Using Protected Attributes to Consider Fairness in Multi-Agent Systems

Gabriele La Malfa[1,*], Jie M. Zhang[1], Michael Luck[2] and Elizabeth Black[1]

[1]UKRI Centre for Doctoral Training in Safe and Trusted AI, King's College London, London, WC2B 4BG
[2]University of Sussex, Brighton, BN1 9RH

Abstract

Fairness in Multi-Agent Systems (MAS) has been extensively studied, particularly in reward distribution among agents in scenarios such as goods allocation, resource division, lotteries, and bargaining systems. Fairness in MAS depends on various factors, including the system's governing rules, the behaviour of the agents, and their characteristics. Yet, fairness in human society often involves evaluating disparities between disadvantaged and privileged groups, guided by principles of Equality, Diversity, and Inclusion (EDI). Taking inspiration from the work on algorithmic fairness, which addresses bias in machine learning-based decision-making, we define *protected attributes* for MAS as characteristics that should not disadvantage an agent in terms of its expected rewards. We adapt fairness metrics from the algorithmic fairness literature—namely, *demographic parity*, *counterfactual fairness*, and *conditional statistical parity*—to the multi-agent setting, where self-interested agents interact within an environment. These metrics allow us to evaluate the fairness of MAS, with the ultimate aim of designing MAS that do not disadvantage agents based on protected attributes.

Keywords

Fairness, bias, Multi-Agent Systems (MAS)

1. Introduction

Multi-Agent Systems (MAS) consist of agents interacting with each other and their surrounding environment to achieve their individual or shared goals. The achievement of an agent's goals may depend on the actions it takes, the actions of other agents, the environment they are situated in, and the rules that govern the MAS. Similarly, fairness in MAS depends on multiple factors. Fairness can be influenced by agents' decision-making processes, as evidenced by research in reinforcement learning focused on developing fair and efficient policies [1]. It can also hinge on mechanism design, as seen in scenarios like goods allocation games [2] or cake-cutting problems [3], where rules can ensure fair reward distribution among agents. Additionally, fairness can be affected by things like an agent's utility [4, 5] or their priority in accessing resources [6, 7], among others.

In human societies, fairness is often defined in terms of characteristics that should not disadvantage an individual or group, such as age, race, disability or gender. For example, in the UK Equality Act 2010[1] these are identified as *protected characteristics*, and UK law states that individuals cannot be discriminated against on the basis of these. These protected characteristics typically define subgroups of the population who have historically been disadvantaged in particular situations, such as age discrimination in the workplace, unequal access to healthcare or barriers in education for people with disabilities and gender disparities in political representation, among others. Driven by the bias that often exists in the training data as a result of these systemic inequalities, machine learning approaches often produce biased results (e.g., discrimination in credit market [8] or justice [9, 10] algorithms); there is a growing body of work (often referred to as *algorithmic fairness*) that aims to identify and mitigate such bias by applying a range of fairness metrics that compare the outcomes achieved by what is identified as advantaged and disadvantaged subgroups of the population (see, e.g., [11, 12] for a review).

AEQUITAS 2024: Workshop on Fairness and Bias in AI | co-located with ECAI 2024, Santiago de Compostela, Spain
*Corresponding author.
gabriele.la_malfa@kcl.ac.uk (G. La Malfa); jie.zhang@kcl.ac.uk (J. M. Zhang); michael.luck@sussex.ac.uk (M. Luck); elizabeth.black@kcl.ac.uk (E. Black)

[1] https://www.gov.uk/guidance/equality-act-2010-guidance

Taking inspiration from the UK Equality Act 2010, we define the concept of *protected attributes* within a multi-agent system, which are any attributes that have been deemed should not disadvantage an agent in terms of its performance within that system. For example, consider a multi-agent setting that includes both artificial agents in the form of autonomous vehicles and human agents who drive their own cars; we may want to ensure that the human agents are not disadvantaged in such a setting. We adapt the following fairness metrics from the algorithmic fairness literature to our multi-agent setting.

