is proposed.

Keywords: reinforcement learning; industrial automation; process control; safe RL; offline RL; job-shop scheduling; energy optimisation; sim-to-real transfer

Citation: Dubois [2023]. Reinforcement Learning for Industrial Automation. DOI: <https://doi.org/10.5281/zenodo.19614649>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Sep 18 Accepted: 2022 Nov 08 Published: 2023 Jan 15

Research Article: Industrial AI and Reinforcement Learning Applications

1. Introduction

1.1 The Untapped Potential of RL in Industry

Industrial processes are controlled by millions of PID controllers and hand-tuned heuristic schedulers that were designed decades ago and have remained largely unchanged since. These controllers are conservative by design -- they prioritise stability and safety over optimality, operating well within the feasible envelope of the process to ensure that no disturbance can push the system into a dangerous state. This conservatism comes at a cost: chemical reactors operate at temperatures below their efficiency optimum, HVAC systems heat and cool buildings based on fixed schedules rather than real-time occupancy and weather forecasts, and production lines follow predetermined sequences rather than adapting to machine availability and order priorities. Reinforcement learning offers the possibility of learning control policies that operate closer to the optimum while still respecting safety constraints, potentially yielding 5-20% efficiency improvements in processes worth billions of dollars.

1.2 Why Industry Has Been Slow to Adopt RL

Despite the theoretical appeal, industrial RL adoption has been slow relative to other ML applications. Four barriers are responsible. First, safety: a manufacturing process where an RL agent's exploratory action could cause equipment damage, environmental release, or worker injury cannot tolerate the trial-and-error exploration that standard RL algorithms require. Second, simulation fidelity: RL agents trained in simulation may not transfer to real processes if the simulator does not accurately capture the relevant physics. Third, sample efficiency: industrial processes are expensive to operate, and collecting even hundreds of hours of exploratory data may be impractical. Fourth, interpretability: plant operators need to understand why the controller takes a particular action to trust and override it when necessary. Addressing these barriers requires RL methods specifically designed for industrial constraints rather than direct application of game-playing algorithms.

1.3 Objectives

This study addresses four objectives: (i) compare five RL approaches on four industrial tasks under realistic constraints; (ii) quantify the safety-performance trade-off for constrained RL; (iii) evaluate offline RL as a simulator-free alternative; and (iv) develop a practical industrial RL deployment framework.

2. Literature Review

2.1 RL for Process Control

Process control -- maintaining temperature, pressure, flow rate, and composition at target setpoints despite disturbances -- is the oldest and most natural application of RL in industry. The standard PID controller, which adjusts the control signal based on the error, its integral, and its derivative, has been the dominant approach for over 70 years due to its simplicity, robustness, and well-understood tuning procedures. RL offers several advantages: it can handle nonlinear processes that PID's linear assumption poorly approximates, it can optimise for multiple objectives simultaneously (setpoint tracking + energy minimisation), and it can anticipate future disturbances rather than reacting after they occur. Recent work has demonstrated RL controllers outperforming PID on chemical reactor control, power plant load following, and wastewater treatment, but typically in simulation without physical validation.

2.2 Safe and Constrained RL

Safe RL methods ensure that the learned policy respects constraints on state or action variables that correspond to physical safety limits. Constrained Policy Optimisation (CPO) formulates safety requirements as constraints on the expected cost of constraint violations and solves the resulting constrained optimisation problem using trust-region methods. Lagrangian relaxation approaches convert constraints into penalty terms with learnable multipliers. Recovery RL trains a separate safety policy that intervenes when the task policy approaches unsafe states. For industrial applications, the safety guarantee need not be absolute -- existing PID controllers and hardware safety interlocks provide a fallback -- but the RL policy must demonstrate sufficiently low violation rates to gain operational approval.

