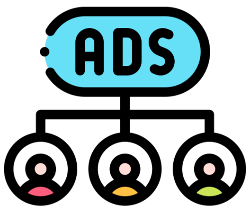
Fair and Efficient Social Decision-Making

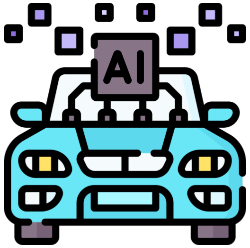
CSCI 699

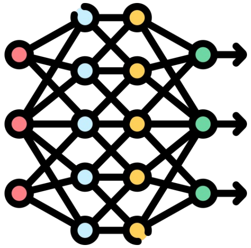
Fairness in Ml and AI

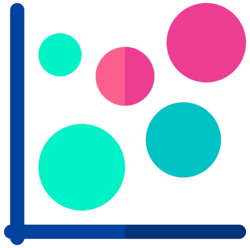
Evi Micha

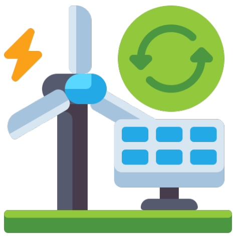
Why fairness?



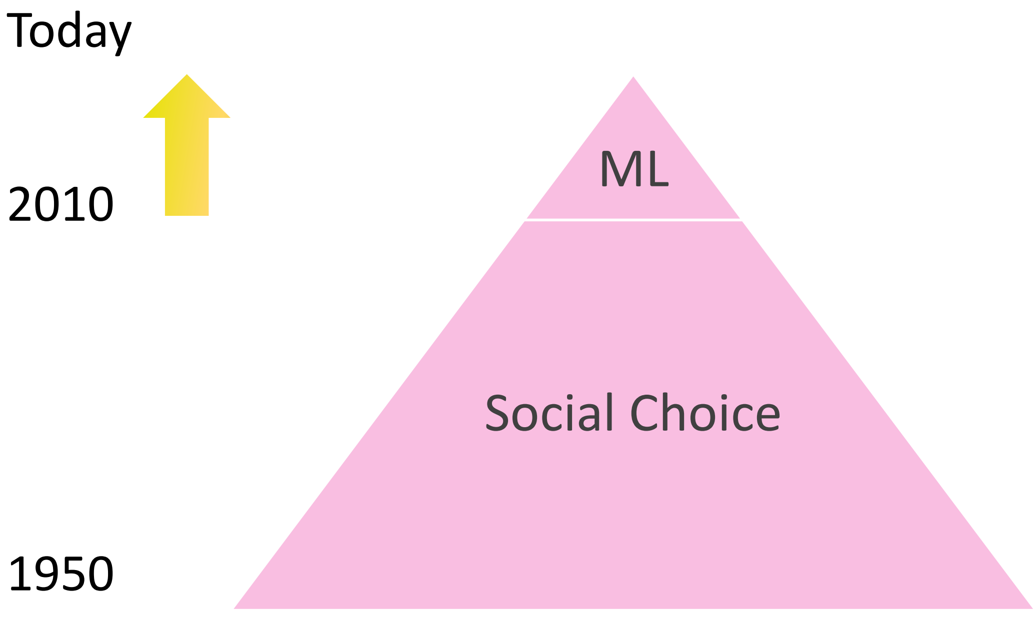


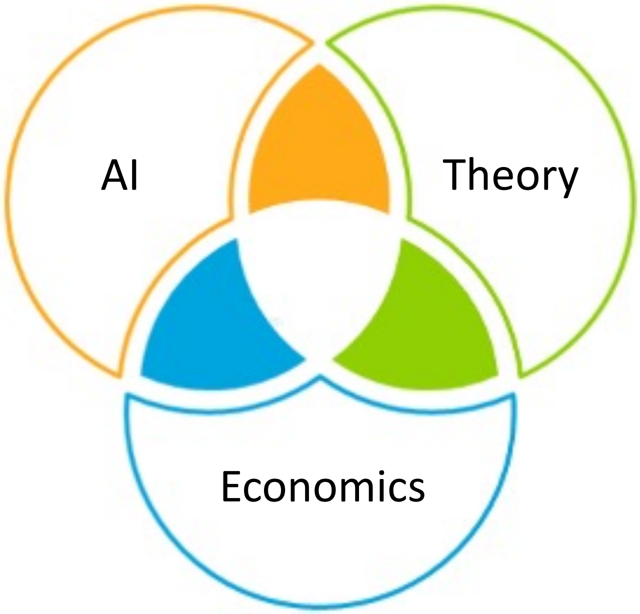






Fairness research





- Envy-freeness
  - Classification, recommender systems, clustering
- Nash social welfare
  - Multi-armed bandits, rankings, classification
- Core
  - Federated learning, clustering

