Mechanism-Informed Learning for Fair Division

Extended Abstract

Ryota Maruo
Kyoto University
Kyoto, Japan
mryota@ml.ist.i.kyoto-u.ac.jp

Tomohiko Yokoyama
The University of Tokyo
Tokyo, Japan
tomohiko_yokoyama@mist.i.u-tokyo.ac.jp

Ayumi Igarashi
The University of Tokyo, RIKEN AIP
Tokyo, Japan
igarashi@mist.i.u-tokyo.ac.jp

Koh Takeuchi
Kyoto University, RIKEN AIP
Kyoto, Japan
takeuchi@i.kyoto-u.ac.jp

ABSTRACT

Fair division provides a simple yet powerful framework for modeling fairness in resource allocation. While existing literature typically assumes complete information about preferences, many practical scenarios involve incomplete preferences, posing challenges in computing fair allocations. In this paper, we propose mechanism-informed preference learning, a framework that integrates neural networks with differentiable approximations of classical fair division mechanisms—adjusted winner, round-robin, and moving-knife—to estimate fair allocations from incomplete preferences. Experiments on real-world household chore preference data show that our mechanism-informed framework achieves fairer allocations, compared to methods without mechanism information.

KEYWORDS

Fair Division; Machine Learning

ACM Reference Format:
Ryota Maruo, Tomohiko Yokoyama, Ayumi Igarashi, and Koh Takeuchi. 2026. Mechanism-Informed Learning for Fair Division: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/HTOP9978

1 INTRODUCTION

Fair division of indivisible resources is a fundamental problem. Fairness notions include envy-freeness (EF) [8], proportionality (PROP) [13], and their relaxations EF1 [2, 4] and PROP1 [7], which are guaranteed by mechanisms such as the adjusted winner (AW) [2, 3], the round-robin (RR) [6], and the moving-knife (MK) [1, 2]. While theoretical methods assume that agents have complete preferences over all items, practical applications often involve incomplete preferences due to time limits, cognitive constraints, or the difficulty of evaluating numerous alternatives [5, 12].

In this paper, we study the problem of computing fair allocations of indivisible items from incomplete preferences. Given a dataset of complete preferences, we propose a mechanism-informed preference learning (MIPL), a framework that minimizes violations of fairness on the complete preference of the allocations computed from estimated preferences. MIPL uses self-supervised learning [9] to generate incomplete preferences and trains a preference estimator that minimizes fairness violations through differentiable approximations of fair-division mechanisms.

We validate MIPL using a new real-world dataset for household chore preferences from 2,000 participants in Japan. We simulate incomplete preferences by multiple missing methods, and show that MIPL yields fairer allocations from incomplete preferences.

2 PROBLEM SETTING

Fair Division. We study the fair division of indivisible chores among agents $N = [n]$ and chores $M = [m]$. Each agent i has a non-negative disutility vector $\boldsymbol{v}_i = (v_{i,1}, \dots, v_{i,m}) \in \mathbb{R}^m_{\geq 0}$, and a *profile* is the matrix $V \in \mathbb{R}^{n\times m}_{\geq 0}$ containing all $\boldsymbol{v}_i$'s. A profile is *normalized* if $\sum_j v_{i,j} = 1$ for all i. An (fractional) *allocation* is a matrix $X \in [0,1]^{n\times m}$ with $\sum_i x_{i,j} = 1$ for all j, and is *integral* if $x_{i,j} \in \{0,1\}$. Agent i's disutility is $v_i(\mathbf{x}_i) = \sum_j v_{i,j} x_{i,j}$. An allocation X is *envy-free (EF)* if $v_i(\mathbf{x}_i) \leq v_i(\mathbf{x}_{i'})$ for all i, i'. As EF may not exist for indivisible chores, we use *EF1*: for every i, i' with $\mathbf{x}_i \neq \emptyset$, there exists $j \in \mathbf{x}_i$ such that $v_i(\mathbf{x}_i \setminus \{j\}) \leq v_i(\mathbf{x}_{i'})$. Similarly, X is *proportional (PROP)* if $v_i(\mathbf{x}_i) \leq \frac{1}{n} v_i(M)$ for all i, and *PROP1* if the inequality holds after removing one chore from $\mathbf{x}_i$. A *mechanism* $\mathcal{M}$ maps a normalized profile to an allocation. We consider AW and RR, which guarantee EF1, and MK, which guarantees PROP1.