- *Demographic parity* – Agents with and without protected attributes should obtain the same expected rewards.
- *Counterfactual fairness* – In both a factual and a counterfactual scenario, where the only difference is whether the protected attributes hold for an agent, agents should obtain the same expected rewards.
- *Conditional statistical parity* – Within a group of agents characterised by a legitimate factor influencing rewards, agents with and without protected attributes should obtain the same expected rewards.

We are able to evaluate different MAS according to these metrics, with the ultimate aim of designing fairer MAS (for example, by configuring the environment in which agents operate to optimise for fairness). Such an approach is inspired by other works outside MAS, such as designing accessible buildings [13] or safe urban environments [14]. Further studies explore environment configurations to optimise rescue operations and autonomous vehicle planning [15, 16]. However, none of them deal with fairness. Hence, we hope this research can offer valuable insights into domains beyond MAS.

To summarise, the contributions of this paper are as follows. We introduce *protected attributes* to MAS – characteristics that should not impact an agent's expected rewards, all other things being equal. We adapt the concepts of *demographic parity*, *counterfactual fairness* and *conditional statistical parity* from the algorithmic fairness literature to the MAS context. The future aim of this work is to use these metrics to evaluate and optimise MAS for fairness.

Motivating example. In future urban environments, we may see vehicles operated by humans and vehicles operated by AI undertaking journeys within the same road network. These human and AI agents navigate city streets to reach their destinations, with the rewards they receive dependent on things like time taken and cost of journey. AI-driven vehicles excel by analysing traffic data in real-time, optimising routes, and communicating with other AI vehicles, providing them with an advantage over the human agents in the system, who are generally less efficient at route optimisation and less well-equipped to coordinate with other road users. To mitigate this advantage of AI agents, we might consider altering the road infrastructure, for example, by providing dedicated lanes for human-controlled vehicles.

2. Related work

Fairness has attracted the attention of Game Theory and MAS researchers for decades alongside psychologists and economists [17, 18, 19, 20]. Factors such as the rules that govern the system can influence fairness in MAS. For instance, this can be seen in the Ultimatum Game, where fairness is influenced by the dynamics between proposers and responders [21, 22, 23]. In goods allocation or cake-cutting games, the rules depend on the type of good being allocated, for example, whether they are divisible or indivisible, goods or chores [24, 25], and fairness depends on the distribution of goods among the agents [2, 3, 7, 26, 27].

Agent behaviour can also influence fairness. Fair behaviours often balance the rewards collected by the community and individuals. For example, Zhang and Shah [28] propose a minimum reward for the worst-performing agent while improving the overall rewards of the whole community of agents. However, fairness and reward optimisation can be in tension, and compromises must be made regarding one of the two sides. Jiang and Lu [29] propose a two-step solution consisting of a single policy for each agent based on fair and optimal rewards, with a controller agent who decides which sub-policies

to implement to maximise environmental rewards and fairness. Other works [30, 31, 32] implement fair optimisation policies within cooperative multi-agent systems, aiming to integrate individualistic and altruistic behaviours. Grupen *et al.* [33] introduce a new measure of team fairness, demonstrating how maximising team rewards in cooperative MAS can lead to unfair outcomes for individual agents.

In contrast to these works, which do not distinguish agents that may be particularly disadvantaged within a system, we consider fairness across agents who do or do not possess protected attributes. We adapt demographic parity [34, 35], counterfactual fairness [35] and conditional statistical parity [36] fairness metrics from the algorithmic fairness literature to the MAS setting.

3. Preliminaries

A multi-agent system consists of multiple decision-making agents who act and interact in an environment to achieve their goals. A *multi-agent system* $S = (E, e_o, Ac, P, At, At^{pr}, \tau)$ is characterised by: the set of possible *environment states* E; the starting state e_0; the set of available *actions* that may be performed by an agent in the environment Ac (including a null action); a *population* $P = \{a_1, \ldots, a_n\}$ of *agents*; the *attributes* $At = \{at_1, \ldots, at_m\}$ available to the agents in P; the *protected attributes* $At^{pr} \subset \{at_1, \ldots, at_m\}$; and the non-deterministic *state transformer function* $\tau : E \times Ac_1 \times \ldots \times Ac_n \to E \times [0, 1]$ that specifies the probability distribution over the possible resulting states that can occur when each agent in the population performs an action (where the possible null action reflects that an agent chooses not to act).

