Towards Fairness In Reinforcement Learning

Alexandra Cimpean
Vrije Universiteit Brussel
Brussels, Belgium
ioana.alexandra.cimpean@vub.be

Pieter Libin
Vrije Universiteit Brussel
Brussels, Belgium
pieter.libin@vub.be

Youri Coppens
Vrije Universiteit Brussel
Brussels, Belgium
youri.coppens@vub.be

Catholijn Jonker
Technische Universiteit Delft
Delft, The Netherlands
c.m.jonker@tudelft.nl

Ann Nowé
Vrije Universiteit Brussel
Brussels, Belgium
ann.nowe@vub.be

ABSTRACT

Automated decision support systems, based on reinforcement learning, are increasingly important across a wide range of complex problem settings that consider individuals or groups of individuals. To use such systems in the real world, fairness of treatment is an essential trait to allow stakeholders to make informed decisions that balance the performance-fairness trade-off. In this work, we propose a universal framework to establish fairness in reinforcement learning agents, with regards to multiple fairness notions. To this end, we formulate sequential fairness notions in function of groups and individuals. First, we present a Markov decision process that is explicitly aware of individuals and group arrangements. Next, we formalise fairness notions in terms of this extended Markov decision process, by maintaining a history of states, actions, rewards, and ground truth feedback. Based on this formalism we classify distinct fairness settings and identify key research challenges to implement this reinforcement learning framework.

KEYWORDS

reinforcement learning, automated decision support, fairness framework, trustworthy AI

1 INTRODUCTION

Fair and balanced automated decision support is essential, to avoid discrimination or favouritism towards individuals and groups. This is crucial in a wide array of applications, such as finance [15], job hiring [24, 25], epidemic mitigation [3, 7, 14] and fraud detection [19]. Fair decision support systems allow stakeholders to make informed decisions taking into account an appropriate performance-fairness trade-off. This is important, as advice that is proposed by a virtual agent potentially impacts individuals and groups. Therefore, it is vital to study this matter to enable a wider acceptance of algorithms that support decision makers. As fairness requirements depend on the problem context and the decision maker's preferences, a framework should be capable of dealing with multiple fairness notions, that encompass the ethical considerations of the problem domain. Consequently, it is important to develop a framework that considers fairness based on sensitive features (e.g., race and gender) and their combinations.

Previous work mainly focused on supervised learning techniques that operate on a given dataset [5, 6, 8, 17, 18]. However, automated decision problems are typically sequential. Furthermore, the setting evolves over time and as such a reinforcement learning (RL) approach is warranted [4]. This means that we must deal with the impact of short-term decisions on long term performance. RL enables an agent to learn a policy by interacting with an environment [27]. At each time t, the agent observes the state s_t of the environment and decides on an action a_t to take, for which it receives a reward r_t and observes the next state s_{t+1}. The agent learns through trial and evaluation by repeatedly interacting with the environment, where it must carefully balance exploration and exploitation to reach an optimal policy [27]. Additionally, the agent may need to deal with stochastic and non-stationary environments where it must adapt its behaviour to maintain its performance.

In a supervised classification setting, the ground truth is known and it is used to train the model. Based on this ground truth, a confusion matrix is derived to reflect on the correctness of the model's predictions. By definition, reinforcement learning agents do not have a priori access to a ground truth, as the agent collects data while interacting with an environment. Therefore, actions taken by the agent cannot be classified to be correct or false, which impedes the use of fairness notions that rely on a confusion matrix. As most fairness notions rely on the ground truth, they are only applicable when feedback regarding this ground truth can be collected from the environment [17].

It is important to emphasise that this ground truth is different from the reward signal in a reinforcement learning setting. While the reward signal may indicate how suitable an action is given a state, it does not conclusively specify whether the action was correct or false. As with the reward, feedback concerning the ground truth can be sparse or delayed, providing limited feedback during most agent-environment interactions. To illustrate this, consider the example of job hiring, where we receive delayed feedback as the candidate can only be evaluated after working for some time. Moreover, candidates can only be evaluated if they are hired and not when they are declined.