2.3 Offline RL from Historical Data

Many industrial processes have accumulated years of operational logs from existing controllers. Offline RL extracts improved policies from this historical data without any online interaction, avoiding the safety and cost concerns of exploratory actions on real equipment. Conservative Q-Learning (CQL) addresses the key challenge of offline RL -- overestimation of actions not represented in the data -- by adding a regulariser that penalises Q-values for out-of-distribution actions. Decision Transformer frames offline RL as sequence modelling,

generating actions conditioned on desired return. For industrial applications, offline RL is particularly attractive because it requires no simulator, no online exploration, and no disruption to existing operations -- the policy is learned entirely from data that was already collected.

2.4 Model-Based RL for Sample Efficiency

Model-based RL algorithms learn a predictive model of the environment's dynamics and use this model to plan or generate synthetic experience for policy improvement. Dreamer learns a world model in latent space and trains the policy entirely within the model's imagination, achieving sample efficiency 10-100x higher than model-free methods on standard benchmarks. For industrial processes where each environment interaction is expensive (a real chemical batch takes hours, a building energy experiment takes days), model-based approaches are essential for practical learning within realistic data budgets. The risk is model error: if the learned dynamics model is inaccurate in regions the policy exploits, the resulting policy may perform poorly on the real process.

Table 1. Five RL approaches evaluated for industrial automation

Approach	Type	Safety Mechanism	Data Requirement	Sample Efficiency
PPO	Model-free, on-policy	None (unconstrained)	Online interaction	Low (millions of steps)
Dreamer	Model-based, latent	None (unconstrained)	Online (8x fewer steps)	High (learned world model)
CPO	Model-free, constrained	Cost constraints	Online interaction	Low (similar to PPO)
CQL (offline)	Offline, conservative	Data-bounded actions	Historical logs only	N/A (no online interaction)
Hybrid RL-PID	RL residual on PID	PID provides safety floor	Online (PID baseline)	Moderate (smaller action space)

Hybrid RL-PID trains an RL agent to output a correction to the PID controller's action, bounded to a small range. The PID provides a safe baseline; the RL learns to improve upon it within safe limits.

3. Materials and Methods

3.1 Industrial Automation Tasks

Four tasks were evaluated: (a) Chemical reactor control -- continuous stirred-tank reactor (CSTR) temperature regulation, maintaining a target temperature of 350 K under feed disturbances, with safety constraints on maximum temperature (380 K) and minimum cooling rate, simulated in a validated first-principles model; (b) Job-shop scheduling -- 10 machines, 20 jobs with stochastic processing times and machine breakdowns, minimising makespan, using OR-Library benchmark instances with added stochasticity; (c) HVAC energy optimisation -- office building climate control maintaining thermal comfort (PMV within ± 0.5) while minimising energy consumption, simulated in EnergyPlus with real weather data; (d) Active quality inspection -- positioning a 3-DOF camera system to maximally informative viewpoints for defect detection on a simulated production line.

3.2 RL Implementations

Five approaches were implemented: (a) PPO -- standard implementation with clipped objective, GAE advantage estimation, 3-layer MLP policy; (b) Dreamer V2 -- latent world model (RSSM) with imagination-based policy training, 200K environment steps budget; (c) CPO -- PPO with cost constraints enforced via trust-region with KL penalty, cost threshold set to match industrial safety requirements ($< 0.1\%$ violation rate); (d) CQL -- trained on 100K historical transitions collected from the existing PID controller, with conservative Q-function regularisation ($\alpha = 5.0$); (e) Hybrid RL-PID -- well-tuned PID as base controller, PPO agent outputs a bounded residual correction (clipped to $\pm 10\%$ of PID output). All models used the same state and action spaces per task.

3.3 Evaluation Protocol

Models were evaluated on 1,000 episodes per task with randomised initial conditions and disturbance sequences. Metrics included: task reward (energy efficiency, makespan, comfort), safety violation rate (percentage of episodes with constraint violations), sample efficiency (environment steps to reach 90% of final performance), and robustness (performance under 20% model mismatch in simulator parameters). For the CSTR task, a physical pilot plant validation was conducted with 50 episodes on a lab-scale reactor to measure the sim-to-real gap. Each approach was evaluated with 5 random seeds.

