Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p26104-cd

Deep Reinforcement Learning for Life-cycle Maintenance Planning of Large Infrastructure Networks

Christos Lathourakis
Digital Built Environment Department, TNO, Delft, The Netherlands. E-mail: christos.lathourakis@tno.nl

Andrés Martínez
Reliable Structures Department, TNO, Delft, The Netherlands. E-mail: andres.martinezcolan@tno.nl

This paper proposes a Deep Reinforcement Learning (DRL) framework designed to derive an optimal sequence of maintenance actions for large infrastructure networks with interacting assets. The DRL agent learns an adaptive policy that maps the system's belief state, i.e., a probability distribution over the true network condition, to optimal maintenance actions. The agent is trained using Proximal Policy Optimization (PPO) with a centralized belief and action architecture. Furthermore, the concept of curriculum learning is employed, where the agent is progressively trained on environments of increasing complexity, transferring the acquired knowledge of the simpler configurations to the subsequent harder ones. The framework is tested on Amsterdam's quay wall network, demonstrating significantly lower expected life-cycle costs compared to heuristic-based plans.

Keywords: Deep Reinforcement Learning, Infrastructure Maintenance, Proximal Policy Optimization, Curriculum Learning, Multi-component Deteriorating Systems

1. Introduction

Maintaining large infrastructure networks requires balancing near-term investment against the long-term risk of failures and malfunctions. Infrastructure deterioration is uncertain due to aging, environmental exposure, and increasing operational demands, while inspection data are often sparse. Optimal maintenance planning under such conditions can be posed as a stochastic optimization problem, where interventions are scheduled over a planning horizon to minimize expected life-cycle costs. However, simply aggregating asset-level optimization outputs is often invalid due to complex interactions and shared constraints. Yet, the alternative of network-level optimization faces the curse of dimensionality, becoming computationally intractable as the number of assets and actions increases.

A wide range of methods has been proposed to address maintenance planning under uncertainty, and the aforementioned underlying challenges. Classical threshold-based approaches, e.g. Time-Based Maintenance (TBM), Condition-Based Maintenance (CBM), typically rely on heuristic rules or expert judgement, and lack an explicit treatment of long-term optimality Ahmad and Kamaruddin (2012), Frangopol et al. (1997). More advanced methods formulate maintenance planning as a stochastic optimization problem, often using Markov Decision Processes (MDPs) or Partially Observable Markov Decision Processes (POMDPs) to model deterioration and imperfect information Corotis et al. (2005), Robelin and Madanat (2007). Dynamic programming-based solutions have been successfully applied to individual assets or small systems van den Boomen et al. (2019), Kuhn (2010), but their applicability diminishes as the state and action spaces grow.

To tackle this issue several studies decompose the problem into asset-level optimization followed by system-level scheduling or prioritization Mendoza-Lugo et al. (2024). Although such decomposition improves scalability, it neglects feedback between assets and may result in globally suboptimal solutions, particularly when network-level constraints or externalities dominate decision-making. Alternatively, multi-objective optimization techniques, such as genetic algorithms, have been used to generate Pareto-optimal maintenance plans balancing cost and re-

liability Frangopol and Liu (2007). While informative, these approaches typically do not yield a unique optimal policy and struggle to capture the sequential nature of maintenance decisions.

In recent years, Deep reinforcement learning (DRL) has emerged as a promising framework for tackling high-dimensional, sequential decision-making problems under uncertainty. By learning policies directly from simulated interactions, DRL circumvents the need for explicit enumeration of state–action spaces and can naturally incorporate long-term planning objectives. As demonstrated in Andriotis and Papakonstantinou (2019), Morato et al. (2023), DRL agents can effectively navigate high-dimensional state spaces to identify policies that outperform traditional heuristics.

In the current work, a DRL agent with a centralized state and action architecture is trained using Proximal Policy Optimization (PPO). To ensure this approach scales to large networks, a curriculum learning strategy is employed, in which the agent is progressively trained on more complex configurations, transferring knowledge from simple to more challenging scenarios. The proposed framework is tested on a case study involving the quay wall network management in Amsterdam.