Recent work on fairness in RL has focused on single fairness notions in application-specific solutions [2, 11, 12, 23, 26, 29] and typically relies on reward shaping [2, 16]. However, such an approach does not suffice for real-world decision support problems as the desired performance-fairness trade-off cannot a priori be defined by stakeholders. Moreover, certain problem settings require multiple fairness notions to be taken into account simultaneously. We therefore argue that a multi-objective reinforcement learning approach is warranted.

Proc. of the Adaptive and Learning Agents Workshop (ALA 2023), Cruz, Hayes, Wang, Yates (eds.), May 29-30, 2023, Online, https:// ala2023.github.io/. 2023.

2 FAIRNESS FRAMEWORK

We describe the various components of the fairness framework, along with their requirements and suitability regarding distinct problem settings. As the presence of the ground truth is required for some fairness notions, it must be either obtained through feedback or approximated based on previous interactions.

To introduce fairness notions in an RL context, we illustrate them based on three running examples. The first example concerns job hiring, where the aim is to hire highly qualified candidates while limiting bias towards sensitive features. The second is an epidemic mitigation example, that aims at imposing contact reductions in an efficient yet fair way. The third example covers fraud detection, where fraudulent transactions must be cleverly flagged, taking into account that verification requires human effort. These running examples provide an initial overview of distinct fairness concerns, to indicate current challenges of implementing a framework for fair RL algorithms.

We highlight that RL can be used both directly or indirectly in the context of real-world problems. On the one hand, in the epidemic mitigation example, a detailed simulator is used to train an agent, after which the learned policies can be studied by public health experts [1, 30]. On the other hand, in a fraud detection setting the agent may learn directly in the real world to flag suspicious transactions.

2.1 Fairness history

A sequential decision process can be formally defined as a Markov Decision Process (MDP) [27], consisting of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a set of rewards $\mathcal{R}$ and a transition function $p : \mathcal{S}\times\mathcal{R}\times\mathcal{S}\times\mathcal{A} \rightarrow [0, 1]$ describing the probability of a next state s_{t+1} and reward r_t given the current state s_t and action a_t. We extend this standard MDP to an fMDP to encode a feedback signal f_t, that concerns which was the correct action $\hat{a}_t$ or rather an indication if the chosen action a_t was correct at time t. Note that this feedback is optional and can be partial, sparse or delayed.

We introduce the following notation regarding individuals and groups. $\mathcal{I}_t$ refers to the set of individuals involved in the decision process at time t and we use i_t to refer to an individual of that set. In the job hiring example, $\mathcal{I}_t$ refers to the set of candidates who applied for the job at time t and on which a decision (i.e., hire or reject the applicant) should be made. In the epidemic mitigation example, $\mathcal{I}_t$ refers to the entire population when deciding on who to impose contact restrictions. We refer to the set of all individuals involved in the decision process from the start $t = 0$ up to time T as $\mathcal{I}^T$.

We define $\mathcal{G}_{g,t}$ as the individuals of $\mathcal{I}_t$ that make up group g. We refer to all individuals involved in the decision process until time T, that belong to group g, as $\mathcal{G}_g^T$. For ease of notation, we assume that groups are predefined and can be empty. In the job hiring example, $\mathcal{G}_g^T$ refers to the group of men or women, who applied for a job before time T. For the epidemic mitigation example, $\mathcal{G}_g^T$ refers to an age group for which the RL agent must decide whether or not to impose contact restrictions.