Advantages

- Key advantages of social choice fairness criteria

- Broadly defined
  - Often depend only on the definition of *who* the agents are and *what* their preferences are
  - Applicable to any setting as long as you define these two pieces of information

- They respect the preferences of the agents to whom we wish to be fair
  - As a consequence, they are often defined beyond just binary decisions

- Notions such as the core achieve group fairness to all possible groups
  - No need to pre-specify the groups
  - The strength of the guarantee scales automatically with the group size and cohesiveness, without having to subjectively choose free parameter values

Envy-Freeness in ML

Classification

- Model
  - Population of individuals given by a distribution D over X
    - Individual i represented using data point $x_i \in X$
  - Classifier $f: X \rightarrow Y$ maps every individual to a classification outcome

- Types of classification outcomes
  - Hard binary classification: $Y = \{0,1\}$
  - Hard multiclass classification: $|Y| = p > 2$
  - Soft binary classification: $Y = [0,1]$
  - Soft multiclass classification: $Y \in \mathbb{R}^p, p > 2$

Classification

- Objective of the principal: minimize the loss $\mathbb{E}_{x\sim D}\left[\ell\left(x, f(x)\right)\right]$
  - If $f(x)$ is a distribution, $\ell\left(x, f(x)\right) = \mathbb{E}_{y\sim f(x)}[\ell(x, y)]$

- Utility function $u: X\times Y \to \mathbb{R}_{\geq 0}$
  - Utility to individual i is $u\left(x_i, f(x_i)\right)$

- Fairness is often modeled as a constraint that uses the utility function u

Individual Fairness

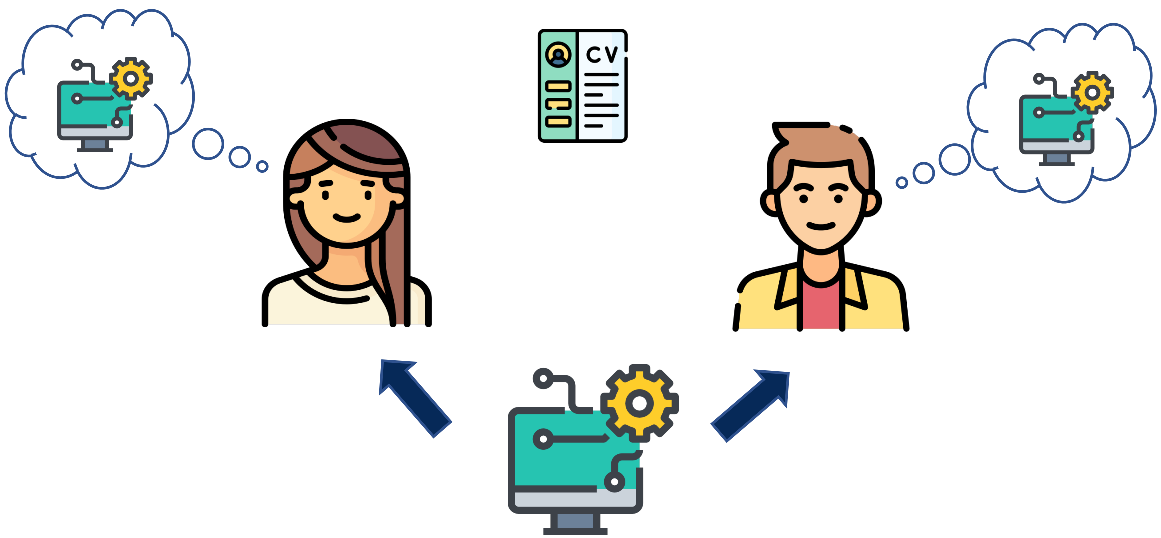
[Dwork, Hardt, Pitassi, Reingold, Zemel, 2012]

“Similar individuals should be treated similarly”

Classifier f is individual fair if:

