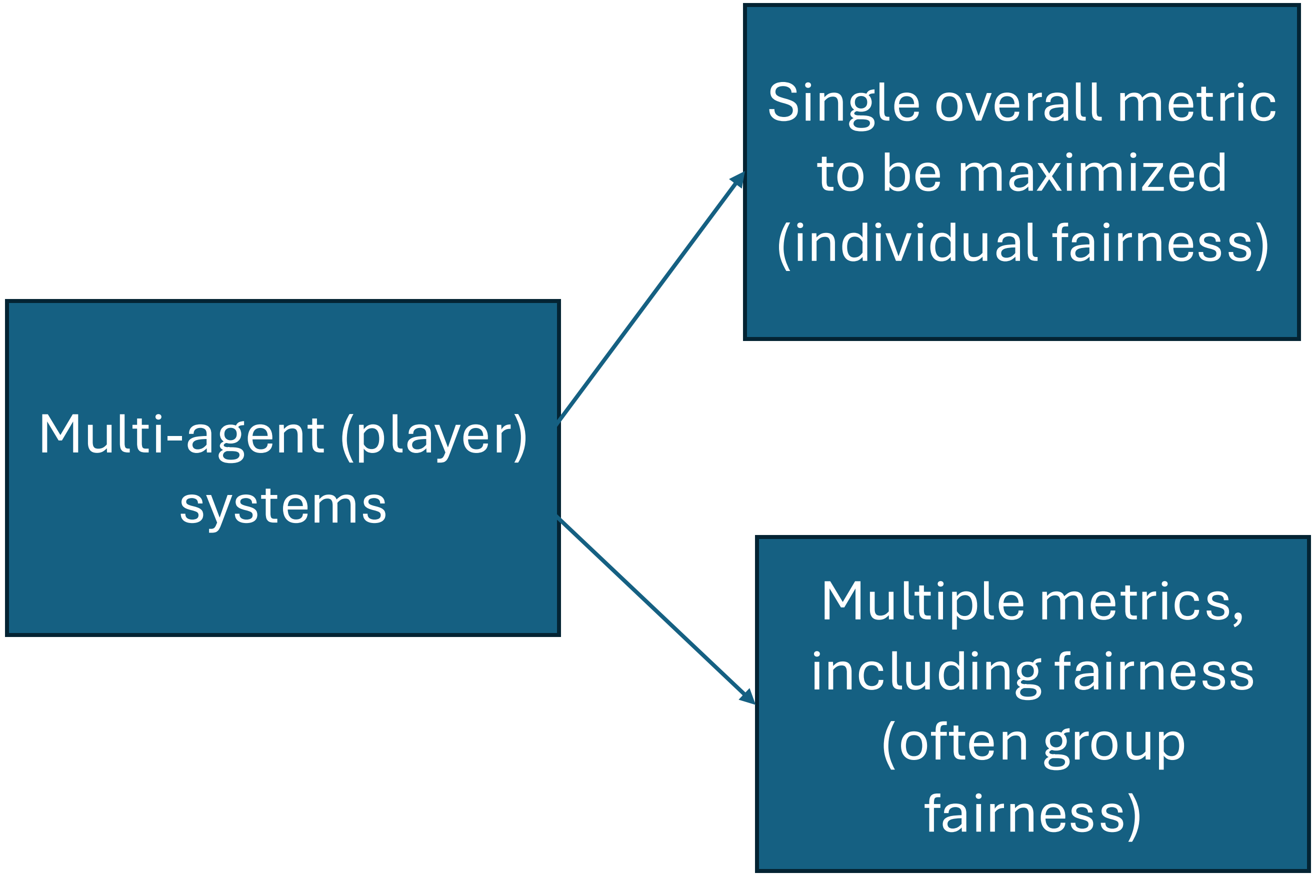
Multi-agent with multi-objective RL

Yingqian Zhang

5 August

Fairness in multi-agent decision-making



Make sure solution/policy is fair for all

Make sure some group statistic (e.g. sd) is good

Fairness in multi-agent decision-making problems

- The system aims to maximize one single performance metric, e.g., allocating bandwidths, optimizing waiting time of roads/drivers with traffic light, minimizing distance in parcel delivery
- Typically, the system's objective is aligned with users' utility, and a utilitarian objective (social welfare) is generally adopted:

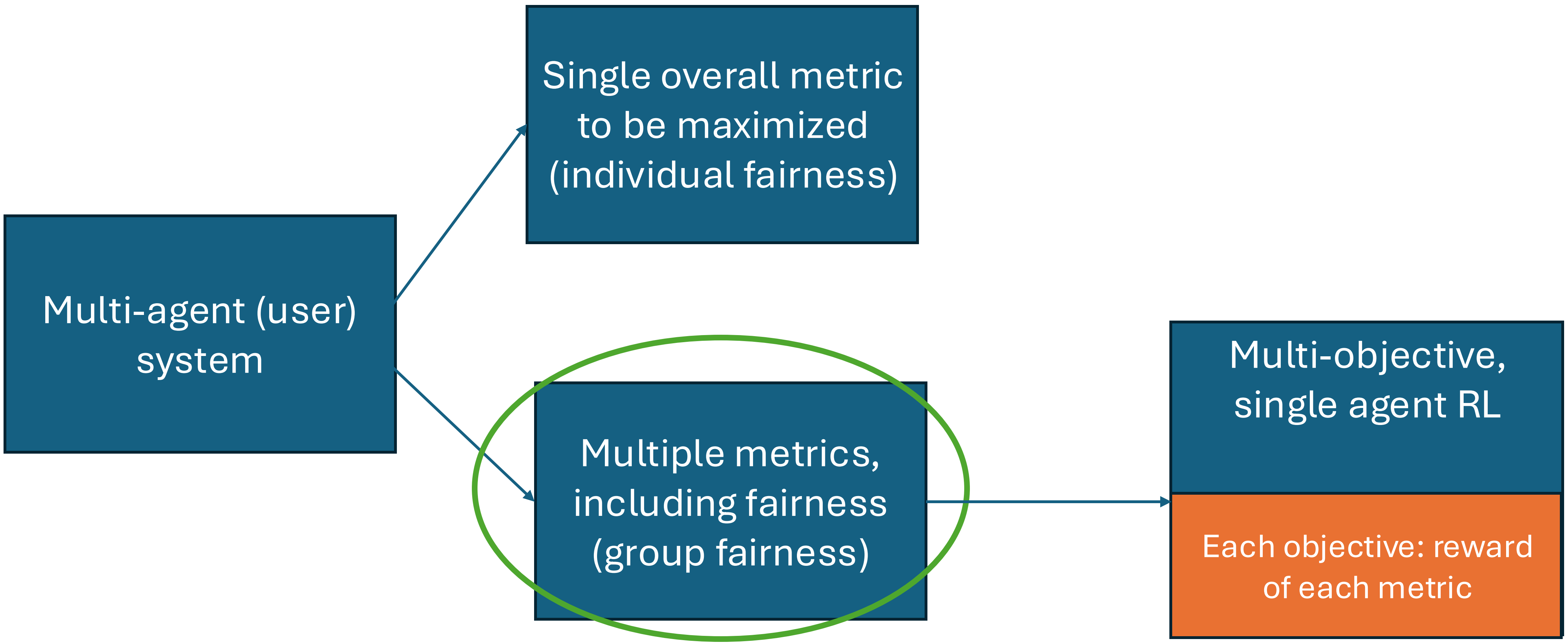
$$v = v_1 + v_2 + \cdots + v_n$$

- Individual fairness: the total utility should be distributed to users in a fair way -> natural to model in RL

Fairness in multi-agent decision-making problems (2)

- The system aims to maximize one or more performance metrics
- Example: Human-robot collaboration in order picking in warehouses
  - Decision: assigning human pickers to robots
  - System objective: to maximize pick rate (min picking time)
  - Human pickers' workload is influenced by the decision but not directly aligned with the system objective
  - The system needs to optimize for two different metrics (pick rate and work load fairness)
- Group fairness often used
  - statistical parity in the decisions
  - less preferred than individual fairness but easier to model in RL

Example: multi-objective fair RL in practice



TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

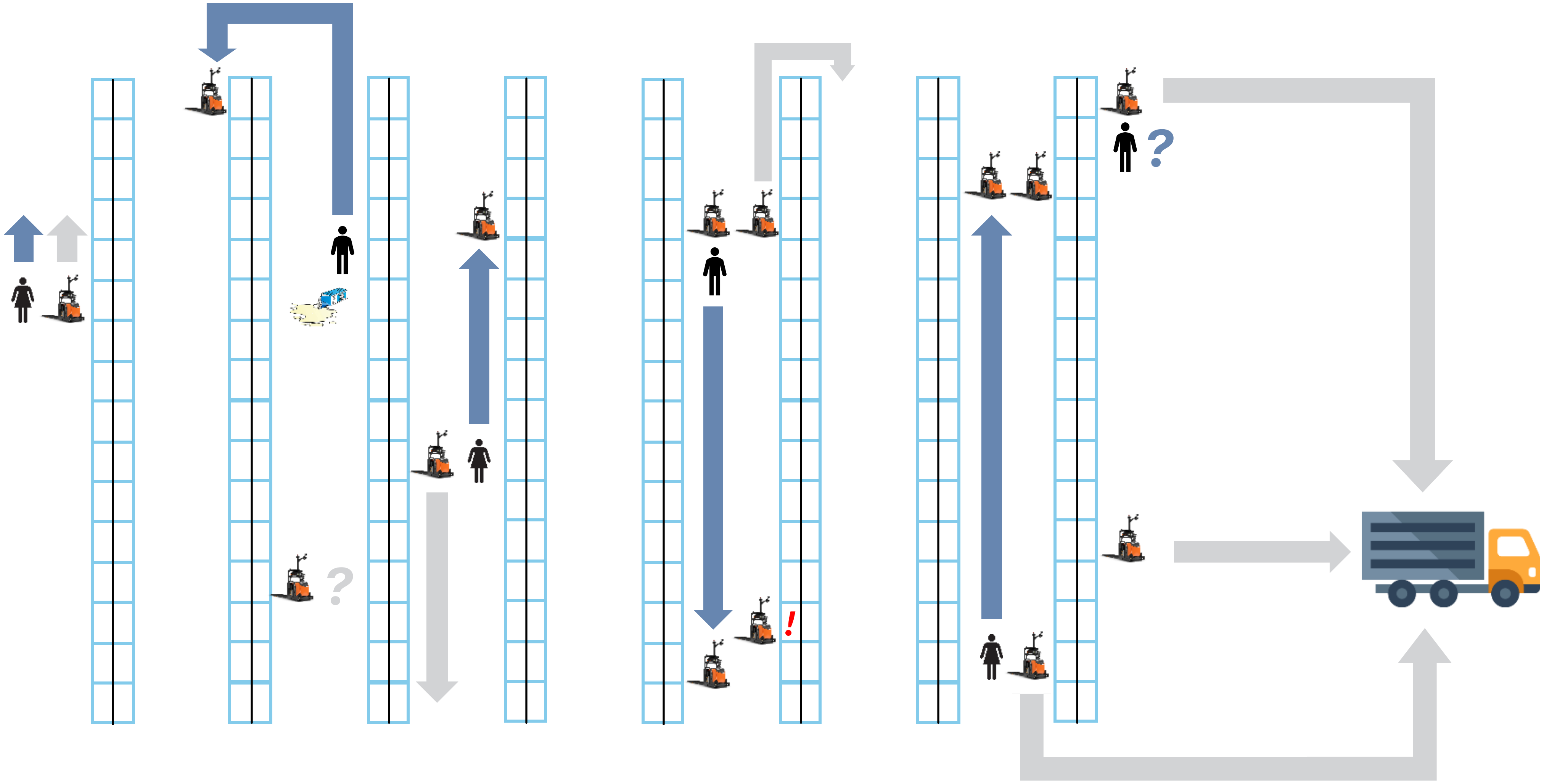
VANDERLANDE

Learning efficient and fair policies for collaborative human-robot order picking

Smit, I. G., Bukhsh, Z., Pechenizkiy, M., Alogariastos, K., Hendriks, K., & Zhang, Y. (2024). Learning Efficient and Fair Policies for Uncertainty-Aware Collaborative Human-Robot Order Picking. arXiv.org.

