Bias and Performance Disparities in Reinforcement Learning for Human–Robot Interaction

Zoe Evans
King's College London
London, UK
zoe.a.evans@kcl.ac.uk

Matteo Leonetti
King's College London
London, UK
matteo.leonetti@kcl.ac.uk

Martim Brandão
King's College London
London, UK
martim.brandao@kcl.ac.uk

***Abstract*—Bias has been shown to be a pervasive problem in machine learning, with severe and unanticipated consequences, for example in the form of algorithm performance disparities across social groups. In this paper, we investigate and characterise how similar issues may arise in Reinforcement Learning (RL) for Human–Robot Interaction (HRI), with the intent of averting the same ramifications. Using an assistive robotics simulation as a case study, we show that RL for HRI can perform differently across models with different waist circumferences. We show this behaviour can arise due to representation bias—unbalanced exposure during training—but also due to inherent task properties that may make assistance difficult depending on physical characteristics. The findings underscore the need to address bias in RL for HRI. We conclude with a discussion of potential practical solutions, their consequences and limitations, and avenues for future research.**

***Index Terms*—reinforcement learning, fairness, ethics**

I. Introduction

Machine Learning algorithms have recently been shown to reproduce harmful bias, whether through their decisions [10], [17] or through *performance disparities* [2], [4]—where performance disparities are differences in performance levels, such as accuracy, on different groups of people. Reinforcement Learning (RL) is a popular learning-based approach to robot control in Human–Robot Interaction (HRI) [1], [6], [11], but issues of bias in this context are not yet understood. If similar issues of performance disparities were to arise in HRI, they could lead to strong safety and technological-acceptance issues. For example, if an assistive robot could only perform assistive tasks to an acceptable or safe level on people that are similar to those the robot was trained on, it could lead certain social groups to be unable to use assistive robots, or to be at high risk of an accident when interacting with them. While issues of bias in supervised learning have often been identified and addressed only at a late stage in the development process, our goal is to avoid this pitfall and anticipate issues of bias in RL early on, especially before deployment in HRI production.

We use an assistive robotics case study to characterise the problem of performance disparities in RL for HRI, and to discuss potential bias mitigation approaches and their limitations.

Our contributions are as follows: 1) We experimentally demonstrate that RL for HRI can perform worse on minority groups; and that uniform exposure does not necessarily lead to unbiased performance. 2) We discuss the social and technical implications of performance disparities in HRI, potential mitigation strategies, and the limitations of those strategies in practice.

II. Related Work

As machine learning techniques have grown in use and impact, so has the scrutiny of the ethical implications of their decisions [12], such as biased and discriminatory outputs [10], [17]. Examples of these biases have been highlighted by many AI ethics researchers, for instance in racially biased gender recognition [4], gender biased hiring algorithms [17], and age bias in pedestrian detection [2] to name a few.

Machine learning bias can be thought of in different ways. For example, Buolamwini and Gebru [4] focus on performance differences across different groups, and Hundt et al. [10] focus on the propagation of socially undesirable stereotypes through robot actions. Researchers have shown that bias in the output of machine learning models can be traced to problems at the various stages of model development [21]. For example, they show it can arise due to historical bias (when our data reflects biased structures in the real world), representation bias (when certain groups are underrepresented in a dataset), measurement bias (when measured data is a proxy for ideal data, and has correlations or deficiencies related to social groups), amongst others [21]. We will consider bias in terms of different performance levels for different groups, where performance is measured, as in RL, in reward.

From the growing body of work in AI ethics, efforts have been made to also apply these concepts to embodied systems, such as robotics and autonomous vehicles. Hundt et al. [10] show how robots using CLIP [19] to make image-based decisions can propagate racist stereotypes. Also in robotics and computer vision, research has shown that pedestrian detection algorithms miss children in images twice as often as they miss adults [2], which is related to a lack of presence of children in training sets (i.e. representation bias). Bias has also been shown to creep into the performance of mobile robot planners when they take into account spatial statistics, such as population density for disaster response [3]. In social robotics, gender bias can occur through decisions of when to back-channel in conversation [18] due to imbalanced datasets and reductive rules for deciding robot cues/actions. Differently from the work described above, we focus on investigating the issue of bias in RL when applied to HRI systems.