Given the fMDP, we assume that a state s_t provided to the RL agent encodes the individuals $\mathcal{I}_t$ and groups $\mathcal{G}_t$ involved in the decision at time t. Furthermore, the agent's action a_t encodes the decision impacting the involved individuals and groups, and the feedback f_t specifies the correctness of that decision. We use the following notation to connect $\mathcal{I}_t$ and $\mathcal{G}_t$ to s_t, a_t and f_t:

$$\mathcal{I}_t[s_t],\ \mathcal{I}_t[a_t],\ \mathcal{I}_t[f_t] \tag{1}$$

$$\mathcal{G}_t[s_t],\ \mathcal{G}_t[a_t],\ \mathcal{G}_t[f_t] \tag{2}$$

To define fairness over time, e.g., to consider new individuals applying for a job, a history of encountered states and chosen actions needs to be maintained, with regards to the impacted individuals and groups. Given an fMDP, we define a history $\mathcal{H}^T$ until time T of past interaction tuples and their feedback regarding the ground truth:

$$\mathcal{H}^T = \{s_t, a_t, r_t, f_t\}_{t=0}^T \tag{3}$$

We define the encountered states and selected actions from history $\mathcal{H}^T$ until time T respectively as $\mathcal{H}_S^T$ and $\mathcal{H}_A^T$. We refer to feedback regarding the correctness of the action as $\mathcal{H}_f^T$. Following from the definitions in Equations 1 and 2, $\mathcal{H}_S^T$, $\mathcal{H}_A^T$ and $\mathcal{H}_f^T$ contain all information available regarding groups $\mathcal{G}^T$ and individuals $\mathcal{I}^T$.

2.2 Fairness notions

We formally define a fairness notion $\mathcal{F}$ as a power set $\mathcal{P}$ over $\mathcal{G}^T$ groups (Equation 4) and $\mathcal{I}^T$ individuals (Equation 5), given the history of encountered states $\mathcal{H}_S^T$, chosen actions $\mathcal{H}_A^T$ and feedback $\mathcal{H}_f^T$ until time T:

$$\mathcal{F} : \mathcal{P}(\mathcal{G}^T) \times \mathcal{P}(\mathcal{H}_S^T) \times \mathcal{P}(\mathcal{H}_A^T) \times \mathcal{P}(\mathcal{H}_f^T) \hookrightarrow \mathbb{R} \tag{4}$$

$$\mathcal{F} : \mathcal{P}(\mathcal{I}^T) \times \mathcal{P}(\mathcal{H}_S^T) \times \mathcal{P}(\mathcal{H}_A^T) \times \mathcal{P}(\mathcal{H}_f^T) \hookrightarrow \mathbb{R} \tag{5}$$

The fairness notion $\mathcal{F} \leq 0$ is defined as the negative absolute difference in treatment between groups or individuals. The closer $\mathcal{F}$ is to zero, the less of a difference there is in treatment between the groups or individuals. When $\mathcal{F} = 0$, the agent has achieved exact fairness with respect to the given fairness notion. While $\mathcal{F}$ may be intractable due to limitations of defining exact fairness [11], we propose to approximate it with $\hat{\mathcal{F}}$. For a future fairness objective, $\mathcal{F}$, and by extension its approximation $\hat{\mathcal{F}}$ provide a starting point for a reward signal that can be used with a multi-objective RL approach.

The availability of a ground truth and as a consequence the confusion matrix[1] impacts which fairness notions can be calculated for a given scenario. Consider the group fairness notion *statistical parity* [5], where the probability of receiving the preferable treatment of the agent ($\mathcal{H}_A^T = 1$) should be the same across groups g and h:

$$\begin{aligned}\mathcal{F} = -|\mathrm{P}(\mathcal{G}_g^T[\mathcal{H}_A^T] = 1|\mathcal{G}_g^T[\mathcal{H}_S^T]) - \\ \mathrm{P}(\mathcal{G}_h^T[\mathcal{H}_A^T] = 1|\mathcal{G}_h^T[\mathcal{H}_S^T])|\end{aligned} \tag{6}$$

Statistical parity requires that $(TP + FP)/(TP + FP + FN + TN)$ is equal for both groups g and h. Because this fairness notion focuses on equal acceptance rate across groups, it can be expressed without

[1]The confusion matrix is defined as a two-dimensional table comparing predictions of a model to the actual values. In the case of binary actions (e.g., hire or reject an applicant) it specifies the number of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN).

