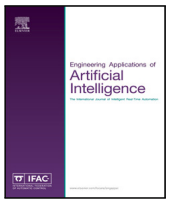
Engineering Applications of Artificial Intelligence 181 (2026) 115524

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Deep reinforcement learning in applied control: Challenges, analysis, and insights

Klinsmann Agyei, Pouria Sarhadi *, Daniel Polani

University of Hertfordshire, School of Physics, Engineering and Computer Science, Hatfield, AL10 9AB, Hertfordshire, United Kingdom

ARTICLE INFO

Keywords:
Deep reinforcement learning
Data-driven model-free control
Applied control
Control benchmarks

ABSTRACT

Deep reinforcement learning (DRL) algorithms have shown increasing promise for autonomous decision-making and continuous control, yet their practical value relative to established control methods remains difficult to assess under realistic control-engineering conditions. Existing evaluations are often narrow, application-specific, or based on inconsistent assumptions. As a result, it remains unclear when learned controllers are suitable alternatives to conventional designs. This paper presents a systematic assessment framework that compares four prominent DRL controllers with a classical control baseline across a diverse set of applied control problems, including non-minimum phase dynamics, flexible mechanical systems, nonlinear marine control, and aerial robotics. The framework standardises modelling assumptions, reward design, controller tuning, computational budgets, performance metrics, robustness tests, and deployment conditions. This allows fair and reproducible analysis across methods and applications. The evaluation covers tracking accuracy, settling time, overshoot, control effort, actuator saturation, disturbance rejection, model uncertainty, robustness margins, and sim-to-real transfer. The results show that performance depends strongly on the algorithmic structure, system dynamics, and operating conditions. Learned controllers provide advantages in some cases, particularly for nonlinearities, constraints, and uncertain operating regimes. Classical control remains competitive in accuracy, simplicity, and reliability for several benchmark conditions. Under combined saturation and unmodelled dynamics, the learned controllers maintained stable operation where the classical baseline became unstable. Real-time deployment on a quadrotor platform further demonstrated stable sim-to-real transfer. The study establishes a rigorous and reproducible evidence base for assessing the practical suitability of these data-driven control algorithms in applied control. It clarifies the trade-offs between learning-based and conventional control. This addresses a critical gap that cannot be easily inferred from isolated, system-specific studies.

1. Introduction

Automatic control theory has undergone more than a century of development in modern history, with a myriad of real-life applications in various industries that extend into our daily lives (Maxwell, 1868; Minorsky, 1922; Bennett, 2002). Most controllers have traditionally been designed based on models or detailed knowledge of the system. Despite many ups and downs throughout its history, the successful deployment of model-based controllers remains unquestionable. Nevertheless, a renewed interest has emerged in data-driven, model-free controllers, motivated by ambitious new goals (Saviolo and Loianno, 2023; Khaki-Sedigh, 2023; Tang and Daoutidis, 2022). These objectives are not aimed at surpassing the already resolved problems of model-based controllers, but rather at achieving broader ambitions:

1. Enhancing the generalisability of controllers beyond existing robust and adaptive methods, which are applied to specific model structures and their associated uncertainties;
2. Addressing problems that cannot be easily modelled by conventional methods, particularly in the development of autonomous and intelligent control systems where decision-making is central (i.e., increasing the autonomy levels of algorithms Antsaklis, 2020; Vagia et al., 2016).

Such challenges include scenarios where control objectives are either poorly defined or cannot be properly specified, necessitating a shift from 'how' to 'what' to control (Chen, 2024), or even towards intrinsic motivation to intelligently determine the most appropriate goal (Tiomkin et al., 2024), potentially realising human-level decision-making and control. Recent advances in Artificial Intelligence (AI), Machine Learning (ML), and their subset, Deep Reinforcement Learning (DRL), have demonstrated significant progress, sometimes achieving superhuman performance, particularly in gaming and simulation environments. These advancements have been more prominent in areas

* Corresponding author.
E-mail addresses: k.agyei@herts.ac.uk (K. Agyei), p.sarhadi@herts.ac.uk (P. Sarhadi), d.polani@herts.ac.uk (D. Polani).

https://doi.org/10.1016/j.engappai.2026.115524
Received 14 December 2025; Received in revised form 3 June 2026; Accepted 22 June 2026
Available online 2 July 2026

such as text, voice, and image processing, particularly demonstrated with the emergence of Large Language Models (LLMs). However, this progress has not yet been convincingly replicated in the control of dynamical systems, although some promising results have been reported. In the field of DRL, several solutions, developed through a paradigm distinct from classical control theory, have attracted considerable attention in the literature, with encouraging results in various control and autonomy applications (Kendall et al., 2019; Eschmann et al., 2024; Hadi et al., 2024). The development of DRL algorithms has seen remarkable advancement, beginning with DQN's breakthrough performance on Atari games by DeepMind (Mnih et al., 2013). For continuous control problems, DDPG (Lillicrap et al., 2015) pioneered an actor–critic approach with deterministic policy gradients, whilst TD3 (He and Hou, 2020) enhanced stability by addressing overestimation bias. PPO (Schulman et al., 2017) introduced a stochastic policy approach, and TD-MPC2 (Hansen et al., 2023) merged model predictive control principles with deep learning for improved long-term performance. TD-MPC2 (Hansen et al., 2023) extends TD-MPC by integrating decoder-free latent-space with a unified architecture that demonstrates scalability and generalisation capabilities across diverse control tasks.

Nonetheless, these approaches' success has not yet translated into widespread industrial adoption (Tang et al., 2025), unlike the enduring fame of the Proportional–Integral–Derivative (PID) controller or, subsequently, the Model Predictive Control (MPC) approach (Hägglund and Guzmán, 2024). Regardless of the complexities of tuning and implementation, a key challenge is the lack of a clear understanding of the capabilities and limitations of DRL approaches when applied to the operational demands of control, especially in a quantified manner. Existing approaches face several real-world limitations that have hindered this understanding. First, most DRL controllers are developed and tested exclusively in simulation, with little consideration of performance degradation during sim-to-real transfer due to unmodelled dynamics, sensor noise, and actuator imperfections. Second, evaluation under idealised conditions, perfect models, no actuator constraints, no measurement noise, means the performance are rarely reproducible in deployment. Third, the absence of standardised, control-oriented benchmarks means that different studies use incompatible metrics, making cross-algorithm comparison unreliable. Some efforts have been made to address this gap through qualitative surveys, such as in Buşoniu et al. (2018) or in specific process control applications focused mainly on tracking performance. However, a systematic and rigorous quantified analysis, as presented in this paper, has not yet been carried out. There have been a number of recent efforts, primarily in the process control community, aimed at quantitatively assessing the performance of these approaches (Lawrence et al., 2022; Dutta and Upreti, 2023; Joshi et al., 2023; Oh, 2024). Nevertheless, as thoroughly reviewed in Section 3, these studies remain limited in scope, and this paper aims to fill that gap by providing a more systematic and comprehensive evaluation. It is important to note that the value of DRL does not rest on outperforming classical controllers in standard regulation tasks, where well-tuned model-based methods remain highly competitive. Rather, its potential lies in scenarios where classical methods face fundamental limitations: systems that are difficult to model, highly nonlinear, or subject to unpredictable constraints and uncertainties that cannot be easily anticipated at design time. Understanding precisely where DRL delivers genuine advantages and where it does not is therefore the central motivation of this study.

This paper addresses a critical research question: How do leading model-free DRL algorithms perform when confronted with real-world control challenges, including actuator constraints, time delays, parameter uncertainties, and external disturbances? We evaluate four well-established DRL algorithms DDPG, TD3, PPO, and TD-MPC2, and a model-based LQR controller, all trained with the same reward(cost) function, on four distinct benchmark problems selected for their representative challenges: a non-minimum phase system, the two-mass spring system, a nonlinear autonomous underwater vehicle model, and real-time implementation on a quadrotor drone.

The primary contributions of this work are as follows. It should be noted that whilst the idea of evaluating DRL through classical control metrics has precedent in the literature, the contributions of this paper lie in the systematic consolidation and empirical rigour of the evaluation rather than in the introduction of new algorithms or theoretical developments.

First, we establish a comprehensive benchmarking methodology that evaluates DRL algorithms using metrics directly relevant to control engineering practice, building on the standardised criteria proposed in Sarhadi (2025). Specifically, the framework includes settling time, rise time, overshoot, integral of squared error (ISE), integral of time-weighted absolute error (ITAE), integral of absolute control effort (IACE), integral of absolute control effort rate (IACER), peak control value, gain margin, and delay margin, collectively covering transient response, control effort, and robustness dimensions essential for industrial deployment. Whilst individual metrics have appeared in prior studies, no existing work applies this complete metric set consistently across multiple algorithms and benchmarks.

Second, unlike previous studies that evaluate algorithms under idealised conditions with perfect models and no practical constraints, our approach incorporates actuator amplitude and rate saturations, measurement noise, external disturbances, time delays, and model uncertainties simultaneously. This reveals performance degradation patterns and failure modes not apparent in idealised simulations, providing crucial insights for reliability assessment.

Third, we provide systematic assessment across four carefully selected benchmark problems representing fundamentally different control challenges: non-minimum phase dynamics, resonant mechanical systems, nonlinear marine vehicle control, and real-time aerial robotics. This diversity enables identification of which algorithmic paradigms suit which classes of control problem, providing practitioners with actionable guidance.

Finally, beyond simulation, we provide concrete evidence of real-world viability through hardware deployment on the Crazyflie platform, quantifying sim-to-real performance degradation and demonstrating that DRL controllers can achieve stable operation despite training-deployment mismatches.

It should be emphasised that the conclusions drawn in this study are limited to the evaluated benchmark systems, operational conditions, and experimental framework considered herein, and should not be interpreted as universal performance claims applicable to all control scenarios or DRL methods. Nonetheless, similar results may be expected in comparable engineering systems.

2. Control algorithms and state-of-the-art

In this section, we present the control methods evaluated in our benchmarking study, including a classical optimal control baseline (LQI) and four deep reinforcement learning approaches (DDPG, TD3, PPO, and TD-MPC2). The selection of these specific algorithms is motivated by their representation of fundamentally different paradigms: DDPG and TD3 cover deterministic off-policy learning, PPO represents stochastic on-policy learning, and TD-MPC2 exemplifies model-based planning. Together they span from deterministic policy learning to stochastic policy learning, and from model-free to hybrid model-based approaches.

2.1. Justification for algorithm selection

The continuous control landscape includes numerous prominent algorithms such as Soft Actor–Critic (SAC) (Haarnoja et al., 2018a), Trust Region Policy Optimisation (TRPO) (Schulman et al., 2015), Advantage Actor–Critic (A2C) (Mnih et al., 2016), and more recent developments like Randomised Ensembled Double Q-learning (REDQ)

(Chen et al., 2021) and Dropout Q-Functions (DroQ) (Hiraoka et al., 2021). Our selection criteria prioritised algorithms that represent fundamentally different approaches to continuous control whilst demonstrating practical viability through widespread adoption and successful applications.

The selected algorithms are therefore not intended to represent the entirety of modern reinforcement learning methods for control, but rather a carefully chosen subset spanning the principal paradigms most relevant to continuous-control applications. The objective of this selection strategy is to enable an interpretable and practically manageable comparative analysis whilst still capturing key methodological differences between deterministic, stochastic, and model-based DRL approaches. Similarly, several prominent alternatives, including SAC, TRPO, REDQ, and hybrid learning-based MPC methods, were not excluded due to lack of relevance or performance, but rather to avoid excessive overlap between algorithmic families and to maintain a focused evaluation framework within the scope of the present study.

DDPG (Lillicrap et al., 2015) serves as the foundational algorithm for deterministic policy learning in continuous spaces, establishing the actor–critic framework that underlies many subsequent developments. Its inclusion is essential as it represents the baseline from which improvements are measured. Whilst DDPG has known limitations, understanding its performance characteristics provides insight into the fundamental capabilities and constraints of deterministic deep reinforcement learning. Its relatively simple implementation and widespread availability in standard libraries (Raffin et al., 2021) also make it accessible for practitioners, justifying its continued relevance despite more sophisticated alternatives.

TD3 (Fujimoto et al., 2018) directly addresses DDPG's critical weaknesses through targeted algorithmic improvements, making it the natural evolution of deterministic policy learning. The algorithm's modifications, twin critics, delayed updates, and target smoothing are specifically designed to combat overestimation bias and training instability that undermine value-based methods. Empirical studies consistently demonstrate TD3's superior performance over DDPG across diverse continuous control tasks (Fujimoto et al., 2018), whilst maintaining similar computational requirements. This combination of improved reliability and practical efficiency has led to TD3 becoming the de facto standard for deterministic policy learning in applied settings.