Robots are being developed to solve tasks in social domains, such as healthcare [7], [14], [16], and RL is a popular approach

to robot control in such domains [1], [6], [11]. Presently, work has begun in characterising some of the potential issues of bias in RL. Whittlestone et al. [23] and Gajane et al. [8] delve into the societal implications and fairness concerns surrounding reinforcement learning (RL). Whittlestone et al. emphasise challenges in deep RL, including ethics, governance, lack of human oversight, safety, reliability, and the potential unintended consequences of reward function design. Gajane et al. survey fairness in RL, highlighting issues such as the interpretability and explainability of RL policies and a lack of focus on societal fairness. Both works underscore the need for addressing these challenges to ensure fair and transparent RL algorithms, although they do not investigate bias in the form of performance disparities.

III. Background

We briefly introduce the notation we are going to use for policies and what they maximise in RL, and the simulation environment on which our experiments are based.

A. Reinforcement Learning

We assume the standard reinforcement learning formulation as an episodic Markov Decision Process (MDP) $M = \langle S, A, T, R, \gamma \rangle$, where S is the set of states, A is the set of actions, T is the transition function, R is the immediate reward, and $0 \leq \gamma \leq 1$ is the discount factor. The agent acts according to a policy $\pi(a \mid s)$, where $a \in A$, $s \in S$, with the aim of maximising its return, that is, the cumulative expected reward G_t from the initial state at time $t = 0$ over a horizon H: $G_0 = E_\pi \left[\sum_{t=0}^{H} \gamma^t R_{t+1} \right]$. Episodes terminate after H actions and the environment is reset to an initial state.

B. Simulation Environment

To study performance disparities, we use an environment where robots collaborate with people with different characteristics. We modified Assistive Gym [6], which is a publicly available Reinforcement Learning environment for training robots in daily assistive-care tasks, such as bed bathing [6]. In its original implementation, human models in Assistive Gym match the average height and waist circumference of the US male [6], [22]. Assistive Gym also provides a representation for the policy π, and it can use RLlib's [15] implementation of SAC [9] for training, as used for our experiments.

IV. Research Questions

We investigate whether and how "representation bias" leads to performance disparities across groups in RL for HRI. In the rest of the paper, we will refer to "representation bias" to mean the under-representation, in training episodes, of groups of human participants with common characteristics.

We pose the following research questions:

- RQ1. Does imbalanced exposure to different groups during training lead to performance disparities between those groups? For example, training mainly on people with an average waist circumference, and then deploying on groups with below-average waist circumference.
- RQ2. Does balanced exposure to different groups during training lead to equal performance across groups?

V. Methodology

Answering each research question requires training RL policies on human models sampled from one distribution, and evaluating the policies on another distribution. We adapted Assistive Gym so that, in each training or evaluation episode, human models are sampled from a normal distribution of waist circumference—instead of both being fixed to the average values of the US Male population. We chose waist circumference as the varying human model parameter because there are published statistics [20], and they can be easily introduced in Assistive Gym. At every training episode the human model is thus spawned with a waist-circumference value sampled from a normal distribution. Two examples of such models in the simulation environment are shown in Figure 1.

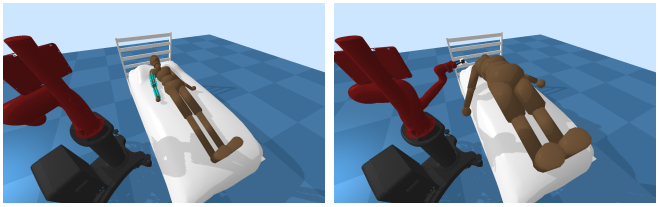


Fig. 1: The Bed Bathing environment of Assistive Gym, where a set of contact points needs to be touched (washed) with a sponge. Adapted so human models are spawned with waist circumference parameters drawn from a distribution.