2. Background

2.1. *Partially Observable Markov Decision Processes (POMDPs)*

Maintenance planning under uncertainty can be formulated using Markov Decision Processes (MDPs), and the interaction of its main components, i.e. the agent and the environment. During each decision step t, the agent observes the system's state, $s_t \in \mathcal{S}$, and chooses an action, $a_t \in \mathcal{A}$. Based on this state-action pair, the decision maker receives a reward $R_t(s_t, a_t) \in \mathcal{R}$ and the environment transitions to the subsequent state s_{t+1} according to probabilistic transition dynamics. A policy π defines a mapping between the state- and action-spaces with the objective of minimizing expected cumulative life-cycle cost.

In real-world infrastructure systems, however, the true condition of assets is rarely directly observable. Inspections and monitoring data provide partial and noisy information, while deterioration processes are subject to both epistemic and aleatoric uncertainties. This motivates the use of Partially Observable Markov Decision Processes (POMDPs), in which decision-making is performed over a belief $b_t \in \mathcal{B}$ which represents a probability distribution over the true system states. The belief state is updated in a Bayesian manner as new observations become available.

While POMDPs provide a rigorous and unified mathematical framework for risk-informed maintenance planning, their exact solution becomes computationally intractable for large, multi-component infrastructure systems due to the exponential growth of the state, action, and belief spaces. This limitation, the so-called curse of dimensionality, renders classical dynamic programming and point-based POMDP solvers impractical for real-world applications, motivating the need for more advanced approaches that scale well to networks composed of many interacting assets.

2.2. *Deep Reinforcement Learning (DRL)*

Reinforcement Learning (RL) provides a model-free framework for sequential decision-making, in which an agent learns a control policy through repeated interaction with an environment. Rather than relying on explicit transition and observation models, the agent improves its behavior through trial-and-error, guided by observed rewards.

Traditional RL methods rely on tabular representations and therefore struggle with high-dimensional or continuous state spaces, as commonly encountered in infrastructure networks. Deep Reinforcement Learning (DRL) addresses this limitation by using deep neural networks to approximate value functions and policies, effectively shifting the optimization problem from explicit state-action enumeration to learning network parameters θ. Actor-critic methods are particularly well-suited for such problems, as they combine a policy network (actor) with a value function estimator (critic), enabling stable learning in stochastic environments.

In this work, a centralized actor-critic architecture is adopted, as introduced in Andriotis and Papakonstantinou (2019). A centralized representation enables coordinated decision-making across

multiple assets while explicitly capturing system-level effects. Training is performed using Proximal Policy Optimization (PPO) Schulman et al. (2017), a trust-region policy gradient algorithm that stabilizes learning by limiting the magnitude of policy updates through a clipped surrogate objective. Experiences are collected in a rollout buffer memory and reused for batch updates, improving sample efficiency and robustness.

Assuming that the actions for the multiple interacting components are conditionally independent, a decomposed policy structure is employed, allowing the agent to output coordinated actions across assets. From a mathematical point of view, this enables the expression of the policy $\pi_\theta(\underline{a_t} \mid b_t)$ as the product of n distinct policies, each referring to an individual component. This approach maintains scalability while preserving sensitivity to global network states.

$$\begin{aligned}\pi_\theta(\underline{a_t} \mid b_t) &= \pi_\theta(a_t^1 \mid b_t)\, \pi_\theta(a_t^2 \mid b_t) \ldots \pi_\theta(a_t^n \mid b_t) \\ &= \prod_{i=1}^{n} \pi_\theta(a_t^i \mid b_t) \end{aligned} \tag{1}$$

2.3. *Curriculum Learning*

Training DRL agents directly on complex, high-dimensional environments often leads to unstable learning or convergence to suboptimal policies. Curriculum learning addresses this challenge by structuring the training process as a sequence of tasks of increasing complexity. The agent is first trained on an easier problem and subsequently transferred to more challenging ones, retaining previously acquired knowledge.

In the context of infrastructure maintenance planning, curriculum learning can be naturally implemented by gradually increasing the number of assets, the strength of component interactions, or uncertainty levels in deterioration and observations. This approach enables the agent to first learn fundamental maintenance patterns before being exposed to full-scale network problems. Previous studies have demonstrated the effectiveness of curriculum learning in improving convergence speed and final policy performance in complex DRL applications Narvekar et al. (2020).

In this work, curriculum learning is employed as a key scalability mechanism, enabling efficient training of DRL agents for large infrastructure networks that would otherwise be computationally prohibitive to address directly.

3. Methodology

This section presents the unified DRL framework developed for network-level life-cycle maintenance planning. Building on the concepts introduced in Section 2, the infrastructure network is modeled as a partially observable, multi-component environment in which coordinated maintenance decisions must be made under uncertainty and long-term cost considerations.