The exclusion of SAC (Haarnoja et al., 2018a), despite its popularity, merits explanation. SAC employs entropy regularisation to learn stochastic policies that maximise both expected return and policy entropy, theoretically providing better exploration and robustness. However, comparative studies have shown that whilst SAC excels in sparse-reward and multi-modal tasks, its performance in traditional control problems with dense rewards often matches or slightly underperforms relative to TD3 (Cai et al., 2023). Furthermore, SAC's entropy tuning introduces an additional hyperparameter that requires careful calibration for control applications where exploration–exploitation balance differs from typical RL benchmarks. Given that PPO already represents stochastic policies in our evaluation and TD3 covers the twin-critic architecture, including SAC would provide limited incremental insight whilst increasing experimental complexity. It is also worth noting that the quadratic reward structure shared across all DRL algorithms does not constrain them to replicate linear behaviour. The results under operational challenges, particularly the Rohrs uncertainty test in Section 4.4.3, provide evidence of this: whilst the LQI controller fails under combined actuator saturations and unmodelled dynamics, all DRL algorithms maintain stable operation and achieve near-zero steady-state error. This fundamentally nonlinear behaviour, emerging from policies trained on the same quadratic cost function as LQI, confirms that the algorithms discover genuinely nonlinear control strategies rather than approximating linear responses.

PPO (Schulman et al., 2017) represents the on-policy, stochastic policy paradigm, offering fundamentally different theoretical guarantees and practical trade-offs compared to off-policy methods. Its selection over TRPO (Schulman et al., 2015) reflects practical considerations: whilst TRPO provides stronger theoretical guarantees through KL-divergence constraints, PPO achieves comparable performance with significantly lower computational cost and implementation complexity (Engstrom et al., 2019). PPO's clipped objective provides an elegant approximation to TRPO's trust region that proves more robust to hyperparameter choices in practice. The algorithm's success across diverse domains—from robotic control (Andrychowicz et al., 2020) to game playing (Berner et al., 2019) demonstrates its versatility and reliability, making it the natural choice for representing on-policy methods.

TD-MPC2 (Hansen et al., 2023) exemplifies the integration of model-based planning with deep reinforcement learning, representing a fundamentally different paradigm from pure model-free approaches. TD-MPC2's decoder-free architecture and unified learning framework address key limitations of earlier model-based methods, particularly the computational overhead of pixel-based planning and the instability of separate model and policy learning. Recent benchmarks demonstrate TD-MPC2's state-of-the-art performance across diverse control tasks (Hansen et al., 2024), validating its position as the leading model-based approach for practical control applications.

We deliberately excluded pure model-based methods without reinforcement learning components, such as Learning-based MPC (Hewing et al., 2020) or Gaussian Process MPC (Kamthe and Deisenroth, 2018), as these represent a different paradigm that warrants separate investigation. Similarly, evolutionary strategies (Salimans et al., 2017) and derivative-free optimisation methods (Mania et al., 2018), whilst showing promise for control applications, operate on fundamentally different principles that would complicate cross-algorithm comparison.

The four selected algorithms are considered sufficient representatives of the continuous DRL design space for the conclusions drawn in this paper. Specifically, the evaluation supports conclusions about the relative performance of deterministic versus stochastic policies, on-policy versus off-policy learning, and model-free versus model-based planning under classical control metrics.

2.2. Linear Quadratic Integral Control (LQI)

The LQI controller serves as our classical baseline, providing an optimal solution for linear systems with complete model knowledge. The selection of LQI over other classical controllers such as PID or MPC is deliberate and justified on several grounds. First, LQI represents a well-understood optimal control solution that provides theoretical performance guarantees for linear systems, making it a meaningful upper bound for classical linear control performance. Second, the LQI cost function $J = \int (x^\top Q x + u_c^\top R u_c) dt$ provides a natural and transparent basis for defining the DRL reward function, ensuring all algorithms optimise the same underlying objective and enabling a fair cross-paradigm comparison. Third, LQI requires full model knowledge, making it the strongest possible classical baseline for systems where such knowledge is available, and thus any DRL advantage observed over LQI is meaningful. PID controllers, whilst more widely deployed industrially, lack the systematic multi-state optimisation framework necessary for fair comparison with DRL methods that optimise over full state vectors. MPC, whilst arguably a stronger baseline for constrained problems, introduces additional design complexity and computational requirements that would complicate the isolation of algorithmic performance differences. The selected LQI weights ($Q = 10I$, $R = 1$) were kept identical across all benchmark systems to maintain consistency and avoid system-specific tuning that would favour the classical baseline. The motivation for including LQI stems from its widespread industrial adoption and well-understood theoretical properties (Anderson and Moore, 2007). By augmenting the standard LQR formulation with integral action, LQI addresses the fundamental limitation of steady-state tracking error that affects other classical controllers. This augmentation transforms the original plant system $\dot{x}_p = A_p x_p + B_p u$ through the

incorporation of an integrator state, yielding the augmented system:

$$\dot{x} = Ax + Bu_c \tag{1}$$

where the augmented state vector $x = \begin{bmatrix} x_p^\top & \int e\,dt \end{bmatrix}^\top$ includes both plant states and the integral of tracking error, with system matrices appropriately constructed to preserve the original dynamics whilst incorporating integral action.

The control law $u_c = -Kx$ emerges from solving the algebraic Riccati equation that minimises the quadratic cost function $J = \int (x^\top Qx + u_c^\top Ru_c)dt$. This formulation provides several advantages that make it an ideal baseline: guaranteed stability margins for the nominal system, systematic tuning through weight matrix selection, and optimal performance within the linear quadratic framework. However, LQI faces inherent limitations when confronting nonlinear dynamics, actuator constraints, and model uncertainties, precisely the challenges where DRL approaches may demonstrate advantages. To ensure fair comparison across all algorithms, we employ the negative of the LQI cost as the reward function for training DRL controllers, thereby aligning all methods towards the same underlying objective. This is a deliberate and important methodological choice: by negating the quadratic cost $J = \int (x^\top Qx + u_c^\top Ru_c)dt$, the DRL reward signal directly mirrors the LQI optimisation objective, ensuring that all algorithms are incentivised to minimise the same weighted combination of tracking error and control effort rather than pursuing algorithm-specific reward structures that would compromise the fairness of the comparison. It is emphasised that the selected LQI weights do not necessarily represent optimal tuning for each individual system, but rather serve as a uniform and transparent classical baseline for comparative evaluation.

2.3. Deep Deterministic Policy Gradient (DDPG)

DDPG (Lillicrap et al., 2015) emerged from the recognition that value-based methods like DQN cannot directly handle continuous action spaces without discretisation, which becomes computationally intractable for high-dimensional control problems. The algorithm's key innovation lies in combining deterministic policy gradients with deep function approximation, enabling direct optimisation of continuous control policies. This approach leverages the insight that deterministic policies can be more sample-efficient than stochastic alternatives when exploration is handled separately through action noise.

The algorithm employs an actor–critic architecture where the actor network $\mu_\phi(s)$ learns a deterministic mapping from states to actions, whilst the critic network $Q_\theta(s,a)$ estimates the expected return. The actor parameters are updated by following the gradient of expected return:

$$\nabla_\phi J \approx \mathbb{E}_{s\sim D}\left[\nabla_a Q_\theta(s,a)\Big|_{a=\mu_\phi(s)} \nabla_\phi \mu_\phi(s)\right] \tag{2}$$

This formulation enables efficient learning by backpropagating the critic's value gradient through the deterministic actor, avoiding the high variance associated with stochastic policy gradients.

The critic is trained using temporal difference learning with experience replay and target networks to stabilise learning:

$$L(\theta) = \mathbb{E}_{(s,a,r,s')\sim D}\left[\left(r + \gamma Q_{\theta'}(s', \mu_{\phi'}(s')) - Q_\theta(s,a)\right)^2\right] \tag{3}$$

where θ' and ϕ' denote slowly-updated target network parameters that provide stable regression targets.

2.4. Twin delayed deep deterministic policy gradient (TD3)

TD3 (Fujimoto et al., 2018) was developed specifically to address the overestimation bias and brittleness that afflict DDPG. The algorithm's name reflects its three key innovations: twin critics to mitigate overestimation, delayed policy updates to reduce per-update error, and target policy smoothing to regularise the critic. These modifications, whilst conceptually simple, dramatically improve learning stability and final performance.

The twin critic architecture maintains two independent Q-networks and uses the minimum of their estimates for policy updates:

$$y = r + \gamma \min_{i=1,2} Q_{\theta_i'}(s', \tilde{a}) \tag{4}$$

This pessimistic estimate reduces the overestimation bias that accumulates through temporal difference bootstrapping. The theoretical foundation rests on the observation that whilst any single critic may overestimate due to function approximation error, the minimum of two independent estimates provides a lower bound that prevents runaway overestimation.

Delayed policy updates address the coupling between actor and critic learning rates. By updating the actor less frequently than the critics, typically every second iteration, TD3 allows value estimates to stabilise before policy improvement. This decoupling prevents the actor from overfitting to transient critic errors during early training when value estimates are least reliable.

Target policy smoothing adds controlled noise to target actions:

$$\tilde{a} = \mu_{\phi'}(s') + \text{clip}(\epsilon, -c, c), \quad \epsilon \sim \mathcal{N}(0, \sigma) \tag{5}$$

This regularisation technique prevents the critic from learning sharp peaks in the value landscape that the actor might exploit, instead encouraging smooth value functions that generalise better.

2.5. Proximal Policy Optimisation (PPO)

PPO (Schulman et al., 2017) emerged from the challenge of achieving stable policy gradient learning without the computational expense of trust region methods like TRPO (Schulman et al., 2015). The algorithm's fundamental insight is that constraining policy updates through a simple clipped objective can achieve similar stability guarantees whilst maintaining computational efficiency suitable for real-time control applications. Unlike DDPG and TD3's deterministic policies, PPO employs stochastic policies that naturally balance exploration and exploitation without requiring explicit noise injection.

The algorithm's core innovation lies in its clipped surrogate objective:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \tag{6}$$

where $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)$ represents the probability ratio between updated and current policies. This formulation prevents destructively large policy updates by clipping the probability ratio when it exceeds the bounds $[1-\epsilon, 1+\epsilon]$, effectively creating a trust region without expensive second-order optimisation.

PPO's stochastic policy $\pi_\theta(a|s)$ offers several advantages for control applications. The natural exploration through policy stochasticity eliminates the need for hand-tuned exploration schedules, whilst the ability to represent multi-modal action distributions enables more flexible control strategies. The on-policy nature of PPO, whilst potentially less sample-efficient than off-policy methods, provides more predictable learning dynamics that are easier to debug and tune in practice.

2.6. Temporal Difference Learning for Model Predictive Control 2 (TD-MPC2)

TD-MPC2 (Hansen et al., 2023) represents a fundamental shift from purely model-free approaches by integrating learned dynamics models with planning-based control. The algorithm addresses the sample inefficiency of model-free methods whilst avoiding the compounding errors that challenge pure model-based approaches. By learning and planning in a compact latent space rather than the raw state space, TD-MPC2 achieves computational efficiency suitable for real-time control whilst maintaining representational flexibility for complex dynamics.

The algorithm learns an encoder that maps observations to latent representations $z_t = e_\phi(s_t)$, within which it simultaneously learns

three components that enable effective planning. The dynamics model $z_{t+1} = f_\phi(z_t, a_t)$ predicts state transitions in latent space, avoiding the complexity of pixel-level or high-dimensional state prediction. The reward model $\hat{r}_t = r_\phi(z_t, a_t)$ estimates immediate rewards without requiring explicit reward engineering. The value function $V_\psi(z_t)$ provides terminal value estimates that extend the planning horizon beyond computational limits.

Planning occurs through trajectory optimisation using the Cross-Entropy Method (CEM):

$$\mathbf{a}^*_{0:H-1} = \arg\max_{\mathbf{a}_{0:H-1}} \sum_{i=0}^{H-1} \gamma^i r_\phi(z_{t+i}, a_{t+i}) + \gamma^H V_\psi(z_{t+H}) \tag{7}$$

This formulation combines the accuracy of short-horizon model-based planning with the long-term awareness provided by value function bootstrapping, mitigating the compounding error problem that limits pure model-based methods.

TD-MPC2's decoder-free architecture eliminates the computational overhead and training instability associated with reconstructing high-dimensional observations, focusing network capacity on control-relevant features. Unlike reconstruction-based models that learn to predict full observations from latent representations, the decoder-free design discards the generative component entirely, retaining only the latent representations necessary for reward prediction, dynamics modelling, and planning. The unified architecture with shared representations across all components enables efficient learning and inference, making real-time control feasible even with limited computational resources.

2.7. State-of-the-art progress in DRL assessment for control applications

Several studies have applied reinforcement learning to control problems, but a comprehensive evaluation of model-free DRL algorithms' performance on standardised control benchmarks remains a significant gap in the literature. As summarised in Table 1, existing work has primarily focused on specific applications or limited algorithm comparisons without systematic assessment across multiple performance criteria.