Order Picking: crucial component of warehouse operation

a sequential decision problem



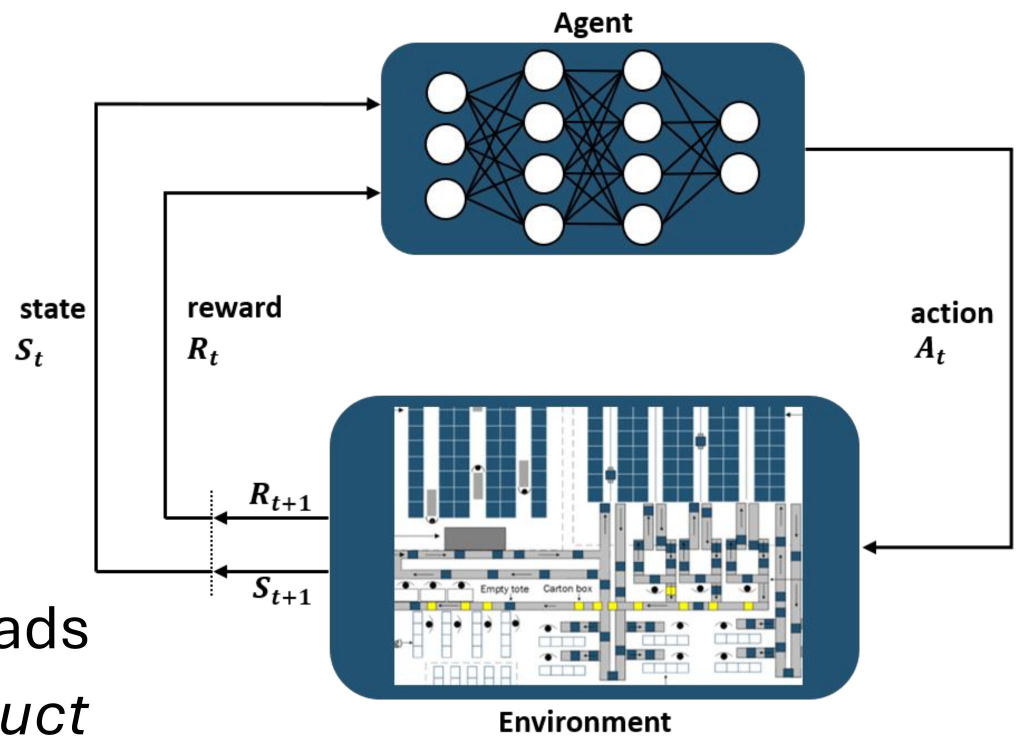
Robot Leading:

Picruns are assigned to AMRs;
AMR moving to a picking location;
A human picker is assigned to AMR;
Repeat

Multi-objective optimization problem

- Develop a 'picker optimizer' in human-robot collaborative picking using RL
- Decision: *Allocate human order pickers to incoming orders/AMRs*

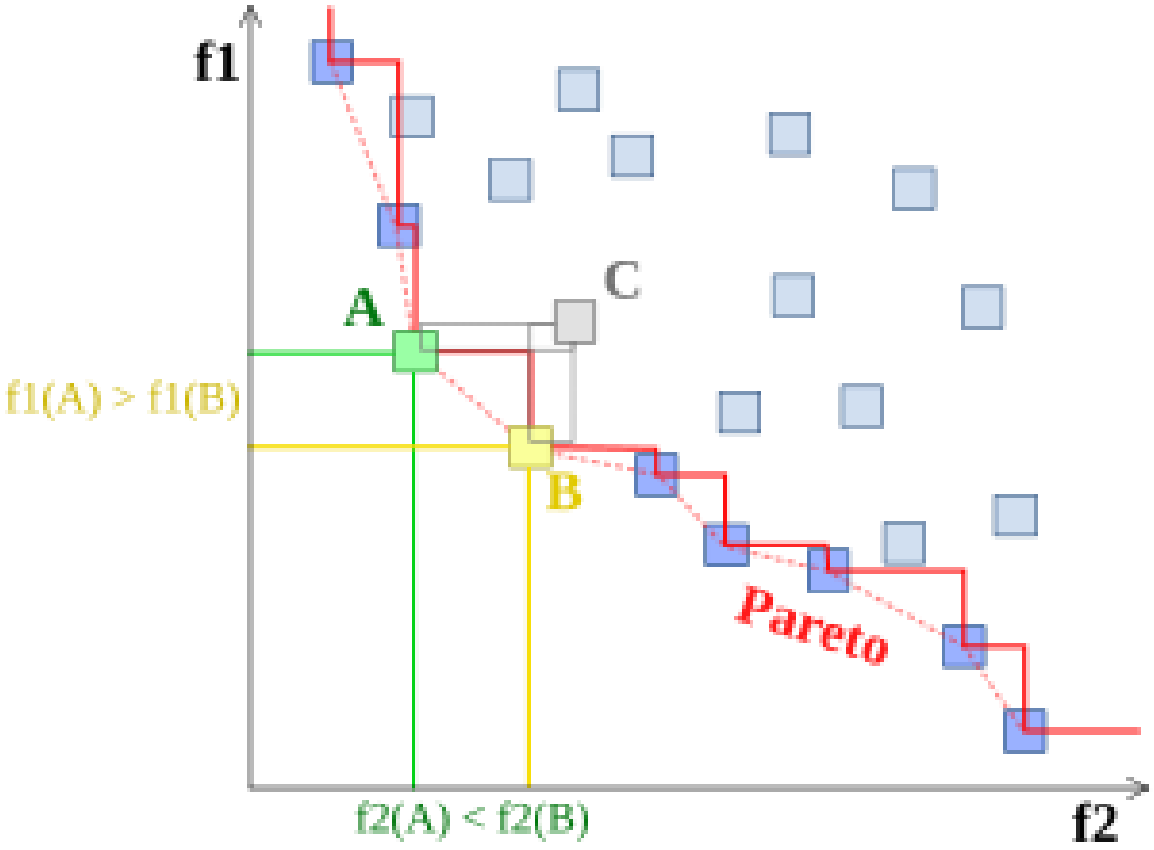
- Optimization objectives
  - Max pick rate → Nr. of picked orders per hour
  - Fairness: Ergonomic regulations: lifting workloads
    - *minimize standard deviation of carried product masses of pickers*



A typical multi-objective optimization problem

Multiple policies (non-dominated) solutions

- The Pareto Front is the set of non-dominated solutions. For each solution (policy in an RL problem) on the Pareto Front, no other solution has a better value for all objectives, called Pareto efficiency



States: features related to pick rate

Current picker information	
Location	Whether the picker is currently at the node.
Picker distance	Provides the distance between picker and the node through warehouse paths.
AMR(s) information	
Location	Whether the AMR is currently at the node.
# of AMRs going	Number of AMRs currently going towards the node.
Destination distance	Minimum travel distance of AMRs with this node as their destination or -10 if none are traveling in towards the node.
Expected time until next destination	Sum of estimated travel time to current destination, pick time at destination and time until the next destination. Value of -10 if no AMR goes for the next pickrun, otherwise AMR with minimum travel time is selected.
Expected time until two-step ahead	Same as expected time until next destination feature but compute the estimates for two-step ahead AMR destination.
# of AMRs within same aisle	AMRs going to a destination within the same aisle as the considered node.
# of AMR waiting	AMRs currently waiting in the same aisle as the considered node.
Picker positioning in the system	
Location	Indicate if any picker other than the picker being assigned is at this node.
Minimum travel distance	Minimum distance to this node among all pickers having this node as destination. If none, the value is -10.
# of pickers	Number of pickers going to a destination within the same aisle as the considered node.
Distance of other pickers	Minimum distance of any other picker to its current destination plus the distance from its current destination to the considered node.
Expected time of other pickers	Similar to the above, but considering the expected time, including expected picking time at the current destination.
Node region information	
Aisle distance from origin	How far the aisle of this node is from the origin, scaled by the warehouse size.
Node depth within aisle	How far toward the beginning or end of the aisle a node is located, scaled by the aisle length.
Node neighborhood features	
Closest next destination distances	Closest and 2nd closest distance to the next destinations of the AMRs going to this node. 0 if no AMRs or last node in the pickrun.
Closest distances to two-step ahead.	Same as above but for the closest two-step ahead destination.
Closest distance to pickers	Minimum distances from this node to the other nodes that are currently the destination of any of the pickers.
Distances to closest unserved AMRs	Distances to the closest and 2nd closest other nodes that are the destination of an AMR and where no picker is already going.

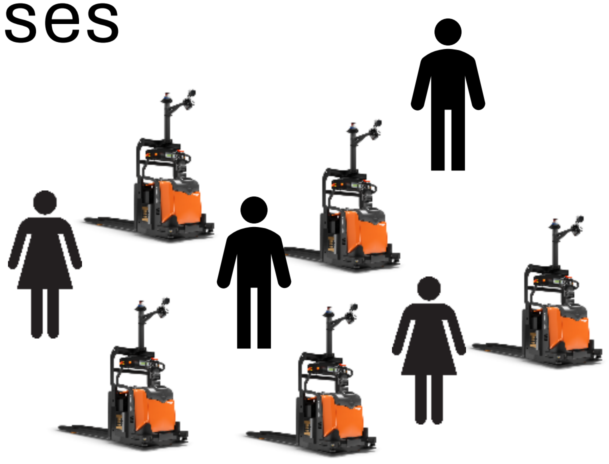
Table 2: List of state space features related to efficiency.

States: features related to workload fairness

Node specific workload information	
Current picker workload	Total mass in kilograms that the picker at this node has picked subtracted by the mean workload of all pickers.
Next picker workload	Same as above when the picker destination is the considered node.
Item weight	Mass in kilograms of a single item stored at the node.
Waiting AMR workload	Mass of the items that must be loaded on the waiting AMRs at this location.
Destination AMRs workload	Mass of the items that must be loaded on the AMRs that are going to this location but are not yet there.
Closest picker workloads	Total masses carried by the two closest pickers to this node in terms of expected arrival time, subtracted by the mean picker workload.
Distributional workload information	
Picker total workload	Workload in kilograms of the controlled picker subtracted by the mean picker workload.
Other picker workloads	Minimum, 25^{th} and 75^{th} percentile, maxixmum workload of all pickers, subtracted by the mean picker workload.

