14th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP14
Dublin, Ireland, July 9-13, 2023

Interpretation and analysis of deep reinforcement learning driven inspection and maintenance policies for engineering systems

Pablo G. Morato
Postdoct. Researcher, Dept. of Wind Energy, Tech. Univ. of Denmark, Roskilde, Denmark

Konstantinos G. Papakonstantinou
Associate Professor, Dept. of Civil & Environmental Engineering, The Pennsylvania State Univ., University Park, USA

Charalampos P. Andriotis
Assistant Professor, Faculty of Architecture and the Built Environment, Delft Univ. of Technology, Delft, Netherlands

Nandar Hlaing
PhD Candidate, ANAST, Dept. of ArGEnCo, Univ. of Liege, Liege, Belgium

Athanasios Kolios
Professor, Dept. of Wind Energy, Tech. Univ. of Denmark, Roskilde, Denmark

ABSTRACT: The application of Deep Reinforcement Learning (DRL) for the management of engineering systems has shown very promising results in terms of optimality and scalability. The interpretability of these policies by decision-makers who are so far mostly familiar with traditional approaches is also needed for implementation. In this work, we address this topic by providing a comprehensive overview of POMDP- and DRL-based management policies, along with simulation-based implementation details, for facilitating their interpretation. By mapping a sufficient statistic, namely a belief state, to the current optimal action, POMDP-DRL strategies are able to automatically adapt in time considering long-term sought objectives and the prior history. Through simulated policy realizations, POMDP-DRL-based strategies identified for representative inspection and maintenance planning settings are thoroughly analyzed. The results reveal that if the decision-maker opts for an alternative, even suboptimal, action other than the one suggested by the DRL-based policy, the belief state will be accordingly updated and can still be used as input for the remainder of the planning horizon, without any requirements for model retraining.

1. INTRODUCTION

Managing engineering systems is vital for ensuring societal progress, enhancing quality of life, and maximizing economic returns. However, the uncertainty associated with the prediction of deterioration mechanisms experienced by engineering systems poses a challenge to making informed maintenance decisions. Information from inspections and/or monitoring can be collected in order to support decision-making, yet recurrent costs are then incurred. With the objective of minimizing engineering systems life-cycle costs while still ensuring an appropriate level of safety, inspection and maintenance (I&M) actions should be timely planned, considering, among others, reliability and cost metrics. In recent years, increasing attention has been devoted to the development of risk-based inspection and maintenance planning methods, assisting decision makers with life-cycle plans, maintenance interventions, and data collection.

Existing risk-based I&M methods can be categorized according to their approach to solving the corresponding complex decision-making under uncertainty and imperfect information problem. By simplifying the global I&M decision problem to a local optimization of predefined decision rules, some methods are successful in identifying static policies for planning I&M actions, usually based on optimized time intervals or damage/reliability thresholds (Luque and Straub, 2019). These methods can be intuitive, but their optimality mainly depends on the designer's experience in defining heuristic decision rules, while delimiting their search in a restricted policy subspace. On the other hand, methods based on Markov Decision Processes (MDPs) and Partially Observable Markov Decision Processes (POMDPs) offer a principled mathematical approach for decision-making under uncertainty, globally solving the I&M stochastic optimization problem (Papakonstantinou and Shinozuka, 2014). By selecting the current action as a function of the belief state, i.e., a sufficient statistic corresponding to the dynamically updated action-observation history, POMDPs generate adaptive I&M policies that account for updated information and previously scheduled maintenance actions. As recently demonstrated in the literature, POMDP-based policies outperform conventional and state-of-the-art heuristic-based strategies (Morato et al., 2022).