knowledge of the ground truth. Other fairness notions require that the ground truth is (partially) known, such as *equal opportunity*

$$\mathcal{F} = -|\mathrm{P}(\mathcal{G}_g^T[\mathcal{H}_A^T] = 1|\mathcal{G}_g^T[\mathcal{H}_f^T] = 1, \mathcal{G}_g^T[\mathcal{H}_S^T]) - \mathrm{P}(\mathcal{G}_h^T[\mathcal{H}_A^T] = 1|\mathcal{G}_h^T[\mathcal{H}_f^T] = 1, \mathcal{G}_h^T[\mathcal{H}_S^T])| \tag{7}$$

where $\mathcal{H}_f^T = 1$ is the correct action as specified by the feedback regarding the ground truth. Equal opportunity requires that the recall or true positive rate $TP/(TP+FN)$ is equal across groups and is consequently independent of FP. However, in order to calculate it, we require a (partial) ground truth which informs us about TP and FN. In the job hiring example, this requires knowing how qualified a job candidate is to calculate the confusion matrix. One example of a setting where a partial ground truth is available is fraud detection, where transactions flagged as fraudulent are manually checked and provide the number of TP and FP. In contrast, there is no information on unflagged transactions which consequently does not support fairness notions relying on FN or TN unless random checks would be performed, or when individuals complain about fraud cases in their experience.

Ensuring people are treated fairly, with regards to all groups they are a part of, is achieved by ensuring all their groups are treated fairly with regards to each other. If the interest is that the individual itself receives fair treatment, then individual fairness notions should be used instead.

Individual fairness notions aim to treat similar individuals similarly [5]. Given two individuals i_t and j_t, we assume a distance $d(i_t, j_t)$ between the individuals. Given the probability distributions M_i and M_j over the actions for i_t and j_t respectively, and a distance metric $D(M_i||M_j)$ between these probability distributions, individual fairness requires that:

$$\forall i_t, j_t \in \mathcal{I}_t : \quad D(M_i||M_j) \leq d(i_t, j_t) \tag{8}$$

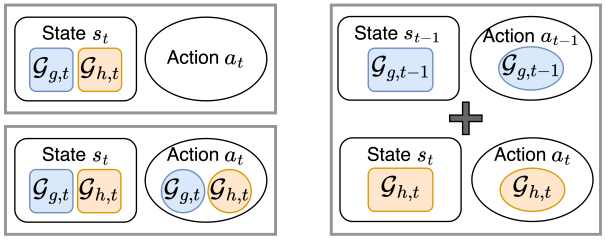
As group fairness notions aim to treat groups that differ by a set of sensitive features similarly, they cannot detect unfairness at an individual level, as all attributes except the sensitive ones are ignored [5]. Similarly, individual fairness notions lack the ability to ensure fairness between groups. Ideally, an RL agent conforms to a collection of both group and individual fairness notions to manage this trade-off, which can be managed using a multi-objective learning approach. We refer to the work of Hayes et al. for an overview of multi-objective reinforcement learning [10].

By formulating fairness notions in terms of the history defined in the previous section, we establish a formal way to reason about fairness notions as reward functions. Yet, as maintaining the full history will prove computationally intractable for most real-world applications, a major challenge remains to construct approximative fairness notions. One research direction is to consider a sliding window approach, where the history is kept for a fixed or dynamic number of steps [21]. Another path is to explore the use of distinct neural sub-networks for the different fairness notions.

2.3 Fairness in sequential decision making

Defining fairness in a sequential setting requires knowledge on how fairness notions can be defined given the agent-environment interactions. Consider the epidemic control example, where an agent must decide how to impose contact reductions each day for an entire country [22]. Throughout the day, all individuals participate in different contact pools such as work, school or community. Therefore, the agent aims to enforce appropriate contact reductions, such that everyone in the population is subjected to similar restrictions.