Rewards

- Pick rate efficiency: Penalty on time that passes



$$R_t^{\text{efficiency}} = \tau_{t-1} - \tau_t$$

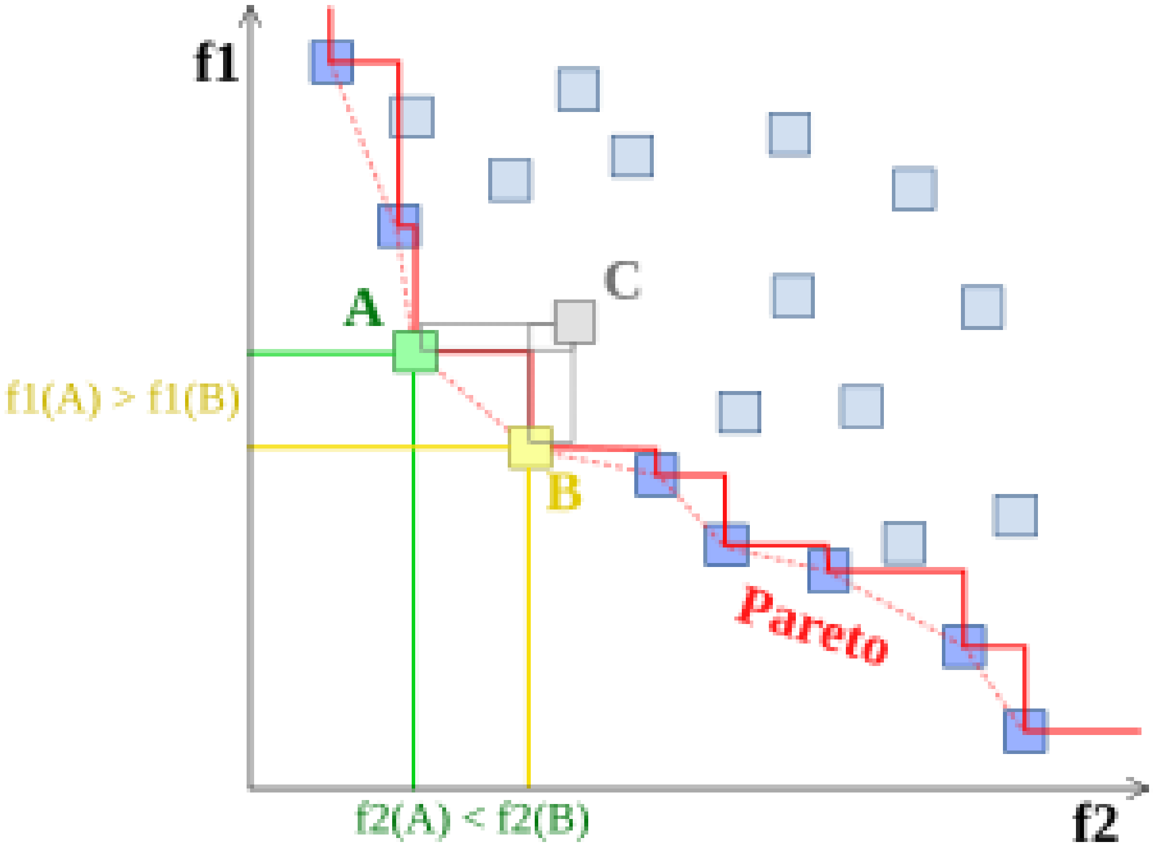
- Fairness
  - Minimize standard deviation of carried product masses
    penalty on increase in standard deviation

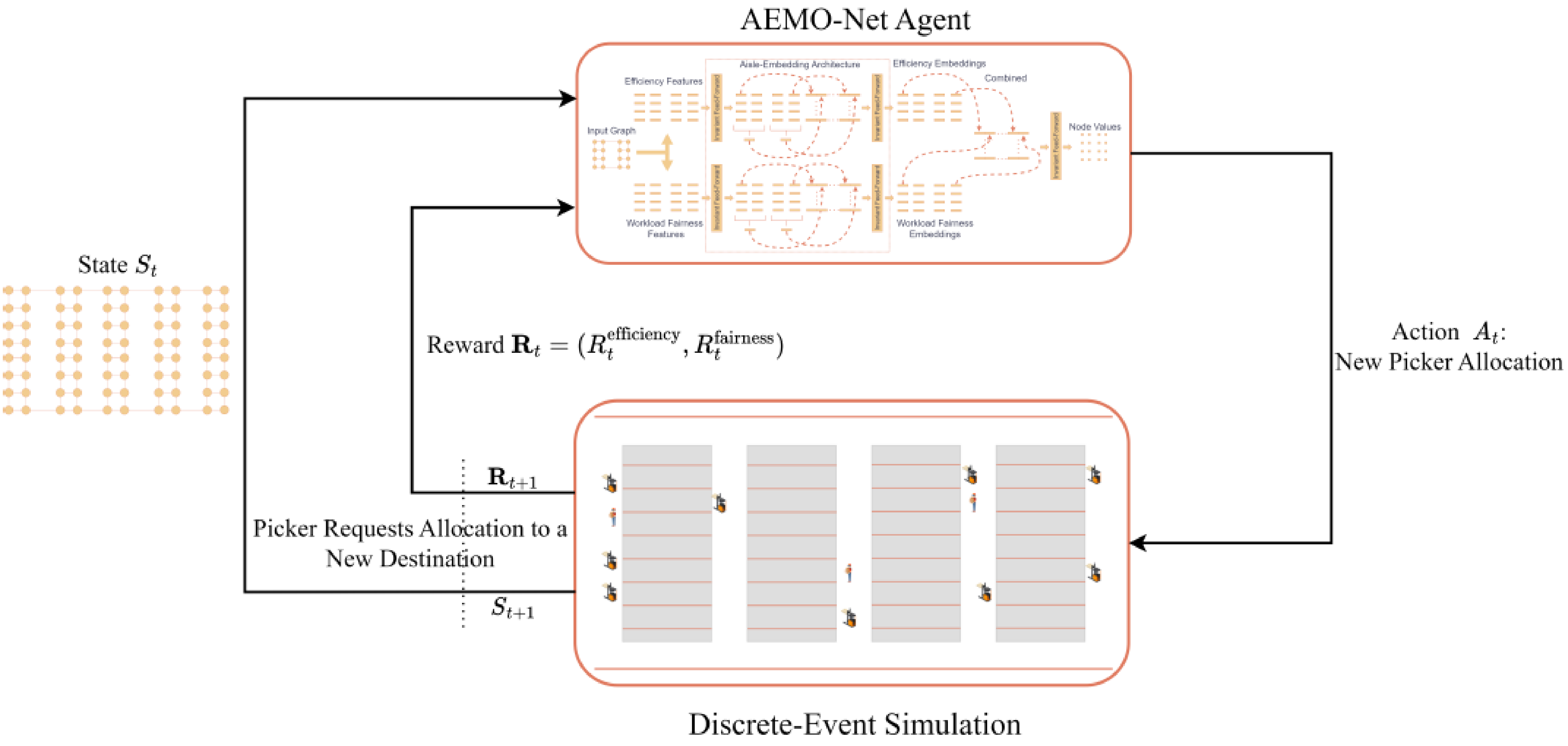
$$R_t^{\text{fairness}} = \sigma\left(W_{1,t-1}, \dots, W_{|\mathcal{K}|,t-1}\right) - \sigma\left(W_{1,t}, \dots, W_{|\mathcal{K}|,t}\right)$$

A typical multi-objective optimization problem

Multiple policies (non-dominated) solutions

- The Pareto Front is the set of non-dominated solutions. For each solution (policy in an RL problem) on the Pareto Front, no other solution has a better value for all objectives, called Pareto efficiency





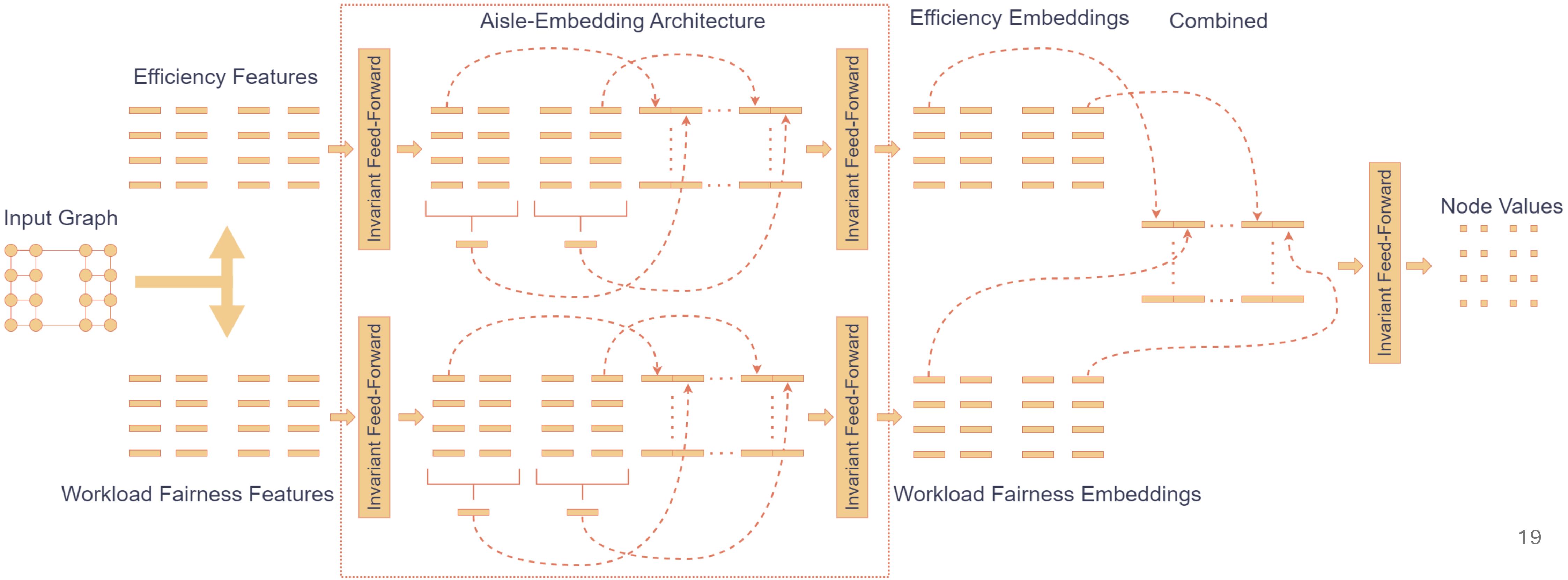
AEMO-Net Agent
Efficiency Features
Aisle-Embedding Architecture
Efficiency Embeddings
Combined
Input Graph
Node Values
Workload Fairness Features
Workload Fairness Embeddings
State S_t
Reward $\mathbf{R}_t = (R_t^{\text{efficiency}}, R_t^{\text{fairness}})$
Action A_t:
New Picker Allocation
$\mathbf{R}_{t+1}$
Picker Requests Allocation to a New Destination
S_{t+1}
Discrete-Event Simulation