Additional computational complexities arise when dealing with multi-component engineering systems due to the fact that I&M policies should be optimally identified in high-dimensional state, action, and observation spaces. While still relying on dynamic programming-based POMDP principles, multi-agent Deep Reinforcement Learning (DRL) methods address the aforementioned computational challenge by approximating policies and/or discounted long-term cost metrics with neural networks (Andriotis and Papakonstantinou, 2019). DRL approaches have been applied for the management of multi-component engineering systems (Morato et al., 2023), providing substantial cost savings compared to other state-of-the-art I&M methods. Along with optimality and scalability benefits, DRL methods can also generate I&M policies for settings under budget and/or safety constraints, additionally supporting decision makers to meet specific safety requirements and enabling optimal resource allocation (Andriotis and Papakonstantinou, 2021).

While the benefits offered by DRL policies have been clearly demonstrated for a wide range of data collection and intervention planning engineering applications in terms of optimality and scalability, better interpretability by decision-makers (e.g., operators, designers, and other stakeholders) accustomed to traditional calendar- and/or condition-based policies is also a practical need. In this work, we thoroughly describe the fundamentals of POMDP- and DRL-based management policies, propose simulation-based methods for facilitating their interpretation, and investigate their intrinsic adaptability and safety properties. In addition, we are looking into the effects of manual user interventions at the time of deployment, for any possible reason, on the action sequence prescribed by the learned policies. Overall, with this investigation, we aim to analyze and demonstrate the flexibility, safety, and interpretability of POMDP-DRL agent-based policies, towards accelerating adoption in real-world settings and improving the understanding of AI-driven decisions for engineering systems management.

2. ANALYSIS AND INTERPRETATION OF AGENT-BASED POLICIES

2.1. POMDP-based policies

A partially observable Markov decision process (POMDP) can be defined as a 7-tuple $\langle S, A, O, T, Z, R, \gamma \rangle$ controlled stochastic process in which an agent acts under uncertainty and imperfect information. At every decision step, the agent reasons based on the current belief $\mathbf{b}$, i.e., the probability distribution over states $s \in S$, takes an action $a \in A$, and then the state randomly transitions to state $s' \in S$, according to a stochastic transition model, $T(s,a,s') := p(s'|s,a)$. At that point, the agent perceives the subsequent state following an observation model $Z(o,s',a) := p(o \mid s',a)$, and finally receives a reward $R(s,a)$, discounted to its present value via the factor γ. The decision-making problem corresponding to the optimal in-

spection and maintenance planning for engineering systems can be adequately formulated as a POMDP, in which the agent reasons in a stochastic environment (i.e., a probabilistic deterioration model) and under imperfect information (i.e., measurement uncertainty associated with relevant inspection techniques).

A POMDP policy ($\pi : \mathbf{B} \rightarrow A$) prescribes actions as a function of the current belief, with the main objective of identifying the optimal policy, $\pi^*(\mathbf{b})$, that maximizes the value function $V(\mathbf{b})$, i.e., the discounted sum of expected rewards. In a structural reliability context, this is often translated to the minimization of the discounted total expected cost, $\mathbf{E}[c_T]$. The cost model can be defined as:

$$c_T = c_{ins} + c_{rep} + c_F \tag{1}$$

where c_{ins} and c_{rep} refer to inspection and repair costs, respectively, and the failure risk corresponds to the failure probability weighted by the failure consequences, i.e., $c_F = p_F c_f$. Note that following formal POMDP terminology, costs are defined as negative rewards.

Optimal POMDP policies $\pi^*(\mathbf{b})$ can be efficiently identified via point-based solvers, parametrized as a function of a finite set of α-vectors, each of which is associated with a specific action. The decision-maker (e.g., operator, designer, etc.) then selects inspection and/or maintenance actions according to the current belief state. At each decision point, the α-vector and corresponding action that maximize the value function $V^*(\mathbf{b})$ are chosen:

$$V^*(\mathbf{b}) = \max_{\alpha \in \Gamma} \sum_{s \in S} b(s)\alpha(s) \tag{2}$$

The expected total cost can be thus simply computed as the weighted sum of the expected total cost corresponding to the specific α-vector at deterioration state s, and the probability of being in that state $b(s)$. After taking action a and collecting observation o, the belief $\mathbf{b}$ is updated via Bayes' rule:

$$b(s') = \frac{p(o \mid s', a)}{p(o \mid \mathbf{b}, a)} \sum_{s \in S} P(s' \mid s, a) b(s) \tag{3}$$