Suppose in our epidemic control example, that each week the agent encounters the different age groups that make up the population. Then each week, the agent chooses contact reductions for the respective age groups. Then at each time t, given an observed state s_t and chosen action a_t, given $\mathcal{G}_t$ groups, a group fairness notion can be defined if s_t contains all respective groups $\mathcal{G}_t[s_t]$ and the chosen action a_t represents the action taken towards each group $\mathcal{G}_t[a_t]$. Figure 1a visualises the possible scenarios with regards to the available action, which can be an action over all groups $\mathcal{G}_t$, or a specific action for each group g specifically. Note that if individuals are defined within the state representation, then Equation 2 can be defined by grouping individuals in $\mathcal{I}_t$ under their respective groups $\mathcal{G}_t$.



(a) (b)

Figure 1: Scenarios where group fairness can be calculated. (a) All groups $\mathcal{G}_t$ are encountered at each time t. Top: action a_t is an action over all groups $\mathcal{G}_t$. Bottom: action a_t encodes a specific action for each group g. (b) All groups $\mathcal{G}_t$ are encountered over a history until time T. The + symbol indicates a union over states and actions.

Next in the epidemic control example, consider that the agent only encounters certain age groups on a weekly basis, which could be the case when youngsters have school vacations at different times, as is the case in The Netherlands [20]. In this instance, youngsters participate more in the community during the week than adults. Then a sufficiently long time horizon must be considered to encounter all age groups. Concretely, if the state s_t contains only information on a subset $\mathcal{B}_t \subset \mathcal{G}_t$ of the respective groups, a fairness notion can only be defined when considering multiple timesteps of encountered groups $\mathcal{B}^T$ to contain sufficient information about all impacted $\mathcal{G}_t$ groups for time t:

$$\mathcal{G}_t[s_t] = \bigcup_{g \in \mathcal{B}^T} \mathcal{G}_g^T[\mathcal{H}_S^T] \tag{9}$$

Similarly, we require multiple timesteps if the action a_t does not define the action for all groups:

$$\mathcal{G}_t[a_t] = \bigcup_{g \in \mathcal{B}^T} \mathcal{G}_g^T[\mathcal{H}_A^T] \tag{10}$$

If individuals are defined within the state representation of the environment, Equations 9 and 10 can be extended to consider cases

where a subset of individuals is encountered. Figure 1b visualises the scenario where only a subset of the groups is available at each time t, requiring a history of timesteps in order to express group fairness notions.

Following up on the same epidemic control example, when the agent encounters all individuals of the population each week, then individual fairness notions can be calculated for the imposed contact reductions. To define an individual fairness notion for $\mathcal{I}_t$ individuals at time t, given an observed state s_t and a chosen action a_t, we require that $\mathcal{I}_t[s_t]$ and $\mathcal{I}_t[a_t]$ are defined. Figure 2a visualises the scenarios where individual fairness can be calculated at each time t. Note that the action can be fine-grained for each individual or coarse-grained over their respective age groups or contact pools.

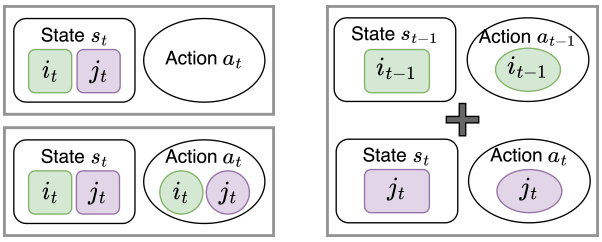


Figure 2: Scenarios where individual fairness can be calculated. (a) All individuals $\mathcal{I}_t$ are encountered at each time t. Top: action a_t is an action over all individuals $\mathcal{I}_t$. Bottom: action a_t encodes a specific action for each individual i_t. (b) All individuals $\mathcal{I}_t$ are encountered over a history until time T. The + symbol indicates a union over states and actions.