All our experiments use the Bed Bathing environment of Assistive Gym, where the robot receives reward for each point it touches (i.e. washes) on a human model lying on a bed with a sponge. To answer RQ1-2 we trained multiple RL policies, each trained on a specific distribution of human waist-circumference. We trained the policies on GPU nodes of the a high-performance computing cluster, for up to 6,000,000 timesteps each. Each evaluation of a policy is made over 200 episodes, which corresponds to 20 seconds of robot behaviour, which we empirically verified to be a long enough horizon for the robot to complete the task or make no further progress. To alleviate issues with randomness in RL training and evaluation, each trial of training and evaluation was repeated three times, and results are therefore averages over 600 evaluation episodes (3 policies, 200 episodes per policy).

VI. Results

In this section, we report on the distributions, range of parameters, and results for the experiments to answer each research question. In each experiment, we measured performance in terms of number of contact points touched (i.e. washed) by the robot on the human model's arm.

All plots for all experiments show the average contact points hit across all trials. The error bars represent 95% confidence intervals of the reported averages. Statistical significance tests were obtained through the Student's t-test.

A. RQ1: Does imbalanced exposure to different groups during training lead to performance disparities?

To answer RQ1, we ran three experiments in which the human's height was fixed at the US Adult Male average, and

the waist circumferences (WC) were drawn in training from normal distributions with different means. In the first experiment, WC had a mean of 43.7cm. In the second experiment, WC had a mean of 87.43 (US Adult Male [20]). In the third experiment, WC had a mean of 122.4cm. We used a standard deviation of 12.99 in all cases. Then, we evaluated the resulting policies in the whole range of possible WC values.

Figure 2 shows that the policies performed better on groups that had more exposure during training: the policy trained on smaller-average WC had a performance peak at WC≈60, the policy trained on average WC peaked at WC≈96, and the policy trained on larger-average WC peaked at WC≈105. The WC performance peaks are not exactly equal to the means of the distribution, but there is a shift in performance with the increase of distribution mean.

Within each experiment, task performance dropped drastically between top and bottom-performing groups. When the policy was trained with a WC mean of 43.7cm, there was a 95% difference between the best-performing model (WC: 61.2cm) and the worst-performing model (WC: 122.4cm). When the policy was trained with a WC mean of 87.4cm, there was an 86.6% difference between the best (WC: 96.2cm) and the worst-performing model (WC: 43.7cm). And when the policy was trained with a WC mean of 122.4cm, there was a 97% percentage difference between the best (WC: 104.9) and the worst performing model (WC: 43.7cm). The differences between the best performing and worst performing models were statisticall significant in all 3 experiments ($p < 0.0001$).

B. RQ2: Does balanced exposure to different groups during training lead to equal performance across groups?

RQ1 showed that policies performed best on groups that had more exposure during training. Therefore, in RQ2 we investigated whether uniform training can be used as a solution to this problem. We fixed the height of the human model to the mean, and trained on a uniform distribution of waist circumference in the range 26.2cm–122.4cm.

Our results show there is an 80.2% difference between the best performing group (WC: 69.94cm) and worst performing group (WC: 122.4cm), and this difference is statistically significant $p < 0.0001$. Groups with the largest and smallest waist circumference performed worse than those with the average waist circumference, despite being seen with equal probability during training.

Therefore, our experiments show that uniform training does not necessarily solve the performance disparity problem. We hypothesize that differences observed are due to task properties that make it harder for the robot to perform well for small and large waist circumference groups, though further experiments are required in order to better characterize this dependency.