$$\forall x, y \in N, \qquad D\big(f(x), f(y)\big) \leq d(x, y)$$

$D(p, q)$ measures some distance between two allocations p, q



Individual Fairness

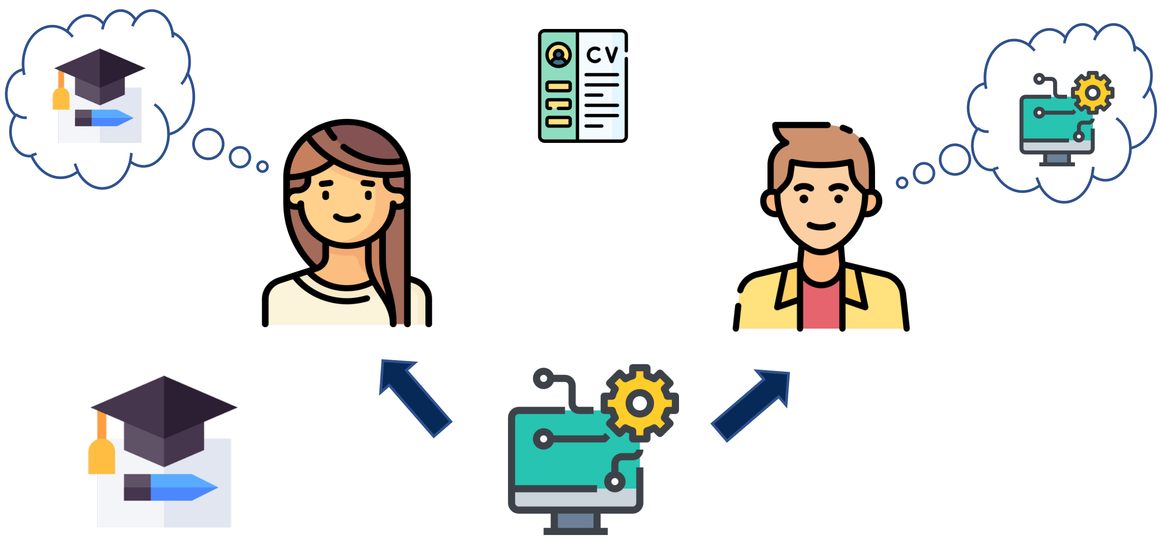
[Dwork, Hardt, Pitassi, Reingold, Zemel, 2012]

"Similar individuals should be treated similarly"

Classifier f is individual fair if:

$$\forall x, y \in N, \qquad D\big(f(x), f(y)\big) \le d(x, y)$$

$D(p, q)$ measures some distance between two allocations p, q



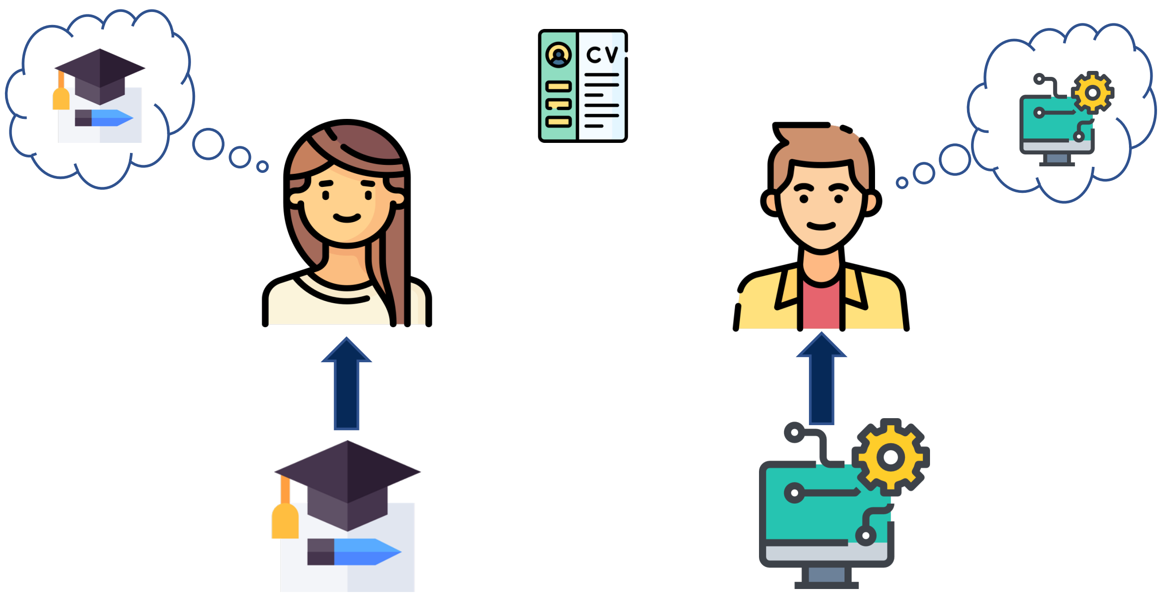
Envy-Freeness

[Balcan, Dick, Noothigattu, Procaccia, 2019]

"Equal individuals shouldn't envy each other"

Classifier f is envy-free if:

$$\forall x, y \in N,\ u_x\big(f(x)\big) \geq u_x\big(f(y)\big)$$



Envy-Freeness

[Balcan, Dick, Noothigattu, Procaccia, 2019]

- Space X of individuals
- Space Y of outcomes
- Utility function $u: X \times Y \to [0,1]$
- **Goal:** Find a classifier h: $X \to Y$ that is envy free and subject to that minimizes the loss
- Does the optimal deterministic classifier incur a loss that is very close to that of the optimal randomized classifier?