At each decision step, the agent observes a centralized belief state, based on which it selects a coordinated set of maintenance actions for all components. An actor–critic architecture is selected, which is trained using PPO. The actor network outputs a stochastic policy that maps the centralized belief state to a joint action across all components, while the critic network estimates the corresponding state value function, $V(b_t)$, i.e. the expected return starting from belief b_t over all possible actions. The aforementioned neural network structures are illustrated in Figure 1.

To address the scalability challenges associated with large infrastructure networks, curriculum learning is incorporated into the training procedure. Rather than training directly on the full-scale problem, the agent is initialized on simplified environments with a reduced number of components and progressively exposed to more complex tasks.

Formally, let $\mathcal{T} = \{T_1, T_2, \ldots, T_M\}$ denote a sequence of training tasks of increasing complexity, where each task T_i is characterized by a network size n_{T_i} such that

$$n_{T_1} < n_{T_2} < \ldots < n_{T_M}.$$

For each task T_i, the agent is trained until convergence, yielding optimized network parameters $\theta^*_{T_i}$. These parameters are then used to initialize the networks for the subsequent task, i.e.

$$\theta^0_{T_{i+1}} \leftarrow \theta^*_{T_i},$$

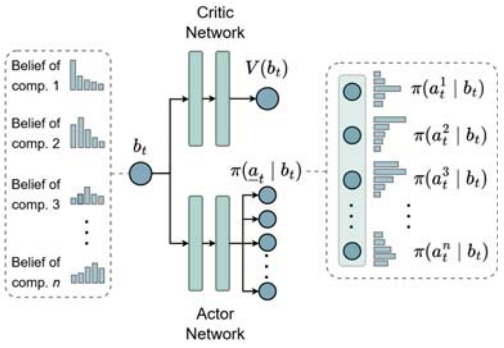

Fig. 1. Schematic representation of the Actor and Critic Networks

thereby transferring learned representations and decision patterns across tasks.

4. Case study: Historic quay walls in Amsterdam

4.1. *Background*

The Municipality of Amsterdam faces significant challenges in maintaining its historic quay wall infrastructure, with around 600 kilometers of quay walls within its scope, many of which are over 100 years old. These structures suffer critical degradation due to a combination of increasing traffic loads, material aging, and a legacy of underinvestment, creating a need for large-scale renewal. The risks associated with this deterioration were illustrated by the sudden collapse of the Grimburgwal quay in 2020 Korff et al. (2022). Managing these assets is further complicated by significant uncertainty regarding their geometry and geotechnical conditions, as well as the logistical and economic impact of interventions within the city. A comprehensive overview of Amsterdam's quay wall network can be found in Luongo et al. (2025).

Previous research by Swaalf et al. (2024) formulated this maintenance problem using MDPs. However, to manage the size of the state space, the optimization was performed at the asset level. This decoupling limits the solution to a collection of local optima rather than a coordinated global strategy. On the other hand, van Remmerden et al. (2025) demonstrated that DRL can effectively handle the maintenance problem, but restricted the scope to a single quay wall with multiple components. This leaves a gap for approaches that benefit from the scalability of DRL to solve the optimization problem at network level.

4.2. *Problem formulation*

4.2.1. *Scope*

The considered quay wall environment focuses on a 20-year planning horizon with annual time steps. The study encompasses the Prinsengracht/Leidsestraat area in the center of Amsterdam, consisting of 32 quay wall segments grouped into 11 quay wall groups, with a total length of $1,983$ meters. A quay wall group is defined as the continuous wall section between two bridges on one side of the canal. Each group may contain multiple segments due to historical repairs or varying deterioration conditions.

4.2.2. *State and action spaces*

Each quay wall segment is modeled as a discrete-state system representing its structural condition. Four deterioration states are defined according to the Municipality of Amsterdam's ARK (Amsterdamse Risicobeoordeling Kademuren) methodology Neijzing and Wesstein (2022), corresponding to increasing levels of structural damage. A terminal *failed* state is introduced to represent structural collapse. Hence, the deterioration state space is discretized into 5 distinct values, i.e. [s_1, s_2, s_3, s_4, s_5].

The action space consists of five discrete maintenance interventions:

- regular management (or do nothing)
- parking restriction
- traffic restriction
- renovation
- new construction

The first three actions are passive or mitigating interventions, as they do not directly alter the structural state but influence the deterioration process by reducing external loads. In contrast, *renovation* and *new construction* are active interventions directly improving the structural condition.