Learning Problem. We aim to compute fair allocations from incomplete preferences. During training, we are given a small offline dataset of complete preference vectors $\{v_1, v_2, \dots\}$, obtained (e.g., via surveys) from a population disjoint from the inference-time agents. Given this dataset, our goal is to construct a procedure that, at inference time, takes an incomplete profile $\tilde{V}$ and outputs an allocation that is fair with respect to the unobserved complete profile V. We assume that incomplete reports at inference follow the same generative model as those in the training data.

3 OUR LEARNING FRAMEWORK

We develop learning-based methods for computing fair allocations from incomplete preferences. We consider (i) direct preference learning, which imputes missing entries and then applies a mechanism,

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/HTOP9978

Table 1: The EF loss of allocations on the Agent-wise setting ($n = 2$).

	$n = 2, p = 0.7$			$n = 2, p = 0.9$		
	$\|S\| = 100$	$\|S\| = 1000$	$\|S\| = 10000$	$\|S\| = 100$	$\|S\| = 1000$	$\|S\| = 10000$
AW (Random)	0.094 ± 0.091	0.094 ± 0.091	0.094 ± 0.091	0.097 ± 0.087	0.097 ± 0.087	0.097 ± 0.087
AW (Mean)	0.025 ± 0.048	0.025 ± 0.048	0.025 ± 0.048	0.069 ± 0.078	0.069 ± 0.078	0.069 ± 0.078
MK (Random)	0.2 ± 0.15	0.2 ± 0.15	0.2 ± 0.15	0.18 ± 0.13	0.18 ± 0.13	0.18 ± 0.13
MK (Mean)	0.085 ± 0.084	0.085 ± 0.084	0.085 ± 0.084	0.1 ± 0.099	0.1 ± 0.099	0.1 ± 0.099
RR (Random)	0.06 ± 0.067	0.06 ± 0.067	0.06 ± 0.067	0.073 ± 0.073	0.073 ± 0.073	0.073 ± 0.073
RR (Mean)	0.026 ± 0.045	0.026 ± 0.045	0.026 ± 0.045	0.053 ± 0.061	0.053 ± 0.061	0.053 ± 0.061
EEF1NN	0.081 ± 0.079	0.11 ± 0.1	0.15 ± 0.13	0.1 ± 0.091	0.2 ± 0.15	0.56 ± 0.18
EEF1NN (MIPL)	0.074 ± 0.079	0.069 ± 0.077	0.094 ± 0.11	0.094 ± 0.087	0.1 ± 0.093	0.18 ± 0.14
AW (Direct)	0.013 ± 0.033	0.011 ± 0.029	0.011 ± 0.03	0.052 ± 0.07	0.04 ± 0.058	0.042 ± 0.061
AW (MIPL)	**0.0091 ± 0.025**$^{***}_{***}$	**0.0083 ± 0.025**$^{***}_{**}$	**0.0081 ± 0.024**$^{***}_{*}$	**0.037 ± 0.051**$^{***}_{***}$	**0.035 ± 0.052**$^{***}_{*}$	**0.037 ± 0.052**$^{***}_{*}$
MK (Direct)	0.066 ± 0.076	0.062 ± 0.07	0.059 ± 0.067	0.078 ± 0.084	0.072 ± 0.081	0.072 ± 0.079
MK (MIPL)	0.05 ± 0.064$_{***}$	0.05 ± 0.064$_{***}$	0.049 ± 0.064$_{***}$	0.064 ± 0.076$_{***}$	0.063 ± 0.076$_{***}$	0.062 ± 0.074$_{***}$
RR (Direct)	0.019 ± 0.037	0.017 ± 0.038	0.017 ± 0.039	0.046 ± 0.057	0.04 ± 0.052	0.041 ± 0.056
RR (MIPL)	0.017 ± 0.036$^{***}_{*}$	0.015 ± 0.033***	0.017 ± 0.035***	0.041 ± 0.052$^{***}_{**}$	0.04 ± 0.051***	0.037 ± 0.051$^{***}_{*}$

and (ii) *mechanism-informed preference learning* (MIPL), which explicitly incorporates fairness of the resulting allocation.