Since beliefs are dynamically updated, POMDP policies are inherently adaptive as opposed to static decision rules, e.g., calendar- or condition-based maintenance approaches. Within a POMDP framework, decisions are influenced over time by prior information as well as previously taken actions, and each policy realization is hence tailored to its particular action-observation sequence. POMDP adaptive strategies can, however, be easily interpreted by simulating a set of policy realizations, from which meaningful statistics can be computed and outstanding trends can be easily identified. The decision maker can gain practical insights by, for instance, generating and analyzing an action histogram, where the percentage of expected actions is represented over the decision horizon. While policy realizations can be simulated from the initial belief, many other scenarios can be similarly examined starting from beliefs at varying decision horizon points and/or conditional to specific action-observation sequences, as the learned alpha-vectors cover the entire belief space. In a structural reliability context, one can additionally explore the expected failure probability resulting from a computed POMDP-based policy by simulating the evolution of the belief states associated with the subspace of failure states.

Additionally, POMDP-based policies provide flexibility to the decision maker if an alternative action rather than the one suggested by the policy should be taken due to practical, economic, or any other reasons, at any specific decision step. In that case, the remaining actions can still be selected by following the original policy, while the expected total cost, and corresponding economic regret, resulting from the user-defined action can be straightforwardly computed through a Bellman backup operation, as:

$$V(\mathbf{b}) = \sum_{s \in S} b(s) R(s, a) + \gamma \left[\sum_{o \in O} p(o \mid \mathbf{b}, a) \cdot V(\mathbf{b}_{s'}) \right] \tag{4}$$

where $\mathbf{b}$ and $\mathbf{b}_{s'}$ correspond to the current and updated beliefs, respectively, and $R(s,a)$ stands for the reward associated with the action taken.

2.2. Multi-agent DRL-based policies

Finding I&M policies for multi-component engineering systems is a challenging computational

problem, as explained in Section 1. While still stemming from POMDP dynamic programming principles, multi-agent deep reinforcement learning (DRL) methods offer additional scalability benefits, being capable of efficiently identifying optimal I&M strategies for very high-dimensional systems. In order to do that, they approximate policies, value functions, or both, with neural networks. When dealing with multi-component systems, adjusting the networks' weights at the training stage according to system metrics is key in order to encourage collaborative behavior among agents, all seeking a common objective, e.g., minimization of the system costs and risks (Morato et al., 2023).

In general, multi-agent DRL methods can be categorized as value-, policy-based, or actor-critic methods. The former parametrizes the action-value function, and at the deployment stage, the policy, $\pi(\mathbf{a}|\mathbf{b})$, selects the actions that lead to the maximum action-value function, $Q(a,\mathbf{b})$, assuming independent control behavior among $l \in N_C$ component agents at the current belief state, $\mathbf{b}$:

$$\pi(a^{(l)}|\mathbf{b}) = \arg\max_{a^{(l)}} Q_l(a^{(l)},\mathbf{b}) \tag{5}$$

Alternatively, the policy itself can be parametrized by component networks, directly selecting the action at the current belief state, $\mathbf{b}$:

$$\pi(\mathbf{a}|\mathbf{b}) = \prod_{l=1}^{N_C} \pi_l(a^{(l)}|\mathbf{b}) \tag{6}$$

Combining both approaches, actor-critic DRL methods approximate components' policies with actor networks steered by value function predictions generated from a critic network. This enables a global cooperative behavior as actors can be trained based on a system metric estimated by the critic. At the deployment stage, however, actions are directly selected from the surrogated policy in both policy-based and actor-critic methods. A detailed overview of actor-critic DRL methods for managing high-dimensional engineering systems can be found in Andriotis and Papakonstantinou (2019).