When only a portion of the individuals is encountered each time step, then we can only calculate individual fairness notions when we maintain a history of interactions. An example for epidemic control is performing PCR tests for children at different times during the week, to prevent new infections by closing the exposed classes or schools [28]. In this case, a fair agent should balance over time which classes are closed to not cause certain children to miss out on more education than others. If state s_t does not contain all $\mathcal{I}_t$ individuals but rather a subset $C_t \subset \mathcal{I}_t$, an individual fairness notion can be defined over multiple timesteps so that all affected individuals C^T are encountered:

$$\mathcal{I}_t[s_t] = \bigcup_{i \in C^T} i^T[\mathcal{H}_S^T] \tag{11}$$

$$\mathcal{I}_t[a_t] = \bigcup_{i \in C^T} i^T[\mathcal{H}_A^T] \tag{12}$$

Figure 2b visualises the scenario where individual fairness can be expressed over multiple timesteps. Note how both group and individual fairness notions can be expressed if the encountered states contain all necessary information about the respective groups and individuals. Regardless of whether the action was specifically assigned to them, their group or the entire population, we can compare the action which affects them to calculate fairness notions.

Figure 3 visualises the agent-environment interactions and how the state, actions, rewards and feedback are maintained by our fairness framework. Note that environments can be stochastic and non-stationary, which impacts how and when individuals and groups are encountered.

Depending on the setting, it could be more important to check fairness notions against the impact of the agent's action rather than against the action itself. We discussed actions with regards to the applicability of fairness notions, however both the immediate and estimated effect follow similar rules as information about them must also be available in the agent-environment interactions. An example of using the impact of an action concerns social contact reduction in the epidemic setting, where the population must reduce contacts to prevent that hospitals' intensive care units are overwhelmed. One possible way of looking at fairness is to require that different age groups are treated equally when it comes to imposed restrictions. However, if the ultimate goal is to reduce severe disease, fairness notions can be considered from the perspective of reducing the number of severely ill for each age group equally.

2.4 Learning and exploration

In the previous sections, we assume that the states in the history encompass all groups $\mathcal{G}^T$ and individuals $\mathcal{I}^T$ necessary to compute the relevant fairness notions. However, to meet this assumption, the relevant states need to be encountered, which is highly dependent on how the agent interacts with the environment. To establish this, we need an appropriate exploration strategy that ensures that sufficient information is collected about all groups $\mathcal{G}^T$ and individuals $\mathcal{I}^T$. On the one hand, to guarantee optimality, this exploration strategy will need to collect information on groups and individuals as broadly as possible. On the other hand, to keep the process computationally tractable, the exploration strategy should be effective and targeted. Note that, as we aspire to settings that aim to support decision makers, we can learn and evaluate policies in simulated environments, prior to deploying them in the real world. This facilitates a model-based reinforcement learning loop that could mitigate the hurdle of computationally intensive exploration strategies.

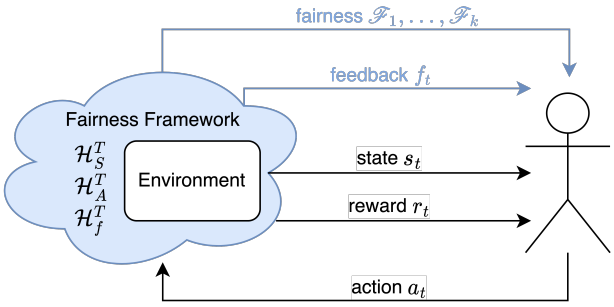


Figure 3: Agent-environment interaction with our fairness framework. The agent interacts with the underlying environment through the fairness framework, to calculate fairness notions given the history of interactions.

3 OUTLOOK

We propose a framework for exploring the use of fairness notions in RL. In this framework, we establish a formulation of fairness notions that can be used as additional reward signals following a multi-objective learning approach. Based on this formulation, we classify distinct fairness settings grounded in real-world problems.

Generalising fairness notions to continuous actions presents an interesting venue to extend fairness to a wider array of problem settings. In the field of regression, algorithms produce a scalar value rather than a discrete action from a predefined set. Consequently, regression compares actions based on how much they differ and can detect correlations between the action and one or more sensitive features [13], which makes it an interesting approach for comparing actions in RL.