The integration of DRL into control applications has progressed through distinct phases. Early work focused on discrete action spaces, with DQN demonstrating breakthrough performance on Atari games (Mnih et al., 2013). For continuous control, DDPG (Lillicrap et al., 2015) established the actor–critic framework that underlies many subsequent developments, whilst TD3 (Fujimoto et al., 2018) and SAC (Haarnoja et al., 2018a) addressed its stability limitations. More recently, model-based approaches such as TD-MPC2 (Hansen et al., 2023) have demonstrated improved sample efficiency by integrating learned dynamics models with planning. Alongside these algorithmic developments, several adjacent research directions have emerged that also aim to address the deployment gap between RL and real-world control. Offline RL methods (Levine et al., 2020) seek to learn from pre-collected datasets without environment interaction, which is attractive for safety-critical systems where online exploration is costly. Distributional RL (Bellemare et al., 2017) models the full return distribution rather than its expectation, offering richer uncertainty information. Hybrid learning-based MPC approaches (Hewing et al., 2020; Kamthe and Deisenroth, 2018) combine the interpretability and constraint satisfaction of MPC with the flexibility of learned models. Whilst these directions show promise, their evaluation in applied control settings using classical control-engineering metrics remains limited, which is precisely the gap this paper addresses.

While Tavakkoli et al. (2024) compares a model-free DDPG controller against a classical LQI controller for non-minimum phase systems under various conditions, their analysis is limited to a single DRL algorithm and does not explore the full spectrum of control challenges present in diverse benchmark problems. Similarly, Dutta and Upreti (2022) evaluates multiple actor–critic methods but focuses primarily on process nonlinearities without comprehensive analysis of control signal characteristics and robustness metrics that are critical for practical implementation.

A more critical examination of existing studies reveals important methodological limitations. Dutta and Upreti (2023) apply DDPG to nonlinear process control but evaluate performance using only ISE and RMSE, providing no information on control effort, robustness margins, or actuator constraints. Joshi et al. (2023) introduce a twin-actor SAC framework for batch processes and demonstrate improved disturbance rejection, yet their evaluation remains confined to domain-specific scenarios without broader generalisability. Oh (2024) provide one of the more rigorous comparisons by evaluating DDPG, TD3, and SAC against MPC for chemical processes, but their performance assessment relies on cost function values rather than classical control metrics such as rise time or gain margin, limiting interpretability for control engineers. Tang et al. (2025) survey DRL in robotics but focus on success rates and sample efficiency, metrics that are largely irrelevant to continuous regulation problems. Collectively, these studies confirm that whilst progress has been made in specific domains, no existing work provides a unified, multi-algorithm evaluation across diverse benchmark problems using a consistent and comprehensive set of control-engineering metrics under realistic operational conditions. This paper addresses that gap directly.

Our work addresses this research gap by providing a systematic, control-oriented evaluation of four leading DRL algorithms across multiple benchmark problems using a comprehensive set of standardised performance criteria. Unlike previous studies that evaluate algorithms using generic reinforcement learning metrics or limited control performance indicators, our approach presents a unified framework for assessment that considers a wide spectrum of characteristics relevant to control engineers.

3. The analysis framework

Regardless of the design methodology, for a controller to become operational successfully in the real world, three main aspects of metrics must be satisfied (Sarhadi, 2025): (1) setpoint tracking performance, (2) control signal feasibility or energy consumption, and (3) robustness. During the design stage, controllers are typically developed with specific objectives that may target one or more metrics within these aspects, as reflected in the cost function, stability analysis, or overall control objective. However, once implementation begins, the controller must rigorously satisfy all these aspects; otherwise, re-tuning or redesign will be required, undermining sustainability in the design process. The final consideration is real-time implementability, which no longer poses a significant challenge thanks to the availability of sufficient computational power.

Moreover, the controller should be either designed or at least tested against operational challenges encountered in real-world conditions, such as disturbances and noise, actuation constraints, dynamic limitations including non-minimum phase behaviour or instability, and model uncertainties. Even a model-free controller is typically developed within a simulated environment before transitioning to the real world, in order to reduce the number of required trials and to improve safety and cost-effectiveness, factors that are often jointly considered. These considerations are not only essential for control design but also extend to higher-level functions such as decision-making in autonomous systems. Such challenges are well addressed in classical control, owing to the availability of tools such as frequency domain techniques. However, they are less thoroughly handled in nonlinear systems. As a result, the capabilities and performance boundaries of emerging DRL-based controllers require further investigation. Our evaluation methodology is grounded in a performance assessment across the three fundamental categories outlined above, employing the following classes of metrics to meticulously analyse system behaviour:

Table 1
Overview of reinforcement learning applications in control.

Ref.	Application	Solution/Algorithm(s)	States	Reward function	Evaluation metrics
Dutta and Upreti (2023)	Nonlinear process control (CSTR, pH)	DDPG	Partial system state	A β -weighted composite reward, minimising absolute tracking error and control effort variation	Integral Squared Error(ISE), Root Mean Squared Error (RMSE)
Joshi et al. (2023)	Batch process control	SAC, DDPG	Partial system state	Absolute tracking error scaled by time	Setpoint tracking error, Disturbance rejection performance
Spielberg et al. (2019)	Self-driving industrial processes (HVAC)	DPG	Partial system state	Negative sum of absolute errors (SAE)	Integral Absolute error in case studies
Bao et al. (2021)	Learning performance in process control	DDACP	Partial system state	Weighted quadratic performance index balancing tracking errors with control	Average Returns
Dutta and Upreti (2022)	Survey of actor–critic methods	Comparison of DDPG, TD3, SAC, PPO, TRPO	Partial system state	Sparse reward on β -weighted performance index	Integral Absolute Error (IAE), Settling time
Oh (2024)	RL vs. MPC for chemical processes	DDPG, TD3, SAC	Partial system state	Negative of the MPC-style stage cost	Mean and standard deviation of the cost over multiple Monte-Carlo runs
Tavakkoli et al. (2024)	Setpoint tracking of non-minimum phase systems	DDPG	Partial system state	Negative LQI cost function	IAE; Overshoot; Settling time
Tang et al. (2025)	Robotics	DDPG, PPO, SAC, TRPO, QR-DQN	Partial system state	Task-specific rewards	Success rate; Sample efficiency
Haarnoja et al. (2018b)	Continuous control	SAC	Full system state	Task reward augmented with entropy maximisation term	Average return; Sample efficiency; Stability across random seeds
Recht (2019)	Continuous control	Policy gradient, Actor–critic	Full/Partial system state	Quadratic cost analogies (LQR-style objectives)	Stability analysis, Control performance insights

1. **Tracking Metrics**: These quantify how accurately the controller follows the reference signal, including both transient behaviour and steady-state performance, as well as its ability to reject disturbances. Metrics include rise time (t_r), maximum overshoot (M_p) (or undershoot, M_u, where applicable), settling time (t_s), steady-state error (e_{ss}), integral of squared error (ISE), and integral of time-weighted absolute error (ITAE).
2. **Control Signal Quality**: These measure characteristics of the control signal that directly affect energy consumption, actuator wear, and implementation feasibility. The criteria in this category include the integral of absolute control effort (IACE), the integral of absolute control effort rate (IACER), and the maximum control value ($u_{\max}$).
3. **Robustness Characteristics**: These assess the controller's ability to maintain stability and performance in the presence of uncertainties, disturbances, and parameter variations. Relevant metrics include gain margin (GM) and delay margin (DM).

This is based on the understanding that a controller achieving perfect tracking, yet requiring excessive control effort or lacking robustness to uncertainties, holds limited practical value. The selection of these specific metrics is justified as follows. Rise time, overshoot, and settling time are the canonical transient response metrics used in industrial controller specification and acceptance testing, providing direct interpretability for control engineers. ISE penalises large instantaneous errors and is particularly sensitive to poor transient behaviour, whilst ITAE emphasises persistent errors that occur later in the response, providing complementary views of tracking quality that neither metric captures alone. IACE quantifies total energy consumption, which is directly relevant to actuator wear and operational cost, whilst IACER captures control signal roughness that can excite unmodelled high-frequency dynamics and reduce actuator lifespan. Gain margin and delay margin were selected as robustness metrics because they provide frequency-domain insight into stability reserves that is directly comparable across algorithms and immediately interpretable in engineering terms, unlike stochastic robustness metrics. Together these metrics provide a multi-dimensional characterisation of controller behaviour that cannot be reduced to any single scalar without loss of important information. Table 1 presents a comparison of existing research in DRL control analysis, highlighting limitations in evaluation scope. The evaluation criteria are summarised in Table 2, and detailed definitions can be found in Sarhadi (2025). Most studies in this field either offer only qualitative assessments (Buşoniu et al., 2018), rely on visual comparisons, or restrict their evaluation to tracking metrics such as the Integral of Absolute Error (IAE).

To address these limitations, our analysis approach encompasses all three performance categories using the metrics in Table 2, with detailed definitions available in Sarhadi (2025).

For consistency throughout this paper, the following notation is adopted. The reference signal is denoted $r(t)$, the system output $y(t)$, and the tracking error $e(t) = r(t) - y(t)$. The control input is denoted u_c. The plant state vector is $x_p \in \mathbb{R}^n$ and the augmented state vector including the integral state is $x \in \mathbb{R}^{n+1}$. All integral performance metrics are evaluated over the simulation horizon T. Stability margins are computed from the open-loop transfer function $G(j\omega)$ at the phase crossover frequency ω_{pc} and gain crossover frequency ω_{cg} respectively.

This multi-faceted evaluation approach enables identification of fundamental performance trade-offs inherent in different control strategies, providing engineers with quantitative guidance for selecting algorithms based on application-specific priorities such as speed, accuracy, energy efficiency, or robustness requirements. These metrics apply equally to regulation and tracking problems, as demonstrated in our multi-step reference experiments, with transient metrics evaluated at each reference change and integral metrics capturing cumulative tracking performance throughout the trajectory.

4. Simulation studies

This section presents simulation results and analysis for our benchmark problems, comparing the performance of DDPG, TD3, PPO, and TD-MPC2 algorithms across multiple scenarios and evaluation criteria.

4.1. Simulation setup

The simulation environment was implemented in Python using PyTorch for the deep learning components and Control Systems Library

Table 2
Performance criteria for comprehensive controller evaluation.

Metric	Mathematical definition	Engineering significance	Unit
t_r	Rise time (10% to 90% of reference)	Speed of initial response	s
M_p	Maximum percentage overshoot	Peak transient above reference	%
t_s	Settling time to 2% of final value	Time to reach steady-state	s
e_{ss}	$\lim_{t\to\infty}\lvert r(t)-y(t)\rvert$	Steady-state tracking accuracy	–
ISE	$\int_0^T e^2(t)\,dt$	Penalises large tracking errors	–
ITAE	$\int_0^T t\,\lvert e(t)\rvert\,dt$	Emphasises persistent late errors	–
IACE	$\int_0^T \lvert u(t)\rvert\,dt$	Total control effort and energy	–
IACER	$\int_0^T \left\lvert \frac{du}{dt}\right\rvert dt$	Control signal smoothness	–
u_{max}	$\max_t \lvert u(t)\rvert$	Peak control demand	–
GM	$20\log_{10}\left(\frac{1}{\lvert G(j\omega_{pc})\rvert}\right)$	Stability robustness to gain variations	dB
DM	$\frac{PM}{\omega_{cg}}$	Robustness to time delays	s

for the dynamic systems simulation. For reproducibility and consistency across benchmarks, we maintained standardised neural network architectures and optimisation procedures if possible whilst adapting hyperparameters for each specific control task. Hyperparameter choices were kept consistent where appropriate and otherwise selected according to the standard recommended settings for each algorithm to support a fair and reproducible comparison.

The architectural choices are justified as follows. A two-layer network with 64 hidden units per layer was selected for all algorithms based on the low-dimensional nature of the state and action spaces in the benchmark problems considered, where deeper or wider networks have been shown to provide no significant benefit whilst increasing training instability (Henderson et al., 2018). ReLU activations were chosen for the hidden layers due to their well-documented advantages in avoiding vanishing gradients and promoting sparse representations (Arulkumaran et al., 2017). Tanh activations were applied at the output layer to bound the action outputs within the normalised control range $[-1, 1]$, ensuring compatibility with the actuator constraints of each benchmark system. The discount factor $\gamma = 0.99$ was selected to encourage long-horizon planning appropriate for regulation tasks where sustained tracking performance is required. Separate learning rates for actor (10^{-4}) and critic (10^{-3}) networks follow the established practice of using a faster critic update to provide stable value estimates before policy improvement (Lillicrap et al., 2015). These choices represent well-established defaults in the continuous control literature and were kept consistent across algorithms to ensure that observed performance differences reflect algorithmic properties rather than architectural advantages.

Optimiser and weight initialisation

All trainable parameters were optimised using the Adam optimiser (Kingma and Ba, 2014) with default momentum coefficients $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. Network weights were initialised using uniform initialisation (Glorot and Bengio, 2010) for all hidden layers, with biases initialised to zero. The output layer of each actor network was initialised with weights drawn from a uniform distribution $\mathcal{U}(-3\times 10^{-3}, 3\times 10^{-3})$ to ensure near-zero initial actions, preventing large initial control signals from destabilising the plant.

Reward function

To maintain a fair cross-paradigm comparison, all DRL algorithms were trained using the negative of the LQI cost function as the reward signal:

$$r_t = -\left(\mathbf{x}_t^\top Q\,\mathbf{x}_t + u_t^\top R\,u_t\right) \tag{8}$$

where $\mathbf{x}_t$ is the augmented state vector (plant states plus integral of tracking error) and u_t is the control input at timestep t. This formulation directly aligns the RL optimisation objective with the LQI cost, ensuring observed performance differences reflect algorithmic properties rather than reward shaping differences.

