Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Importance Sampling for Fair Policy Selection*

Shayan Doroudi[1,3], **Philip S. Thomas**[2] and **Emma Brunskill**[3]
[1] Carnegie Mellon University
[2] University of Massachusetts Amherst
[3] Stanford University
shayand@cs.cmu.edu, pthomas@cs.umass.edu, ebrun@cs.stanford.edu

Abstract

We consider the problem of off-policy policy selection in reinforcement learning: using historical data generated from running one policy to compare two or more policies. We show that approaches based on importance sampling can be *unfair*—they can select the worse of two policies more often than not. We then give an example that shows importance sampling is systematically unfair in a practically relevant setting; namely, we show that it unreasonably favors shorter trajectory lengths. We then present sufficient conditions to theoretically guarantee fairness. Finally, we provide a practical importance sampling-based estimator to help mitigate the unfairness due to varying trajectory lengths.

1 Introduction

In this paper, we consider the problem of *off-policy policy selection*: using historical data generated from running one policy to compare two or more policies. Off-policy policy selection methods can be used, for example, to decide which policy should be deployed when two or more batch reinforcement learning (RL) algorithms suggest different policies or when a data-driven policy is compared to a policy designed by a human expert. The *importance sampling* (IS) estimator, a statistical technique that was introduced to the RL community by [Precup *et al.*, 2000], can give unbiased estimates of the performance of policies using historical data. Importance sampling lies at the foundation of many policy selection and policy search algorithms [Mandel *et al.*, 2014; Levine and Koltun, 2013; Thomas *et al.*, 2015b]. The primary contribution of this paper is that we show that importance sampling is often *unfair* when used for policy selection: when comparing two policies, the worse of the two policies may be returned more than half the time.

Moreover, we show that using importance sampling for policy selection can be unfair in practically relevant settings. In particular, we show that IS can favor policies that produce shorter trajectories. Although IS is an unbiased estimator, this unfairness arises because policy selection involves taking a maximum over estimated quantities. Depending on the distribution of estimates, importance sampling may systematically favor policies that are worse in expectation.

We then present two new approaches for avoiding unfairness when using importance sampling for policy selection. First, we give sufficient conditions under which using importance sampling for policy selection is fair, and provide algorithms that guarantee a related notion of safety. We also describe how our approach to guaranteeing fairness and safety is related to the notions of power analysis and statistical hypothesis testing. Although of theoretical interest, the reliance of this approach on conservative concentration inequalities limits its practicality. Thus, in our second approach, we introduce a new practical IS-based estimator that lacks the theoretical properties of our first approach, but which can help mitigate unfairness due to differing trajectory lengths.

Although there is significant literature surrounding reducing the variance and mean squared error of off-policy policy evaluation methods that use IS-based estimators [Powell and Swann, 1966; Dudík *et al.*, 2011; Jiang and Li, 2016; Thomas and Brunskill, 2017], other challenges associated with off-policy policy selection, such as fairness, have not been explored in the literature. Our notion of fairness differs from the ethical notion of fairness in machine learning (e.g., ensuring machine learning algorithms do not make decisions that discriminate against certain populations). However, our notion of fairness could be of relevance to such settings as a policy selection algorithm may unfairly favor policies that are discriminatory, such as myopic policies when making hiring decisions [Jabbari *et al.*, 2017]; developing this relevance further would be an interesting line of future work. Similar notions of fairness have recently been proposed in the online RL setting, where an algorithm is fair if it never takes a worse action with higher probability than a better action [Jabbari *et al.*, 2017]. This is similar to our notion of fairness in the offline RL setting, where an algorithm is fair if it does not choose a worse policy with higher probability than the best candidate policy. However, the issues surrounding fairness in the two settings are different due to the different nature of each setting. By introducing a notion of fairness for policy selection and highlighting some limitations of existing IS-based approaches, we hope to motivate further work on developing practical and fair policy selection algorithms.

*This is an abridged version of our paper by the same name that appeared in *Uncertainty in Artificial Intelligence* (UAI) 2017 [Doroudi *et al.*, 2017].