In the aforementioned DRL methods, each component agent receives the belief state as an input following either a centralized or decentralized scheme. In the former, agents are informed about the belief state of all agents in the system, while in the latter, each agent only receives local information, i.e., its own belief state. In any case, the above-mentioned approaches select actions as a function of the belief state, yielding substantial benefits compared to other static optimization methods. Influenced by the experienced action-observation sequence from all components, DRL policies are also intrinsically adaptive. As for the case of POMDP-based policies, one can straightforwardly interpret the strategy by observing summary statistics over policy realizations, which can be computed through simulation-based methods (e.g., Monte Carlo). This is further illustrated in the numerical experiments (Section 3) for the case study of a 9-out-of-10 system subject to fatigue deterioration.

DRL-based policies also offer flexibility at the practical implementation stage. If the decision maker opts for an alternative user-defined action rather than the one suggested by the original policy, for any possible reason, the subsequent actions can still be selected without the need for retraining. In that case, the component beliefs should be simply updated according to the newly chosen actions and collected observations, and used for the next step. The regret associated with an alternative policy can be also numerically calculated, providing further insights to the decision maker.

3. NUMERICAL EXPERIMENTS

3.1. Optimal inspection and maintenance planning for a structural component

We analyze here a POMDP-based strategy aimed at managing an offshore wind turbine structural component subject to fatigue over a 20-year horizon by optimally scheduling, at each time step t, *do-nothing*, *inspect*, or *repair* actions. To probabilistically characterize the fatigue deterioration evolution, the annual crack growth is described through Paris' law:

$$d_{t+1} = \left[\left(1-\frac{m}{2}\right) C_{FM} Y^m S_R^m \pi^{m/2} n + d_t^{1-m/2}\right]^{2/(2-m)} \tag{7}$$

where the crack depth is denoted as d, propagating according to crack growth parameters $ln(C_{FM}) \sim$

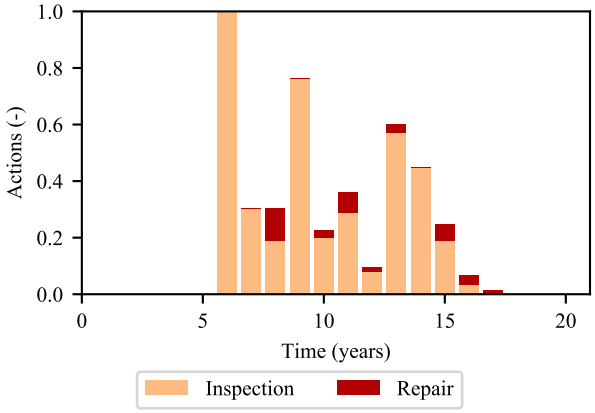

Figure 1: Action histogram based on 10,000 policy realizations from the examined POMDP-based policy.

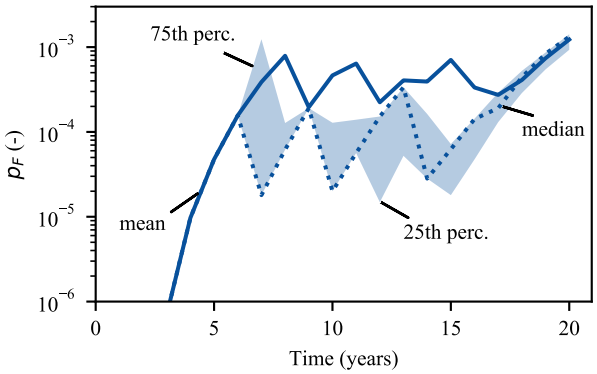

Figure 2: Component failure probability summary statistics over time corresponding to the analyzed POMDP-based policy.

$\mathrm{N}(\mu = -27.7903, \sigma = 0.3473)$ and $m = 3$, over $n = 10^6$ annual cycles. The applied cyclic loading is driven by the expected stress range $S_R = q\Gamma(1 + 1/h)$, described through parameters q and h, and accounting for local effects via the geometric factor $Y \sim \mathrm{LN}(\mu = 1, \sigma = 0.1)$. At the first time step, the crack size is specified as $d_0 \sim \mathrm{Exp}(\mu = 0.1235$ mm), becoming the initial belief, and a component failure corresponds to the event in which the crack size exceeds a critical value, $d_c = 16$ mm.