Within the overarching topic of ethics, work on explainable AI focuses on making algorithms interpretable and provides explanations for their decisions [9]. While explainability aims to provide transparency with regards to an agent's decisions and policy, fairness focuses on whether or not the agent makes decisions which conform to expected impartial treatment. We argue that fairness is an equally important aspect to focus on to work towards ethical AI. To truly build a fair decision support system, we envision the need to combine fairness notions with explainable reinforcement learning, such that fairness can be taken into account when explaining policies to the decision maker.

ACKNOWLEDGMENTS

Alexandra Cimpean receives funding from the Fonds voor Wetenschappelijk Onderzoek (FWO) via fellowship grant 1SF7823N. Pieter Libin gratefully acknowledges support from FWO postdoctoral fellowship 1242021N, FWO grant G059423N, and the Research council of the Vrije Universiteit Brussel (OZR-VUB via grant number OZR3863BOF). Catholijn Jonker's work is supported by the National Science Foundation (NWO) under grant number 1136993. Ann Nowé and Youri Coppens acknowledge support from FWO grant G062819N.

REFERENCES

[1] Steven Abrams, James Wambua, Eva Santermans, Lander Willem, Elise Kuylen, Pietro Coletti, Pieter Libin, Christel Faes, Oana Petrof, Sereina A. Herzog, Philippe Beutels, and Niel Hens. 2021. Modelling the early phase of the Belgian COVID-19 epidemic using a stochastic compartmental model and studying its implied future trajectories. *Epidemics* 35 (2021), 100449.

[2] Jingdi Chen, Yimeng Wang, and Tian Lan. 2021. Bringing fairness to actor-critic reinforcement learning for network utility optimization. In *IEEE Conference on Computer Communications*. IEEE Press, Vancouver, BC, Canada, 1–10.

[3] Alexandra Cimpean, Timothy Verstraeten, Lander Willem, Niel Hens, Ann Nowé, and Pieter Libin. 2023. Evaluating COVID-19 vaccine allocation policies using Bayesian m-top exploration. *arXiv preprint arXiv:2301.12822* (2023), 26.

[4] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. 2020. Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Barcelona, Spain, 525–534.

[5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. ACM, New York, NY, USA, 214–226.

[6] Cynthia Dwork, Christina Ilvento, Guy N. Rothblum, and Pragya Sur. 2020. Abstracting Fairness: Oracles, Metrics, and Interpretability. In *1st Symposium on Foundations of Responsible Computing*. Curran Associates, Inc., 16.

[7] Ezekiel J. Emanuel, Govind Persad, Adam Kern, Allen Buchanan, Cécile Fabre, Daniel Halliday, Joseph Heath, Lisa Herzog, R. J. Leland, Ephrem T. Lemango, Florencia Luna, Matthew S. McCoy, Ole F. Norheim, Trygve Ottersen, G. Owen Schaefer, Kok-Chor Tan, Christopher Heath Wellman, Jonathan Wolff, and Henry S. Richardson. 2020. An ethical framework for global vaccine allocation. *Science* 369, 6509 (2020), 1309–1312.

[8] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016), 16.

[9] Bryce Goodman and Seth Flaxman. 2017. European union regulations on algorithmic decision making and a "right to explanation". *AI Magazine* 38, 3 (2017), 50–57.

[10] Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. 2022. A practical guide to multi-objective reinforcement learning and planning. In *AAMAS* (2022/04/13), Vol. 36. 26.

[11] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in Reinforcement Learning. In *ICML*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. PMLR, Sydney, Australia, 1617–1626.

[12] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2016. Fairness in Learning: Classic and contextual bandits. *Advances in Neural Information Processing Systems* 29 (2016), 325–333.

[13] Junpei Komiyama, Akiko Takeda, Junya Honda, and Hajime Shimao. 2018. Nonconvex Optimization for Regression with Fairness Constraints. In *ICML*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholm, Sweden, 2737–2746.