2 Background

We consider sequential decision making settings in stochastic domains. In such domains, an agent interacts with the environment, and in doing so, it generates a *trajectory*, $\tau \triangleq (O_0, A_1, R_1, O_1, A_2, R_2, \ldots, A_T, R_T, O_T)$, which is a sequence of observations, actions, and rewards, with trajectory length T. The observations and rewards are generated by the environment according to a stochastic process—such as a Markov decision process (MDP) or partially observable Markov decision process (POMDP)—that is unknown. The agent chooses actions according to a stochastic policy π, which is a conditional probability distribution over actions A_t given the partial trajectory $\tau_{1:t-1} \triangleq (O_0, A_1, R_1, O_1, A_2, R_2, \ldots, O_{t-1})$ of prior observations, actions, and rewards. The value[1] of a policy π, V^π, is the expected sum of rewards when the policy is used: $V^\pi \triangleq \mathbb{E}_{\tau\sim\pi} \sum_{t=1}^{T} R_t$, where $\tau \sim \pi$ means that the actions of τ are sampled according to π. The agent's goal is to find and execute a policy with a large value.

In this paper, we consider offline (batch) reinforcement learning (RL) where we have a batch of data, called historical data, that was generated from some known behavior policy π_b. We are interested in the problem of batch policy selection: identifying a good policy for use in the future. This typically involves policy evaluation or estimating the value of a policy π_e using the historical data that was generated from the behavior policy π_b. If $\pi_e = \pi_b$ this is known as *on-policy policy evaluation*. Otherwise it is known as *off-policy policy evaluation*.

2.1 Importance Sampling

In this paper, we focus on off-policy policy evaluation and selection. Specifically, we focus on estimators that use importance sampling. Model-based off-policy estimators tend to have lower variance than IS-based estimators, but at the cost of being biased and asymptotically incorrect (not consistent estimators of V^π) [Mandel *et al.*, 2014]. In contrast, IS-based estimators can provide unbiased estimates of the value. There has been significant interest in using IS-based techniques in RL for policy evaluation [Precup *et al.*, 2000; Jiang and Li, 2016; Thomas and Brunskill, 2016], as well as growing recent interest in using it for policy selection [Mandel *et al.*, 2014] and the related problems of policy search and policy gradient optimization [Jie and Abbeel, 2010; Levine and Koltun, 2013; Thomas *et al.*, 2015b; Wang *et al.*, 2016].

The IS estimator is given by:

$$\hat{V}^{\pi_e}_{\text{IS}} \triangleq \frac{1}{n} \sum_{i=1}^{n} w_i \sum_{t=1}^{T_i} R_{i,t}$$

where

$$w_i \triangleq \prod_{t=1}^{T_i} \frac{\pi_e(a_{i,t}|\tau_{i,1:t-1})}{\pi_b(a_{i,t}|\tau_{i,1:t-1})}$$

[1]We use the term value to mean the average of the value function over all states or the expected return of a policy (sometimes represented by $J(\pi)$ or $\rho(\pi)$).

Notice that in the on-policy case (i.e., when $\pi_e = \pi_b$), we obtain the standard Monte Carlo estimator:

$$\hat{V}^{\pi_e}_{\text{MC},n} \triangleq \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T_i} R_{i,t}$$

The IS estimator is an unbiased and strongly consistent estimator of V^{π_e} if $\pi_e(a|\tau_{1:t-1}) = 0$ for all actions, a, and partial trajectories, $\tau_{1:t-1}$, where $\pi_b(a_t|\tau_{1:t-1}) = 0$. However, $\hat{V}^{\pi_e}_{\text{IS}}$ often has high variance. The *weighted importance sampling* (WIS) estimator,

$$\hat{V}^{\pi_e}_{\text{WIS}} \triangleq \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \sum_{t=1}^{T_i} R_{i,t}$$

is a variant of the IS estimator that is also strongly consistent and often has much lower variance but is *not* an unbiased estimator of V^{π_e}.