State observation

All algorithms receive the full augmented state as the observation. For each benchmark, the observation vector $\mathbf{o}_t$ consists of the plant state $x_p \in \mathbb{R}^n$, the tracking error $e_t = r - y_t$, and the accumulated integral of error $\int_0^t e\,d\tau$. This is consistent with the LQI augmented state formulation, ensuring all algorithms have access to the same information as the classical baseline.

Exploration and warmup

For off-policy algorithms (DDPG, TD3, TD-MPC2), training begins with a warmup phase of 5000 timesteps during which actions are sampled uniformly at random from the action space to pre-populate the replay buffer before any gradient updates are performed. TD3 uses zero-mean Gaussian exploration noise $\mathcal{N}(0, 0.2)$ clipped to $[-0.5, 0.5]$ during training, and adds target policy smoothing noise $\mathcal{N}(0, 0.2)$ clipped to $[-0.5, 0.5]$ to target actions during critic updates. TD-MPC2 augments its MPPI-planned actions with Gaussian perturbation noise $\mathcal{N}(0, 0.3)$ during training. PPO, as an on-policy algorithm, does not use a replay buffer; instead, stochasticity is inherent in its Gaussian policy $\pi_\theta(a|s) = \mathcal{N}(\mu_\theta(s), \sigma^2)$ where σ is a learned state-independent parameter. For the classical baseline, the LQI controller was implemented using identical weighting matrices ($Q = 10I$, $R = 1$) across all benchmark systems to ensure consistency and avoid system-specific bias in the baseline controller, where I is the scaled identity matrix. Table 3 presents the hyperparameter configuration for the Non-minimum phase system (NMPs) benchmark, which follows the same structure used across all experiments. Comprehensive hyperparameter settings for all other benchmarks are available in our public repository.[1]

All DRL algorithms were trained for 5 independent runs using random seeds $\{0, 1, 2, 3, 4\}$. The results presented throughout Section 4 correspond to the best-performing checkpoint across these 5 seeds for each algorithm, selected based on the highest mean evaluation return during training. This approach ensures that the reported performance represents the capability of each algorithm rather than a single potentially unrepresentative run.

Initial conditions and preprocessing

All simulations start from rest with zero initial conditions $x_p(0) = \mathbf{0}$. No normalisation is applied to observations or rewards. Actions bounded to $[-1, 1]$ via Tanh output activation are linearly rescaled to the system-specific range $[-u_{max}, u_{max}]$ before being applied to the plant.

4.2. Non-minimum phase CSTR plane

The first system is a Continuous Stirred-Tank Reactor (CSTR) process control benchmark with the following transfer function (Bequette, 2003):

$$G(s) = \frac{-1.135s + 3.199}{s^2 + 4.719s + 5.47} \tag{9}$$

The presence of the unstable zero introduces fundamental performance limitations and undershoot behaviour that significantly challenges control system design (Åström, 2000; Sarhadi, 2025). Despite these challenges being documented in classical control theory, the efficacy of deep reinforcement learning approaches in managing such

Table 3
DRL implementation parameters.

Parameter	DDPG	TD3	PPO	TD-MPC2
Total timesteps	200,000	200,000	200,000	200,000
Batch size	256	256	64	256
Discount factor (γ)	0.99	0.99	0.99	0.99
Target update (τ)	0.005	0.005	N/A	0.001
Actor lr	1e−4	1e−4	1e−4	1e−4
Critic lr	1e−3	1e−3	1e−3	1e−3
Hidden dim	64	64	64	128
Actor net	[64, 64]	[64, 64]	[64, 64]	[64, 64]
Critic net	[64, 64]	[64, 64] x2	[64, 64]	[64, 64]
Expl noise	OU Noise	Gaussian	N/A	Gaussian
Policy noise	N/A	0.2	N/A	N/A
Noise clip	N/A	0.5	N/A	N/A
Policy delay	1	2	N/A	N/A
Buffer size	100,000	100,000	N/A	100,000
Clip param	N/A	N/A	0.2	N/A
Sim dt	0.1 s	0.1 s	0.1 s	0.1 s
Ep length	25 s	25 s	25 s	25 s
Activation	ReLU	ReLU	ReLU	ReLU
Output act	Tanh	Tanh	Tanh	Tanh

Table 4
Performance results for non-minimum phase system.

Metric	LQI	DDPG	TD3	PPO	TD-MPC2
t_r (s)	3.90	2.30	1.70	0.90	0.60
M_p (%)	0.00	0.00	0.10	12.60	34.50
t_s (s)	7.70	6.10	3.30	5.50	8.10
e_{ss}	0.00	0.00	0.00	0.00	0.00
ISE	2.00	1.20	1.10	1.00	1.20
ITAE	5.60	2.80	1.30	1.80	3.80
IACE	39.90	41.70	42.30	43.40	44.40
IACER	1.65	1.35	1.40	2.20	3.50
u_{max}	1.65	1.73	1.81	2.10	2.64
GM	40.00	2.14	2.51	1.90	1.83
DM	1.70	1.61	1.58	0.90	0.7

dynamics remains relatively unexplored (Tavakkoli et al., 2024). Algorithms are trained with parameters in Table 3.

All controllers were implemented using the same environment, which provides a consistent interface for training and evaluating the different algorithms across all benchmark problems. The environment encapsulates system dynamics, reference generation, disturbance application, and performance measurement. For each benchmark system, the appropriate state–space model was integrated within this environment framework.

4.2.1. Performance results

Fig. 1 shows the simulation results for the non-minimum phase system. Table 4 presents the performance metrics for the non-minimum phase CSTR system. All controllers exhibit the characteristic inverse response, with initial movement opposite to the reference direction before convergence.

TD-MPC2 achieves the fastest rise time (0.60 s) but with substantial overshoot (34.50%) and the highest control effort ($u_{max} = 2.64$). PPO demonstrates rapid response (0.90 s) with moderate overshoot (12.60%) and the best ISE performance (1.00) among all controllers. TD3 delivers balanced performance with minimal overshoot (0.10%), excellent settling time (3.30 s), and superior ITAE (1.30), indicating effective error minimisation over time. DDPG shows conservative behaviour similar to TD3 but with slower dynamics.

The baseline LQI controller exhibits zero overshoot but significantly slower response ($t_r = 3.90$ s, $t_s = 7.70$ s) and requires the lowest control effort (IACE = 39.90). All DRL algorithms demand higher control activity than LQI, with IACE values ranging from 41.70 to 44.40, reflecting the trade-off between response speed and actuator demands.

Robustness analysis reveals LQI's superior gain margin (40.00) compared to DRL algorithms (1.83–2.51). Among DRL controllers, TD3 achieves the highest gain margin (2.51) whilst DDPG provides the best delay margin (1.61), suggesting varying robustness characteristics across algorithms.

4.2.2. Robustness under perturbed conditions

Under perturbed conditions with step disturbance (0.20 at $t = 15$ s) and measurement noise ($\sigma = 0.20$ at $t = 20$ s) as shown in Fig. 2, all controllers maintain stability whilst exhibiting distinct recovery characteristics. DRL algorithms demonstrate faster disturbance recovery than LQI, with TD3 returning to within 2% of reference in approximately 3.50 s compared to LQI's 6.00 s. When noise is introduced, TD3 maintains the tightest output bounds whilst TD-MPC2 shows increased sensitivity with larger variations.

The results highlight fundamental trade-offs in non-minimum phase control: aggressive controllers (TD-MPC2, PPO) achieve faster response at the cost of overshoot and control effort, whilst conservative approaches (TD3, DDPG, LQI) prioritise stability. TD3 emerges as the most balanced solution, combining reasonable speed, minimal overshoot, and robust disturbance rejection, critical attributes for practical implementation in chemical process control where both performance and stability are essential.

4.2.3. Analysis of performance trade-offs

The results highlight fundamental trade-offs in controlling non-minimum phase systems. A clear inverse relationship between rise time and overshoot is evident: TD-MPC2 and PPO achieve the fastest responses (0.60 s, 0.90 s) but with significant overshoots (34.50%, 12.60%), whilst TD3 and DDPG maintain minimal overshoot at the cost of slower rise times.

Among DRL controllers, TD3 achieves the highest gain margin (2.51) whilst DDPG provides the best delay margin (1.61), indicating different robustness characteristics across algorithms.

The comparative analysis indicates that controller selection depends on application priorities: PPO for minimal tracking error, TD3 for balanced performance with robustness, TD-MPC2 for fastest response, or LQI for minimal control effort. For most practical applications, TD3's combination of reasonable speed, minimal overshoot, and robust margins provides the most versatile solution. These performance differences can be explained from the architecture of each algorithm, TD3's twin-critic architecture reduces overestimation bias, producing more deliberate and smooth policies that avoid overcorrection on the inverse response, explaining why it excels in ITAE and minimal overshoot. PPO's clipped surrogate objective prevents destructively large policy updates, resulting in conservative responses that naturally limit overshoot at the cost of slower rise times. TD-MPC2's planning horizon enables anticipatory control that commits to aggressive early actions, explaining its fast rise time but substantial overshoot on the non-minimum phase dynamics.

4.3. Two mass spring benchmark problem

The second benchmark considered is the infamous ACC 1992 benchmark problem (Wie and Bernstein, 1992b), Two Mass Spring System(TMS). First introduced in 1992 as a touchstone benchmark in testing control systems, it is important because of the non-collocated actuator and sensor dynamics, making it a prevalent benchmark in control literature (Wie and Bernstein, 1992a). Its dynamics are represented in state–space form as:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\frac{k}{m_1} & \frac{k}{m_1} & 0 & 0 \\ \frac{k}{m_2} & -\frac{k}{m_2} & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{1}{m_1} \\ 0 \end{bmatrix} u \tag{10}$$

[1] https://github.com/Klins101/Performance-Analysis-of-Deep-Reinforcement-Learning-for-Applied-Control-Applications.git.

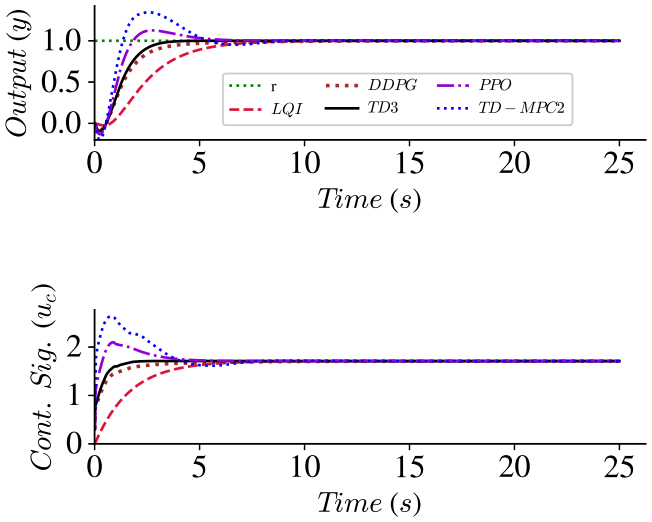


Fig. 1. Results for the NMP system in ideal conditions.

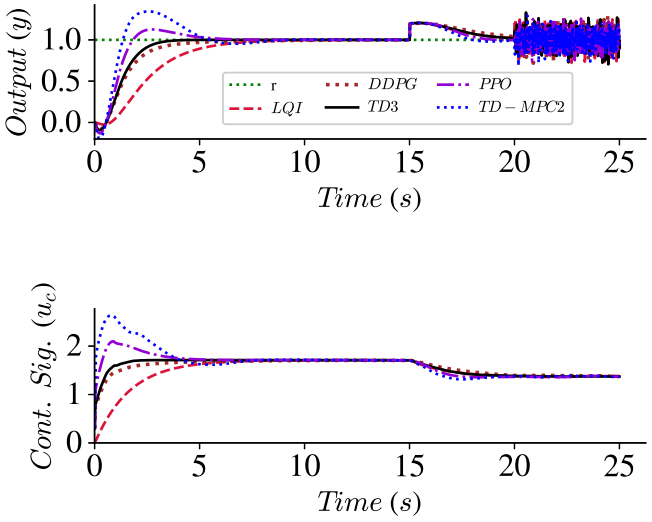


Fig. 2. Results for NMP in the perturbed test.

where x_1 and x_2 are the positions of body 1 and body 2, respectively; x_3 and x_4 are the velocities; u is the control input acting on body 1.

For our nominal system, we set $m_1 = m_2 = 1$ and $k = 1$ with appropriate units. It is challenging and used as a main benchmark in numerous control studies

4.3.1. Performance results

The nominal step response results presented in Fig. 3 reveal distinctly different performance characteristics among the evaluated controllers. The quantitative analysis confirms significant algorithmic differences in handling the system's oscillatory dynamics and resonant modes.

Table 5 presents the nominal performance metrics. LQI achieves the fastest rise time (1.40 s) with moderate overshoot (11.20%), leveraging model-based design for precise dynamics exploitation. However, it requires the highest control effort (IACE = 15.18) whilst maintaining superior control smoothness (IACER = 45.86).