Maintenance actions that require street closures are modeled at the quay wall group level. If any

segment within a group is subject to such an intervention, the entire group is considered unavailable for traffic during that year. This coupling introduces spatial dependencies between segments and reflects the practical constraints of urban planning.

4.2.3. *Transition model*

The deterioration and recovery dynamics of quay wall segments are represented through action-dependent state transition models. Under passive actions, transitions are stochastic and restricted to either remaining in the current state or moving into the next worse state. For the baseline *regular management* action, the deterioration probabilities progress geometrically, i.e. the probability of transitioning from one state to the next doubles with each successive deterioration level. This results in an expected structural lifetime of 100 years from the initial state to failure.

Mitigating actions reduce deterioration rates while preserving the geometric structure of the transition probabilities. Specifically, *parking restriction* and *traffic restriction* extend the expected lifetime to 150 and 200 years, respectively. Active interventions have deterministic effects: *renovation* improves the condition state by one level, whereas *new construction* resets the segment to the initial state. All maintenance actions are assumed to occupy a full annual time step, during which no further state transitions occur.

4.2.4. *Cost structure*

In this optimization problem, a discount factor of $\gamma = 0.99$ is employed. The cost function comprises three components: direct maintenance costs, failure costs, and city-level costs that capture broader urban impacts.

Direct maintenance costs are proportional to quay wall length. *Renovation* interventions incur a cost of €25,000 per linear meter, while *new construction* incurs €50,000 per linear meter. These unit costs are applied uniformly across all segments regardless of their location or condition and represent average investment levels consistent with current renewal practices.

Failure costs are incurred when a segment reaches the failed state. The cost per linear meter is modeled as the *new construction* cost multiplied by a consequence factor ranging from 1 to 4, accounting for differences in quay wall height and functional importance within the canal and road networks. This formulation reflects the increased societal impact of failures at critical locations.

City-level costs capture the broader urban impact of maintenance interventions that require street closures. These costs are derived from traffic simulations performed using the digital twin platform described in Lohman et al. (2023), which compares disrupted and undisrupted network conditions. The simulations quantify vehicle delays, valued at €15 per vehicle-hour, as well as emissions of CO_2, NO_x, $PM_{2.5}$, and PM_{10}, monetized using standard unit cost estimates. Peak-hour simulation results are scaled to annual values using temporal expansion factors, and an additional hindrance factor is applied to capture wider urban disruption effects beyond direct traffic impacts.

4.3. *Benchmark approaches*

The performance of the proposed DRL-based maintenance policy is evaluated against a set of heuristic benchmark strategies reflecting common engineering practice. These strategies define deterministic mappings from observed condition states to maintenance actions. Since the exact state is not directly observable, but rather the belief over states, the decision-making of these benchmarks relies on the state with the highest probability in each decision step. While such methods do not account for long-term system-level optimization, they provide meaningful reference points for assessing the potential of the proposed method.

Table 1 summarizes the heuristic policies considered in this study. The policies range from purely reactive strategies, where interventions are only triggered upon reaching critical or failed states, to more proactive approaches that introduce gradual mitigation or early intervention measures as deterioration progresses.

In addition to the heuristic benchmarks, a pseudo-Markov Decision Process (pseudo-MDP) strategy is considered, where each quay wall segment is treated as an independent component, and an optimal maintenance policy is derived for

Table 1. Heuristic benchmark maintenance policies. Action abbreviations are defined below.

Policy	s_1	s_2	s_3	s_4	s_5
Reactive to failure	–	–	–	–	R
Reactive to critical condition	–	–	–	L	R
Mitigation of critical condition	–	–	–	T	R
Gradual response	–	P	T	L	R
Early intervention	–	T	L	R	R

Legend: – Do nothing; P Close parking; T Close traffic; L Life extension; R Full renovation.

each segment in isolation. The resulting component policies are then superimposed to construct a network-wide maintenance strategy. Although computationally efficient, this approach neglects interdependencies between assets, e.g. shared traffic disruptions and coordinated interventions.

5. Results

The DRL agent was trained using the curriculum learning strategy described in Section 2.3. Training was conducted in two stages: the agent was first trained on an environment of 16 quay walls, after which the optimized neural network parameters θ^* were transferred to initialize training in the more complex 32-quay wall environment. This staged training process enables the agent to progressively acquire increasingly sophisticated maintenance strategies while maintaining training stability. The results presented in what follows refer to the 32-quay wall environment.