Multi-objective learning algorithm

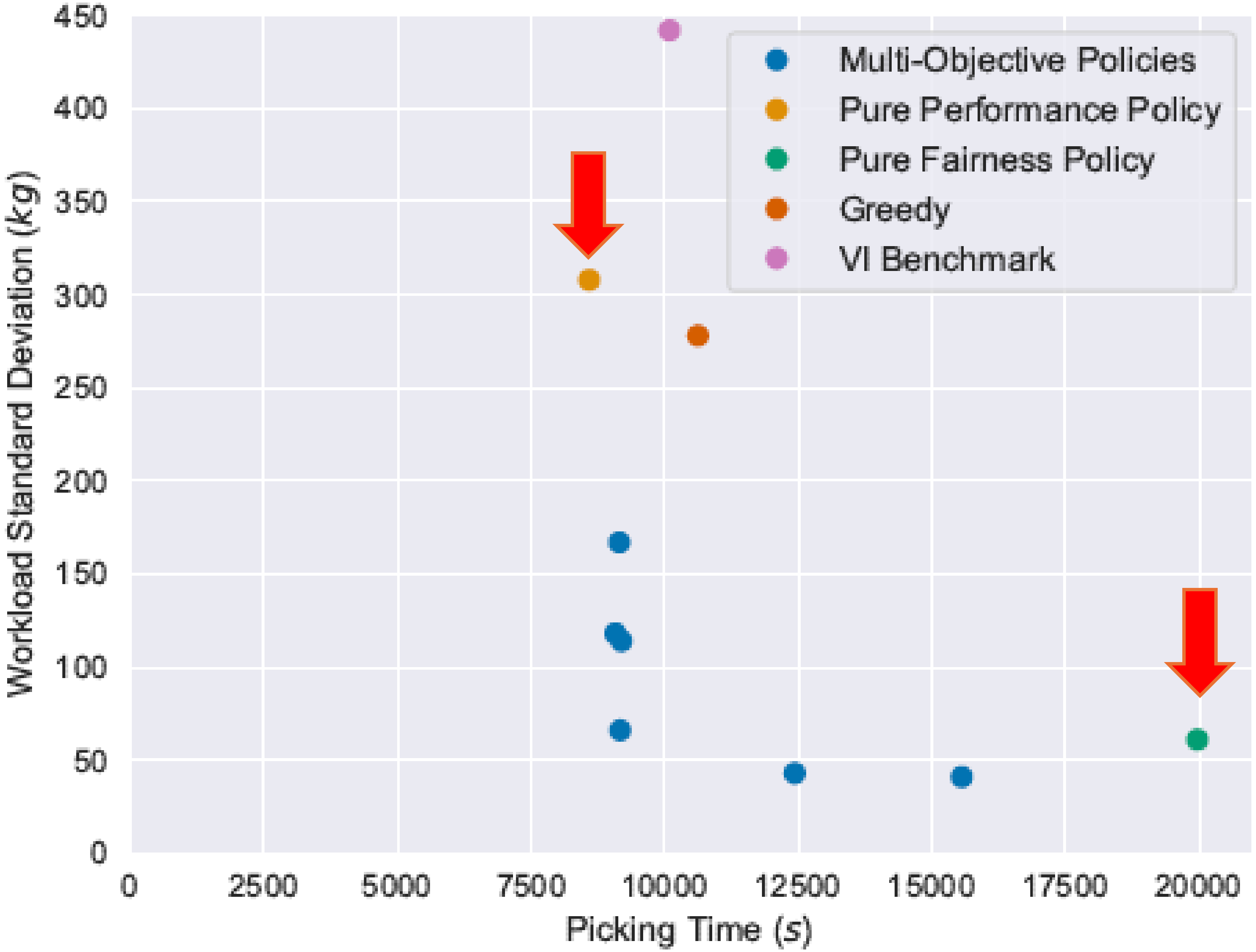
- Extend Proximal Policy Optimization (PPO) by adding an evolutionary component
- (A meta-policy approach, to present non-dominated set)
  - Train initial set of policies on variety of objective weights
  - Evolutionary loop:
    - For each policy, predict which weights can help improve objective the most
    - Select new weights to optimize based on predicted improvement
    - Update policies for several policy-gradient iterations
    - Update Pareto Front

Multi-Objective Aware Network

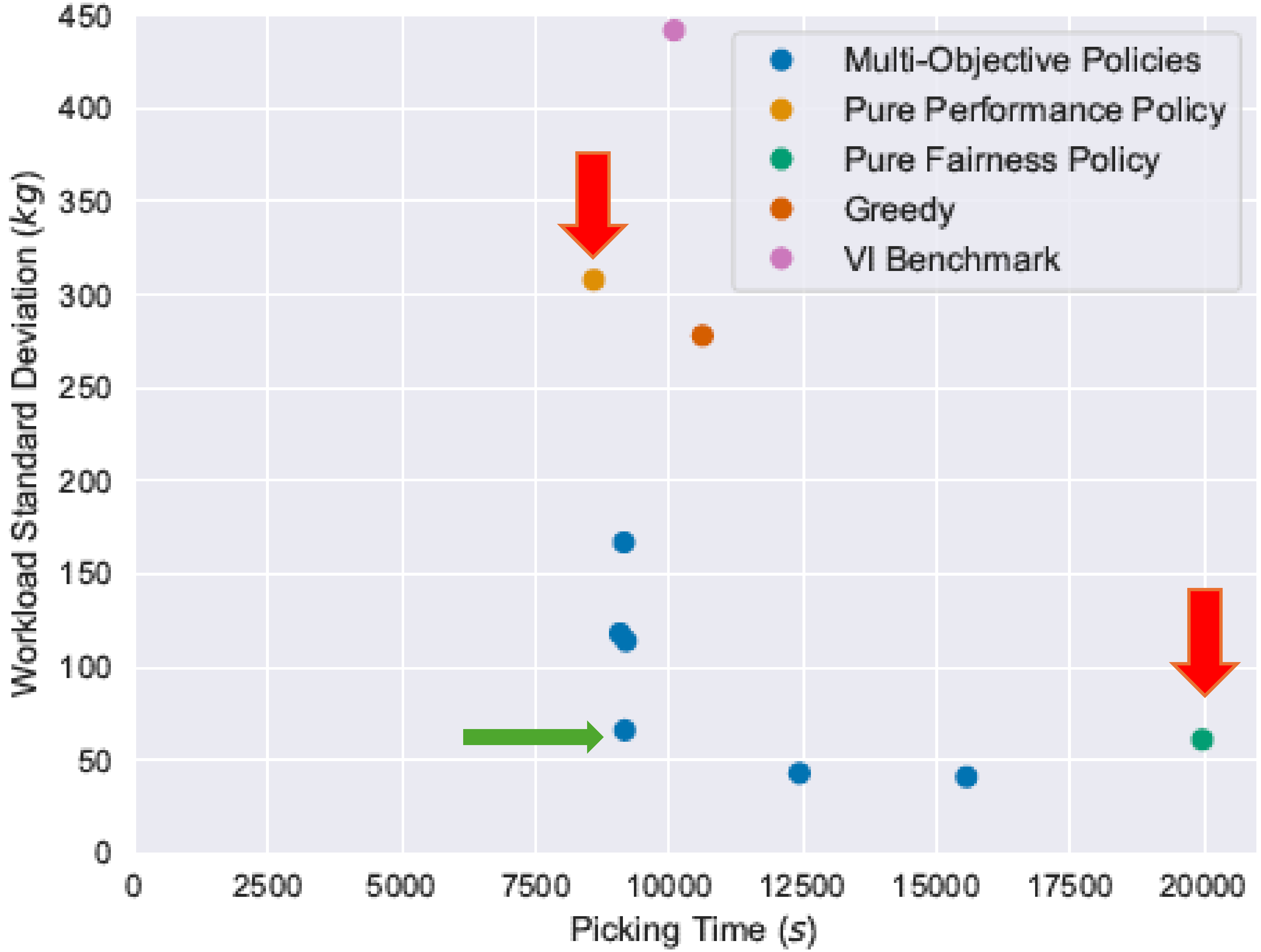
- Node-specific information, distributional information, workload fairness features
- Feature Separation: enable more stable learning



Experiment: *trade-off between fairness and pick rate*

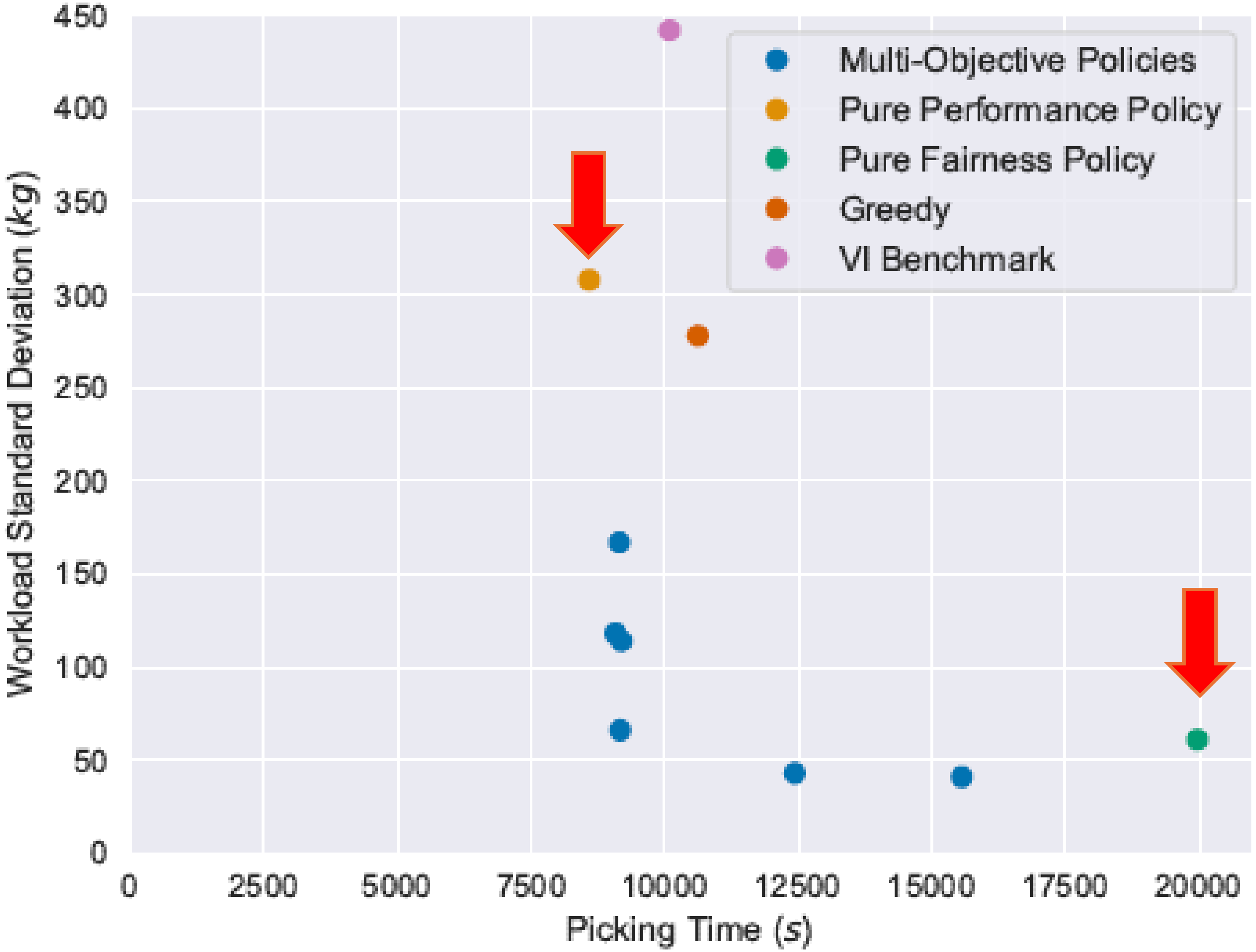


Experiment: *trade-off between fairness and pick rate*



This MORL policy: by sacrificing just 6.7% of pick rate efficiency, it decreases the workload standard deviation by **78.6%**