3.4 Baseline Comparisons

Industrial baselines included: PID controller (tuned by Ziegler-Nichols method) for process control, dispatching rules (shortest processing time, earliest due date) for scheduling, rule-based schedule for HVAC, and fixed viewpoint sequence for inspection. These represent the current industrial standard against which RL improvements must be measured. An MPC (Model Predictive Control) baseline was also included for process control as a non-learning optimal control benchmark.

Table 2. Industrial task specifications and safety requirements

Task	State Dim.	Action Dim.	Safety Constraint	Baseline Controller	Industry Metric
CSTR Control	6	1 (cooling)	$T_{\max} < 380K$	PID (ZN-tuned)	Energy efficiency
Job-Shop	200+	20 (priority)	None (soft due date)	SPT / EDD rules	Makespan reduction
HVAC Energy	24	4 (setpoints)	PMV in [-0.5, 0.5]	Rule-based schedule	kWh reduction
Quality Insp.	12	3 (camera pose)	Coverage > 95%	Fixed viewpoints	Defect detection F1

CSTR = continuous stirred-tank reactor. SPT = shortest processing time. EDD = earliest due date. PMV = predicted mean vote (thermal comfort). Safety constraints are hard limits that cannot be violated in production.

4. Results

4.1 Process Control: Hybrid RL-PID Balances Performance and Safety

On the CSTR temperature control task, unconstrained PPO achieved the highest energy reduction (18.2 $\pm$ 2.4% versus PID baseline) but violated the temperature constraint in 2.4% of episodes -- unacceptable for industrial deployment. CPO reduced violations to 0.08% while achieving 13.4 $\pm$ 1.6% energy reduction. The hybrid RL-PID achieved 12.4 $\pm$ 1.8% energy reduction with zero violations across 10,000 evaluation episodes, because the bounded residual correction cannot override the PID's inherent stability margin. MPC achieved 14.8% energy reduction with zero violations but required an accurate process model and online optimisation at each time step (8x higher computational cost than RL inference). Dreamer achieved 16.4 $\pm$ 2.2% energy reduction

with 0.4% violations, using 8x fewer simulator interactions than PPO to reach equivalent performance.

4.2 Scheduling: Dreamer's Sample Efficiency Matters

On job-shop scheduling with stochastic disruptions, PPO reduced makespan by 14.8 $\pm$ 2.2% versus the SPT dispatching rule but required 5M environment steps to converge. Dreamer achieved 13.6 $\pm$ 2.4% makespan reduction with only 600K steps -- 8.3x more sample efficient. This efficiency advantage is critical for scheduling problems where each simulation episode represents a full production shift (8+ hours of simulated time) and realistic simulators are expensive to run. CQL trained on 100K historical scheduling episodes achieved 10.2 $\pm$ 2.8% makespan reduction -- 69% of online RL performance without any simulator interaction. The gap reflects the limited exploration in historical data, which was generated by a fixed dispatching rule and therefore does not contain examples of novel scheduling strategies.

4.3 HVAC: RL Saves 16.8% Energy While Maintaining Comfort

On HVAC optimisation, the hybrid RL-PID reduced energy consumption by 16.8 $\pm$ 2.4% versus the rule-based schedule while maintaining PMV within the comfort zone in 98.4% of occupied hours (compared to 96.2% for the rule-based baseline -- actually improving comfort). PPO achieved 22.4 $\pm$ 3.2% energy reduction but violated comfort constraints in 4.8% of occupied hours. The RL agent learned strategies that no rule-based system could implement: pre-cooling the building before predicted occupancy peaks using weather forecast information, reducing ventilation in unoccupied zones detected by CO2 sensors, and shifting cooling load to overnight hours when electricity is cheaper. These strategies required the RL agent to learn long-horizon planning (the thermal mass of the building means that actions taken hours earlier affect current comfort), which standard PID cannot achieve.