3 Fair Policy Selection

A policy selection algorithm is given a set of candidate policies and must choose one of the policies for use in the future. Any policy evaluation estimator can be converted to a policy selection algorithm by simply evaluating each policy using the estimator, and then selecting the policy that has the largest estimated value. Thus we can use the Monte Carlo estimator for on-policy policy selection, and we can use the IS or WIS estimators for off-policy policy selection.

There are two natural properties we would like in a batch policy selection algorithm:

- **Consistency**: In the limit as the amount of historical data goes to infinity, the algorithm should always select the policy that has the largest value.
- **Fairness**: With *any* amount of historical data, the probability that the algorithm selects a policy with the largest value should be greater than the probability that it selects any other policy.[2]

Exploring and ensuring the fairness of (IS-based) policy selection algorithms is the focus of this paper. There has been recent interest on combining model-based estimators and IS-based estimators [Jiang and Li, 2016; Thomas and Brunskill, 2016], however as both model-based estimators and (as we will show) IS-based estimators are unfair, it is easy to show that these new estimators must also be unfair. Therefore, we restrict our attention to the standard IS and WIS estimators.

Let $\pi^* = \arg\max_{\pi\in\Pi} V^\pi$ be the policy with the highest value in policy class Π. We can now formally state what it means for a policy selection algorithm to be fair.

Definition 3.1. A policy selection algorithm that chooses policies from a policy class Π is *fair* if whenever the algorithm outputs a policy, the probability that it outputs π^* is at least as large as the probability that it will output any other policy. The algorithm is strictly fair if the probability of outputting policy π^* is strictly greater than the probability of outputting any other policy.

[2]For simplicity, hereafter we assume that there are no two candidate policies that are equally good.

Notice that the probabilistic guarantee in this definition conditions on when the algorithm outputs a policy. This allows for a policy selection algorithm that does not output any policy in cases when it cannot determine which policy is better; a similar approach is taken in [Thomas *et al.*, 2015a; Thomas *et al.*, 2017]. In this paper, we focus on the case where there are only two policies.

4 Unfairness of Importance Sampling

We now give a simple toy example to illustrate that importance sampling is unfair and to give some intuition behind why it is unfair. Suppose we want to use importance sampling to select the better of two one-step policies, π_e and π_b, where we have prior data collected from π_b. Each policy must decide among two actions: a_1, which always gives a reward of 0.1, and a_2, which always gives a reward of 1. Further suppose that π_b chooses a_1 with probability 0.9 and π_e always chooses a_2. Clearly, π_e is a much better policy so a fair policy selection algorithm should choose π_e at least half the time. However, if we draw only a single sample from π_b, we get that with probability 0.9, a_1 is sampled, so $\hat{V}_{\text{IS}}^{\pi_e} = 0$ and $\hat{V}_{\text{IS}}^{\pi_b} = 0.1$. Thus, with probability $0.9 > 0.5$, IS chooses π_b, making IS unfair.

Furthermore, notice that as we decrease the probability that π_b samples a_2, the gap between the performance of the policies increases, yet the probability that IS chooses the right policy only decreases! Of course, comparing two policies based on a single sample from a behavior policy is an unrealistic setting, but the problem still holds for larger numbers of samples. In fact, for any particular number of samples drawn from the behavior policy, we can change the probability that π_b samples a_2 to ensure IS will be unfair.

Intuitively, the unfairness of importance sampling is due to the fact that the importance sampling estimator is often highly skewed, and the further our candidate policies are from the behavior policy, the more skewed the IS estimator becomes (and hence, the "less fair"). We now show how this property manifests itself in a practically relevant setting. In particular, we show that IS can favor policies that are more likely to result in shorter trajectories. However, this is not the only setting where the unfairness of IS might be of practical import. For example, we have also shown that IS can favor myopic policies—policies that obtain less total reward, but which obtain it early in an episode—over policies that are optimal in the longer term (see [Doroudi *et al.*, 2017]).

4.1 IS Favors Shorter Trajectories

In this section, we show that importance sampling can systematically favor policies that assign higher probability to shorter trajectory lengths in domains where the length of each trajectory may vary. This is a problem that could arise in many practical domains, for example domains where a user is free to leave the system at any time, such as a student solving problems in an educational game or a user chatting with a dialogue system. In such systems, a bad policy may cause a user to leave the system sooner resulting in a short trajectory, which makes it particularly problematic that importance sampling can favor policies that assign higher probability to shorter trajectories. The following example shows that importance sampling can favor policies that generate shorter trajectories even when they are clearly worse.