Direct Preference Learning. We create self-supervised training instances $(\tilde{V}_s, V_s)$ by masking entries in complete profiles V_s sampled from the offline dataset. We consider a profile loss

$$\mathrm{Loss}_{\mathrm{pro}}(\boldsymbol{\theta}) = \sum_{s \in S} \|g_{\boldsymbol{\theta}}(\tilde{V}_s) - V_s\|_F^2,$$

where $g_{\boldsymbol{\theta}}$ is a preference estimator and $\|\cdot\|_F$ is the Frobenius norm. At inference, the imputed profile $g_{\hat{\boldsymbol{\theta}}}(\tilde{V})$ is fed into a mechanism $\mathcal{M}$. However, this loss does not take account of the fairness of the resulting allocation under the true preferences.

Mechanism-Informed Preference Learning. To address this, we define a fairness violation function $\mathrm{Loss}_{\mathrm{fair}}$

$$\mathrm{Loss}_{\mathrm{fair},\mathcal{M}}(\boldsymbol{\theta}) = \sum_{s \in S} \ell(\mathcal{M}(g_{\boldsymbol{\theta}}(\tilde{V}_s)); V_s),$$

where we consider ℓ as an EF loss ℓ_{EF}:

$$\ell_{\mathrm{EF}}(\hat{X}; V) = \max_{i,i' \in [n]} \left\{\max\{v_i(\hat{X}_i) - v_i(\hat{X}_{i'}), 0\}\right\}, \tag{1}$$

or a PROP loss ℓ_{PROP}:

$$\ell_{\mathrm{PROP}}(\hat{X}; V) = \max_{i \in [n]} \left\{\max\{v_i(\hat{X}_i) - 1/n, 0\}\right\}. \tag{2}$$

We minimize the hybrid objective with a hyper-parameter $\lambda \in [0, 1]$

$$\mathrm{Loss}_{\mathcal{M}}(\boldsymbol{\theta}) = \lambda \cdot \mathrm{Loss}_{\mathrm{fair},\mathcal{M}}(\boldsymbol{\theta}) + (1 - \lambda) \cdot \mathrm{Loss}_{\mathrm{pro}}(\boldsymbol{\theta}), \tag{3}$$

Since mechanisms such as AW, RR, and MK contain discrete steps and are not differentiable, optimizing the objective (3) is difficult. We therefore introduce a differentiable surrogate $\mathcal{M}_{\mathrm{d}}^{\tau}$ that smoothly approximates $\mathcal{M}$ as $\tau \to 0$, e.g., by replacing sorting used in AW with softsort [11]. We optimize the surrogate hybrid loss

$$\mathrm{Loss}_{\mathcal{M}_{\mathrm{d}}^{\tau}}(\boldsymbol{\theta}) = \lambda\, \mathrm{Loss}_{\mathrm{fair},\mathcal{M}_{\mathrm{d}}^{\tau}}(\boldsymbol{\theta}) + (1 - \lambda)\, \mathrm{Loss}_{\mathrm{pro}}(\boldsymbol{\theta}).$$

and use the original $\mathcal{M}$ at inference to obtain integral allocations.

4 EXPERIMENTS

Dataset. We collected preference data for 33 household chores from 2,000 Japanese participants, and released the dataset[1]. Each participant rated chores by preference (1-3) and required time (1-14); disutilities were computed as the product and normalized. To create self-supervised instances, we split participants into training/validation/test sets and generated complete profiles by sampling $n \in \{2, 5\}$ agents to obtain $|S| \in \{100, 1000, 10000\}$ training profiles and 1,000 each for validation and testing. Incomplete profiles were generated by: (i) Agent-wise: randomly removing a proportion $p \in \{0.7, 0.9\}$ of chores independently for each agent, (ii) Chore-wise: randomly removing the same chores for all agents in a profile, and (iii) Top-t: keeping the $t \in \{5, 10\}$ highest-valued chores.