VII. Implications and Mitigations of Performance Disparities in RL-HRI

A. Implications

When robots are deployed to work with people at scale, it is important that they work well with all types of people. When trained with reinforcement learning, robots may perform better with the types of people they observed more often during training. In the case of Assistive Gym, our experiments show that a reinforcement learning agent trained on people following a population-level distribution of waist circumferences performs worse on people on the margins of that distribution–such as obese individuals. Performance disparities in robots that utilise RL to learn how to interact may have implications around trust, the quality of service of robots, and safety.

In the Assistive Gym environment, we have shown that providing the robot agent with a uniform distribution over different types of people does not necessarily eliminate performance disparities between different types of people. Training robots with a uniform distribution of people is also likely to be practically impossible in the real world, as a uniform distribution of people may be hard to get access to, or may consist an unethical practice if the selection is made across protected characteristics.

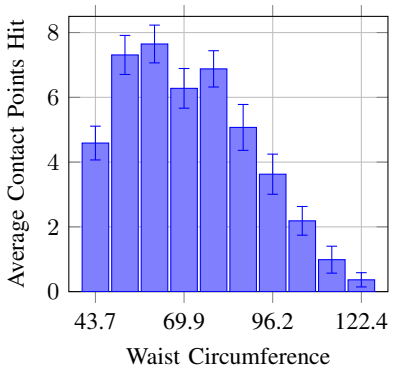
In clinical settings like assistive bathing, a robot may train with adults and later be deployed to assist children. In real-world assistive robotics, it may be natural to make the behaviour depend on the waist circumference of the person, avoiding the particular issues we have identified. However, there may be other, less obvious, features that a robot may inadvertently learn to maximise return for that cannot be anticipated easily during training.

In other contexts, such as the workplace, there may be further consequences to this disparity in return across groups of people. Human–robot collaboration is under active development for warehouses and factories [13]. Here, a robot may end up performing better with, or being safer around, one group of people in the warehouse than with another, leading to feelings of dissatisfaction and unfairness amongst workers [5].

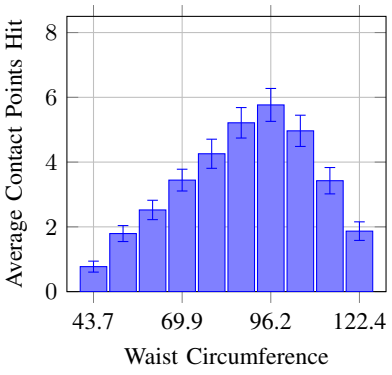
Alternatively, if performance disparities between groups are not attributed to the robot's behaviour, further issues could occur for the part of the workforce that the robot performs worse on. A robot may work well with a majority group in the workplace, with which it was trained, but perform collaborative tasks poorly (more slowly, making costly mistakes more often) with the minority group. It may appear that the robot is performing its tasks well in general and therefore it is the minority group who is under-performing. If the workers' pay is tied to an economic output quota, as is the case in many warehouses, then an unfair financial disparity may also emerge between the majority and minority groups.

A possible adjustment could be to train different policies for different types of people, rather than have one singular policy that works across all groups. For example, we can treat each task with each group of people as a different task entirely. Then, policies can be learnt that maximise the return for each group specifically. However, we would need to know at deployment time the specific features that belong to the person the robot is working with. Accessing a person's features at deployment time might not always be feasible—especially features such as height, weight, race, gender, class, etc.

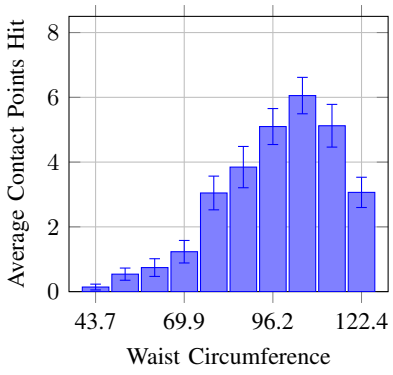
A technical solution to the problem of performance disparities would be to re-weight rewards based on group membership, increasing the rewards of less-seen or lower-return