	$\hat{V}_{\text{MC}}$	$\hat{V}_{\text{IS}}$	$\hat{V}_{\text{WIS}}$
π_X	1.39	0.98	1.98
π_Y	39.52	0.010	0.020

Table 1: Median estimates, out of 100 simulations, of different estimators using 100 samples of π_X and π_Y in the domain in Section 4.1.

Consider the domain given in Figure 1. Now suppose we have data collected from a behavior policy π_b that takes each action with probability 0.5. We want to compare two policies: π_X, which takes action a_X with probability 0.99, and π_Y, which takes action a_Y with probability 0.99. Consider the case where $L = 80$. Clearly π_Y is the better policy, because it incurs a lot of reward when we encounter trajectories of length 80, while only losing out on a small reward when encountering the short trajectories. Table **??** shows the median estimate, out of 100 simulations, of the Monte Carlo estimator, as well as the median IS and WIS estimates using 1000 samples each. We find that while π_Y is, in actuality, much better, IS essentially only weights the shorter trajectories, so the estimates only reflect how well the policies do on those trajectories. WIS simply (almost) doubles the estimates because half of the samples have extremely small importance weights.

So why does this occur? When using IS in settings where trajectories can have varying lengths, the importance weight of shorter trajectories can be much larger than for longer trajectories, because for longer trajectories we are multiplying more ratios of probabilities that are more often smaller than one. This happens even if the policy we are evaluating is more likely to produce longer trajectories than a shorter one (because there are exponentially many longer trajectories and so each individual trajectory has an exponentially smaller weight than an individual short trajectory).

5 Guaranteeing Fairness

Given that importance sampling is not fair in general, we would like to understand under what conditions we can guarantee importance sampling can be used for fair policy selection. Let V_{Max}^{π} be the largest value that could result from policy π and w_{Max}^{π} be the largest importance weight possible for policy π with samples drawn from behavior policy π_b. In what follows, we assume, for simplicity, that the minimum possible value for all policies is 0. Algorithm 1 is a fair policy selection algorithm provided that $\epsilon \leq |V^{\pi_1} - V^{\pi_2}|$ and $\delta \leq 0.5$.

This result can be shown with a simple application of Hoeffding's inequality. Alternatively, we can use other concentration inequalities to obtain fairness conditions of a similar form. Setting $\delta = 0.5$ is sufficient to guarantee fairness, but we can guarantee stronger notions of fairness by choosing some $\delta > 0.5$ (i.e., whenever the algorithm outputs a policy, it outputs the better policy with probability at least $\delta > 0.5$).

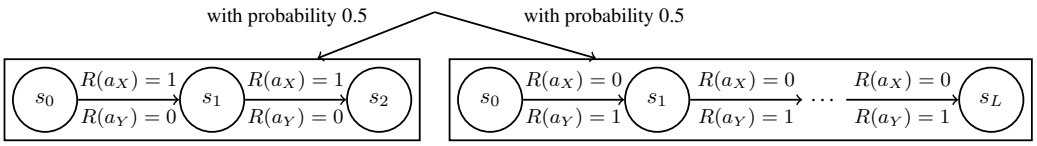

Figure 1: Domain in Section 4.1. The agent is placed uniformly at random in either a chain MDP of length 2 or a chain of length L. At each time step, action a_X deterministically gives a reward of 1 to the agent if the agent is in the chain of length 2 and 0 otherwise, and action a_Y deterministically gives a reward of 1 to the agent if the agent is in the chain of length L and 0 otherwise. Both actions progress the agent along the chain.