Experiment: *trade-off between fairness and pick rate*



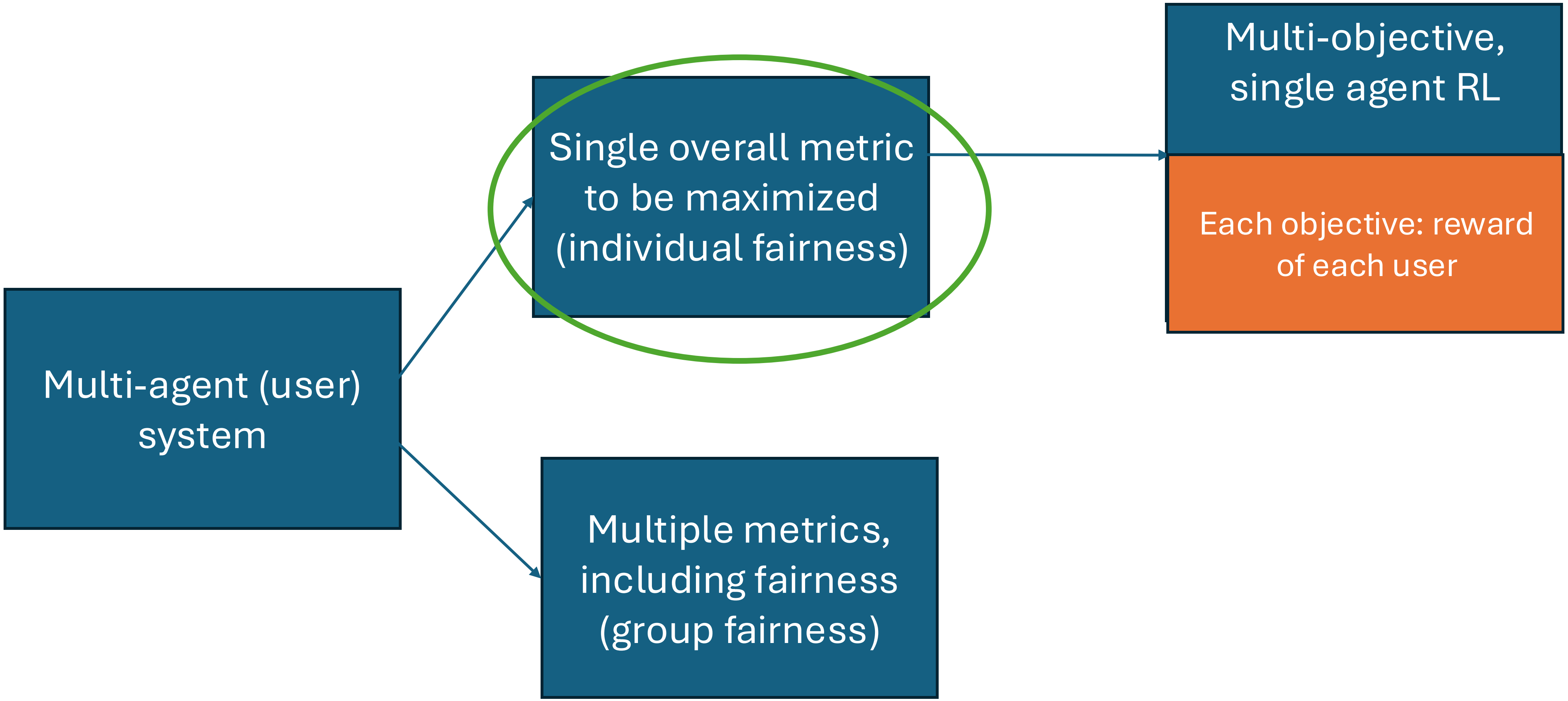
MORL even improves the pure fairness solution!

Fair MORL for collaborative human-robot order picking in warehouses

- Good trade-off between picking time and fairness
  - Explicitly outline achievable trade-offs
  - Simultaneous improvement of picking times and workload fairness
  - Price of fairness is low!

- Is this the best way of modelling and achieving fairness? We don’t know.

Example



Multi-objective, single agent RL
Each objective: reward of each user
Single overall metric to be maximized (individual fairness)
Multi-agent (user) system
Multiple metrics, including fairness (group fairness)

Siddique, U., Weng, P. and Zimmer, M., 2020, November. Learning fair policies in multi-objective (deep) reinforcement learning. ICML.

- Use generalized Gini social welfare function (GGF) to model rewards of

$$\mathrm{GGF}_{\boldsymbol{w}}(\boldsymbol{v}) = \sum_{i=1}^{D} w_i v_i^{\uparrow},$$

- A multi-objective MDP is defined as (D is nr of objectives)
  - Reward: $\boldsymbol{R}_{a,s} \in \mathbb{R}^D$
  - Value function (with discounted reward): $\boldsymbol{V}_{\pi,s} = \mathbb{E}_{\boldsymbol{P}_\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} \boldsymbol{R}_t \mid s\right],$
  - Objective fuction $\operatorname{argmax}_\pi \boldsymbol{J}(\pi)$

  - All take value in $\mathbb{R}^D$

Fair optimization problem

- Integrating GGF with MOMDPs, a fair optimization problem is formulated, which is the problem of determining a policy that generates a fair distribution of rewards to D fixed users

$$\operatorname*{argmax}_{\pi} \mathrm{GGF}_{\boldsymbol{w}}(\boldsymbol{J}(\pi)),$$

Some theretical properities (see paper)

- DQN, A2C and PPO algorithms are adapted
- Traffic light: to learn a controller that optimizes the expected waiting times per road.
- Trade off: worse average waiting times, better fairness (GGF scores)

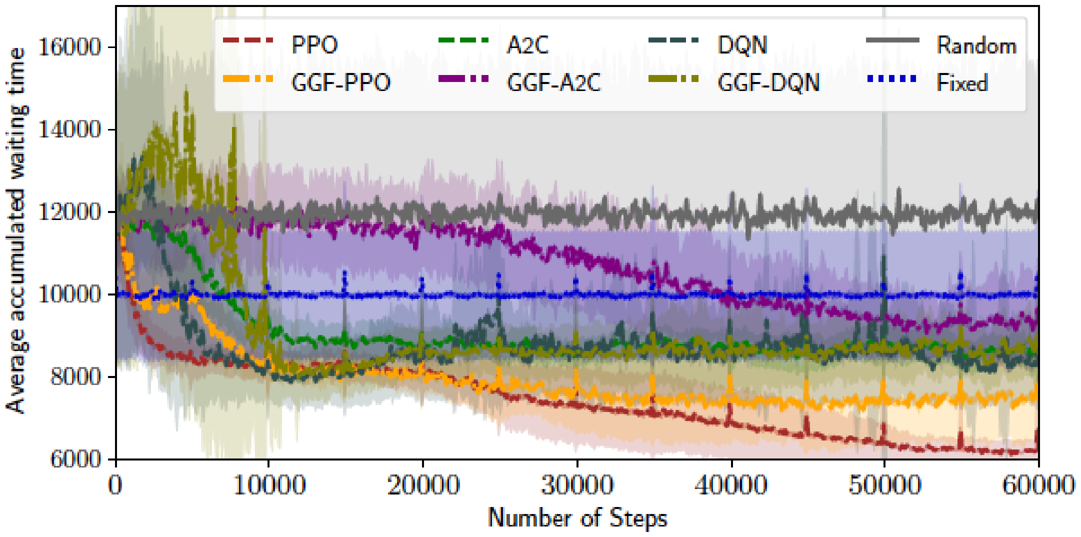


Figure 6. Average waiting times of DQN, A2C, PPO, and their GGF counterparts during learning phase, and those of the fixed and random policies in the TL domain.

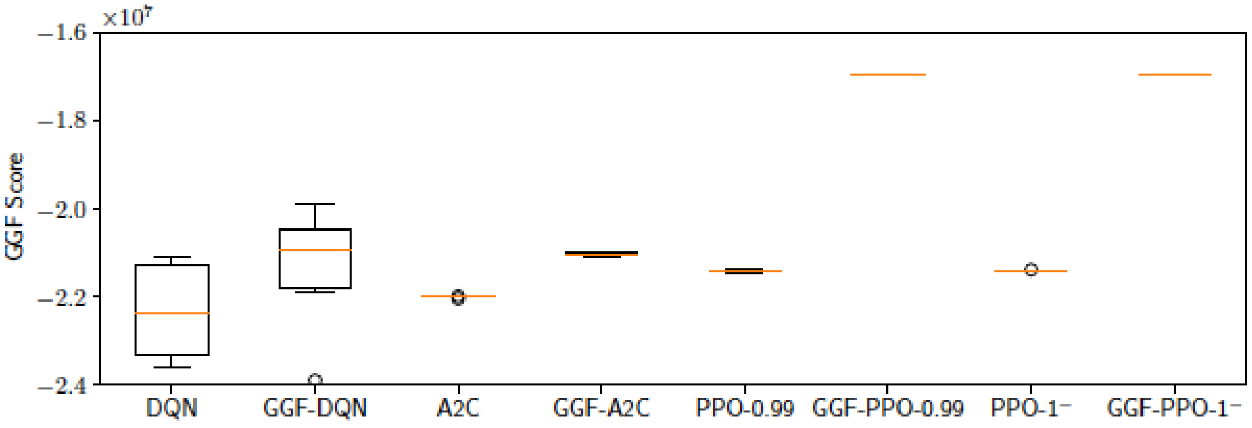


Figure 7. GGF scores of DQN, A2C, PPO, and their GGF versions, with those of PPO and GGF-PPO when γ is close to 1, during the testing phase in the TL domain.

Why fairness?

Why fairness?

- Societal value: responsible and trustworthy AI

  e.g. Zhang, X., Tu, R., Liu, Y., Liu, M., Kjellstrom, H., Zhang, K. and Zhang, C., 2020. How do fair decisions fare in long-term qualification?