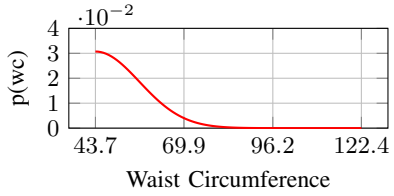
(a) Policy trained on humans with average WC=43.7



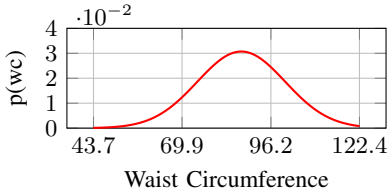
(b) Policy trained on humans with average WC=87.43



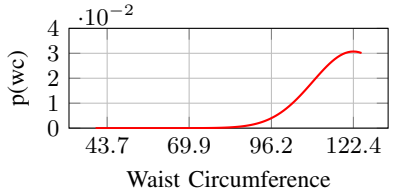
(c) Policy trained on humans with average WC=122.4



(d) Distribution of training population



(e) Distribution of training population



(f) Distribution of training population

Fig. 2: Task performance when interacting with humans of different waist circumference. Three different policies. Differences between performance of the top performing WC and bottom-performing WC are statistically significant in all cases $p < 0.0001$.

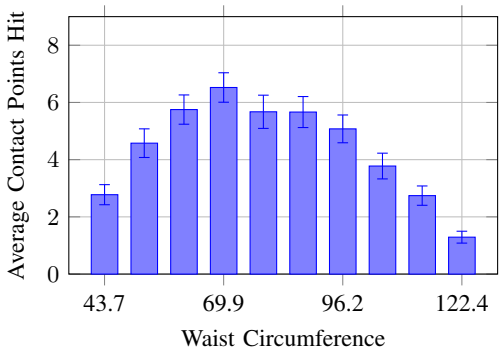


Fig. 3: Task performance when interacting with humans of different waist circumferences. Underlying policy trained on a human model following a uniform distribution of waist circumference.

groups so as to make the policy perform better on those groups. Similarly, we could adjust the learning rate to speed up and give more emphasis to lower-return groups in training. However, this may be difficult to implement in practice, due to the need for tuning weights or exploration parameters and their trade-offs. A higher learning rate may also lead to less predictable actions, as values change more rapidly, and therefore decrease safety.

VIII. Conclusions

We investigated the problem of performance disparities in RL for HRI using Assistive Gym as a case study. We found that robots that learn to interact with humans using RL can perform their tasks worse on people they do not see as often in training. We also found that uniform exposure to people does not necessarily lead to uniform performance, as task difficulty may depend on the physical characteristics of humans.

We then discussed how, in practice, researchers and industry could attempt to mitigate these problems, using different technical and socio-technical approaches, and the potential implications of such approaches. These implications concerned precarious work contracts, direct and indirect discrimination of different groups in hiring or robot access, and technical challenges of bias mitigation in reinforcement learning for robotics.

Some limitations of our study include the fact that we only analyse a Bed Bathing task in Assistive Gym, and a single training algorithm (SAC). In the future, it would be interesting to test whether this phenomenon is more or less pronounced in other types of tasks, training algorithms, and policy network architectures.

Interesting directions of further work include investigating how to create fairer reinforcement learning algorithms for HRI that do not depend on knowing the features at deployment time (i.e, making separate models for separate groups), or do not depend on knowing group features at all. In the future, more research on bias in RL should be conducted, for example on bias encoded in reward functions and simulation environments, bias mitigation algorithms, and social-technical approaches to bias mitigation.

Acknowledgements

This work was supported by the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence [EP/S023356/1].