Envy-Freeness

[Balcan, Dick, Noothigattu, Procaccia, 2019]

- Observation: Envy-freeness is too strong for deterministic classifiers
  - Loss of optimal deterministic EF classifier ≥ 1

Utilities

	1	0	$\boxed{0}$
	$\boxed{1/\gamma}$	1	0

Losses

	1	0	$\boxed{0}$
	$\boxed{0}$	1	1

Envy-Freeness

[Balcan, Dick, Noothigattu, Procaccia, 2019]

- Observation: Envy-freeness is too strong for deterministic classifiers
  - Loss of optimal randomized EF classifier $\leq {}^{1}/_{\gamma}$

Utilities

	1 (1/γ)	0	0 (1-1/γ)
	1/γ	1	0

Losses

	1 (1/γ)	0	0 (1-1/γ)
	0	1	1

Envy-Freeness

[Balcan, Dick, Noothigattu, Procaccia, 2019]

- Space X of individuals
- Space Y of outcomes
- Utility function $u: X \times Y \rightarrow [0,1]$
- A classifier h: $X \rightarrow \Delta(Y)$ is (α, β)-EF if
  - $\Pr_{x,x' \sim P}\big(u(x,h(x)) < u(x,h(x')) - \beta\big) \leq \alpha$
  - where $u(x, \mathrm{h}(\mathrm{X})) = E_{y\sim \mathrm{h}(x)} u(x,y)$
- Learning problem:
  - Access to an unknown distribution P over X and their utility functions
  - Find a (α, β)-EF that minimizes expected loss $E_{x\sim P}[\ell(x,h(x))]$
    - $\ell(x,h(x)) = E_{y\sim \mathrm{h}(x)} \ell(x,y)$
- Theorem (informal): Exponential many samples are needed for generalizing

Preference-Informed IF

[Kim, Korolova, Rothblum, Yona, 2019]

"Similar individuals should be treated similarly"

Classifier f is individual fair if:

$$\forall x, x' \in N, D\big(f(x), f(x')\big) \leq d(x, x')$$

"Equal individuals shouldn't envy each other"

Classifier f is envy-free if:

$$\forall x, x' \in N,\ u_x\big(f(x)\big) \geq u_x\big(f(x')\big)$$

"Similar individuals shouldn't envy each other too much"

Classifier f is PIIF if:

$$\forall x, x' \in N, \exists z \in Y, D\big(z, f(y)\big) \leq d(x, y) \wedge u_x\big(f(x)\big) \geq u_x(z)$$

- PIIF requires that either $f(x)$ satisfies individual fairness with respect to $f(y)$ or x prefers their allocation over some alternative allocation that would have satisfied individual fairness with respect to $f(y)$

- Theorem (informal): Any policy that is either IF or EF is also PIIF

Metric EF

[Kim, Korolova, Rothblum, Yona, 2019]

"Similar individuals shouldn't envy each other too much"

Classifier f satisfies metric α −EF if:

$$\forall x, x' \in N,\ u_x\big(f(x)\big) \geq u_x\big(f(x')\big) - \alpha \cdot d(x, x')$$

- A utility function u is ℓ − Lipschit with respect to D:Δ(Y)×Δ(Y)→$\mathbb{R}_+$ if
$$u\big(f(x), f(x')\big) \leq \ell \cdot D(f(x), f(x'))$$

- Theorem: If u is ℓ − Lipschit , then a PIIF classifier f satisfies *metric* ℓ *−EF*
- Proof:
- Suppose that a policy f satisfies PIIF
- Then, there exists $z \in Y$ such that
- $$\begin{aligned} u_x\big(f(x)\big) &\geq u_x(z) && \text{(Since } f \text{ satisfies PIIF)} \\ &\geq u_x\big(f(y)\big) - \Big(u_x\big(f(y)\big) - u_x(z)\Big) \\ &\geq u_x\big(f(y)\big) - \ell \cdot D(f(y), z) && \text{(from Lipschitness)} \\ &\geq u_x\big(f(y)\big) - \ell \cdot d(y, x) && \text{(Since } f \text{ satisfies PIIF)} \end{aligned}$$

Envy-Freeness ⇒ Recommendations

Envy-Freeness ⇒ Recommendations

[Do, Corbett-Davies, Atif, Usunier, 2023]