If an inspection is conducted, the crack size is observed with measurement uncertainty defined by the Probability of Detection curve, $PoD(d) = 1 - 1/(1 + [d/0.45]^{0.9})$, while the component returns to its initial condition, d_0, after a repair. In terms of costs, inspection and repair actions incur $c_{ins} = 10^3$ and $c_{rep} = 10^4$ monetary units, respectively, whereas a system failure costs $c_F = 10^6$ monetary units, all discounted yearly by a factor of $\gamma = 0.94$. The I&M decision problem is here formulated as a POMDP, which is then solved through the SARSOP point-based solver (Kurniawati et al., 2008). The reader is directed to (Hlaing et al., 2022) for a more detailed description of this case study.

Analyzing a POMDP-based I&M strategy

The resulting POMDP policy is analyzed in a simulation-based environment by running 10,000 policy realizations. Fig. 1 shows the corresponding annual action histogram, representing the percentage of inspection and repair actions, whereas the resultant annual failure probability is illustrated in Fig. 2. We can observe that the first inspection is always performed at year 6 and the first repair action is often planned two years after that. Frequent I&M actions are planned in the middle of the component's lifetime in order to effectively control the failure risk, yet I&M actions are logically less prescribed at the end of the horizon.

In this study, we also investigate the effect of selecting an alternative action rather than the optimal one suggested in the POMDP policy. We consider the following scenario: the optimal POMDP policy is followed up to year 7, but at that point, the decision-maker re-evaluates unforeseen economic restrictions, and she/he cannot follow the POMDP-prescribed repair (Rep) action. Now, the potential alternative actions are: (i) do-nothing (DN), in which case the fatigue deterioration will naturally progress according to the defined transition model, and (ii) inspect (Ins), in which case a crack can either be detected or not.

The total expected cost, $\mathbf{E}[c_T]$, associated with alternative actions is computed from Equation (4), where the value associated with the updated belief, $V(\mathbf{b}_{s'})$, is still estimated from the original POMDP policy. Fig. 3 represents the breakdown of the normalized expected total cost associated with all examined actions with respect to the original one. It can be observed that, in this case, a do-nothing action is the most suboptimal choice, as it results in a high failure risk from avoiding the required maintenance action. Logically, a repair action still needs to be prescribed the following year but no further

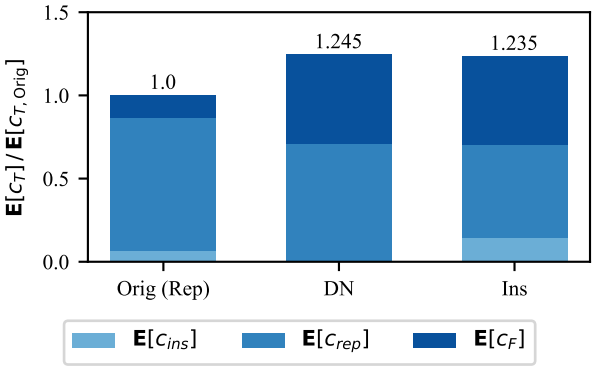

Figure 3: Breakdown of the total expected cost, $\mathbf{E}[c_T]$, resulting from the investigated user-defined actions, all normalized with respect to the original POMDP policy.

inspections are longer planned. Interestingly, an inspection action is less suboptimal, since more information is gathered and there is a probability that the deterioration condition is better than expected, in which case the corresponding inspection outcome and subsequent actions would result in a significant reduction of risk and repair costs, as compared to the do-nothing action related policy sequence.

3.2. Optimal management of a multi-component engineering system

In this second application, we investigate a DRL-based I&M policy identified for optimally managing a 9-out-of-10 system subject to fatigue deterioration. With the objective of minimizing the discounted expected total cost throughout a 30-year horizon, the decision maker can either *do-nothing*, *inspect*, or *repair* any component each year. Component inspections and repairs costs incur $c_{ins}^{(l)} = 1$ and $c_{rep}^{(l)} = 20$ monetary units, respectively, and a system failure results in a cost equivalent to $c_F = 50,000$ monetary units, all discounted in time by a factor $\gamma = 0.95$.