- Economic value
  - Fairness may lead to higher long-term economic value

A case study
Fair Task Allocation in the Port of Rotterdam

Fair task allocation in Port of Rotterdam

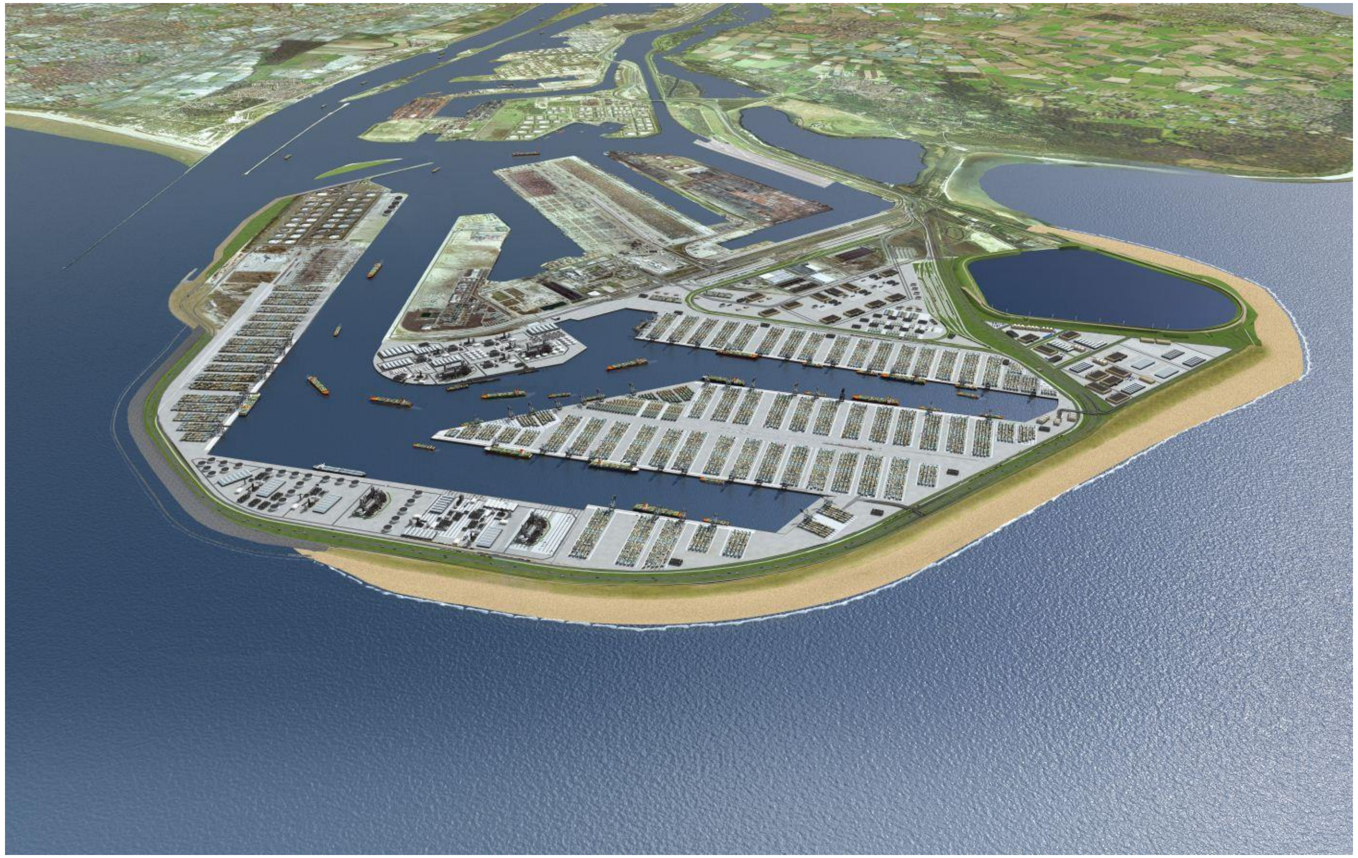
Challenge:

Increasing inter-terminal transport jobs

Solution:

Using existing trucks at the port to do ITT jobs

A task allocation problem



Task allocation problem

- Inputs:
  - Tasks with finite time windows
  - Companies that own trucks
    - agents with available resources during given time periods, incurring costs for doing tasks
- Output: an allocation of tasks among companies with maximized *optimization objectives*
  - number of allocated jobs is maximized
  - total cost is minimized
  - allocation is fair to the participating companies

Which fairness notion?

- Individual fairness is important

so, we first find most fair index, and then optimize cost

• We do not want to add too much computational complexity

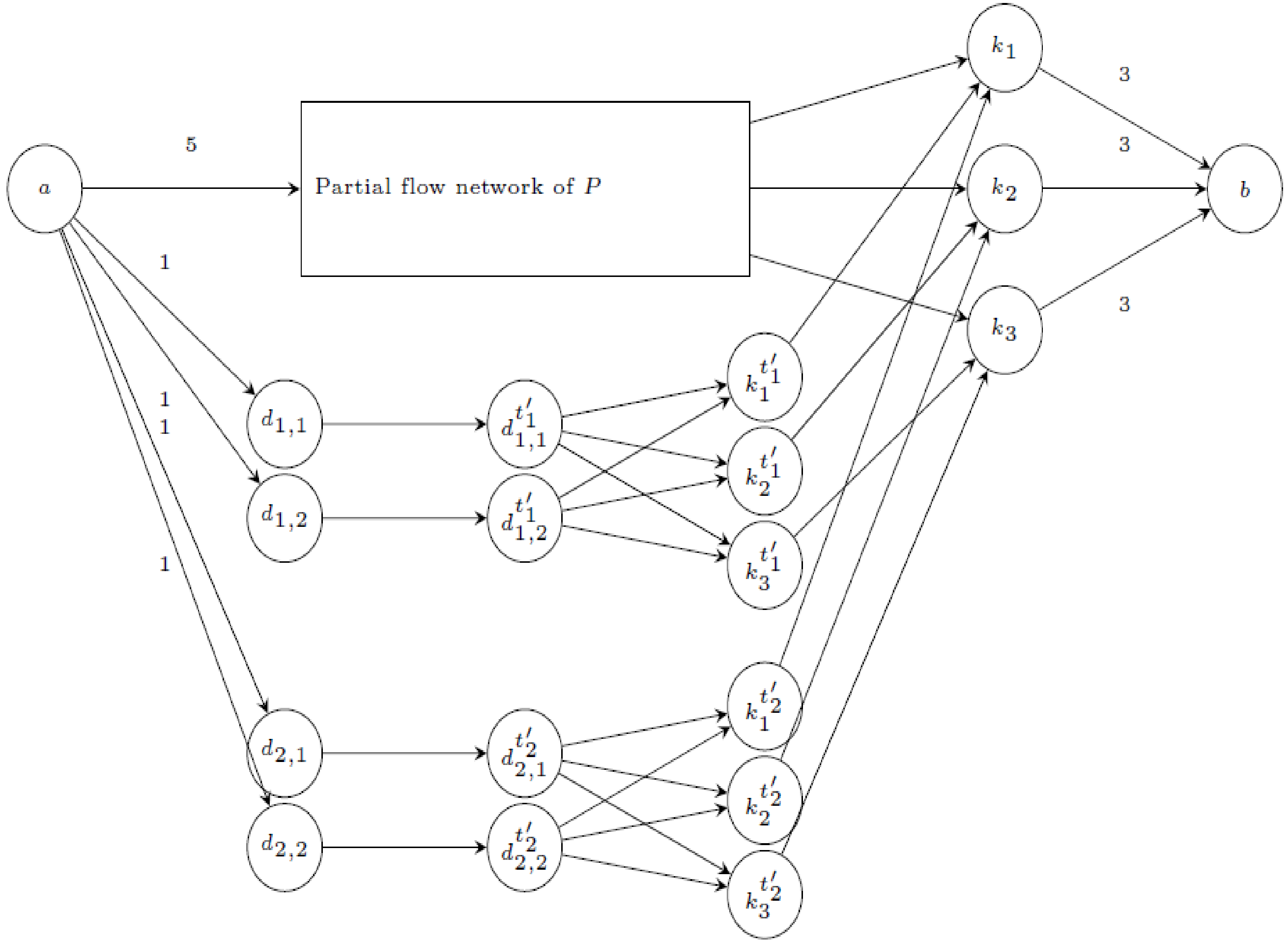
Which fairness notion?

- max-min fairness

- The new algorithm

guarantees optimal fairness, and min cost, and

it stays polynomial!



What is the price of introducing fairness in matching for platform?

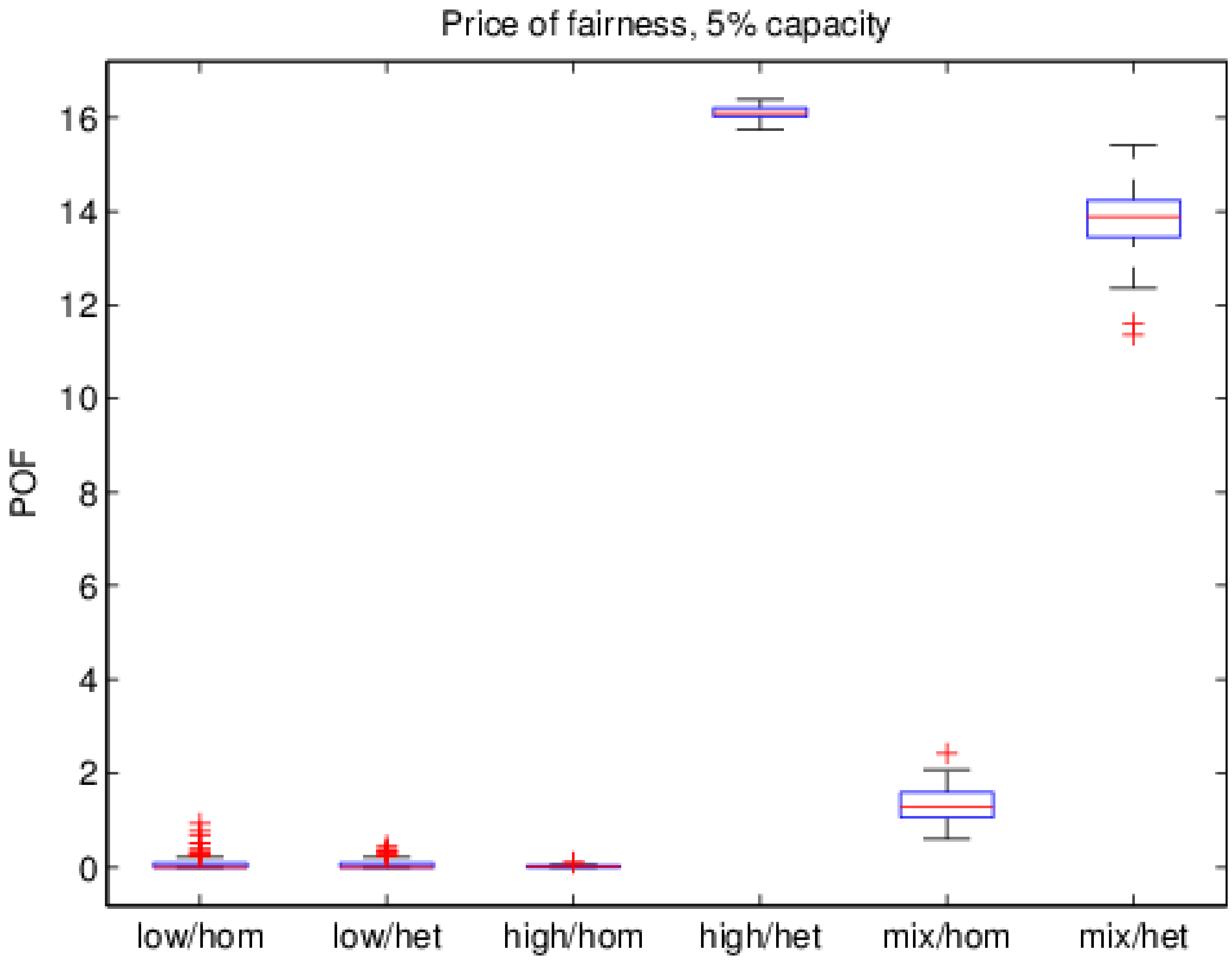
Experiments: one-time matching

- What is the extra cost of using fair matching?