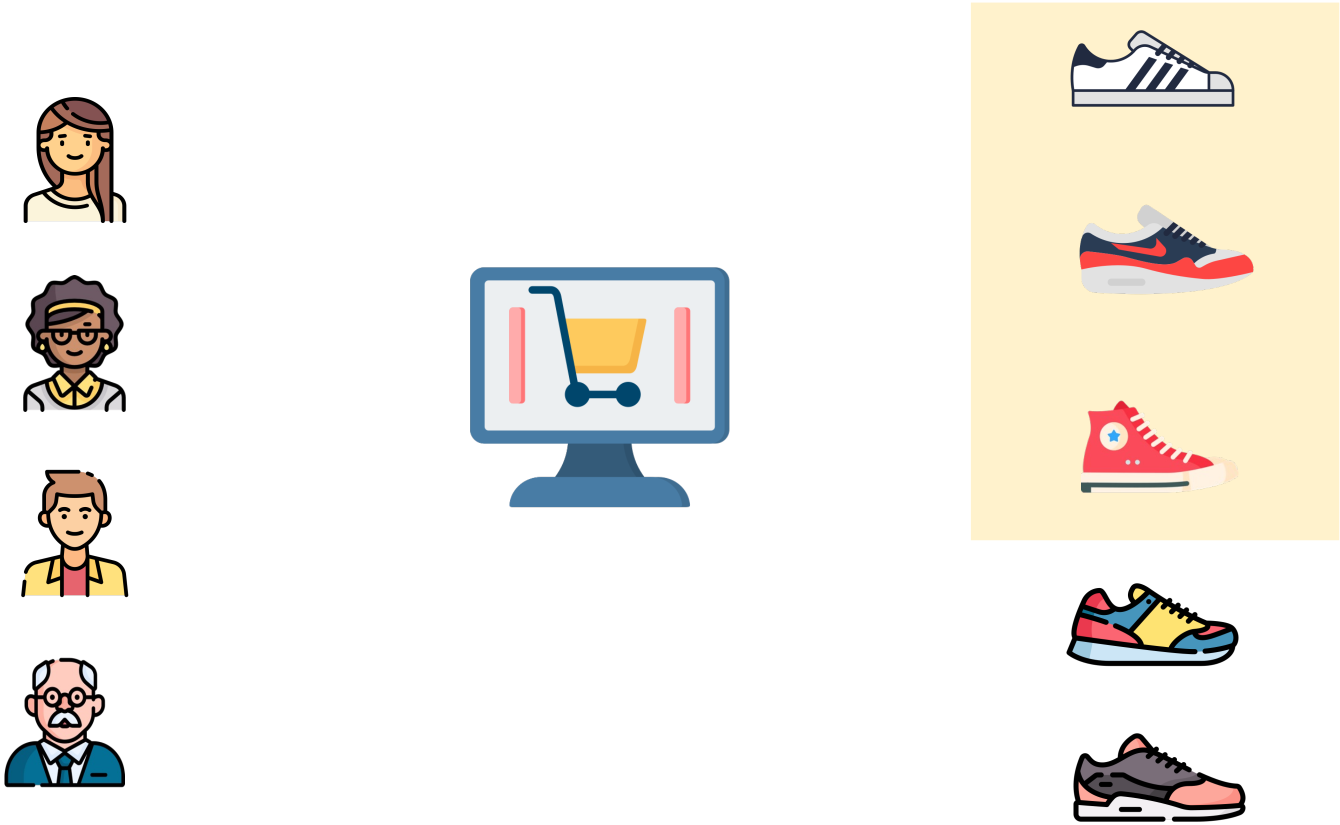
- Model
  - Individuals represented by data points in set X
  - A set items Y
  - A set of contexts C

- Recommendation policy π
  - $\pi_x(y|c)$ = probability of recommending item y to user x given a context c

- Utility function: $u_x(\pi_x) = \mathbb{E}_{c\sim C_m, y\sim\pi_x(\cdot|c)}[v_x(y|c)]$

- Envy-freeness: $\forall x, x' \in X,\ u_x(\pi_x) \geq u_x(\pi_{x'}) - \varepsilon$

Envy-Freeness ⇒ Recommendations



Two-Sided Fairness in Recommendations

[Biswas, Patro, Ganguly, Gummadi, Chakraborty, 2023]

- Many-to-many matching
  - Each user is recommended k products
  - Each product may be recommended to a different number of users

- Relevance of products to users given by $V: X \times Y \to \mathbb{R}$

- Recommendation policy π
  - Each user x is recommended $\pi_x \subseteq Y$ with $|\pi_x| = k$
  - Let π_x^* be the top-k products for user x by relevance

- Utilities
  - Utility to user x given by $u_x(\pi_x) = \frac{\sum_{y \in \pi_x} V(x,y)}{\sum_{y \in \pi_x^*} V(x,y)}$
  - Utility to product y given by $E_y(\pi)$, the number of users y is exposed to

Two-Sided Fairness in Recommendations

[Biswas, Patro, Ganguly, Gummadi, Chakraborty, 2023]