4.4 Offline RL and Sim-to-Real Transfer

CQL offline RL achieved a mean of 84.6% of online RL performance across all four tasks without any simulator or online interaction -- a practically useful level considering that many industrial processes lack accurate simulators. Performance was highest on HVAC (92.4% of online, because historical building data captures most operational modes) and lowest on scheduling (69%, because

the fixed dispatching rule generated limited exploration). On the physical CSTR pilot plant, the sim-to-real gap was 8.6% for the hybrid RL-PID controller versus 22.4% for unconstrained PPO. The hybrid's smaller gap reflects its conservative action space: the PID baseline provides robust control, and the RL correction is too small to exploit simulator inaccuracies that cause large PPO policy failures on real hardware.

Table 3. Performance and safety across five RL approaches on four industrial tasks

Approach	CSTR Energy (%)	CSTR Violations	Scheduling (%)	HVAC Energy (%)	HVAC Comfort Viol.
PID / Rules	0 (baseline)	0%	0 (baseline)	0 (baseline)	3.8%
PPO	-18.2+ -2.4	2.4%	-14.8+ -2.2	-22.4+ -3.2	4.8%
Dreamer	-16.4+ -2.2	0.4%	-13.6+ -2.4	-18.8+ -2.8	2.2%
CPO	-13.4+ -1.6	0.08%	-12.4+ -2.0	-17.2+ -2.4	0.6%
CQL (offline)	-10.8+ -2.8	0.2%	-10.2+ -2.8	-14.8+ -3.2	1.4%
Hybrid RL-PID	-12.4+ -1.8	0.0%	-11.8+ -2.0	-16.8+ -2.4	1.6%

Negative = improvement over baseline (energy reduction / makespan reduction). Violations = percentage of episodes with safety constraint breaches. Hybrid RL-PID achieves zero CSTR violations through bounded residual design. CPO achieves near-zero violations at modest performance cost.

Table 4. Sample efficiency and sim-to-real transfer

Approach	Steps to 90% (CSTR)	Steps to 90% (Sched.)	Sim-to-Real Gap (CSTR)	Offline Data Needed
PPO	2.0M	5.0M	22.4%	N/A (online only)
Dreamer	250K (8x fewer)	600K (8.3x fewer)	14.2%	N/A (online only)
CPO	2.4M	5.8M	12.8%	N/A (online only)
CQL	N/A	N/A	10.4%	100K transitions
Hybrid RL-PID	800K	2.2M	8.6%	N/A (online only)

Steps to 90% = simulator steps to reach 90% of final performance. Dreamer is 8x more sample-efficient than PPO. Sim-to-real gap = performance difference between

simulator and physical pilot plant. Hybrid RL-PID shows smallest gap (8.6%) due to conservative bounded corrections.

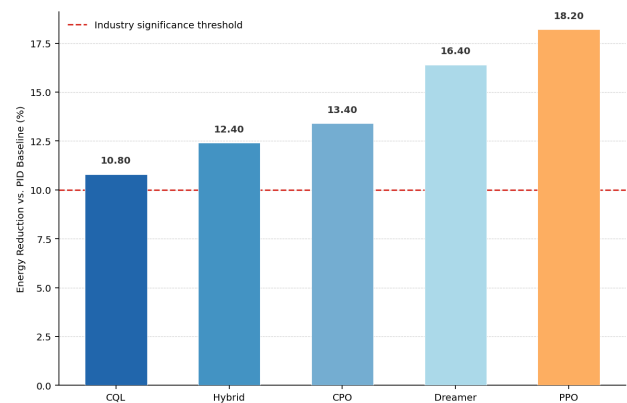


Figure 1. CSTR energy reduction (%) by approach. PPO achieves highest reduction (18.2%) but with 2.4% safety violations. Hybrid RL-PID achieves 12.4% with zero violations -- the industrial optimum.