An *agent* a_x within a multi-agent system $(E, e_o, Ac, P, At, At^{pr}, \tau)$ (where $a_x \in P$) is defined as a tuple $(At_x, Ac_x, \pi_x, \rho_x)$ where: the *attribute evaluation function* $At_x : At \to \{0, 1\}$ specifies which attributes hold true for the agent; $Ac_x \subseteq Ac$ are the *actions available to the agent*; the non-deterministic *policy* $\pi_x : E \to Ac_x \times [0, 1]$ specifies how an agent will act in any given state (represented as a probability distribution over the possible actions); and the *reward function* $\rho_x : E \times E \to \mathbb{R}$ specifies the reward the agent receives for moving between two states.

A *possible run* within a multi-agent system $S = (E, e_o, Ac, P, At, At^{pr}, \tau)$ (where P consists of n agents) is denoted $r = (e_0, (ac_1^1, \ldots, ac_1^n), e_1, \ldots, (ac_j^1, \ldots, ac_j^n), e_j)$ where: for each $a_x \in P$ and for each i such that $0 < i \leq j$, $(ac_i^x, p) \in \pi_x(e_{i-1})$ and $p > 0$; and for each i such that $0 \leq i < j$, $(e_{i+1}, p) \in \tau(e_i, (ac_i^1, \ldots, ac_i^n))$ and $p > 0$. The set of *all possible runs* within a multi-agent system S is denoted $\mathcal{R}^S$.

Let $r = (e_0, (ac_1^1, \ldots, ac_1^n), e_1, \ldots, (ac_j^1, \ldots, ac_j^n), e_j) \in \mathcal{R}^S$ where $S = (E, e_o, Ac, P, At, At^{pr}, \tau)$. We can determine the probability r will occur, denoted $p(r \mid S)$, as follows.

$$p(r \mid S) = \left(\prod_{i=0}^{j-1}\left(\prod_{x=1}^{n} p_x \text{ where } (ac_{i+1}^x, p_x) \in \pi_x(e_i)\right)\right) \cdot \left(\prod_{i=0}^{j-1} p_i \text{ where } (e_{i+1}, p_i) \in \tau(e_i, (ac_{i+1}^1, \ldots, ac_{i+1}^n))\right)$$

For a run $r = (e_0, (ac_1^1, \ldots, ac_1^n), e_1, \ldots, (ac_j^1, \ldots, ac_j^n), e_j)$, the reward achieved by an agent a_x is $Rew(a_x, r) = \sum_{i=1}^{j} \rho_x(e_{i-1}, e_i)$.

The *expected reward* of an agent a_x within a system S, denoted $ExpRew(a_x, S)$, is thus $ExpRew(a_x, S) = \sum_{r \in \mathcal{R}^S} Rew(a_x, r).p(r \mid S)$.

Motivating example, continued. The city traffic consists of a population of cars, each capable of steering, accelerating or braking. Cars also possess attributes like speed or safety features. Cars are either driven by AI or humans, and we consider being driven by humans to be a protected attribute of cars. AI-driven cars can find optimal paths to reach their destination more efficiently than human-driven ones. If we consider agents reaching a hospital, we can foresee fairness problems as AI-driven cars would be advantaged. When the cars act with a specific probability, the environment changes state. Also, each car obtains a reward when reaching its destination. A car's policy is a decision rule based on the state of the crossroads.

4. Fairness in MAS

We define fairness by comparing, in different ways, the rewards gathered by individuals or groups of agents possessing and not possessing protected attributes. We adapt *demographic parity* [34, 35], *counterfactual fairness* [35] and *conditional statistical parity* [36] to MAS.

Demographic parity in MAS is achieved when the expected rewards of agents are not influenced by whether or not they possess protected attributes, all else being equal.