At each time step, t, components deteriorate according to the crack growth law described in Eq. 7, with material parameters $ln(C_{FM}) \sim \mathrm{N}(\mu = -35.2, \sigma = 0.5)$ and $m = 3.5$, an applied stress range $S_R \sim \mathrm{N}(\mu = 70, \sigma = 10\,\mathrm{N/mm^2})$ over $n = 10^6$ annual cycles, and an initial crack size specified as $d_0 \sim \mathrm{Exp}(\mu = 1\ \mathrm{mm})$. In this setting, the geometric factor is considered as a deterministic value, $Y = 1$. If an inspection is conducted, the crack size is partially observed with measurement uncertainty described by $PoD(d) \sim \mathrm{Exp}(\mu = 8)$. A component fails if the crack size exceeds a critical value, $d_c = 20$ mm, and the system fails if any two components fail. The I&M decision-making problem is here encoded as a deterioration rate POMDP, where the non-stationary damage evolution is defined as a function of the deterioration rate. To facilitate inference, the continuous crack size is discretized into $|S_d| = 30$ states. The reader is directed to (Morato et al., 2023) for a more detailed description of the case study. The formulated POMDP is solved via a deep decentralized multi-agent actor-critic scheme (Andriotis and Papakonstantinou, 2019), featuring ten control agents guided by a critic, all parametrized with fully connected neural networks. Specifically, actor and critic networks include two hidden layers of 100 and 200 neurons, respectively. During the training, the learning rate is adjusted from 10^{-4} to 10^{-5} for the actor, and from 10^{-3} to 10^{-4} for the critic. Besides, the exploration noise linearly decreases from an initial 100% random noise to a random noise of 1% over the first 20,000 episodes, remaining constant afterward. At the deployment stage, each component actor indicates the optimal action as a function of all component belief states.

Interpreting a DRL-based I&M strategy

In this specific case, the identified DRL policy is fully described by the trained component actors networks. To analyze the produced strategy in terms of decisions and safety, action histograms and failure probability statistics are computed via Monte Carlo simulations.

The percentage of inspection and repair actions are represented over decision steps for each component in Fig. 4. The decision maker can gain valuable insights by observing when/where the actions are usually planned. The shown policy randomly identified a few of the components to apply initial inspections, as all of them are equally important, with the same deterioration model and beliefs. Without any loss of generality, other training sets we performed showed the same trends, with different components now randomly identified as being initially inspected.