Figure 2 illustrates the training of the DRL agent, with the life-cycle maintenance cost being plotted along training episodes. The best-performing benchmark approaches, introduced in Table 1, are also included in the graph. It is observed that the DRL framework consistently outperforms the heuristic strategies, reaching significantly lower expected maintenance costs. Table 2 reports the total expected maintenance costs over the considered 20-year planning horizon

Another valuable insight is the effect of using curriculum learning, which significantly improves training stability and convergence. Curriculum-trained agents achieved lower expected costs and smoother convergence patterns, compared to agents trained directly on the full complexity task.

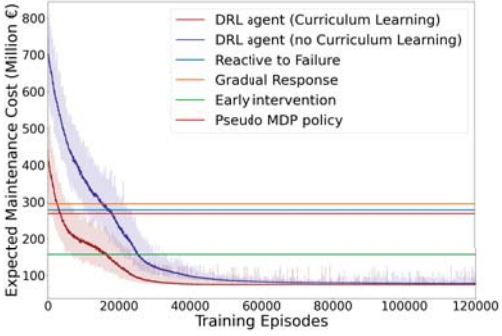

Fig. 2. Training performance of the DRL agent, comparing curriculum learning and cold-start training.

The results indicate that the DRL-based strategy yields a substantial reduction in total expected life-cycle costs compared to all heuristic and pseudo-MDP benchmarks. Even the most proactive heuristic policy, which emphasizes early intervention, remains more than twice as expensive as the learned DRL policy.

As explained in Section 4.2.4, the total maintenance cost consists of three components: direct intervention costs, failure costs and city-level (interaction) costs. Figure 3 presents these sub-costs for the different maintenance strategies.

It is observed that the DRL framework achieves significantly lower values across all three categories, resulting in a markedly lower total cost when compared to the benchmark strategies.

Table 2. Comparison of total expected maintenance costs for the 32-quay wall environment.

	Total expected cost (million €)	**Total expected cost relative to DRL**
Reactive to Failure	278.6	3.69
Reactive to Critical Condition	278.6	3.69
Mitigation of Critical Condition	278.6	3.69
Gradual Response	295.2	3.91
Early Intervention	156.9	2.08
Pseudo MDP policy	268.6	3.56
DRL Agent	**75.5**	**1.00** (baseline)

6. Conclusion

This paper presented a Deep Reinforcement Learning (DRL) framework for life-cycle maintenance planning of large infrastructure networks under uncertainty. By formulating the problem as a partially observable, multi-component decision process and addressing it using Proximal Policy Optimization (PPO) with a centralized actor–critic architecture, the proposed approach overcomes key scalability limitations inherent in traditional optimization and heuristic-based methods.

Curriculum learning proved to be beneficial for stable and efficient training. Application to Amsterdam's historic quay wall network demonstrated that the learned DRL-based policy substantially outperforms typical heuristic maintenance strategies, achieving significant reductions in expected life-cycle costs while accounting for failure risks and urban disruptions.

A potential direction for future research is the development of dimension-independent curriculum learning schemes to further improve scalability to even larger networks. Moreover, the incorporation of explicit constraints, such as annual budget limits and resource availability, can be explored to enhance practical applicability. Finally, the framework could be extended to continuous state representations informed by data-driven and Bayesian deterioration models, as explored in Lathourakis et al. (2023). These extensions will further strengthen the role of DRL as a practical decision-support tool for large-scale infrastructure asset management.

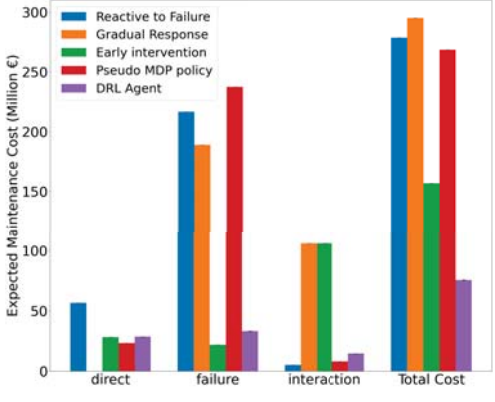

Fig. 3. Breakdown of total expected maintenance costs into direct costs, failure costs, and interaction costs.