References

[1] Akalin, N., Loutfi, A.: Reinforcement learning approaches in social robotics. Sensors **21**(4), 1292 (2021)

[2] Brandao, M.: Age and gender bias in pedestrian detection algorithms. arXiv preprint arXiv:1906.10490 (2019)

[3] Brandao, M., Jirotka, M., Webb, H., Luff, P.: Fair navigation planning: a resource for characterizing and designing fairness in mobile robots. Artificial Intelligence **282**, 103259 (2020)

[4] Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77–91. PMLR (2018)

[5] Claure, H., Chen, Y., Modi, J., Jung, M., Nikolaidis, S.: Multi-armed bandits with fairness constraints for distributing resources to human teammates. In: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. pp. 299–308 (2020)

[6] Erickson, Z., Gangaram, V., Kapusta, A., Liu, C.K., Kemp, C.C.: Assistive gym: A physics simulation framework for assistive robotics. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 10169–10176. IEEE (2020)

[7] Fasola, J., Matarić, M.J.: A socially assistive robot exercise coach for the elderly. Journal of Human-Robot Interaction **1**(1), 1–16 (2010)

[8] Gajane, P., Saxena, A., Tavakol, M., Fletcher, G., Pechenizkiy, M.: Survey on fair reinforcement learning: Theory and practice. arXiv preprint arXiv:2205.10032 (2022)

[9] Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al.: Soft actor-critic algorithms and applications. arXiv preprint arXiv:1812.05905 (2018)

[10] Hundt, A., Agnew, W., Zeng, V., Kacianka, S., Gombolay, M.: Robots enact malignant stereotypes. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 743–756 (2022)

[11] Jakhotiya, Y., Haque, I.: Improving assistive robotics with deep reinforcement learning. arXiv preprint arXiv:2209.02160 (2022)

[12] Jobin, A., Ienca, M., Vayena, E.: The global landscape of ai ethics guidelines. Nature machine intelligence **1**(9), 389–399 (2019)

[13] Krnjaic, A., Steleac, R.D., Thomas, J.D., Papoudakis, G., Schäfer, L., To, A.W.K., Lao, K.H., Cubuktepe, M., Haley, M., Börsting, P., et al.: Scalable multi-agent reinforcement learning for warehouse logistics with robotic and human co-workers. arXiv preprint arXiv:2212.11498 (2022)

[14] Kyrarini, M., Lygerakis, F., Rajavenkatanarayanan, A., Sevastopoulos, C., Nambiappan, H.R., Chaitanya, K.K., Babu, A.R., Mathew, J., Makedon, F.: A survey of robots in healthcare. Technologies **9**(1), 8 (2021)

[15] Liang, E., Liaw, R., Nishihara, R., Moritz, P., Fox, R., Goldberg, K., Gonzalez, J., Jordan, M., Stoica, I.: Rllib: Abstractions for distributed reinforcement learning. In: International conference on machine learning. pp. 3053–3062. PMLR (2018)

[16] Ozkil, A.G., Fan, Z., Dawids, S., Aanes, H., Kristensen, J.K., Christensen, K.H.: Service robots for hospitals: A case study of transportation tasks in a hospital. In: 2009 IEEE international conference on automation and logistics. pp. 289–294. IEEE (2009)

[17] Parasurama, P., Sedoc, J.: Gendered information in resumes and its role in algorithmic and human hiring bias. In: Academy of Management Proceedings. vol. 2022, p. 17133. Academy of Management Briarcliff Manor, NY 10510 (2022)

[18] Parreira, M.T., Gillet, S., Winkle, K., Leite, I.: How did we miss this? a case study on unintended biases in robot social behavior. In: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. pp. 11–20 (2023)

[19] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

[20] Sabo, R.T., Ren, C., Sun, S.S.: Comparing height-adjusted waist circumference indices: the fels longitudinal study. Open journal of endocrine and metabolic diseases **2**(3), 40 (2012)

[21] Suresh, H., Guttag, J.: A framework for understanding sources of harm throughout the machine learning life cycle. In: Equity and access in algorithms, mechanisms, and optimization, pp. 1–9 (2021)

[22] Tilley, A.R., et al.: The measure of man and woman: human factors in design. John Wiley & Sons (2001)

[23] Whittlestone, J., Arulkumaran, K., Crosby, M.: The societal implications of deep reinforcement learning. Journal of Artificial Intelligence Research **70**, 1003–1030 (2021)