Among DRL algorithms, TD3 demonstrates precision with minimal overshoot (0.30%) and the lowest maximum control effort ($u_{\max} = 1.18$). Its twin-critic architecture effectively mitigates overestimation bias, achieving the smoothest control signals (IACER = 10.91) at the cost of slower rise time (2.50 s). PPO provides balanced performance with reasonable transient response (1.90 s rise time, 8.80% overshoot) and moderate control effort, whilst DDPG and TD-MPC2 exhibit more conservative behaviours.

Tracking performance analysis reveals LQI's superiority with the lowest ISE (6.04) and ITAE (10.07), reflecting optimal design benefits. DRL algorithms show higher tracking errors (ISE: 7.13–8.56) but significantly lower control efforts (IACE: 4.52–6.79), highlighting the fundamental trade-off between tracking precision and actuator demands.

4.3.2. Robustness under perturbed conditions

Under perturbed conditions incorporating white noise and external disturbances (Fig. 4), the relative performance characteristics reveal critical insights into algorithm robustness and practical applicability. The dramatic shift in performance rankings under uncertainty demonstrates the importance of evaluating controllers beyond nominal conditions.

Table 6 quantifies performance under disturbances and measurement noise. All controllers experience degradation, with ISE values increasing by 7.6–10.5% for LQI and DRL algorithms. LQI maintains the best tracking (ISE = 6.5) but requires increased control effort (IACE rising from 15.18 to 19.6).

TD3 exhibits superior control smoothness under perturbations (IACER = 43.3), significantly outperforming other algorithms. This

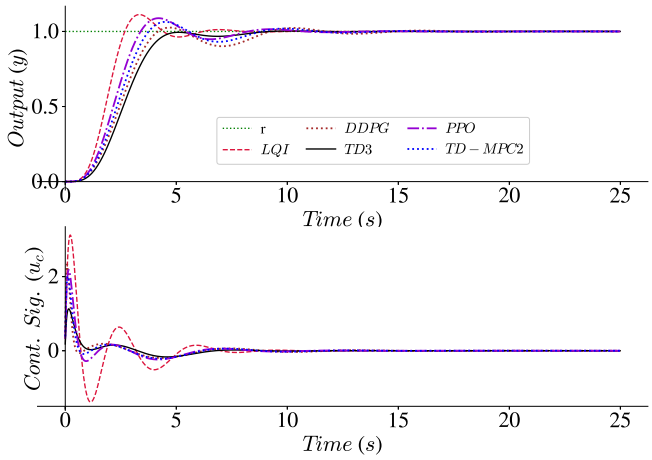


Fig. 3. Step response comparison for TMS system showing nominal performance characteristics.

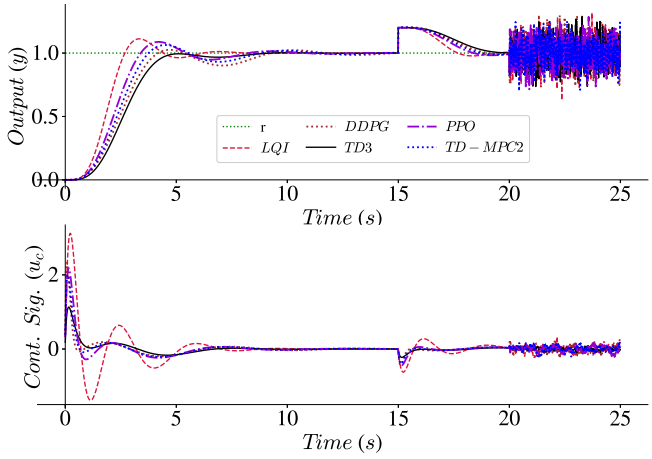


Fig. 4. TMS system performance under perturbed conditions with disturbances and measurement noise.

Table 5
Performance metrics for TMS system under nominal conditions.

Metric	LQI	DDPG	TD3	PPO	TD-MPC2
t_r (s)	1.40	2.30	2.50	1.90	2.10
M_p (%)	11.20	2.60	0.30	8.80	6.40
t_s (s)	5.70	10.70	7.90	7.70	8.30
e_{ss}	0.00	0.00	0.00	0.00	0.00
ISE	6.04	7.92	8.56	7.13	7.58
ITAE	10.07	24.91	18.24	16.02	19.76
IACE	15.18	5.73	4.52	6.79	5.84
IACER	45.86	21.34	10.91	22.37	19.15
u_{max}	3.14	2.32	1.18	2.26	2.03
GM (dB)	21.15	1.43	2.85	1.95	2.45
DM (s)	0.17	0.25	0.52	0.35	0.48

Table 6
Performance metrics for TMS system under perturbed conditions.

Metric	LQI	DDPG	TD3	PPO	TD-MPC2
ISE	6.5	8.5	9.0	7.5	8.0
ITAE	70.5	93.7	89.3	81.0	86.0
IACE	19.6	8.1	5.9	8.9	7.8
IACER	106.2	109.1	43.3	80.7	86.5

robustness, combined with the lowest control effort (IACE = 5.9), validates the twin-critic architecture's effectiveness for uncertain environments. PPO demonstrates balanced robustness with moderate degradation across all metrics, whilst DDPG shows the highest sensitivity to disturbances.

4.3.3. Analysis of performance trade-offs

The TMS benchmark reveals distinct control philosophies: LQI prioritises tracking accuracy through high control effort, achieving superior nominal performance but requiring substantial actuator activity. TD3 optimises control economy and smoothness, sacrificing response speed for precision and robustness critical attributes for mechanical systems with actuator constraints.

Robustness analysis confirms LQI's superior gain margin (21.15 dB) compared to DRL algorithms (1.43–2.85 dB), whilst TD3 achieves the best delay margin (0.52 s) among DRL controllers. These metrics suggest complementary robustness properties: classical control excels in parameter variations whilst TD3 handles time delays more effectively.

The results demonstrate that controller selection depends on application priorities. LQI suits applications demanding precise tracking with available actuator capacity, TD3 excels where control smoothness and economy are paramount, whilst PPO offers reliable balanced performance. These differences could lie in the algorithms' update rules. TD3's delayed policy updates allow value estimates to stabilise before policy improvement, producing smooth control signals that avoid exciting the resonant mode of the two-mass spring system, explaining its superior IACER. LQI's full model knowledge allows it to exploit the system dynamics directly, achieving faster rise time but at substantially higher control cost. PPO's on-policy learning continuously updates from fresh trajectory data, providing consistent performance across the oscillatory dynamics without accumulating overestimation errors.

Table 7
Model parameters and state variables used in AUV dynamics.

Param./Var.	Description	Value	Param./Var.	Description	Value
m (kg)	Vehicle mass	30.50	x_G (m)	CG position (long.)	0.00
I_z (kg m^2)	Yaw inertia	3.45	$Y_{\dot{v}}$ (kg m/rad)	Added Mass (sway)	−35.50
Y_{uv} (kg/m)	Body lift force and fin lift	−28.60	$Y_{v\|v\|}$ (kg/m)	Cross-flow Drag (sway)	−1310.00
Y_{ur} (kg/rad)	Added mass cross term and fin lift	5.22	$Y_{r\|r\|}$ (kg m^2/rad)	Cross-flow drag (yaw)	0.63
$Y_{uu\delta_r}$ (kg/m rad)	Fin lift force	21.37	$N_{\dot{v}}$ (kg m)	Added mass (sway-yaw)	1.93
$N_{\dot{r}}$ (kg m^2/rad)	Added mass (yaw)	−4.88	$N_{v\|v\|}$ (kg)	Cross-flow drag (sway)	−3.18
$N_{r\|r\|}$ (kg m^2/rad^2)	Cross-flow drag (yaw)	−94.00	$N_{uu\delta_r}$ (kg/rad)	Body and fin lift and munk moment	−17.50
N_{uv} (kg)	Fin lift moment (surge-sway)	−24.00	N_{ur} (kg m/rad)	Added mass cross term and fin lift	−2.00
u_0 (m/s)	Forward speed (constant)	1.50	δ_r (rad)	Rudder angle (input)	NA
v (m/s)	Sway velocity (state)	NA	r (rad/s)	Yaw rate (state)	NA
ψ (rad)	Heading angle (state)	NA			

4.4. Nonlinear Autonomous Underwater Vehicle (AUV)

The AUV benchmark represents the complex nonlinear yaw dynamics of the REMUS AUV (Prestero, 2001) incorporating realistic hydrodynamic effects including quadratic damping, velocity-dependent coefficients, and cross-coupling between sway and yaw motions. The system dynamics are governed by:

$$\begin{cases} (m - Y_{\dot{v}})\dot{v} + (mx_G - Y_{\dot{r}})\dot{r} = \left(Y_{v|v|}|v| + Y_{uv}u_o\right) v \\ \quad + \left(Y_{r|r|}|r| + (Y_{ur} - m)u_o\right) r + Y_{uu\delta_r}u_o^2\delta_r \\ (mx_G - N_{\dot{v}})\dot{v} + (I_z - N_{\dot{r}})\dot{r} = \left(N_{v|v|}|v| + N_{uv}u_o\right) v \\ \quad + \left(N_{r|r|}|r| + (N_{ur} - mx_G)u_o\right) r + N_{uu\delta_r}u_o^2\delta_r \\ \dot{\psi} = r \end{cases} \tag{11}$$

All parameters and variables of this model are defined in Table 7 and are based on empirically validated REMUS AUV from Prestero (2001). This benchmark represents a challenging control problem on the realistic nonlinear model of an AUV with inherent non-BIBO stability, added mass effects, and significant nonlinearities that require active stabilisation, making it representative of real-world maritime control scenarios. To comprehensively evaluate controller performance, we incorporate realistic operational constraints including actuator amplitude and rate saturations ($|\delta_r| \leq 20$, $|\dot{\delta}_r| \leq 30$), unmodelled dynamics representing worst-case uncertainties, and environmental disturbances typical of maritime operations. These conditions test the algorithms' ability to handle the practical limitations encountered in autonomous underwater vehicle deployment beyond idealised simulation scenarios.

4.4.1. Performance analysis

Fig. 5 reveals distinctly different performance characteristics across all evaluated algorithms, with significant variations in transient response and control aggressiveness. The quantitative results demonstrate the algorithms' varying degrees of adaptation to the complex hydrodynamic dynamics inherent in marine vehicle control.

TD3 exhibits the most aggressive response characteristics with the fastest rise time (0.3 s) and rapid convergence, though this performance comes at the cost of significant overshoot (8.6%) and the highest control effort demands.

DDPG achieves similarly rapid response with competitive rise time (0.5 s) and substantial overshoot (10.1%), demonstrating effective handling of the nonlinear dynamics through aggressive control strategies. PPO delivers more conservative performance with moderate rise time (1.3 s) and overshoot (9.9%), reflecting its inherent policy stability mechanisms that provide robustness at the expense of transient aggressiveness.

TD-MPC2 exhibits notably conservative response characteristics with slower rise time (1.4 s) compared to other DRL methods, though achieving excellent overshoot control (3.9%) that approaches the classical baseline. This behaviour reflects the model-based approach's tendency towards cautious control strategies when operating with learned dynamics models, prioritising stability over aggressive performance (see Table 8).

Table 8
Performance results for nonlinear AUV system.

Metric	LQI	DDPG	TD3	PPO	TD-MPC2
t_r (s)	2.3	0.5	0.3	1.3	1.4
M_p (%)	5.5	10.1	8.6	9.9	3.9
t_s (s)	6.6	4.5	2.8	3.9	3.5
e_{ss}	0.00	0.00	0.00	0.00	0.00
ISE	15.2	2.1	1.7	6.5	7.4
ITAE	31.2	7.1	3.3	9.6	8.2
IACE	11.0	2.9	3.9	1.5	1.3
IACER	10.0	43.9	65.5	11.1	20.3
u_{max}	0.5	1.9	2.7	0.6	0.6
GM	25.30	16.20	15.85	16.45	13.32
DM	0.69	0.35	0.32	0.38	0.42

The control effort analysis reveals significant disparities in actuator demands across algorithms. LQI maintains exceptionally smooth control signals with minimal peak demand (0.5), reflecting its optimal design principles but at the cost of sluggish response characteristics. Conversely, DRL algorithms exhibit substantially higher control activity, with TD3 demanding maximum control authority (2.7) and the highest control rate variations (IACER = 65.5), indicating aggressive policy behaviour that may challenge actuator limitations in practical implementations.

4.4.2. Disturbance and noise rejection

Table 9 presents controller performance under perturbed conditions incorporating measurement noise and external disturbances. All controllers exhibit minimal performance degradation, with ISE values increasing by factors of 1.04–1.12 compared to nominal conditions.

LQI shows minimal degradation with ISE increasing from 15.2 to 15.8, whilst maintaining moderate control effort (IACE = 13.2) amongst all controllers. The DRL algorithms demonstrate varying robustness characteristics: TD3 achieves the best tracking performance (ISE = 1.9) despite disturbances, whereas TD-MPC2 exhibits slightly higher degradation with ISE increasing from 7.4 to 7.6.