[14] Pieter J. K. Libin, Arno Moonens, Timothy Verstraeten, Fabian Perez-Sanjines, Niel Hens, Philippe Lemey, and Ann Nowé. 2021. Deep Reinforcement Learning for Large-Scale Epidemic Control. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, Yuxiao Dong, Georgiana Ifrim, Dunja Mladenić, Craig Saunders, and Sofie Van Hoecke (Eds.). Springer International Publishing, Cham, 155–170.

[15] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. In *35th International Conference on Machine Learning*, Vol. 80. PMLR, Stockholm, Sweden, 3150–3158.

[16] Weiwen Liu, Feng Liu, Ruiming Tang, Ben Liao, Guangyong Chen, and Pheng Ann Heng. 2020. *Balancing Between Accuracy and Fairness for Interactive Recommendation with Reinforcement Learning*. Vol. 12084 LNAI. Springer International Publishing, Cham. 155–167 pages.

[17] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. On the applicability of ML fairness notions. *SIGKDD Explor. Newsl.* (2020), 32.

[18] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (2021), 35 pages.

[19] Dennis Soemers, Ann Nowé, Tim Brys, Kurt Driessens, and Mark Winands. 2018. Adapting to Concept Drift in Credit Card Transaction Data Streams Using Contextual Bandits and Decision Trees. *AAAI* 32, 1 (2018), 7831–7836.

[20] Government of the Netherlands. 2023. Setting school holiday dates. https://www.government.nl/topics/school-holidays/setting-school-holiday-dates

[21] Javier Ortiz Laguna, Angel García Olaya, and Daniel Borrajo. 2011. A Dynamic Sliding Window Approach for Activity Recognition. In *User Modeling, Adaption and Personalization*, Joseph A. Konstan, Ricardo Conejo, José L. Marzo, and Nuria Oliver (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 219–230.

[22] Mathieu Reymond, Conor F Hayes, Lander Willem, Roxana Rădulescu, Steven Abrams, Diederik M Roijers, Enda Howley, Patrick Mannion, Niel Hens, Ann Nowé, and Pieter Libin. 2022. Exploring the pareto front of multi-objective covid-19 mitigation policies using reinforcement learning. *arXiv preprint arXiv:2204.05027* (2022), 25.

[23] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A Rodriguez-Aguilar. 2021. Guaranteeing the Learning of Ethical Behaviour through Multi-Objective Reinforcement Learning. *ALA* (2021), 9.

[24] Candice Schumann, Samsara N. Counts, Jeffrey S. Foster, and John P. Dickerson. 2019. The Diverse Cohort Selection Problem. In *AAMAS*. 601–609.

[25] Candice Schumann, Jeffrey S. Foster, Nicholas Mattei, and John P. Dickerson. 2020. We need fairness and explainability in algorithmic hiring. In *AAMAS*. 1716–1720.

[26] Umer Siddique, Paul Weng, and Matthieu Zimmer. 2020. Learning fair policies in multiobjective (Deep) reinforcement learning with Average and Discounted Rewards. *ICML* 119 (13–18 Jul 2020), 8864–8874.

[27] Richard S. Sutton, Andrew G. Barto, and et al. 2018. *Reinforcement Learning : An Introduction*. MIT Press. 526 pages.

[28] Andrea Torneri, Lander Willem, Vittoria Colizza, Cécile Kremer, Christelle Meuris, Gilles Darcis, Niel Hens, and Pieter JK Libin. 2022. Controlling SARS-CoV-2 in schools using repetitive testing strategies. *eLife* 11 (2022), 23.

[29] Paul Weng. 2019. Fairness in reinforcement learning. *arXiv preprint arXiv:1907.10323* (2019), 5.

[30] Lander Willem, Steven Abrams, Pieter J.K. Libin, Pietro Coletti, Elise Kuylen, Oana Petrof, Signe Møgelmose, James Wambua, Sereina A. Herzog, Christel Faes, Philippe Beutels, and Niel Hens. 2021. The impact of contact tracing and household bubbles on deconfinement strategies for COVID-19. *Nature Communications* 12, 1 (2021), 1–9.