Definition 1 (Demographic Parity). *Let $S = (E, e_o, Ac, P, At, At^{pr}, \tau)$ be a system and let $at^{pr} \in At^{pr}$ be the protected attribute under consideration. Demographic parity is satisfied for at^{pr} in S if and only if: for all $a_x, a_y \in P$, if $At_x(at^{pr}) = 1$, $At_y(at^{pr}) = 0$, and for all $at' \in At \setminus \{at^{pr}\}$, $At_x(at') = At_y(at')$, then $ExpRew(a_x, S) = ExpRew(a_y, S)$.*

Where demographic parity is not satisfied for a particular protected attribute, we can measure the extent to which this is the case, denoted $DemPar(at^{pr}, S)$, as follows.

$$DemPar(at^{pr}, S) = \sum_{\substack{a_x, a_y \in P \text{ such that } At_x(at^{pr})=1,\ At_y(at^{pr})=0, \\ \text{and for all } at' \in At \setminus \{at^{pr}\}, At_x(at') = At_y(at')}} ExpRew(a_x, S) - ExpRew(a_y, S) \tag{1}$$

Note that if demographic parity holds for at^{pr} in S then $DemPar(at^{pr}, S) = 0$.

Counterfactual fairness in MAS is achieved when the expected rewards of agents remain the same in both a factual and a counterfactual world, where in the latter, we change the protected attribute of the agents while keeping all other elements the same.

Definition 2 (Counterfactual Fairness). *Let $S = (E, e_o, Ac, P, At, At^{pr}, \tau)$ be a system where $P = \{(At_1, Ac_1, \pi_1, \rho_1), \ldots, (At_n, Ac_n, \pi_n, \rho_n)\}$, and let $at^{pr} \in At^{pr}$ be the protected attribute under consideration. Let $S' = (E, e_o, Ac, P', At, At^{pr}, \tau)$ be the counterfactual of S such that $P' = \{(At'_1, Ac_1, \pi_1, \rho_1), \ldots, (At'_n, Ac_n, \pi_n, \rho_n)\}$ where for all i such that $1 \leq i \leq n$: if $At_i(at^{pr}) = 0$, then $At'_i(at^{pr}) = 1$; if $At_i(at^{pr}) = 1$, then $At'_i(at^{pr}) = 0$; and for all $at \in At \setminus \{at^{pr}\}$, $At_i(at) = At^a_i prime(at)$. Counterfactual fairness is satisfied for at^{pr} in S if and only if: for all $a_x = (At_x, Ac_x, \pi_x, \rho_x) \in P$, for all $a'_x = (At'_x, Ac_x, \pi_x, \rho_x) \in P'$, $ExpRew(a_x, S) = ExpRew(a'_x, S')$. Where counterfactual fairness is not satisfied, we can measure the extent to which this is the case, denoted $CountFair(at^{pr}, S)$, as follows.*

$$CountFair(at^{pr}, S) = \sum_{a_x \in P \text{ such that } At_x(at^{pr})=1} ExpRew(a_x, S) - ExpRew(a'_x, S') \tag{2}$$

Note that if counterfactual fairness holds for at^{pr} in S then $CountFair(at^{pr}, S) = 0$.

Conditional statistical parity in MAS is achieved when the expected rewards of agents are not influenced by whether or not they possess protected attributes when conditioning on a legitimate factor, assuming all other elements are the same. A legitimate factor is an attribute that has been identified as one that may legitimately affect an agent's reward.

Definition 3 (Conditional Statistical Parity). *Let $S = (E, e_o, Ac, P, At, At^{pr}, \tau)$ be a system, let $LF \subseteq (At \setminus At^{pr})$ be the set of legitimate factors, and let $at^{pr} \in At^{pr}$ be the protected attribute under consideration. Conditional statistical parity is satisfied for at^{pr} with LF in S if and only if: for all $a_x, a_y \in P$, if $At_x(at^{pr}) = 1$, $At_y(at^{pr}) = 0$, $At_x(at^{lf}) = At_y(at^{lf}) = 1$ for all $at^{lf} \in LF$, and for all $at' \in At \setminus \{at^{pr}\}$, $At_x(at') = At_y(at')$, then $ExpRew(a_x, S) = ExpRew(a_y, S)$.*

Where conditional statistical parity is not satisfied, we can measure the extent to which this is the case, denoted $CondSP(at^{pr}, LF, S)$, as follows.

$$CondSP(at^{pr}, LF, S) = \sum_{\substack{a_x, a_y \in P \text{ such that } At_x(at^{pr})=1,\ At_y(at^{pr})=0,\\ At_x(at^{lf})=At_y(at^{lf})=1 \text{ for all } at^{lf} \in LF,\\ \text{and for all } at' \in At \setminus \{at^{pr}\},\ At_x(at')=At_y(at')}} ExpRew(a_x, S) - ExpRew(a_y, S) \tag{3}$$

Note that if conditional statistical parity holds for at^{pr} with LF in S then $CondSP(at^{pr}, LF, S) = 0$.