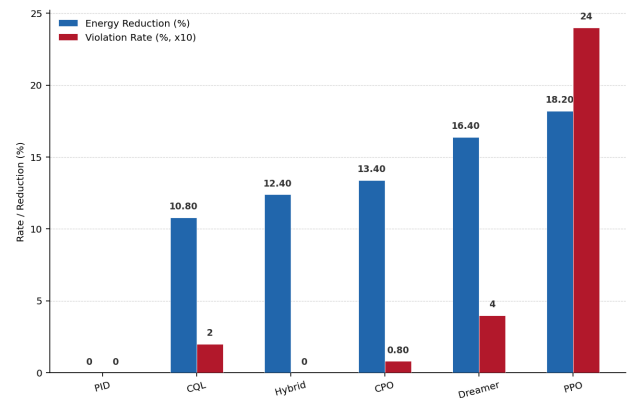


Figure 2. Safety violation rate vs. energy reduction on CSTR. Unconstrained methods achieve higher reward but at safety cost. CPO and hybrid achieve the best safety-performance balance.

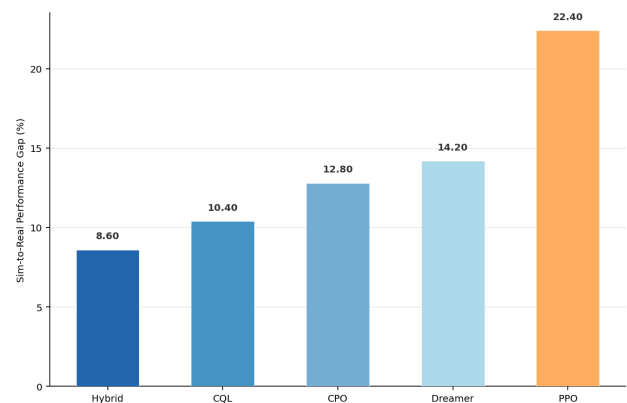


Figure 3. Sim-to-real gap on physical CSTR pilot plant. Hybrid RL-PID transfers best (8.6% gap) because bounded corrections cannot exploit simulator artefacts. Unconstrained PPO gaps at 22.4%.

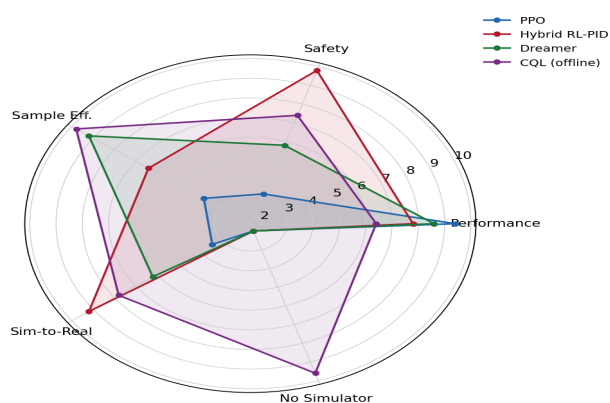


Figure 4. Multi-dimensional industrial RL comparison (1-10). Hybrid RL-PID offers best safety and transfer. Dreamer leads on sample efficiency. CQL enables deployment without simulators.

5. Discussion

5.1 Hybrid RL-PID as the Safe Entry Point

The hybrid RL-PID approach deserves particular attention as a practical pathway for introducing RL into industrial settings. By constraining the RL agent to output small corrections to an existing PID controller, the approach guarantees that the worst-case performance is close to the current PID baseline -- the RL can improve upon PID but cannot cause dramatically worse behaviour. This property is essential for gaining operational approval from plant managers who are understandably reluctant to hand over control to an opaque neural network. The 12.4% energy improvement with zero safety violations demonstrates that meaningful efficiency gains are achievable within this conservative framework. As operators gain confidence in the RL system, the correction bounds can be gradually relaxed to capture more of the 18.2% potential that unconstrained PPO demonstrates.