The control activity metrics reveal significant increases across all algorithms, with IACER values exceeding 140 for most DRL controllers. TD-MPC2 shows the highest control rate activity (IACER = 149.9), indicating aggressive compensation strategies that may challenge actuator bandwidth limitations. PPO maintains moderate performance degradation (ISE = 6.7) with intermediate control effort, suggesting a balanced robustness-performance trade-off. These results as shown in Fig. 6 highlight the inherent challenge of maintaining smooth control under uncertainty, regardless of the control paradigm employed.

4.4.3. Performance evaluation under operational challenges

To assess controller performance under realistic operational constraints, a comprehensive assessment incorporating actuator saturations, model uncertainties, and external perturbations was conducted. The reference amplitude is increased to 10 to activate actuator saturation limits of $|\delta_r| \leq 20$ and rate constraints of $|\dot{\delta}_r| \leq 30$/s. Model uncertainty is introduced through unmodelled dynamics $G_u(s) = 225/(s^2 +$

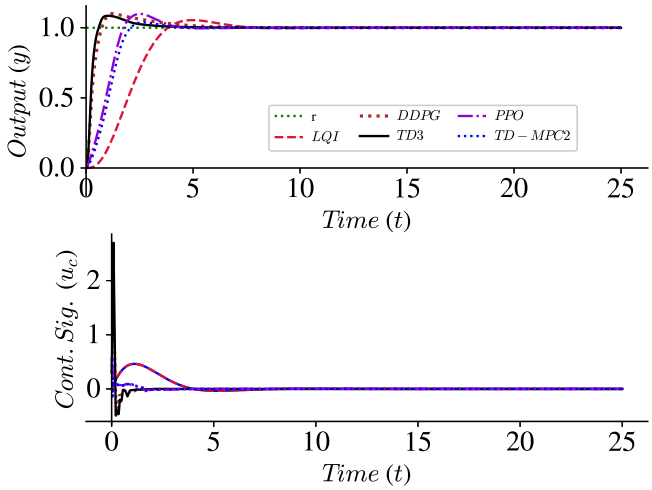


Fig. 5. Step response comparison for nonlinear AUV system.

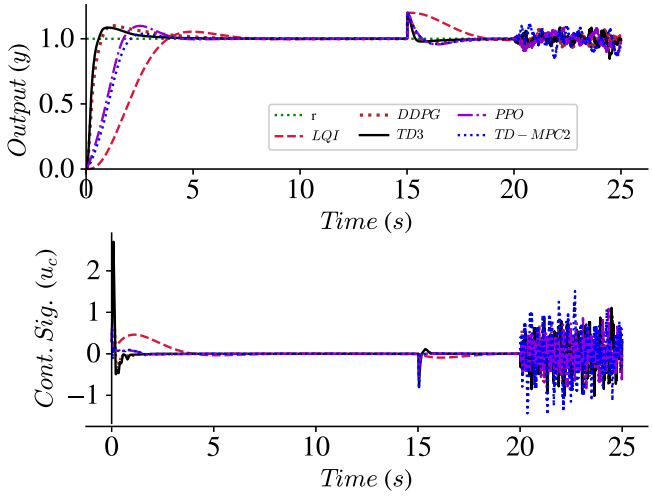


Fig. 6. AUV performance under perturbed conditions with measurement noise and external disturbances.

Table 9
Performance metrics for AUV system under perturbed conditions.

Metric	LQI	DDPG	TD3	PPO	TD-MPC2
ISE	15.8	2.3	1.9	6.7	7.6
ITAE	107.3	52.8	48.1	72.8	71.0
IACE	13.2	15.3	20.0	14.6	25.3
IACER	12.4	141.2	143.8	142.9	149.9

$12s + 225$), representing Rohrs' example (Rohrs et al., 2003), a worst-case scenario known to challenge adaptive controllers. External perturbations include a step disturbance of 2 at $t = 15$ s and measurement noise with $\sigma = 0.05$ at $t = 20$ s.

Fig. 7 illustrates the controller responses under these challenging conditions. The LQI controller exhibits sensitivity to the windup phenomenon, resulting in unbounded oscillations. In contrast, DDPG and TD3 show aggressive initial responses with overshoots of 57.87% and 46.47% respectively, but successfully converge with minimal steady-state error. PPO and TD-MPC2 adopt conservative strategies, limiting maximum control efforts to 8.58 and 12.48 respectively.

Table 10 quantifies the performance differences. TD3 achieves optimal tracking (ISE = 3.74, ITAE = 2.66), whilst PPO prioritises control economy (IACE = 2.02) with superior stability margins (GM = 7.55 dB, DM = 0.37 s). Notably, the DRL algorithms were not trained for these extreme conditions, yet demonstrate impressive adaptability in handling simultaneous saturations, unmodelled dynamics, and disturbances, validating their implicit robustness acquired through stochastic training environments.

Table 10
Controller simulation results, under operational challenges for the AUV System.

Metric	DDPG	TD3	PPO	TD-MPC2
t_r (s)	1.32	1.34	2.14	2.07
M_p (%)	57.87	46.47	23.28	20.00
t_s (s)	17.69	23.12	24.98	24.58
e_{ss}	0.04	0.05	0.10	0.05
ISE	4.82	3.74	4.57	3.73
ITAE	3.02	2.66	5.18	2.95
IACE	2.44	3.51	2.02	2.94
IACER	139.84	280.14	174.32	333.24
$u_{\max}$	20.00	19.57	8.58	12.48
GM	2.80	3.15	7.55	3.28
DM	0.16	0.11	0.37	0.22

4.4.4. Performance tradeoffs and practical implications

The AUV benchmark reveals fundamental engineering tradeoffs between transient performance, control authority requirements, and robustness to operational constraints. Under nominal conditions, TD3 achieves exceptional settling times and error minimisation at the cost of

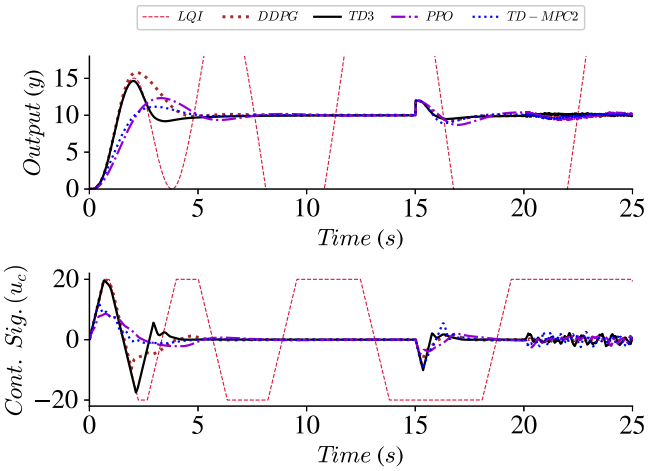


Fig. 7. Controller performance comparison with operational challenges.

substantial control effort. However, the operational challenges assessment reveals a critical distinction: whilst classical LQI control provides conservative nominal performance, it exhibits catastrophic failure under combined saturations and model uncertainty (Rohrs test), highlighting fundamental limitations of linear control for constrained nonlinear systems.

DRL algorithms demonstrate resilience across all operating conditions. DDPG provides an optimal compromise between aggressive performance and robustness, maintaining stable operation even under severe operational constraints. PPO's conservative strategy, initially appearing suboptimal in nominal conditions, proves advantageous under challenging scenarios with superior stability margins (GM = 7.55 dB). TD-MPC2's model-based approach achieves balanced performance, though with increased control activity under perturbations.

The stark contrast between nominal and constrained operation underscores a critical insight for maritime applications: robustness to operational constraints outweighs nominal performance metrics. It should be noted that the Rohrs test result should be interpreted as evidence of implicit robustness acquired through stochastic training rather than as a claim of superiority over dedicated robust control methods, which were not the focus of this benchmarking study. The DRL algorithms' ability to implicitly handle saturations, unmodelled dynamics, and disturbances without explicit anti-windup or robust control modifications represents a paradigm shift from classical approaches. This capability proves essential for autonomous marine systems operating in unpredictable oceanic environments where actuator limits, model uncertainties, and environmental disturbances are unavoidable rather than exceptional conditions.

4.5. Statistical analysis and training characteristics

To complement the deterministic performance evaluation presented in the preceding sections, this subsection reports training characteristics and statistical performance metrics for the nonlinear AUV benchmark across 5 independent random seeds. The AUV was selected as the representative system due to its nonlinear dynamics and the diversity of algorithmic behaviours observed, making it the most informative benchmark for assessing training variability and computational feasibility.

Table 11 reports computational cost and inference latency. PPO is the most computationally efficient with a training time of 243.3 ± 3.3 s and the lowest inference latency of 0.041 ± 0.002 ms, whilst TD-MPC2 requires significantly longer training (3425.2 ± 64.6 s) due to its planning procedure. All algorithms achieve inference latencies well below 1 ms, confirming real-time deployability across all evaluated methods. DDPG and TD3 offer the best balance between computational cost and performance, whilst TD-MPC2's higher cost is justified by its superior constraint handling demonstrated in the operational challenges test.

Table 11
Computational cost and runtime (mean ± std, 5 seeds).

Metric	DDPG	TD3	PPO	TD-MPC2
Training time (s)	485.6 ± 21.2	487.1 ± 22.1	243.3 ± 3.3	3425.2 ± 64.6
Inference latency (ms)	0.048 ± 0.000	0.048 ± 0.002	0.041 ± 0.002	0.087 ± 0.003
Parameter count	$18\,564 \pm 0$	$27\,910 \pm 0$	9219 ± 0	$73\,095 \pm 0$

Table 12
Training stability — Reward variance (mean ± std, 5 seeds).

Metric	DDPG	TD3	PPO	TD-MPC2
Reward variance	82.88 ± 58.78	41.43 ± 18.32	172.02 ± 50.95	657.08 ± 266.94

Table 12 reports training stability through reward variance across seeds. TD3 demonstrates the most stable training (41.43 ± 18.32) followed by DDPG (82.88 ± 58.78), confirming that the twin-critic architecture and delayed updates produce more consistent learning dynamics. TD-MPC2 exhibits the highest reward variance (657.08 ± 266.94), reflecting the sensitivity of model-based planning to random initialisation of the learned dynamics model. PPO shows moderate variance (172.02 ± 50.95) consistent with its on-policy learning dynamics.

Table 13 reports safety during learning through divergence count, the number of episodes where the controller produced unstable or unbounded responses during training. DDPG and TD3 exhibit the fewest divergences (105.6 ± 13.4 and 108.2 ± 10.5 respectively), demonstrating safer exploration behaviour. PPO shows substantially more divergences (1095.6 ± 42.8) due to its on-policy stochastic exploration, and TD-MPC2 exhibits the highest divergence count (3972.4 ± 2104.0) reflecting the instability of early model-based planning with poorly initialised dynamics models.

Table 14 reports mean ± standard deviation of classical control metrics across seeds. DDPG and TD3 demonstrate the fastest rise times (0.45 ± 0.08 s and 0.44 ± 0.11 s) with low variance, confirming consistent aggressive response behaviour. PPO achieves the lowest overshoot ($7.90 \pm 0.42\%$) with low variance, reflecting its stable on-policy learning. The low standard deviations across all metrics confirm that the best-checkpoint results reported in the main evaluation are representative of each algorithm's true capability rather than outlier performance.

4.6. Crazyflie quadrotor

The final benchmark test evaluates the real-time implementability of the controllers for the altitude control problem of a quadrotor drone,

Table 13
Safety during learning — divergence count (mean ± std, 5 seeds).

Metric	DDPG	TD3	PPO	TD-MPC2
Divergence count	105.6 ± 13.4	108.2 ± 10.5	1095.6 ± 42.8	3972.4 ± 2104.0

Table 14
Classical control performance metrics (mean ± std, 5 seeds).

Metric	DDPG	TD3	PPO	TD-MPC2
t_r (s)	0.45 ± 0.08	0.44 ± 0.11	1.39 ± 0.08	1.55 ± 0.06
M_p (%)	10.24 ± 2.77	11.07 ± 3.71	7.90 ± 0.42	1.98 ± 2.11
t_s (s)	3.40 ± 0.49	3.51 ± 0.80	3.51 ± 0.62	1.63 ± 2.41
e_{ss}	0.001 ± 0.001	0.000 ± 0.000	0.002 ± 0.001	0.001 ± 0.001

an aspect not considered in such detail in other papers. The Crazyflie 2.1+ nano-quadrotor (Bitcraze, 2021) represents a challenging benchmark, featuring inherent instability, fast dynamics, and strict real-time constraints that mirror industrial UAV control requirements. Fig. 8 presents the experimental configuration enabling direct validation of simulation-trained controllers on physical hardware.