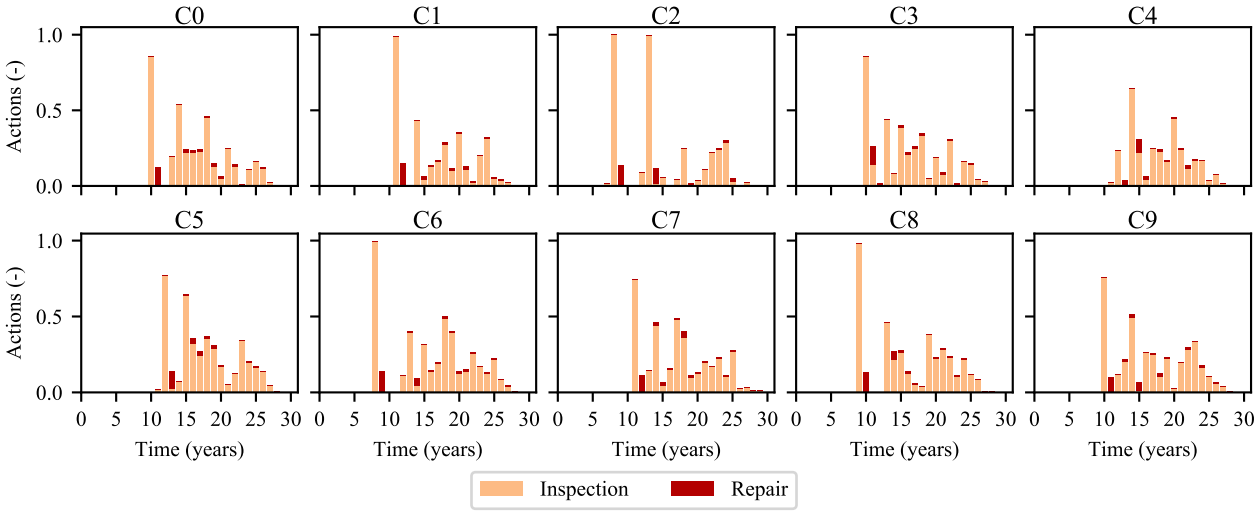

Figure 4: Action histogram showcasing 10,000 realizations simulated from the investigated DRL-based policy.

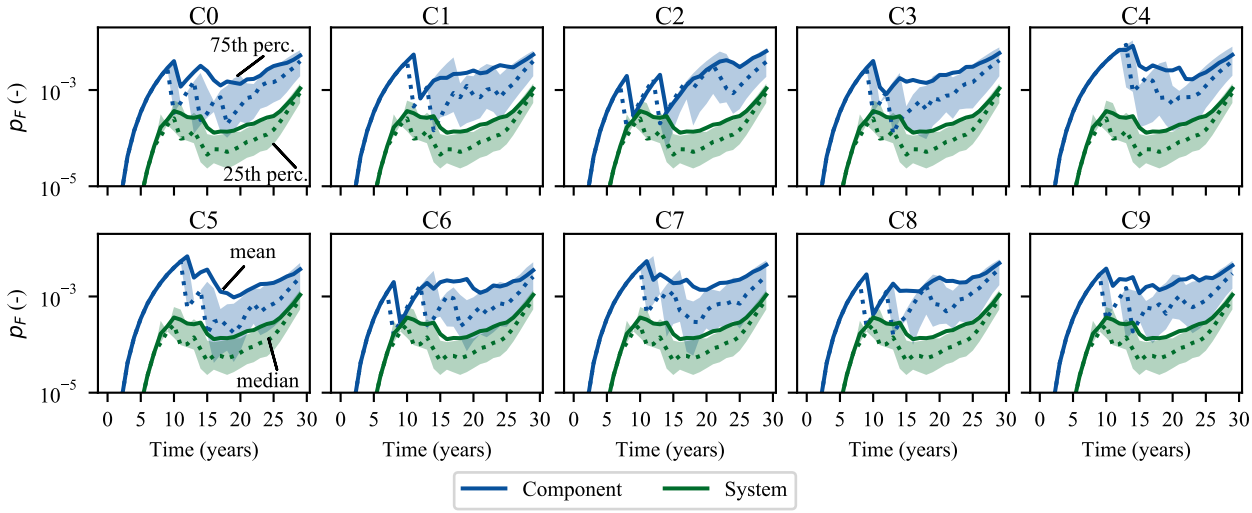

Figure 5: Component and system failure probability summary statistics over time corresponding to the analyzed DRL-based strategy.

In order to provide insights with respect to safety, the expected failure probability is also simulated over 10,000 policy realization and shown in Fig. 5, delimited by 25th and 75th percentiles. Based on this representation, the expected component and system failure probability can be assessed against relevant regulations and can be easily translated into a risk-based metric. By additionally examining the presented percentiles, we can infer that the failure probability is often lower than its expected value. As mentioned before, POMDP- and DRL-based policies also offer additional flexibility to the decision-maker. Let us consider a similar scenario as in the previous example: we are now at year 11 and the decision maker is evaluating the consequences associated with the selection of an alternative user-defined action that year due to a practical restriction that does not allow her/him to follow the DRL-prescribed policy. Still relying on the original DRL policy, the total expected cost associated with the remaining decision steps can be computed by selecting subsequent actions based on the related updated belief states. As a result of this process, Fig. 6 showcases the normalized expected to-