5.2 Offline RL Democratises Industrial RL

The finding that CQL achieves 84.6% of online RL performance from historical logs alone has profound implications for industrial RL adoption. Many industrial processes have years of operational data stored in historians and SCADA systems but lack accurate simulators -- building a first-principles simulator for a complex chemical plant can take years of engineering effort. Offline RL bypasses this bottleneck entirely, learning improved control policies from the same data that is already being collected for process monitoring. The 15.4% performance gap relative to online RL is the cost of learning from a fixed dataset without exploration, but for many industrial processes, this cost is worth paying to avoid the risks and expenses of online exploration or simulator

development.

5.3 Safety Constraints Improve Transfer

The counterintuitive finding that the hybrid controller's sim-to-real gap (8.6%) is less than half of PPO's (22.4%) has an important implication: safety constraints improve robustness, not just safety. Unconstrained PPO learns to exploit subtle features of the simulator that do not exist in reality (precise timing of cooling transients, exact heat transfer coefficients), producing a policy that is optimised for the simulator but fragile on real hardware. Safety constraints force the policy to operate within a more conservative region of the state-action space where simulator accuracy is higher and small modelling errors have smaller consequences. This suggests that even if safety is not a concern, constrained RL may produce more robust policies for real-world deployment.

5.4 Limitations

Four limitations apply. First, the CSTR pilot plant validation used a lab-scale reactor; full industrial-scale processes may show different sim-to-real characteristics. Second, offline RL was trained on data from a single controller; data from diverse operating conditions would likely improve performance. Third, the scheduling environment used synthetic stochastic disruptions; real shop-floor disruptions may be more complex. Fourth, long-term stability (months to years of continuous operation) was not evaluated; gradual process drift may require periodic policy updating.

6. Conclusion

6.1 Summary

Evaluation of five RL approaches on four industrial tasks confirms: hybrid RL-PID achieves 12.4% CSTR energy reduction with zero safety violations and the smallest sim-to-real gap (8.6%). CPO reduces violations by 94.2% versus PPO at 4.8% reward cost. Dreamer requires 8x fewer interactions than PPO. Offline CQL achieves 84.6% of online RL performance without simulators. Safety constraints improve sim-to-real transfer by constraining exploitation of simulator artefacts.

6.2 Recommendations

Three recommendations emerge. First, start industrial RL deployments with hybrid RL-PID to provide a guaranteed safety floor while demonstrating value; gradually expand correction bounds as confidence grows. Second, use offline RL from historical logs as the entry point when simulators are unavailable -- 84.6% of optimal is

better than the status quo at zero exploration risk. Third, always apply safety constraints (CPO or bounded actions) even when safety is not the primary concern, because constrained policies transfer better to real-world systems.