- Two-sided fairness
  - Fairness for users: envy-freeness up to one (EF1)

    $$\forall x, x' \in X, \exists y \in \pi_{x'}: u_x(\pi_x) \geq u_x(\pi_{x'} \setminus \{y\})$$

  - Fairness for products: minimum exposure $\bar{E}$

    $$\forall y \in Y, E_y(\pi) \geq \bar{E}$$

- Theorem: There exists an efficient algorithm that achieves EF1 among all users and the minimum exposure guarantee among at least $m - k$ products
  - The algorithm executes two variations round robin. At the first execution, it ensures EF1 for users and minimum exposure of all products. At the second execution, it ensures that k products are recommended to each use

- Future directions: Fairness to products in terms of the relevance, asymmetric entitlements of users

Two-Sided Fairness in Recommendations

[Freeman, M, Shah, 2021]

- Many-to-many matching
  - Each user is recommended k products
  - Each product is recommended to k users

- Relevance of products to users given by $V: X \times Y \rightarrow \mathbb{R}$

- Recommendation policy π
  - Each user x is recommended $\pi_x \subseteq Y$ with $|\pi_x| = k$
  - Each product y is recommended to $\pi_y \subseteq X$ with $|\pi_y| = k$

- Utilities
  - Utility to user x given by $u_x(\pi_x) = \sum_{y \in \pi_x} V(x, y)$
  - Utility to product y given by $u_y(\pi_y) = \sum_{\mathrm{x} \in \pi_{\mathrm{y}}} V(x, y)$

Two-Sided Fairness in Recommendations

[Freeman, M, Shah, 2021]

- Two-sided fairness
  - Fairness for users: envy-freeness up to one (EF1)

$$\forall x, x' \in X, \exists y \in \pi_{x'}: u_x(\pi_x) \geq u_x(\pi_{x'} \setminus \{y\})$$

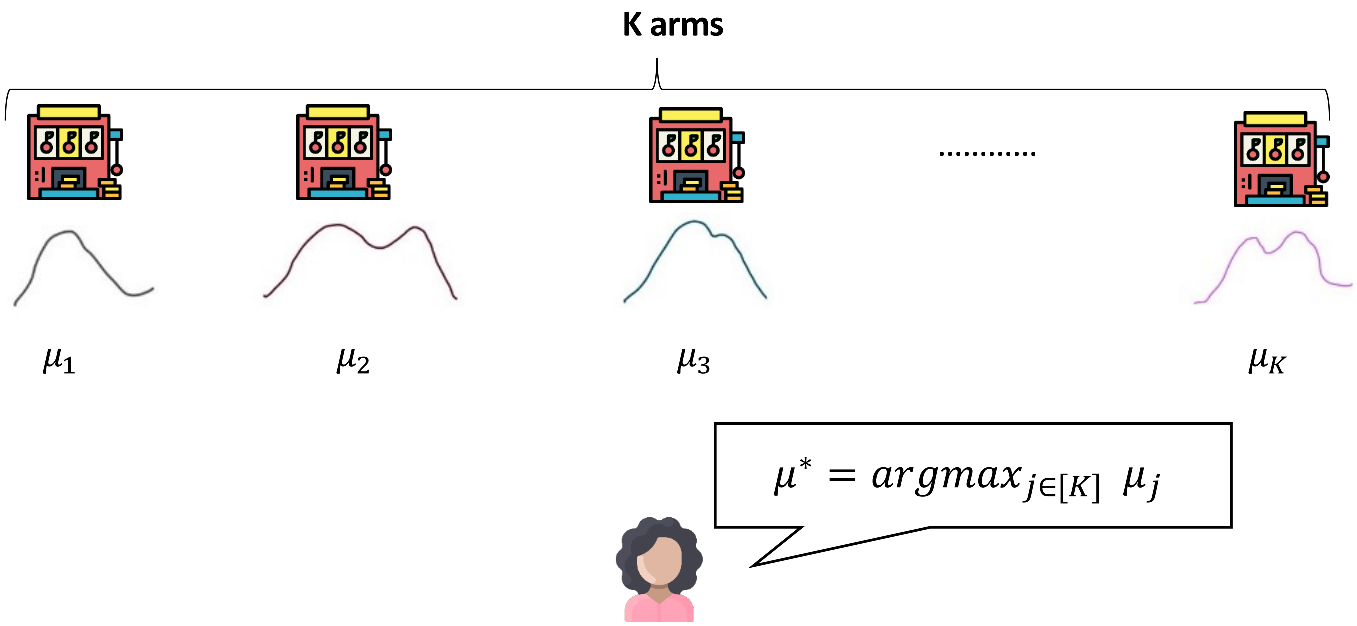
  - Fairness for products: envy-freeness up to one (EF1)

$$\forall \mathrm{y}, \mathrm{y}' \in \mathrm{Y}, \exists x \in \pi_{\mathrm{y}}: u_{\mathrm{y}}\left(\pi_{\mathrm{y}}\right) \geq u_{\mathrm{y}}\left(\pi_{\mathrm{y}'} \setminus \{x\}\right)$$