Acknowledgement

The authors extend their gratitude to the Municipality of Amsterdam, and particularly to the Program Quay Walls and Bridges, for their assistance and provision of essential information for the case study. This research was partially conducted within TNO's "AI for infra" internal knowledge program.

References

Ahmad, R. and S. Kamaruddin (2012). An overview of time-based and condition-based maintenance in industrial application. *Computers & industrial engineering 63*(1), 135–149.

Andriotis, C. P. and K. G. Papakonstantinou (2019). Managing engineering systems with

large state and action spaces through deep reinforcement learning. *Reliability Engineering & System Safety 191*, 106483.

Corotis, R. B., J. Hugh Ellis, and M. Jiang (2005). Modeling of risk-based inspection, maintenance and life-cycle cost with partially observable markov decision processes. *Structure and Infrastructure Engineering 1*(1), 75–84.

Frangopol, D. M., K.-Y. Lin, and A. C. Estes (1997). Life-cycle cost design of deteriorating structures. *Journal of structural engineering 123*(10), 1390–1401.

Frangopol, D. M. and M. Liu (2007). Maintenance and management of civil infrastructure based on condition, safety, optimization, and life-cycle cost. *Structure and Infrastructure Engineering 3*(1), 29–41.

Korff, M., M.-J. Hemel, and D. J. Peters (2022). Collapse of the grimburgwal, a historic quay in amsterdam, the netherlands. *Proceedings of the Institution of Civil Engineers-Forensic Engineering 175*(4), 96–105.

Kuhn, K. D. (2010). Network-level infrastructure management using approximate dynamic programming. *Journal of Infrastructure Systems 16*(2), 103–111.

Lathourakis, C., C. Andriotis, and A. Cicirello (2023). Inference and maintenance planning of monitored structures through markov chain monte carlo and deep reinforcement learning. In *14th international conference on applications of statistics and probability in civil engineering, ICASP14*, Volume 2262, pp. 103339.

Lohman, W., H. Cornelissen, J. Borst, R. Klerkx, Y. Araghi, and E. Walraven (2023). Building digital twins of cities using the inter model broker framework. *Future Generation Computer Systems 148*, 501–513.

Luongo, D., G. Nicodemo, A. Venmans, M. Korff, L. Sartorelli, H. Maljaars, and D. Peduto (2025). The quay walls of amsterdam, netherlands: An approach for collapse risk mitigation at the municipal scale based on multisource monitoring and surveying data. *Journal of Geotechnical and Geoenvironmental Engineering 151*(2), 05024014.

Mendoza-Lugo, M. A., M. Nogal, and O. Morales-Nápoles (2024). Network-level optimisation approach for bridge interventions scheduling. *Structure and Infrastructure Engineering*, 1–17.

Morato, P. G., C. P. Andriotis, K. G. Papakonstantinou, and P. Rigo (2023). Inference and dynamic decision-making for deteriorating systems with probabilistic dependencies through bayesian networks and deep reinforcement learning. *Reliability Engineering & System Safety 235*, 109144.

Narvekar, S., B. Peng, M. Leonetti, J. Sinapov, M. E. Taylor, and P. Stone (2020). Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research 21*(181), 1–50.

Neijzing, L. and R. Wesstein (2022). Amsterdamse risicobeoordeling kademuren - versie 1. Report, Gemeente Amsterdam, Amsterdam.

Robelin, C.-A. and S. M. Madanat (2007). History-dependent bridge deck maintenance and replacement optimization with markov decision processes. *Journal of infrastructure systems 13*(3), 195–201.

Schulman, J., F. Wolski, P. Dhariwal, A. Radford, and O. Klimov (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Swaalf, L., A. Martínez, R. Sinha, and P. de Heer (2024). A decision-making framework for large-scale maintenance in cities: Application to urban quay walls in amsterdam. In *Workshop on Life Cycle Management (LCM 2024)*.

van den Boomen, M., P. L. van den Berg, and A. R. M. Wolfert (2019). A dynamic programming approach for economic optimisation of lifetime-extending maintenance, renovation, and replacement of public infrastructure assets under differential inflation. *Structure and infrastructure engineering 15*(2), 193–205.

van Remmerden, J., M. Kenter, D. M. Roijers, C. Andriotis, Y. Zhang, and Z. Bukhsh (2025). Deep multi-objective reinforcement learning for utility-based infrastructural maintenance optimization. *Neural Computing and Applications*, 1–24.