Conditional statistical parity is demographic parity within subsets of the population characterised by legitimate factors. For example, in algorithmic fairness, such a metric is used to verify whether the probability of predicting re-offence for male and female prisoners is the same for similar age groups, which is the legitimate factor [37].

Motivating example, continued. In the city traffic example, demographic parity would be achieved if the sum of the expected rewards obtained by AI-driven cars and human-driven cars were equal, all other things being equal. In other words, the protected attribute should not affect the expected rewards gathered by the human-driven cars compared to the AI-driven ones. Counterfactual fairness is achieved if the sum of the expected rewards of the cars remains the same in both a factual and a counterfactual world, where in the latter, agents possess the protected attribute (i.e., cars are driven by humans) while keeping all other factors constant. Conditional statistical parity is achieved if the sum of the cars' expected rewards is not influenced by whether or not they possess protected attributes when conditioned on a legitimate factor, e.g., a certain range of speed capacity of the cars, assuming all other elements are the same.

We can use the metrics above to measure fairness of different systems. Our ultimate goal is to optimise systems for these different fairness measures, for example by adjusting the starting state of the environment, or the way the environment responds to the agents' actions.

5. Conclusion and future work

This paper is a first step towards ensuring that certain sub-groups of agents are not disadvantaged in multi-agent systems. We identify *protected attributes*, which are characteristics that should not disadvantage an agent in terms of its expected rewards. Inspired by algorithmic fairness, we adapt *demographic parity*, *counterfactual fairness* and *conditional statistical parity* to analyse fairness in MAS. Our metrics assess fairness from various perspectives in any multi-agent system where expected rewards are applicable. Additional metrics from the algorithmic fairness literature, such as equal opportunity, equalised odds [38], disparate impact [39], or other metrics based on causal reasoning [40, 41] could be adapted to this setting to capture other aspects of fairness. Our methodology applies to MAS, involving both human and AI agents, as motivated by our example. It could also be used to improve the fairness of human societies by modelling these as multi-agent systems and seeing how changes to the system affect the various fairness metrics defined here.

In future work, we plan to analyse these fairness metrics experimentally in different settings, both competitive and cooperative, to find system configurations that enhance fairness. We will use techniques such as Bayesian optimisation [42], evolutionary algorithms [43] and sparse sampling techniques [44] to try to identify system configurations that optimise for the different fairness metrics.

Acknowledgments

This work was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org).

References

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

[2] G. Amanatidis, H. Aziz, G. Birmpas, A. Filos-Ratsikas, B. Li, H. Moulin, A. A. Voudouris, X. Wu, Fair division of indivisible goods: Recent progress and open questions, Artificial Intelligence 322 (2023) 103965. URL: https://www.sciencedirect.com/science/article/pii/S000437022300111X. doi:https://doi.org/10.1016/j.artint.2023.103965.

[3] A. D. Procaccia, Cake cutting: not just child's play, Commun. ACM 56 (2013) 78–87. URL: https://doi.org/10.1145/2483852.2483870. doi:10.1145/2483852.2483870.

[4] U. Endriss, N. Maudet, Welfare engineering in multiagent systems, in: A. Omicini, P. Petta, J. Pitt (Eds.), Engineering Societies in the Agents World IV, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 93–106.

[5] D. Bertsimas, V. Farias, N. Trichakis, The price of fairness, Operations Research 59 (2011) 17–31. doi:10.1287/opre.1100.0865.

[6] S. De Jong, K. Tuyls, K. Verbeeck, N. Roos, Priority awareness: Towards a computational model of human fairness for multi-agent systems, in: K. Tuyls, A. Nowe, Z. Guessoum, D. Kudenko (Eds.), Adaptive Agents and Multi-Agent Systems III. Adaptation and Multi-Agent Learning, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 117–128.