- Theorem: When each side agrees on the ranking of the other side by relevance, a policy that is EF1 w.r.t. both users and products exists and can be computed efficiently
  - Round robin by determining the order carefully
- Open question: Does a policy that is EF1 w.r.t. both sides always exist?
- Future directions: Non-stationary recommendations, different entitlements

Nash Social Welfare in ML

Multi-Armed Bandits

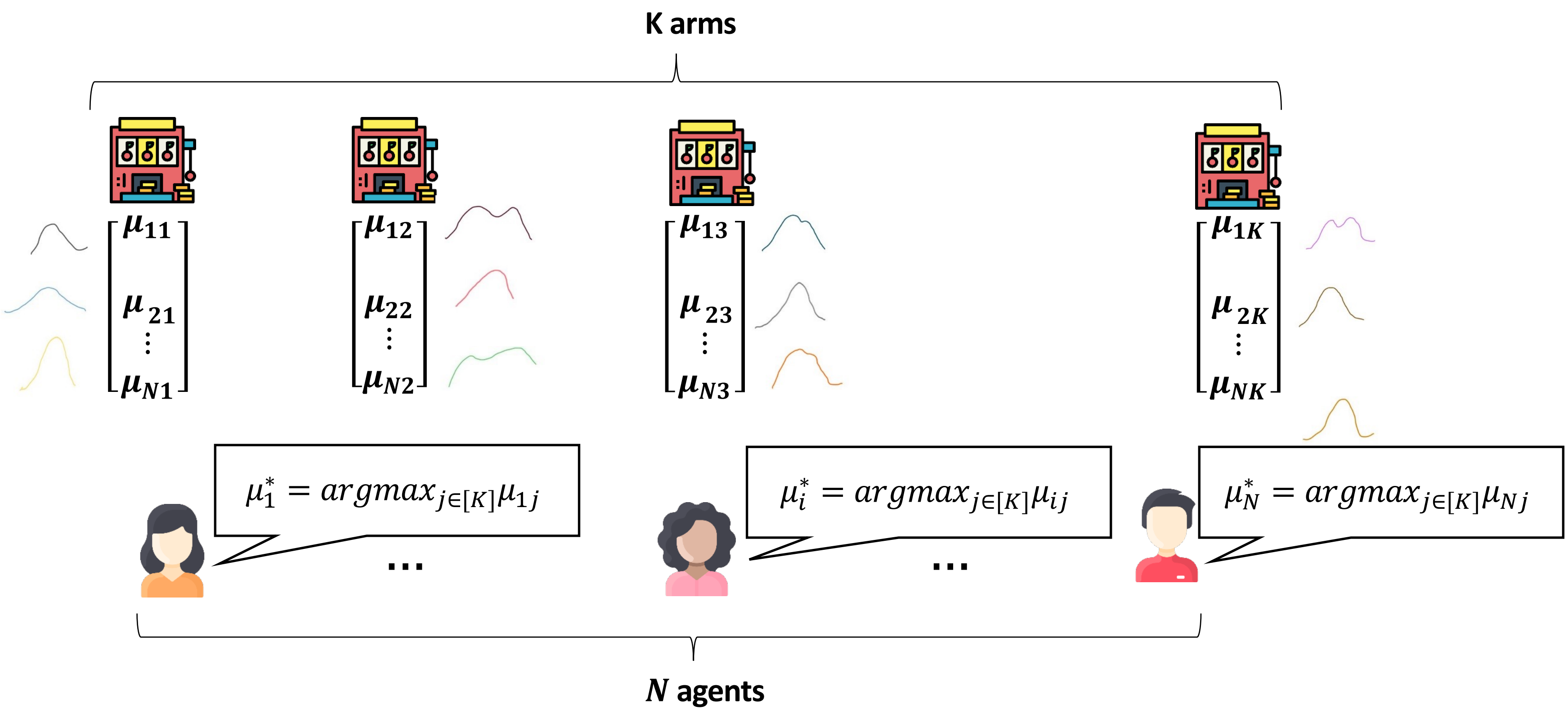


Exploration vs Exploitation

Regret: $R_T = T\mu^* - \sum_{t=1}^{T} \mu(t)$

Multi-Agent Multi-Armed Bandits

[Hossain, M, Shah, 2021]



What is a fair policy?