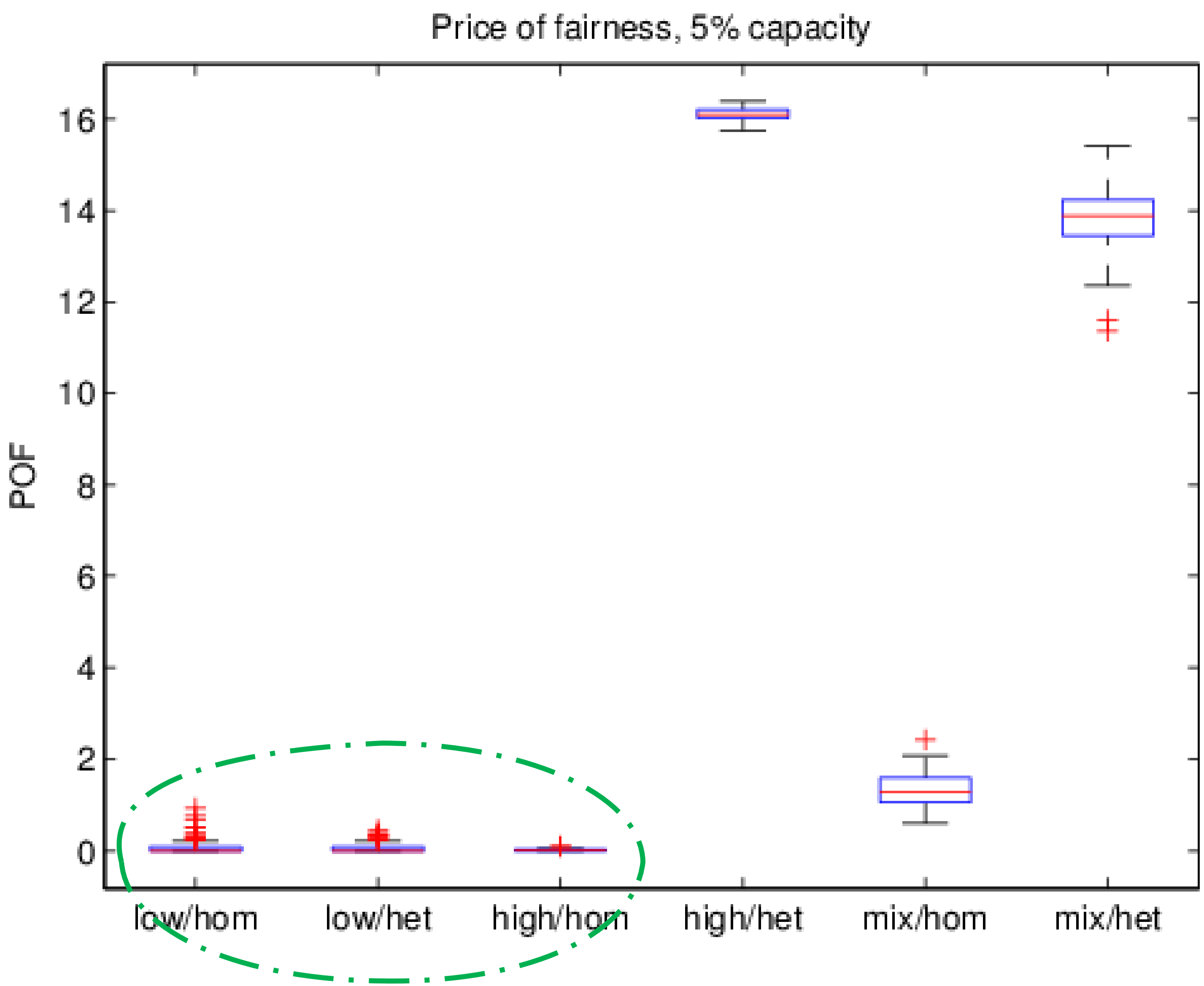
$$\text{Price of fairness} = \frac{\text{total cost of fair policy}}{\text{total cost of myopic policy}} - 1$$

- Testing with different market scenarios

- Price of Fairness for platform



- Price of Fairness for platform



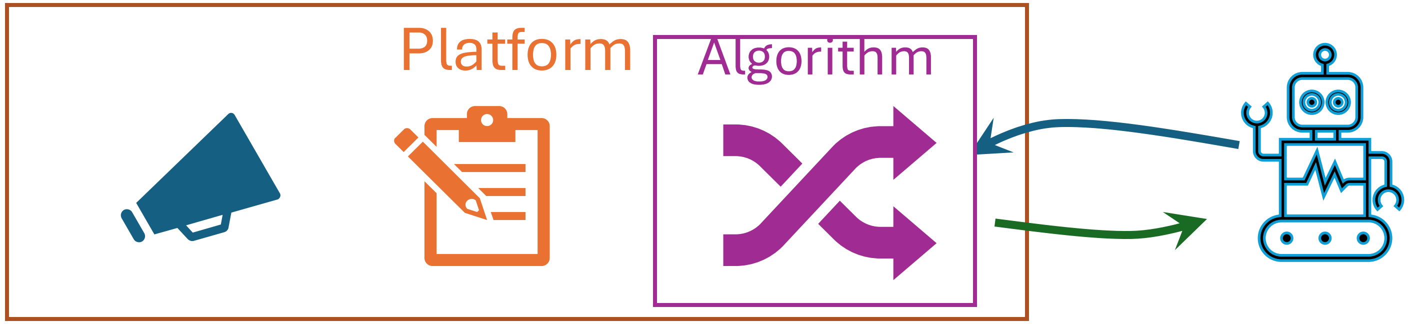
In these scenarios, price of fairness is extremely small

Hypothesis:

Fair matching leads to higher social welfare & higher business value in long-term

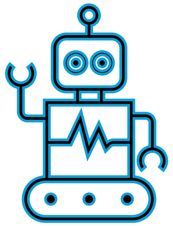
A simulation study

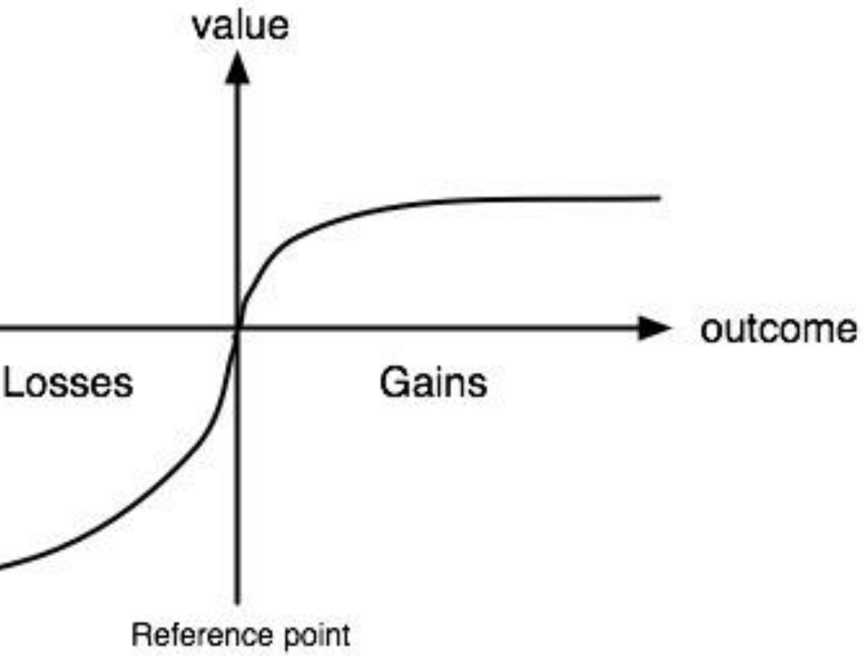
- *Model companies' participation behavior in repeated matching games*
    - Their behaviors are influenced by matching outcomes
    - Their behaviors influence the matching outcome of future rounds



Agent behavioral model

- Agent's behavior (i.e., participation probability) is dependent on experiences in previous rounds.
- Model agent's participation decision using ***prospect (loss-aversion) theory***





Evaluation

- Social welfare = (total value of allocated jobs – total cost)

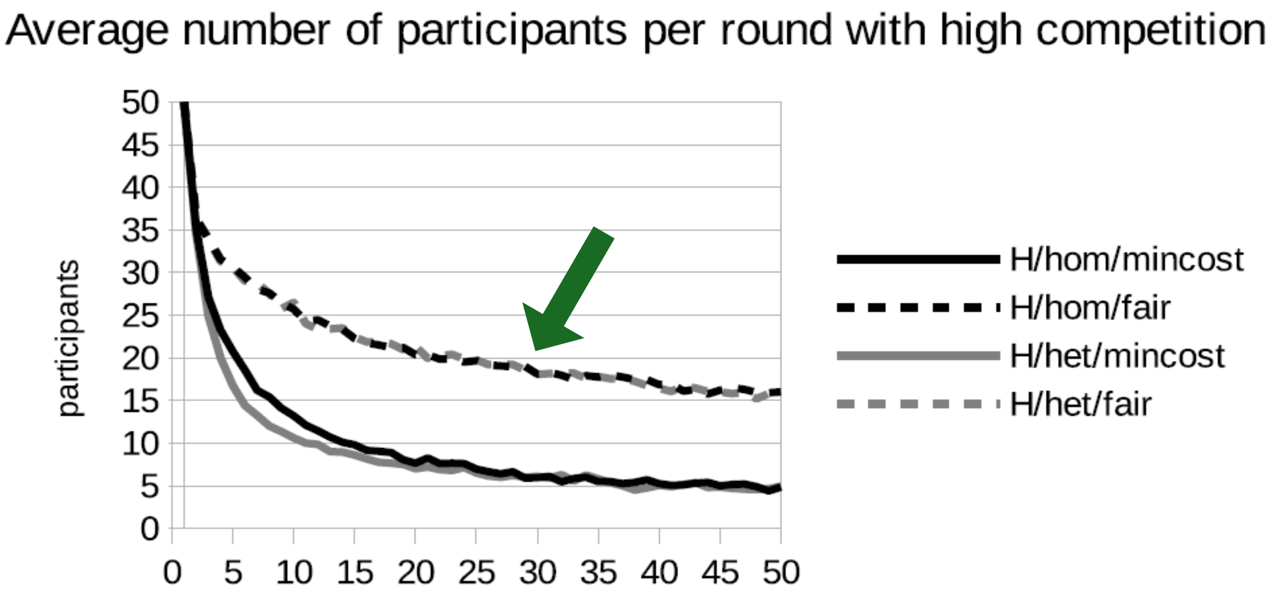
Simulate 50 rounds (i.e. days) of matching