[7] X. Bu, Z. Li, S. Liu, J. Song, B. Tao, Fair division with prioritized agents, in: Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'23/IAAI'23/EAAI'23, AAAI Press, 2023. URL: https://doi.org/10.1609/aaai.v37i5.25688. doi:10.1609/aaai.v37i5.25688.

[8] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, A. Walther, Predictably unequal? the effects of machine learning on credit markets, The Journal of Finance 77 (2022) 5–47. doi:https://doi.org/10.1111/jofi.13090.

[9] "Fair" Risk Assessments: A Precarious Approach for Criminal Justice Reform, Stockholm, Sweden, 2018.

[10] J. Johndrow, K. Lum, An algorithm for removing sensitive information: Application to race-independent recidivism prediction, The Annals of Applied Statistics 13 (2017). doi:10.1214/18-AOAS1201.

[11] B. Hutchinson, M. Mitchell, 50 years of test (un)fairness: Lessons for machine learning, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 49–58. URL: https://doi.org/10.1145/3287560.3287600. doi:10.1145/3287560.3287600.

[12] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, K. Lum, Algorithmic fairness: Choices, assumptions, and definitions, Annual Review of Statistics and Its Application (2021). URL: https://api.semanticscholar.org/CorpusID:228893833.

[13] M. Zallio, P. J. Clarkson, Inclusion, diversity, equity and accessibility in the built environment: A study of architectural design practice, Building and Environment 206 (2021) 108352. URL: https://www.sciencedirect.com/science/article/pii/S0360132321007496. doi:https://doi.org/10.1016/j.buildenv.2021.108352.

[14] J. Thompson, M. Stevenson, J. S. Wijnands, K. A. A Nice, G. DPA, J. Silver, M. Nieuwenhuijsen, P. Rayner, R. Schofield, R. Hariharan, C. N. Morrison, A global analysis of urban design types and road transport injury: an image processing study, The Lancet Planetary Health 4 (2020) e32–e42. URL: https://www.sciencedirect.com/science/article/pii/S2542519619302633. doi:https://doi.org/10.1016/S2542-5196(19)30263-3.

[15] J. Kozůbek, Z. Flasar, I. Dumišinec, Military factors influencing path planning, in: U. Z. A. Hamid, V. Sezer, B. Li, Y. Huang, M. A. Zakaria (Eds.), Path Planning for Autonomous Vehicle, IntechOpen, Rijeka, 2019. URL: https://doi.org/10.5772/intechopen.86421. doi:10.5772/intechopen.86421.

[16] S. Karma, E. Zorba, G. Pallis, G. Statheropoulos, I. Balta, K. Mikedi, J. Vamvakari, A. Pappa,

M. Chalaris, G. Xanthopoulos, M. Statheropoulos, Use of unmanned vehicles in search and rescue operations in forest fires: Advantages and limitations observed in a field trial, International Journal of Disaster Risk Reduction 13 (2015) 307–312. URL: https://www.sciencedirect.com/science/article/pii/S2212420915300364. doi:https://doi.org/10.1016/j.ijdrr.2015.07.009.

[17] E. Fehr, K. M. Schmidt, A theory of fairness, competition, and cooperation, The Quarterly Journal of Economics 114 (1999) 817–868. URL: http://www.jstor.org/stable/2586885.

[18] A. Falk, U. Fischbacher, A theory of reciprocity, Games and Economic Behavior 54 (2006) 293–315. URL: https://www.sciencedirect.com/science/article/pii/S0899825605000254. doi:https://doi.org/10.1016/j.geb.2005.03.001.

[19] S. De Jong, K. Tuyls, K. Verbeeck, Fairness in multi-agent systems, The Knowledge Engineering Review 23 (2008) 153–180. doi:10.1017/S026988890800132X.

[20] M. A. Nowak, K. M. Page, K. Sigmund, Fairness versus reason in the ultimatum game, Science 289 (2000) 1773–1775. doi:10.1126/science.289.5485.1773.