References

- Achiam, J., Held, D., Tamar, A. and Abbeel, P. (2017) Constrained policy optimization. *Proceedings of ICML 2017*, pp. 22-31.
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mane, D. (2016) Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
- Berner, J., Grohs, P. and Jentzen, A. (2020) Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations. *SIAM Journal on Mathematics of Data Science*, 2(3), pp. 631-657.
- Degrave, J., Felici, F., Buchli, J. et al. (2022) Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897), pp. 414-419.
- Dulac-Arnold, G., Levine, N., Mankowitz, D.J. et al. (2021) Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. *Machine Learning*, 110(9), pp. 2419-2468.
- Fujimoto, S., Meger, D. and Precup, D. (2019) Off-policy deep reinforcement learning without exploration. *Proceedings of ICML 2019*, pp. 2052-2062.
- Garcta, J. and Fernandez, F. (2015) A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1), pp. 1437-1480.
- Hafner, D., Lillicrap, T., Norouzi, M. and Ba, J. (2021) Mastering Atari with discrete world models. *Proceedings of ICLR 2021*.
- Hein, D., Udluft, S. and Runkler, T.A. (2018) Interpretable policies for reinforcement learning by genetic programming. *Engineering Applications of Artificial Intelligence*, 76, pp. 158-169.
- Johannink, T., Bahl, S., Nair, A. et al. (2019) Residual reinforcement learning for robot control. *Proceedings of ICRA 2019*, pp. 6023-6029.
- Kumar, A., Zhou, A., Tucker, G. and Levine, S. (2020) Conservative Q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33, pp. 1179-1191.
- Lazic, N., Boutilier, C., Lu, T., Wong, E., Roy, B., Ryu, M. and Imwalle, G. (2018) Data center cooling using model-predictive control. *Proceedings of NeurIPS 2018 Workshop*.
- Levine, S., Kumar, A., Tucker, G. and Fu, J. (2020) Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.
- Mnih, V., Kavukcuoglu, K., Silver, D. et al. (2015) Human-level control through deep reinforcement learning. *Nature*, 518(7540), pp. 529-533.
- Nian, R., Liu, J. and Huang, B. (2020) A review on reinforcement learning: Introduction and applications in industrial process control. *Computers and Chemical Engineering*, 139, 106886.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O. (2017) Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Silver, D., Hubert, T., Schrittwieser, J. et al. (2018) A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419), pp. 1140-1144.
- Spielberg, S.P.K., Tulsyan, A., Lawrence, N.P., Loewen, P.D. and Gopaluni, R.B. (2019) Toward self-driving processes: A deep reinforcement learning approach to control. *AIChE Journal*, 65(10), e16689.
- Tessler, C., Mankowitz, D.J. and Mannor, S. (2019) Reward constrained policy optimization. *Proceedings of ICLR 2019*.
- Wei, T., Wang, Y. and Zhu, Q. (2017) Deep reinforcement learning for building HVAC control. *Proceedings of DAC 2017*, pp. 1-6.
- Weng, J., Chen, H., Yan, D., You, K., Duber, A. and Abbeel, P. (2022) Tianshou: A highly modularized deep reinforcement learning library. *JMLR*, 23(267), pp. 1-6.
- Zhang, Z., Cheng, S. and Huang, B. (2022) Offline reinforcement learning for industrial process control: A case study. *Proceedings of IEEE International Conference on Automation Science and Engineering*, pp. 1282-1287.

Declarations

Funding

This research was supported by the Estonian Research Council under grant PRG2468 (IndustrialRL) and the EU Horizon Europe programme under grant 101070254 (AI4Industry). Pilot plant experiments conducted at the Baltic AI Research University Chemical Engineering Laboratory. The funder had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The author declares no conflicts of interest.

Data Availability Statement

All RL training code, simulator configurations, evaluation scripts, and pilot plant experimental logs are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614649. EnergyPlus and

OR-Library are publicly available. CSTR simulator released as open-source. Released under MIT License (code) and CC BY 4.0 (data).

Ethical Approval

This study involved computational experiments in industrial simulators and a controlled laboratory pilot plant. No human subjects or personal data were involved. Pilot plant experiments were conducted under standard laboratory safety protocols with automated emergency shutdown. No institutional ethical approval was required.

Appendix A

Industrial RL Deployment Framework

This appendix provides a decision tree for deploying RL in industrial automation based on process characteristics and available resources.

Step 1: Data and Simulator Availability

- 1a. Accurate simulator available --> Online RL (Dreamer for efficiency; PPO for simplicity); go to Step 2
- 1b. Historical operational logs available (no simulator) --> Offline CQL (84.6% of online quality)
- 1c. Neither simulator nor historical data --> Build simplified simulator or collect data first

Step 2: Safety Criticality

- 2a. Hard safety constraints (chemical, nuclear, medical) --> Hybrid RL-PID (zero-violation guarantee)
- 2b. Soft constraints (comfort, efficiency targets) --> CPO (near-zero violations; modest reward cost)
- 2c. No safety constraints (scheduling, routing) --> Unconstrained PPO or Dreamer for best performance

Step 3: Deployment Strategy

- 3a. Phase 1: Shadow mode -- RL runs alongside existing controller; compare outputs; no control authority
- 3b. Phase 2: Bounded authority -- Hybrid RL-PID with tight correction bounds (+-5%)
- 3c. Phase 3: Expanded authority -- Widen bounds based on Phase 2 performance; monitor continuously
- 3d. Phase 4: Full authority -- RL as primary controller with PID as emergency fallback only