The complete Crazyflie dynamics comprise a 12-state nonlinear model capturing full 6-DOF motion with complex aerodynamic coupling (Nguyen et al., 2023). However, to evaluate the generalisability and adaptability of the algorithms, we employ a reduced-order linear model that retains key control characteristics while eliminating nonlinearities, to assess whether the controllers can demonstrate simulation-to-real transfer capability. This experiment represents a realistic deployment scenario where the following practical conditions hold: the system state is fully observable via onboard sensors, the training model captures the dominant dynamics sufficiently for policy transfer. The successful transfer of all simulation-trained policies to physical hardware validates these conditions and provides a concrete benchmark for the transfer requirements necessary for real-world DRL deployment.

The vertical motion of this vehicle can be modelled using the Newton's second law: $m\ddot{z} = T - mg$, where z represents altitude, m is the drone mass, g is the gravitational acceleration, and T denotes the total thrust of propellers. The goal is to control altitude z by applying thrust T, with gravity g treated as a constant disturbance compensated by feedforward thrust during takeoff manoeuvre. Therefore, the vertical dynamics of the drone can be represented in state–space form as:

$$\begin{bmatrix}\dot{z}\\ \ddot{z}\end{bmatrix} = \begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}\begin{bmatrix}z\\ \dot{z}\end{bmatrix} + \begin{bmatrix}0\\ \frac{1}{m}\end{bmatrix}T + \begin{bmatrix}0\\ -g\end{bmatrix}, \quad (12)$$

$$y = z.$$

where the state vector comprises the altitude z and the vertical velocity $\dot{z}$, with the mass $m = 0.027$ kg.

4.6.1. Performance validation

The challenging multi-step reference sequence 0 → 0.5 m → 1.0 m → 0.5 m systematically evaluates tracking consistency across operating points whilst revealing the differences between idealised simulation and physical implementation. This comprehensive evaluation protocol provides unique insights into the practical viability of simulation-trained DRL controllers for safety-critical aerial systems.

Figs. 9, 10 presents a comparison validating simulation fidelity whilst highlighting the inevitable complexities of real-world deployment. The simulation results Fig. 9 demonstrate clean tracking performance across all algorithms, with quantitative metrics presented in Table 15. TD3 achieves the most aggressive transient response with rapid settling (2.6 s) despite moderate overshoot (37.5%), whilst LQI provides the most conservative approach with longer settling times (4.9 s) but acceptable overshoot characteristics (17.0%). PPO exhibits the highest overshoot (49.9%) among all controllers, reflecting its stochastic policy exploration, whilst TD-MPC2 demonstrates balanced performance with overshoot (42.6%) and competitive settling behaviour.

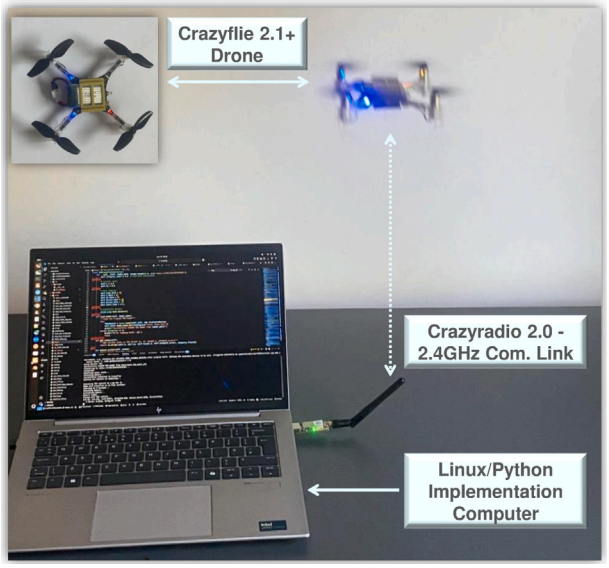


Fig. 8. Experimental setup for real-world validation of trained DRL controllers on the Crazyflie platform.

Table 15
Performance results for Crazyflie multi-step reference tracking (Simulation).

Metric	LQI	DDPG	TD3	PPO	TDMPC2
t_r (s)	0.23	0.20	0.25	0.30	0.21
M_p (%)	17.00	33.20	37.50	49.90	42.60
t_s (s)	4.90	2.90	2.60	4.10	3.30
e_{ss}	0.00	0.00	0.00	0.00	0.00
ISE	0.83	1.18	1.20	2.32	1.52
ITAE	37.11	39.28	33.93	60.31	47.64
IACE	0.79	0.66	0.81	0.83	0.71
IACER	6.92	6.90	7.37	3.98	4.37
u_{max}	0.30	0.40	0.30	0.10	0.20

Table 16
Performance results for Crazyflie multi-step reference tracking (Real hardware).

Metric	LQI	DDPG	TD3	PPO	TDMPC2
t_r (s)	0.30	0.64	0.62	0.60	0.80
M_p (%)	30.00	28.70	21.70	24.40	22.90
t_s (s)	3.60	3.50	3.00	4.30	4.00
e_{ss}	0.00	0.00	0.00	0.00	0.00
ISE	1.60	2.06	1.80	2.00	2.30
ITAE	53.40	63.10	50.10	55.20	58.70
IACE	4.40	3.40	2.60	2.70	2.90
IACER	18.0 0	8.00	6.40	4.3	4.00
u_{max}	0.50	0.20	0.20	0.10	0.10

The control effort analysis reveals distinct characteristics among algorithms. TD3 and LQI demonstrate similar maximum control demands (0.30), whilst PPO exhibits remarkably conservative control usage (0.10) at the cost of increased tracking errors. The integral control effort rates (IACER) indicate that LQI maintains the smoothest control signals (6.92), with TDMPC2 providing competitive smoothness (4.37) through its trajectory optimisation approach.

4.6.2. Real-world implementation results

The experimental validation on actual Crazyflie hardware (Table 16) reveals the inevitable but manageable performance degradation when transitioning from simulation to physical implementation. Most significantly, all algorithms demonstrate increased rise times compared to simulation, with TDMPC2 showing the most pronounced degradation (0.21 s to 0.80 s), suggesting sensitivity to model uncertainties inherent in real-world dynamics.

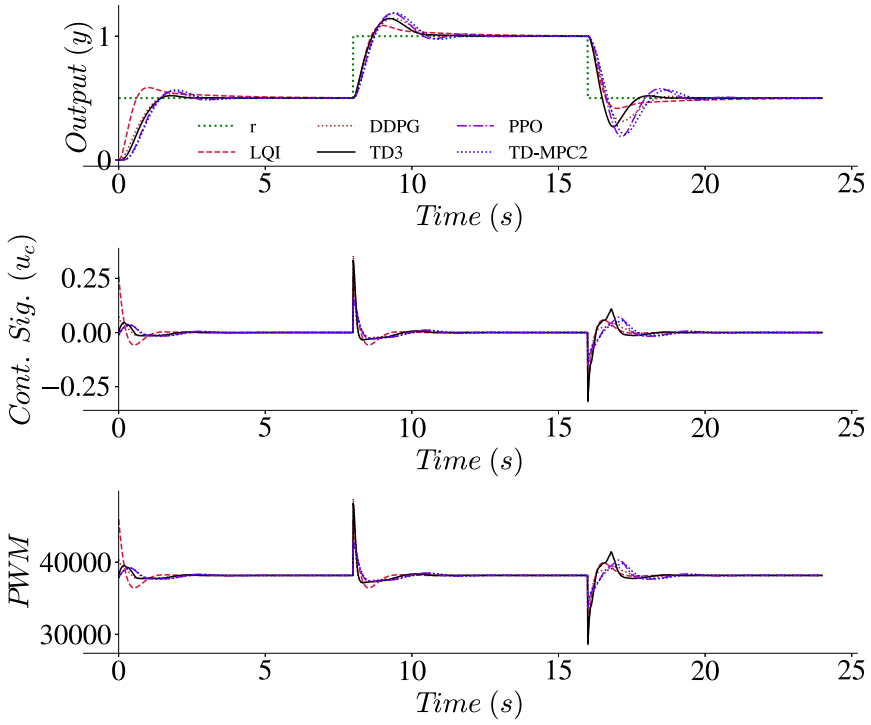


Fig. 9. Multi-step reference tracking comparison (Simulation).

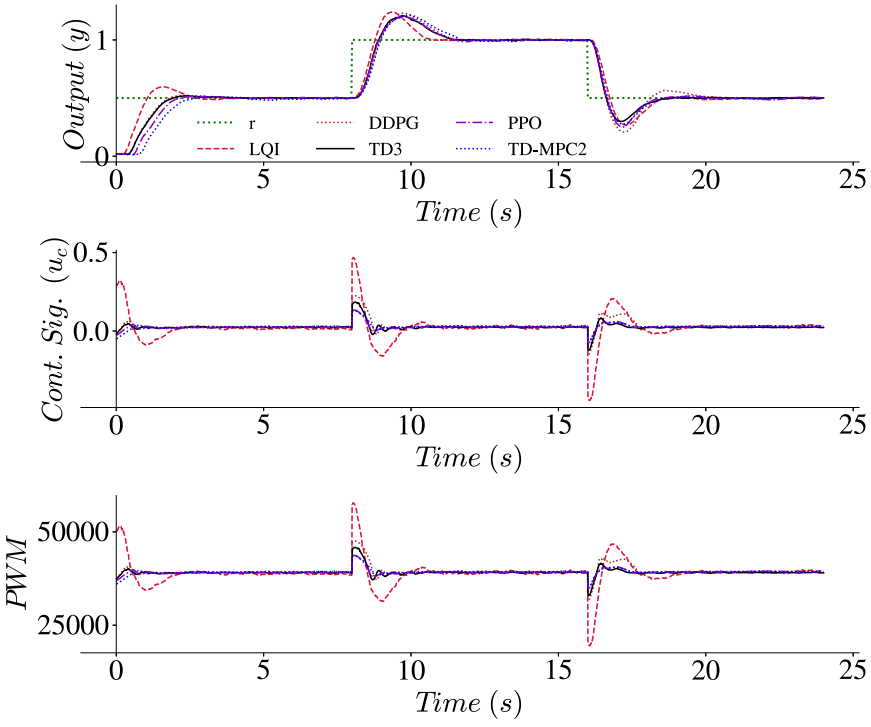


Fig. 10. Multi-step reference tracking comparison (Experimental validation).

Remarkably, the overshoot characteristics demonstrate convergence in real-world implementation, with all algorithms exhibiting similar peak overshoots (21.7%–30.0%) compared to the wider variation observed in simulation (17.0%–49.9%). This convergence suggests that physical constraints, including actuator limitations, sensor noise, and aerodynamic effects, naturally limit aggressive control behaviour, providing an inherent robustness mechanism absent in idealised simulation environments.

The control effort analysis reveals a critical insight: real-world implementation demands significantly higher control activity (IACE increases by 3–5 times across all algorithms) and control rates (IACER increases by 2–3 times). This phenomenon reflects the continuous correction required to compensate for sensor noise, aerodynamic disturbances, and model uncertainties. TD3 maintains the best settling performance (3.0 s) whilst requiring moderate control effort, demonstrating superior practical robustness among DRL approaches.

4.6.3. Sim-to-real transfer analysis

The quantitative comparison between simulation and experimental results provides valuable insights into sim-to-real transfer effectiveness. The most significant degradation occurs in integral time-weighted error metrics (ITAE), increasing by 35%–60% across all algorithms, indicating that whilst fundamental tracking capability is preserved, transient response quality inevitably suffers under real-world conditions.

Critically, all algorithms maintain zero steady-state error in both domains, validating the integral action component of the state representation and confirming the robustness of the fundamental control architecture. The successful completion of the multi-step reference sequence across all controllers demonstrates that simulation-trained policies possess sufficient robustness margins to handle unmodelled dynamics, sensor noise, and environmental disturbances characteristic of practical UAV operations. It is worth noting that rise time degradation in sim-to-real transfer is a well-documented phenomenon across various applications (Tiwari et al., 2026; Zhao et al., 2020) and does not uniquely undermine DRL viability. The successful completion of the multi-step reference sequence by all algorithms on physical hardware confirms that timing requirements were met in practice despite the increased rise times observed relative to simulation.

5. Conclusion

This paper presented several months of investigation into the achievable performance of widely discussed deep reinforcement learning controllers, a topic still lacking structured evaluation in the existing literature. We considered key practical control challenges including noise, disturbances, model uncertainty, actuator rate and amplitude saturation, time delay, and real-time applicability through detailed black-box analyses using tailored evaluation metrics. The study evaluated these controllers on well-known benchmarks: a non-minimum-phase process control example, the challenging ACC two-mass–spring problem, nonlinear REMUS AUV yaw control, and real-time deployment on a quadrotor's altitude controller.

Our analysis reveals that no single algorithm dominates across all performance criteria, emphasising the critical importance of application-specific controller selection based on performance priorities. The quantitative evaluation demonstrates distinct algorithmic strengths that map to specific control requirements. Within the benchmarks evaluated, TD3 emerges as the most consistently reliable DRL approach for applications demanding control smoothness and robustness, achieving superior error minimisation with the best ISE and ITAE performance across multiple benchmarks whilst maintaining the lowest control rate variations. Its twin-critic architecture effectively mitigates overestimation bias, making it particularly suitable for systems with strict actuator wear considerations or where control signal quality directly impacts operational costs. Across the tested benchmarks, for safety-critical applications requiring predictable behaviour under uncertainty, PPO provides the most consistent performance across varying conditions, demonstrating good stability margins and balanced degradation under disturbances. Its stochastic policy formulation naturally handles multi-modal control scenarios and provides inherent exploration without requiring hand-tuned noise schedules, making it ideal for systems operating across diverse operating points.