[21] D. G. Rand, C. E. Tarnita, H. Ohtsuki, M. A. Nowak, Evolution of fairness in the one-shot anonymous ultimatum game, Proceedings of the National Academy of Sciences 110 (2013) 2581–2586. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1214167110. doi:10.1073/pnas.1214167110. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1214167110.

[22] T. Cimpeanu, C. Perret, T. A. Han, Cost-efficient interventions for promoting fairness in the ultimatum game, Knowledge-Based Systems 233 (2021) 107545. URL: https://www.sciencedirect.com/science/article/pii/S0950705121008078. doi:https://doi.org/10.1016/j.knosys.2021.107545.

[23] J.-Y. Kim, K.-M. Lee, Evolution of fairness in the divide-a-lottery game, Scientific Reports 13 (2023). doi:10.1038/s41598-023-34131-w.

[24] H. Aziz, I. Caragiannis, A. Igarashi, T. Walsh, Fair allocation of indivisible goods and chores, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 53–59. URL: https://doi.org/10.24963/ijcai.2019/8. doi:10.24963/ijcai.2019/8.

[25] H. Hosseini, A. Mammadov, T. Was, Fairly allocating goods and (terrible) chores, 2023. arXiv:2305.01786.

[26] H. Aziz, B. Li, S. Xing, Y. Zhou, Possible fairness for allocating indivisible resources, in: Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS '23, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2023, p. 197–205.

[27] B. Li, H. Ma, Double-deck multi-agent pickup and delivery: Multi-robot rearrangement in large-scale warehouses, IEEE Robotics and Automation Letters 8 (2023) 3701–3708. doi:10.1109/LRA.2023.3272272.

[28] C. Zhang, J. A. Shah, Fairness in multi-agent sequential decision-making, in: Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 27, Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/792c7b5aae4a79e78aaeda80516ae2ac-Paper.pdf.

[29] J. Jiang, Z. Lu, Learning fairness in multi-agent systems, 2019. arXiv:1910.14472.

[30] E. Hughes, J. Z. Leibo, M. Phillips, K. Tuyls, E. Dueñez Guzman, A. García Castañeda, I. Dunning, T. Zhu, K. McKee, R. Koster, H. Roff, T. Graepel, Inequity aversion improves cooperation in intertemporal social dilemmas, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 31, Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/7fea637fd6d02b8f0adf6f7dc36aed93-Paper.pdf.

[31] J. X. Wang, E. Hughes, C. Fernando, W. M. Czarnecki, E. A. Duéñez Guzmán, J. Z. Leibo, Evolving intrinsic motivations for altruistic behavior, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2019, p. 683–692.

[32] M. Zimmer, C. Glanois, U. Siddique, P. Weng, Learning fair policies in decentralized cooperative

multi-agent reinforcement learning, 2021. arXiv:2012.09421.

[33] N. A. Grupen, B. Selman, D. D. Lee, Cooperative multi-agent fairness and equivariant policies, 2022. arXiv:2106.05727.

[34] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 214–226. URL: https://doi.org/10.1145/2090236.2090255. doi:10.1145/2090236.2090255.

[35] M. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4069–4079.

[36] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, Algorithmic decision making and the cost of fairness, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 797–806. URL: https://doi.org/10.1145/3097983.3098095. doi:10.1145/3097983.3098095.

[37] R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, Fairness in criminal justice risk assessments: The state of the art, Sociological Methods & Research 50 (2021) 3–44. doi:10.1177/0049124118782533.

[38] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, 2016. arXiv:1610.02413.

[39] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, 2015. arXiv:1412.3756.

[40] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, B. Schölkopf, Avoiding discrimination through causal reasoning, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 656–666.

[41] R. Nabi, I. Shpitser, Fair inference on outcomes, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18, AAAI Press, 2018.

[42] J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms, 2012. arXiv:1206.2944.

[43] P. A. Vikhar, Evolutionary algorithms: A critical review and its future prospects, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), 2016, pp. 261–265. doi:10.1109/ICGTSPICC.2016.7955308.

[44] M. Kearns, Y. Mansour, A. Y. Ng, A sparse sampling algorithm for near-optimal planning in large markov decision processes, in: Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'99, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999, p. 1324–1331.