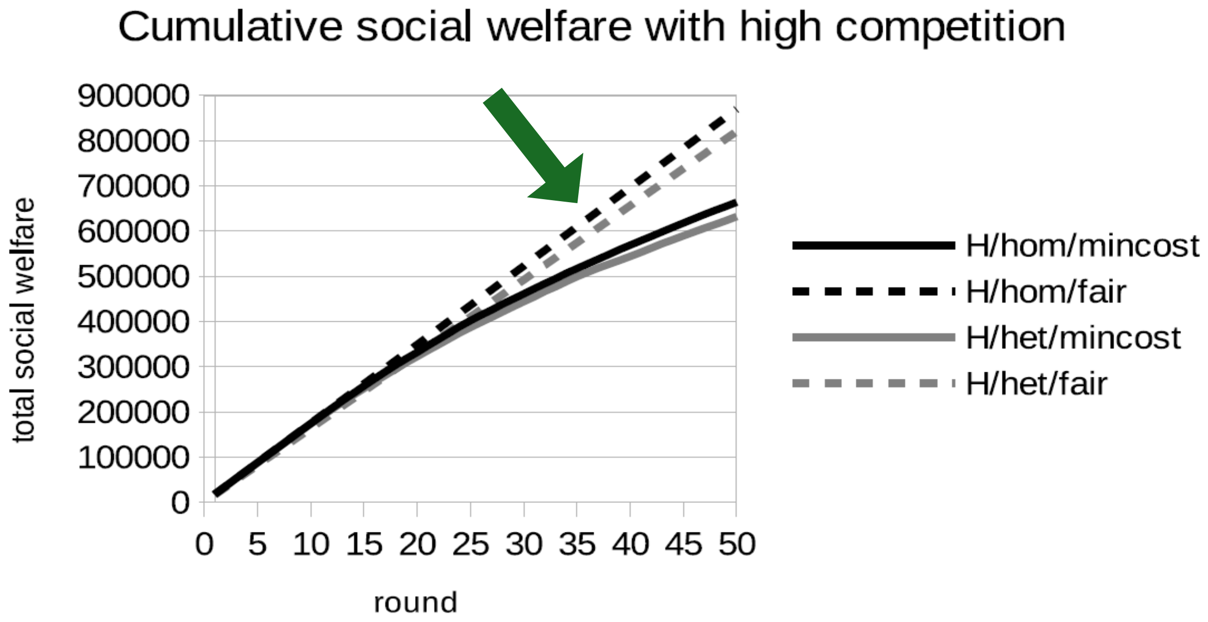
In the long run:

- more allocated jobs
- more participants
- increased social welfare

Fairness leading to

higher economic & social value!





Many work on fair optimization, although not RL

Paper	Measure	Approach	Solution	Domain
Kleinberg et al. (2001)	Max-min fairness	Approximation algorithm	Single	Load balancing
Harks (2005)	Proportional and max-min fairness	Lagrangian optimization	Single	Bandwidth allocation
Pioro (2007)	Max-min fairness	Seq. lexicographic optimization	Single	Bandwidth allocation
Ishida et al. (2006)	Variance			
Pishdad et al. (2010)	Quality of service fairnes			
Koppen et al. (2010)	Max-min fairness			
Meng and Khoo (2010)	Custom fairness measure			
Devarajan et al. (2012)	Jain's fairness index			
Tangpattanakul et al. (2012)	Maximum difference			
Stolletz and Brunner (2012)	Custom fairness constrai			
Escoffier et al. (2013)	α-fairness			
Amaldi et al. (2013)	Max-min fairness			
Bertin et al. (2014)	Custom fairness measure			
Yue and You (2014)	Nash bargaining fairness			
Yaacoub and Dawy (2014)	Max-min and quality of vice fairness			
Dely et al. (2015)	Max-min fairness			
Partov et al. (2015)	Custom fairness measure			
Sawik (2015)	Custom fairness measure			
L. Xu et al. (2015)	Jain's fairness index			

Z. Li et al. (2016)	Max-min fairness	ϵ-constraint method	Multi	Network traffic offloading
Busa-Fekete et al. (2017)	Generalized Gini Index	Online gradient descent	Single	Multi-objective bandits
X. Liu et al. (2017)	Max-min fairness	Evolutionary algorithms	Single	Load balancing
V. H. Nguyen and Weng (2017)	Generalized Gini Index	Primal-dual algorithm	Single	Classic combinatorial optimization
Alabi et al. (2018)	Multiple convex group-fairness measures	Polynomial-time reduction method	Single	General multi-objective optimization
Doi et al. (2018)	Custom and max-min fairness	Decomposition-based metaheuristic	Single	Crew scheduling
Limmer and Dietrich (2018)	Custom fairness measure	Genetic Algorithm	Multi	Dynamic pricing
Arribas et al. (2019)	α-fairness	Heuristic non-convex optimizer	Single	Network optimization
Diao et al. (2019)	Max-min fairness	Iterative algorithm	Single	Data allocation and trajectory optimization
J. Jiang and Lu (2019)	Custom variance-based measure	Hierarchical multi-agent RL	Single	Multi-agent RL
Zhao (2019)	Max-min and quality of service fairness	Alternating optimizatior		
Clausen et al. (2020)	Max-min and leximin fairness, and variance	Genetic algorithm		
Jagtenberg and Mason (2020)	Nash social welfare	MILP and local search		
Kermany et al. (2020)	Custom fairness metric	Genetic algorithm		
Z. Zhang et al. (2020)	Max-min fairness	Multi-objective local sea		
Z. Li et al. (2021)	Max-min fairness	MILP		
Lu and Wang (2021)	Max-min fairness	Alternating optimizatior		
Malencia et al. (2021)	Max-min fairness	Supermodular algorithm		
Munguía-López and Ponce-Ortega (2021)	Nash social welfare and max-min fairness	MILP		
Purushothaman and Nagarajan (2021)	Jain's fairness index	Evolutionary algorithm		

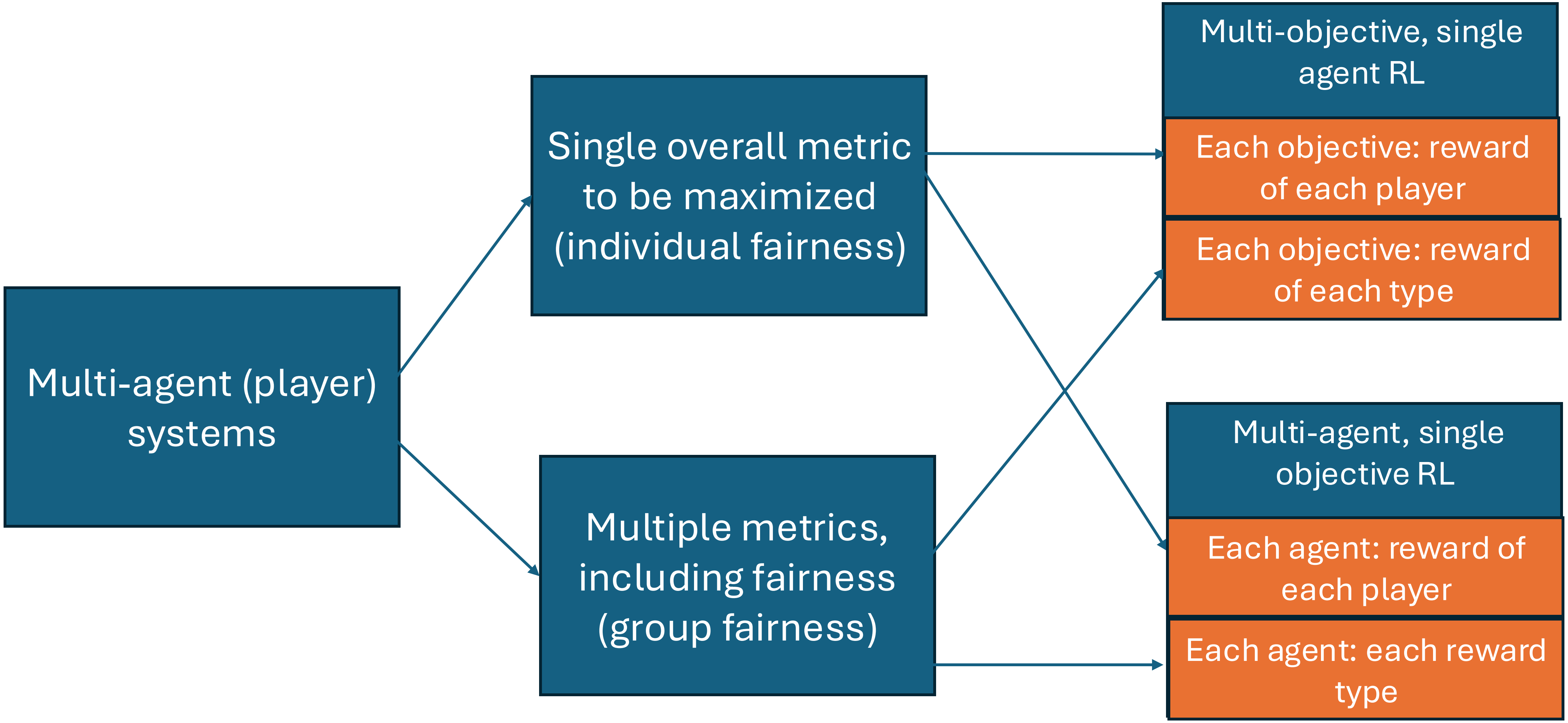
Rahmattalabi et al. (2021)	Multiple group-fairness measures	MILP	Single	Influence maximization
Tang et al. (2021)	Gini coefficient	Genetic algorithm	Multi	Water resource allocation
Zhou et al. (2021)	Variance	Ant colony system algorithm	Multi	Crew scheduling
Zimmer et al. (2021)	Max-min and proportional fairness and Generalized Gini Index	multi-agent RL algorithm	Single	General multi-agent RL
Arribas et al. (2022)	α-fairness	Extremal optimization	Single	Network Optimization
Fan et al. (2022)	Nash social welfare	Q-learning adaptation	Single	Multi-objective classic RL
F. Li et al. (2022)	Custom fairness measure	Genetic algorithm	Multi	Multi-workflow scheduling
Y. Liu, Huangfu, et al. (2022)	Quality of service fairness	Proximal stochastic gradient descent	Single	UAV placement
Kuai et al. (2022)	Max-min fairness	Offline PPO	Single	Virtual network scheduling
Sadiq et al. (2022)	Custom fairness measure	Non-linear marine predator algorithm	Single	Power allocation
Y. Wang et al. (2022)	Maximum difference	Genetic algorithm adaptation	Multi	Virtual power plant profit allocation
Gong and Guo (2023)	Gini coefficient adaptation	Custom genetic approach	Multi	Influence maximization
Y. Jiang et al. (2023)	Custom fairness measure	Genetic algorithm with large neighborhood search	Multi	Airport gate assignment
Wu et al. (2023)	Custom fairness measure	Multiple gradient descent	Multi	Recommender System

Challenge: fairness RL for decision-making

- Lack of overview on
  - Which fairness notions are most appropriate for different problems, which are both meaningful and operationally feasible (computable)

- Modeling fairness: a need for guidelines on how to effectively integrate fairness within the RL paradigm.

Fairness in multi-agent decision-making



Challenge: from computational point of view

- Some fairness notions are easier to be incorporated into existing optimization models/algorithms, e.g., max-min, Jain's index, Nash social welfare measure
- Many not:
  "even with very simple preferences (additive), deciding whether there is a Pareto-efficient and envy-free allocation is computationally very hard"
  - De Keijzer et al., 2009

  also see: *Brandt et al., 2012: computational social choice*
- Solving complex decision-making (NP-hard) problems with RL is still immature