When actuator constraints represent the primary limitation, TD-MPC2 excels through its model-based trajectory optimisation, demonstrating superior performance under saturation conditions as evidenced in the AUV benchmark. The algorithm's ability to anticipate and plan around constraints makes it particularly valuable for systems with strict operational limits or where constraint violation carries significant penalties. For applications prioritising implementation simplicity and computational efficiency, DDPG remains viable despite its known limitations, offering straightforward deployment with acceptable performance for less demanding control tasks. Classical LQI maintains relevance for linear systems where model accuracy is high and computational efficiency is paramount, achieving optimal control effort whilst providing well-understood stability guarantees through systematic design procedures.

All controllers were implemented using the same environment, which provides a consistent interface for training and evaluating the different algorithms across all benchmark problems. The environment encapsulates system dynamics, reference generation, disturbance application, and performance measurement. For each benchmark system, the appropriate state–space model was integrated within this environment framework.

Our findings indicate that DRL algorithms successfully learn control policies offering performance comparable to classical model-based controllers whilst managing real-world challenges including delays, non-minimum-phase behaviour, actuator constraints, and model uncertainties. Notably, these methods handle actuator saturation and windup without explicit anti-windup design, and demonstrate promising sim-to-real transfer capabilities. Despite these successes, the achievable performance does not significantly exceed that of simple state feedback with integral action in standard regulation tasks. The true promise of DRL lies in addressing problems where classical methods face fundamental limitations, and their greatest value may lie in hierarchical control architectures where strategic decision-making augments low-level control. Accordingly, the reported observations should be interpreted as empirical findings derived from the selected benchmark problems and operational challenges considered in this work, rather than as general claims of superiority of DRL over established control methodologies across all applications.

Future work should prioritise developing adaptive algorithm selection frameworks that automatically identify optimal approaches based on system characteristics and performance requirements, given the pronounced algorithmic variations observed across benchmark systems. The significant differences in control efficiency between classical and DRL methods present compelling opportunities for hybrid architectures that intelligently combine mathematical optimality with adaptive learning capabilities.

Whilst this study provides a comprehensive evaluation, one important limitation should be acknowledged. The current evaluation does not address safety related issues, which is a critical consideration for real-world deployment in safety-critical applications where exploration-induced constraint violations may have serious consequences. Establishing formal safety guarantees during training, for example, through constrained reinforcement learning or safe exploration mechanisms, remains an important direction for future work.

CRediT authorship contribution statement

Klinsmann Agyei: Writing – original draft, Visualization, Validation, Software, Methodology, Data curation, Conceptualization. **Pouria Sarhadi:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Daniel Polani:** Writing – review & editing, Validation, Supervision, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Anderson, B.D., Moore, J.B., 2007. Optimal Control: Linear Quadratic Methods, Reprint of the 1990 original edition Dover Publications.

Andrychowicz, O.M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., et al., 2020. Learning dexterous in-hand manipulation. Int. J. Robot. Res. 39 (1), 3–20.

Antsaklis, P., 2020. Autonomy and metrics of autonomy. Annu. Rev. Control. 49, 15–26.

Arulkumaran, K., Deisenroth, M.P., Brundage, M., Bharath, A.A., 2017. Deep reinforcement learning: A brief survey. IEEE Signal Process. Mag. 34 (6), 26–38.

Åström, K.J., 2000. Limitations on control system performance. Eur. J. Control 6 (1), 2–20.

Bao, Y., Zhu, Y., Qian, F., 2021. A deep reinforcement learning approach to improve the learning performance in process control. Ind. Eng. Chem. Res. 60 (15), 5504–5515.

Bellemare, M.G., Dabney, W., Munos, R., 2017. A distributional perspective on reinforcement learning. In: International Conference on Machine Learning. Pmlr, pp. 449–458.

Bennett, S., 2002. A brief history of automatic control. IEEE Control Syst. Mag. 16 (3), 17–25.

Bequette, B.W., 2003. Process Control: Modeling, Design, and Simulation. Prentice Hall Professional.

Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al., 2019. Dota 2 with large-scale deep reinforcement learning. arXiv:1912.06680 URL https://arxiv.org/abs/1912.06680.

Bitcraze, A., 2021. Datasheet crazyflie 2.1 - rev 3. Available from Bitcraze AB.

Buşoniu, L., De Bruin, T., Tolić, D., Kober, J., Palunko, I., 2018. Reinforcement learning for control: Performance, stability, and deep approximators. Annu. Rev. Control. 46, 8–28.

Cai, Y., Zhang, C., Zhao, H., Zhao, L., Bian, J., 2023. Curriculum offline reinforcement learning. In: Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems. pp. 1221–1229.

Chen, W.-H., 2024. Goal-oriented control systems (GOCS): From how to what. IEEE/CAA J. Autom. Sin. 11 (4), 816–819.

Chen, X., Wang, C., Zhou, Z., Ross, K.W., 2021. Randomized ensembled double Q-learning: Learning fast without a model. In: International Conference on Learning Representations. URL https://openreview.net/forum?id=AY8zfZm0tDd.

Dutta, D., Upreti, S.R., 2022. A survey and comparative evaluation of actor-critic methods in process control. Can. J. Chem. Eng. 100 (9), 2028–2056.

Dutta, D., Upreti, S.R., 2023. A reinforcement learning-based transformed inverse model strategy for nonlinear process control. Comput. Chem. Eng. 178, 108386.

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Janoos, F., Rudolph, L., Madry, A., 2019. Implementation matters in deep rl: A case study on ppo and trpo. In: International Conference on Learning Representations.

Eschmann, J., Albani, D., Loianno, G., 2024. Learning to fly in seconds. IEEE Robot. Autom. Lett..

Fujimoto, S., Hoof, H., Meger, D., 2018. Addressing function approximation error in actor-critic methods. In: International Conference on Machine Learning. PMLR, pp. 1587–1596.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, pp. 249–256.

Haarnoja, T., Zhou, A., Abbeel, P., Levine, S., 2018a. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: International Conference on Machine Learning. Pmlr, pp. 1861–1870.

Haarnoja, T., Zhou, A., Abbeel, P., Levine, S., 2018b. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: ICML.

Hadi, B., Khosravi, A., Sarhadi, P., 2024. Adaptive Formation Motion Planning and Control of Autonomous Underwater Vehicles Using Deep Reinforcement Learning. IEEE J. Ocean. Eng. 49 (1), 311–328.

Hägglund, T., Guzmán, J.L., 2024. Give us PID controllers and we can control the world. IFAC-OnLine 58 (7), 103–108.

Hansen, N., Su, H., Wang, X., 2023. TD-MPC2: Scalable, robust world models for continuous control. arXiv preprint arXiv:2310.16828.

Hansen, N., Su, H., Wang, X., 2024. TD-MPC2: Scalable, Robust World Models for Continuous Control. In: Proceedings of the International Conference on Learning Representations. ICLR, URL https://arxiv.org/abs/2310.16828.

He, Q., Hou, X., 2020. WD3: Taming the estimation bias in deep reinforcement learning. In: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence. ICTAI, IEEE, pp. 391–398.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D., 2018. Deep Reinforcement Learning that Matters. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. AAAI, URL https://arxiv.org/abs/1709.06560.

Hewing, L., Wabersich, K.P., Menner, M., Zeilinger, M.N., 2020. Learning-based model predictive control: Toward safe learning in control. Annu. Rev. Control. Robot. Auton. Syst. 3 (1), 269–296.

Hiraoka, T., Imagawa, T., Hashimoto, T., Onishi, T., Tsuruoka, Y., 2021. Dropout q-functions for doubly efficient reinforcement learning. arXiv preprint arXiv:2110.02034.

Joshi, T., Kodamana, H., Kandath, H., Kaisare, N., 2023. TASAC: A twin-actor reinforcement learning framework with a stochastic policy with an application to batch process control. Control Eng. Pract. 134, 105462.

Kamthe, S., Deisenroth, M., 2018. Data-efficient reinforcement learning with probabilistic model predictive control. In: International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1701–1710.

Kendall, A., et al., 2019. Learning to drive in a day. In: 2019 International Conference on Robotics and Automation. ICRA, pp. 8248–8254.

Khaki-Sedigh, A., 2023. An Introduction to Data-Driven Control Systems. John Wiley & Sons.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Lawrence, N.P., Forbes, M.G., Loewen, P.D., McClement, D.G., Backström, J.U., Gopaluni, R.B., 2022. Deep reinforcement learning with shallow controllers: An experimental application to PID tuning. Control Eng. Pract. 121, 105046.

Levine, S., Kumar, A., Tucker, G., Fu, J., 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint arXiv:2005.01643.

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D., 2015. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.

Mania, H., Guy, A., Recht, B., 2018. Simple random search of static linear policies is competitive for reinforcement learning. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Vol. 31, Curran Associates, Inc..

Maxwell, J.C., 1868. On governors. Proc. R. Soc. Lond. (16), 270–283.

Minorsky, N., 1922. Directional stability of automatically steered bodies. J. Am. Soc. Nav. Eng. 34 (2), 280–309.

Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K., 2016. Asynchronous methods for deep reinforcement learning. In: International Conference on Machine Learning. PmLR, pp. 1928–1937.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M., 2013. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

Nguyen, N., Plimpton, J., Storro, H., 2023. Crazyflie 2.1 Quadcopter Nonlinear System Identification. Advisor: Dr. Uri Rogers.

Oh, T.H., 2024. Quantitative comparison of reinforcement learning and data-driven model predictive control for chemical and biological processes. Comput. Chem. Eng. 181, 108558.

Prestero, T., 2001. Development of a six-degree of freedom simulation model for the REMUS autonomous underwater vehicle. In: MTS/IEEE Oceans 2001. an Ocean Odyssey. Conference Proceedings (IEEE Cat. No. 01CH37295). Vol. 1, IEEE, pp. 450–455.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., Dormann, N., 2021. Stable-baselines3: Reliable reinforcement learning implementations. J. Mach. Learn. Res. 22 (268), 1–8, URL https://jmlr.org/papers/v22/20-1364.html.

Recht, B., 2019. A tour of reinforcement learning: The view from continuous control. Annu. Rev. Control. Robot. Auton. Syst..

Rohrs, C., Valavani, L., Athans, M., Stein, G., 2003. Robustness of continuous-time adaptive control algorithms in the presence of unmodeled dynamics. IEEE Trans. Autom. Control 30 (9), 881–889.

Salimans, T., Ho, J., Chen, X., Sidor, S., Sutskever, I., 2017. Evolution strategies as a scalable alternative to reinforcement learning. arXiv preprint arXiv:1703.03864.

Sarhadi, P., 2025. On the standard performance criteria for applied control design: PID, MPC or Machine Learning Controller?. arXiv preprint arXiv:2503.14379.

Saviolo, A., Loianno, G., 2023. Learning quadrotor dynamics for precise, safe, and agile flight control. Annu. Rev. Control. 55, 45–60.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., Moritz, P., 2015. Trust region policy optimization. In: International Conference on Machine Learning. PMLR, pp. 1889–1897.

Schulman, J., et al., 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Spielberg, S., Tulsyan, A., Lawrence, N.P., Loewen, P.D., Bhushan Gopaluni, R., 2019. Toward self-driving processes: A deep reinforcement learning approach to control. AIChE J. 65 (10), e16689.

Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Martín-Martín, R., Stone, P., 2025. Deep reinforcement learning for robotics: A survey of real-world successes. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39, pp. 28694–28698.

Tang, W., Daoutidis, P., 2022. Data-driven control: Overview and perspectives. In: 2022 American Control Conference. ACC, IEEE, pp. 1048–1064.

Tavakkoli, F., Sarhadi, P., Clement, B., Naeem, W., 2024. Model free deep deterministic policy gradient controller for setpoint tracking of non-minimum phase systems. In: 2024 UKACC 14th International Conference on Control. CONTROL, IEEE, pp. 1–6.

Tiomkin, S., Nemenman, I., Polani, D., Tishby, N., 2024. Intrinsic Motivation in Dynamical Control Systems. PRX Life 2 (3), 033009.

Tiwari, R., Khapre, S., Singh, A., 2026. Reinforcement learning in robotic systems: A review on sim-to-real transfer. Robot. Auton. Syst. 105327.

Vagia, M., Transeth, A.A., Fjerdingen, S.A., 2016. A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? Appl. Ergon. 53, 190–202.

Wie, B., Bernstein, D.S., 1992a. Benchmark problems for robust control design. J. Guid. Control Dyn. 15 (5), 1057–1059.
Wie, B., Bernstein, D.S., 1992b. Benchmark problems for robust control design (1992 ACC version). In: 1992 American Control Conference. IEEE, pp. 2047–2048.
Zhao, W., Queralta, J.P., Westerlund, T., 2020. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In: 2020 IEEE Symposium Series on Computational Intelligence. SSCI, IEEE, pp. 737–744.