Algorithm 1 Fair Policy Selection

Require: $\pi_1, \pi_2, V^{\pi_1}_{\text{Max}}, V^{\pi_2}_{\text{Max}}, \epsilon, \delta$
$\tau_1, \tau_2, \ldots, \tau_n \sim \pi_b$
if $w^{\pi_1}_{\text{Max}} V^{\pi_1}_{\text{Max}} + w^{\pi_2}_{\text{Max}} V^{\pi_2}_{\text{Max}} \leq \epsilon \sqrt{\frac{2n}{\ln 1/\delta}}$ **then**
return $\max(\hat{V}^{\pi_1}_{\text{IS}}, \hat{V}^{\pi_2}_{\text{IS}})$
else
return `No Fair Comparison`
end if

Notice that this result tells us that as long as neither policy is too far from the behavior policy in terms of the largest possible importance weight, then we can guarantee fairness, which intuitively makes sense; we can only fairly compare policies that are similar to the behavior policy. However, how far the policies are allowed to stray will also depend on how different the values of the policies are from each other. This is a quantity we do not know, so we must pick an ϵ where either we think $\epsilon \geq |V^{\pi_1} - V^{\pi_2}|$ or we are comfortable with the possibility of selecting a policy whose value is ϵ worse than that of the better policy. Thus ϵ can be thought of as a hypothetical effect size as would be encountered in hypothesis testing.

6 Practical Fairness: Varying Trajectory Lengths

While Algorithm 1 provides a way to guarantee fairness, it is based on a concentration inequality that is naturally quite loose in most cases, and would likely result in often returning `No Fair Comparison`. Practically, it would be desirable to have algorithms that can provide fair comparisons more often. As a first step in this direction, here we discuss a heuristic approach to policy selection for domains where we have varying trajectory lengths, as seen in Section 4.1. The reason for focusing on this particular aspect of unfairness is because it is systematic (potentially arising in any domain where trajectories vary in length), yet it seems like there should be a way to correct for the systematic preference towards shorter trajectories in practically relevant domains. The idea we propose here is to compute an IS-based estimate for each trajectory length individually and then recombine the estimates to get a new estimate. We propose using the following estimator, which we refer to as the *Per-Horizon Weighted Importance*

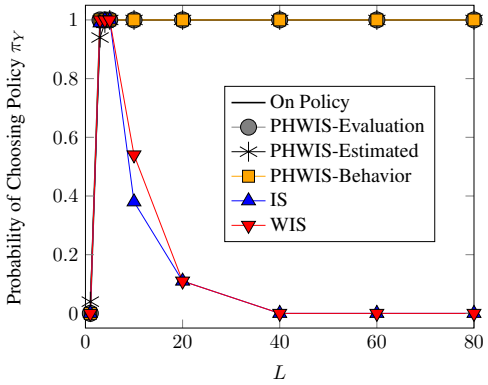

Figure 2: Probability of various estimators choosing π_Y over π_X for different values of L in the domain given in Section 4.1 with 1000 trajectories drawn from the uniform random behavior policy. For each estimator, the probability of outputting π_Y was estimated using 100 independent estimates.

Sampling (PHWIS) estimator:

$$\hat{V}_{\text{PHWIS}} = \sum_{l \in L} W_l \underbrace{\frac{1}{\sum_{\{i|T_i=l\}} w_i} \sum_{\{i|T_i=l\}} w_i \sum_{t=1}^{T_i} R_{i,t}}_{\text{WIS estimate on } l\text{-length trajectories}}$$

where L is the set of trajectory lengths that appear in the data and W_l is a weight for the relative importance of each trajectory length.

Notice that in the domain in Section 4.1, the length of the trajectories did not depend on the policy that was used to generate them; in such cases, we can use the following weights: W_l (Behavior) $\triangleq \frac{|\{i|T_i=l\}|}{n}$. The weights simply count the proportion of trajectories in our data (i.e., generated by the behavior policy) that have length l. We will refer to PHWIS with this weighting scheme as PHWIS-Behavior. Now we will see how this estimator performs on the domain in Section 4.1 (Figure 1) where $L \in \{1, 3, 5, 10, 20, 40, 60, 80\}$ given 1000 trajectories from the uniform random behavior policy. Figure 2 shows that while IS and WIS are unfair (choose π_X more often than π_Y) when the long trajectories are of length 20, PHWIS-Behavior always chooses the policy that the on-policy Monte Carlo estimator would choose (i.e., π_X when $L = 1$, and π_Y otherwise).