Models and Baselines. We evaluated direct preference learning (Direct) and MIPL with the EF loss (1) or PROP loss (2). We implemented the preference estimator $g_{\boldsymbol{\theta}}$ by multi-layer perceptrons. We selected the hyper-parameters τ and λ using the validation set. We used the following baselines: (i) Random: Missing disutilities were sampled uniformly at random, (ii) Mean: Missing disutilities were imputed by the average of observed disutilities for each agent, and (iii) EEF1NN: A neural network that predicts an allocation from a complete preference [10]. We employed EEF1NN (MIPL) that additionally minimizes the profile loss. Preferences imputed by Random or Mean baselines were fed into the mechanisms.

Results. Due to page limit, we report the EF loss for two agents in the Table 1. All values in these tables represent means and standard deviations on the test set, scaled by a factor of 10. For each column, the lowest mean is in **bold**, and the second-lowest is underlined. Superscript (*) and subscript ($_*$) asterisks represent the Wilcoxon signed-rank test results comparing MIPL with the best baseline and its Direct variant, respectively ($^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$). We confirmed that MIPL computes fairer allocations than baseline methods in most scenarios. While AW (Mean) or RR (Mean) were the best of baselines, AW (MIPL) achieved the EF loss that is 32–70% of the baselines, and also significantly outperformed AW (Direct).

5 CONCLUSION

We proposed mechanism-informed preference learning that integrates neural networks with differentiable approximations of fair division mechanisms to compute fair allocations from incomplete preferences. Experiments show improved fairness over baselines.

ACKNOWLEDGMENTS

This work was supported by JST BOOST, Grant Number JPMJBS2407; JST FOREST Grant Numbers JPMJFR226O; and JST FOREST Program, Grant Number JPMJFR232S.

[1] https://github.com/MandR1215/HouseholdChorePreference_Dataset

REFERENCES

[1] A. Keith Austin. 1982. Sharing a cake. *The Mathematical Gazette* 66, 437 (1982), 212–215.

[2] Haris Aziz, Ioannis Caragiannis, Ayumi Igarashi, and Toby Walsh. 2022. Fair allocation of indivisible goods and chores. *Autonomous Agents and Multi-Agent Systems* 36 (2022), 1–21.

[3] Steven J. Brams and Alan D. Taylor. 1996. *Fair Division: From Cake-Cutting to Dispute Resolution.* Cambridge University Press.

[4] Eric Budish. 2011. The combinatorial assignment problem: Approximate competitive equilibrium from equal incomes. *Journal of Political Economy* 119, 4 (2011), 1061–1103.

[5] Eric Budish, Gérard P. Cachon, Judd B. Kessler, and Abraham Othman. 2017. Course match: A large-scale implementation of approximate competitive equilibrium from equal incomes for combinatorial allocation. *Operations Research* 65, 2 (2017), 314–336.

[6] Ioannis Caragiannis, David Kurokawa, Hervé Moulin, Ariel D. Procaccia, Nisarg Shah, and Junxing Wang. 2019. The unreasonable fairness of maximum Nash welfare. *ACM Transactions on Economics and Computation* 7, 3 (2019), 1–12.

[7] Vincent Conitzer, Rupert Freeman, and Nisarg Shah. 2017. Fair public decision making. In *Proceedings of the 18th ACM Conference on Economics and Computation (EC).* 629–646.

[8] Duncan K. Foley. 1967. Resource allocation and the public sector. *Yale Economic Essays* 7 (1967), 45–98.

[9] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. 2024. A Survey on Self-Supervised Learning: Algorithms, Applications, and Future Trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 12 (2024), 9052–9071.

[10] Shaily Mishra, Manisha Padala, and Sujit Gujar. 2022. EEF1-NN: Efficient and EF1 allocations through neural networks. In *Proceedings of 19th Pacific Rim International Conference on Artificial Intelligence (PRICAI).* 388–401.

[11] Sebastian Prillo and Julian Eisenschlos. 2020. SoftSort: A continuous relaxation for the argsort operator. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, Vol. 119. 7793–7802.

[12] Ermis Soumalias, Behnoosh Zamanlooy, Jakob Weissteiner, and Sven Seuken. 2024. Machine learning-powered course allocation. In *Proceedings of the 25th ACM Conference on Economics and Computation (EC).* 1099.

[13] Hugo Steinhaus. 1949. The problem of fair division. *Econometrica* 17 (1949), 315–319.