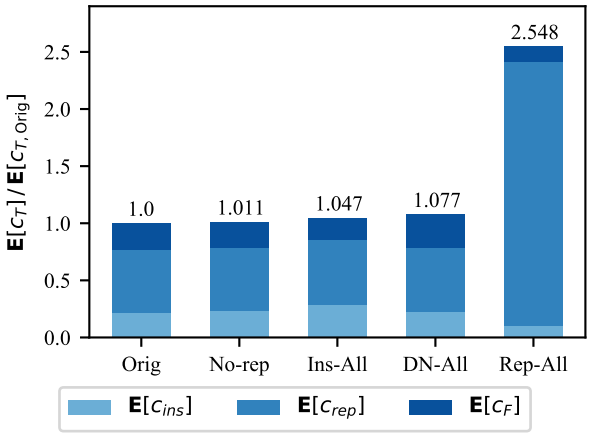

Figure 6: Breakdown of the total expected cost, $\mathbf{E}[c_T]$, resulting from the investigated user-defined actions, all normalized with respect to the original DRL policy.

tal cost with respect to the original policy, when the following actions are taken at year 11: (i) original policy (Orig), (ii) repairs are substituted by inspections (No-rep), (iii) all components are inspected (Ins-All), (iv) no I&M actions are planned for any component (DN-All), and (v) all components are repaired (Rep-All). The results reveal that the cost only increases by 1.1% if repairs are avoided and 4.7% when all components are inspected at that step, whereas a global do-nothing action is 7.7% more costly. Repairing all components is, however, very suboptimal as the cost surges more than 250% compared to the original action.

4. CONCLUSIONS

POMDP-DRL policies prescribe actions as a function of the sufficient statistic that corresponds to the action-observation sequence history, whereby a specific policy realization is tailored to previously taken actions and collected information. This paper shows that POMDP-DRL adaptive strategies can be easily interpreted via simulation methods, providing valuable insights to decision-makers, who may also straightforwardly evaluate the consequences of selecting an alternative action rather than the one suggested by the POMDP-DRL policy, without the need for retraining neural network architectures or value iteration methods.

ACKNOWLEDGEMENTS

The support provided by the TU Delft AI Labs program is gratefully acknowledged.

5. REFERENCES

Andriotis, C. P. and Papakonstantinou, K. G. (2019). "Managing engineering systems with large state and action spaces through deep reinforcement learning." *Reliability Engineering and System Safety*, 191, 106483.

Andriotis, C. P. and Papakonstantinou, K. G. (2021). "Deep reinforcement learning driven inspection and maintenance planning under incomplete information and constraints." *Reliability Engineering & System Safety*, 212, 107551.

Hlaing, N., Morato, P. G., Nielsen, J. S., Amirafshari, P., Kolios, A., and Rigo, P. (2022). "Inspection and maintenance planning for offshore wind structural components: integrating fatigue failure criteria with Bayesian networks and Markov decision processes." *Structure and Infrastructure Engineering*, 18(7), 983–1001.

Kurniawati, H., Hsu, D., and Lee, W. S. (2008). "SARSOP: Efficient point-Based POMDP planning by approximating optimally reachable belief spaces.." *Proceedings of Robotics: Science and Systems*, Switzerland.

Luque, J. and Straub, D. (2019). "Risk-based optimal inspection strategies for structural systems using dynamic Bayesian networks." *Structural Safety*, 76, 68–80.

Morato, P. G., Andriotis, C. P., Papakonstantinou, K. G., and Rigo, P. (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.

Morato, P. G., Papakonstantinou, K. G., Andriotis, C. P., Nielsen, J. S., and Rigo, P. (2022). "Optimal inspection and maintenance planning for deteriorating structural components through dynamic Bayesian networks and Markov decision processes." *Structural Safety*, 94, 102140.

Papakonstantinou, K. G. and Shinozuka, M. (2014). "Planning structural inspection and maintenance policies via dynamic programming and Markov processes. Part I: Theory." *Reliability Engineering and System Safety*, 130, 202–213.