However, note that in cases where different policies may generate trajectories of different lengths (for example, bad policies causing users to dropout sooner), this simple form of weighting might not work too well. Ideally, we would like

to use the following weights: W_l (Evaluation) $\triangleq \Pr(|\tau| = l|\tau \sim \pi_e)$ where π_e is the evaluation policy for which we would like to estimate $\hat{V}^{\pi_e}$. However, we cannot actually compute these weights because we do not know the probability of any trajectory being generated by the evaluation policy. Instead, we propose a heuristic way to obtain weights that behave similarly to W_l (Evaluation) , namely: W_l (Estimated) $\triangleq \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i^{1/T_i} \mathbb{1}(T_i = l)$. We showed that using these weights does well in a toy domain where trajectory lengths are policy-dependent (see [Doroudi *et al.*, 2017] for some motivation behind this estimator and evaluation results). However, we believe further work can be put into finding a good weighting scheme for PHWIS for practically relevant settings.

7 Conclusion

In this paper, we examined the problem of off-policy policy selection and introduced a new property for policy selection algorithms called fairness. We showed that importance sampling is unfair when used for policy selection even though it is an unbiased estimator for policy evaluation. We presented two approaches to deal with this issue, a theoretical solution and a new practical estimator. This is but a first step in tackling the issue of fairness in off-policy policy selection. Our hope is that introducing the notion of fairness for policy selection will result in growing interest in the challenges involved in off-policy policy selection, including how unfairness propagates to policy search methods that optimize over an infinite class of policies.

Acknowledgments

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A130215 and R305B150008 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Dept. of Education. The research was also supported in part by a NSF CAREER grant.

References

[Doroudi *et al.*, 2017] Shayan Doroudi, Philip S. Thomas, and Emma Brunskill. Importance sampling for fair policy selection. In *Uncertainity in Artificial Intelligence*. Association of Uncertainty in Artificial Intelligence, 2017.

[Dudík *et al.*, 2011] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, pages 1097–1104. Omnipress, 2011.

[Jabbari *et al.*, 2017] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *International Conference on Machine Learning*, pages 1617–1626. PMLR, 2017.

[Jiang and Li, 2016] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

[Jie and Abbeel, 2010] Tang Jie and Pieter Abbeel. On a connection between importance sampling and the likelihood ratio policy gradient. In *Advances in Neural Information Processing Systems*, pages 1000–1008. Curran Associates, Inc., 2010.

[Levine and Koltun, 2013] Sergey Levine and Vladlen Koltun. Guided policy search. In *International Conference on Machine Learning*, pages 1–9. PMLR, 2013.

[Mandel *et al.*, 2014] Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *Internatonal Conference on Autonomous Agents and Multi-Agent Systems*, pages 1077–1084. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[Powell and Swann, 1966] Michael J. D. Powell and J. Swann. Weighted uniform sampling—a Monte Carlo technique for reducing variance. *IMA Journal of Applied Mathematics*, 2(3):228–236, 1966.

[Precup *et al.*, 2000] Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *International Conference on Machine Learning*, pages 759–766. Morgan Kaufman, 2000.

[Thomas and Brunskill, 2016] Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR, 2016.

[Thomas and Brunskill, 2017] Philip S. Thomas and Emma Brunskill. Importance sampling with unequal support. In *AAAI*, pages 2646–2652. Association for the Advancement of Artificial Intelligence, 2017.

[Thomas *et al.*, 2015a] Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *AAAI*, pages 3000–3006. Association for the Advancement of Artificial Intelligence, 2015.

[Thomas *et al.*, 2015b] Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, pages 2380–2388. PMLR, 2015.

[Thomas *et al.*, 2017] Philip S. Thomas, Bruno Castro da Silva, Andrew G Barto, and Emma Brunskill. On ensuring that intelligent machines are well-behaved. *arXiv preprint arXiv:1708.05448*, 2017.

[Wang *et al.*, 2016] Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Remi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*, 2016.