Multi-Agent Multi-Armed Bandits

[Hossain, M, Shah, 2021]

- Distribution $p = [p_1, \dots, p_K]$ gives expected reward $\sum_{j=1}^{K} p_j \cdot \mu_{ij}$ to agent iå
- Maximizing welfare functions
  a) Utilitarian welfare $\sum_{i=1}^{N} \sum_{j=1}^{K} p_j \cdot \mu_{ij}$
  b) Egalitarian welfare $\min_{i \in N} \sum_{j=1}^{K} p_j \cdot \mu_{ij}$
  c) Nash welfare $\prod_{i=1}^{N} \sum_{j=1}^{K} \boldsymbol{p_j} \cdot \boldsymbol{\mu_{ij}}$

$p_1^a = 1$ $\quad \begin{bmatrix}1\\1\\0\end{bmatrix} \quad \begin{bmatrix}0\\0\\1\end{bmatrix} \quad$ $p_2^a = 0$

$p_1^b = 1/2$ $\quad$ $p_2^b = 1/2$

$p_1^c = 2/3$ $\quad$ $p_2^c = 1/3$

- Regret: $R_T = NSW(p^*, \mu) - \sum_{t=1}^{T} NSW(p(t), \mu)$

Explore First

[Hossain, M, Shah, 2021]

Exploration

- Pull each arm L times
- Calculate $\hat{\mu}_{ij} = \sum_{t=1}^{L} \frac{X_{ij}^{t}}{L}$

Exploitation

- $\hat{p} = argmax_p NSW(p, \hat{\mu})$

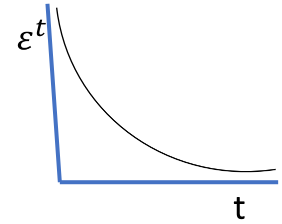
- When $L = \widetilde{\Theta}(N^{2/3}K^{-2/3}T^{2/3})$, then $\boldsymbol{E}[\boldsymbol{R_T}] = \widetilde{\boldsymbol{O}}(\boldsymbol{N^{2/3}K^{1/3}T^{2/3}})$
- When $L = \widetilde{\Theta}(N^{1/3}K^{-1/3}T^{2/3})$, then $\boldsymbol{E}[\boldsymbol{R_T}] = \widetilde{\boldsymbol{O}}(\boldsymbol{N^{1/3}K^{2/3}T^{2/3}})$

ϵ-Greedy

[Hossain, M, Shah, 2021]

For t=1, 2… do

- Toss a coin with success probability ε^t



If success **do**

Exploration

- *pull arm j*
- $j \leftarrow j + 1 \ mod \ n$

Else do

Exploitation

- Calculate $\hat{\mu}_{ij}^t = \sum_{s=1}^{t} \frac{X_{ij}^s}{n_{ij}^t}$
- $p^t = argmax_p NSW(p, \hat{\mu}^t)$

- When $\varepsilon^t = \widetilde{\Theta}(N^{2/3}K^{1/3}t^{-1/3})$, then $\boldsymbol{E}[\boldsymbol{R_T}] = \widetilde{\boldsymbol{O}}(\boldsymbol{N^{2/3}K^{1/3}T^{2/3}})$
- When $\varepsilon^t = \widetilde{\Theta}(N^{1/3}K^{2/3}t^{-1/3})$, then $\boldsymbol{E}[\boldsymbol{R_T}] = \widetilde{\boldsymbol{O}}(\boldsymbol{N^{1/3}K^{2/3}T^{2/3}})$

Upper Confidence Bound (UCb)

[Hossain, M, Shah, 2021]

For t=1, 2… do

- Calculate $\hat{\mu}_{ij}^t = \sum_{s=1}^{t} \frac{X_{ij}^s}{n_{ij}^t}$
- $Calculate\ UCB(p) = NSW(p, \hat{\mu}^t) + \alpha^t \cdot \sum_{j=1}^{K} p_j \cdot \sqrt{\frac{\log(NKt)}{n_{ij}^t}}$
- $p^t = argmax_p UCB(p)$

- When $\alpha^t = \mathrm{N}$, then $\boldsymbol{E}[\boldsymbol{R_T}] = \widetilde{\boldsymbol{O}}(\mathbf{NK}\boldsymbol{T}^{\mathbf{1/2}})$
- When $\alpha^t = \widetilde{\Theta}(N^{1/2}K^{1/2})$, then $\boldsymbol{E}[\boldsymbol{R_T}] = \widetilde{\boldsymbol{O}}(\boldsymbol{N}^{\mathbf{1/2}}\boldsymbol{K}^{\mathbf{3/2}}\boldsymbol{T}^{\mathbf{1/2}})$

Upper Confidence Bound (UCb)

[Jones, Nguyen, Nguyen, 2023]

For t=1, 2… do

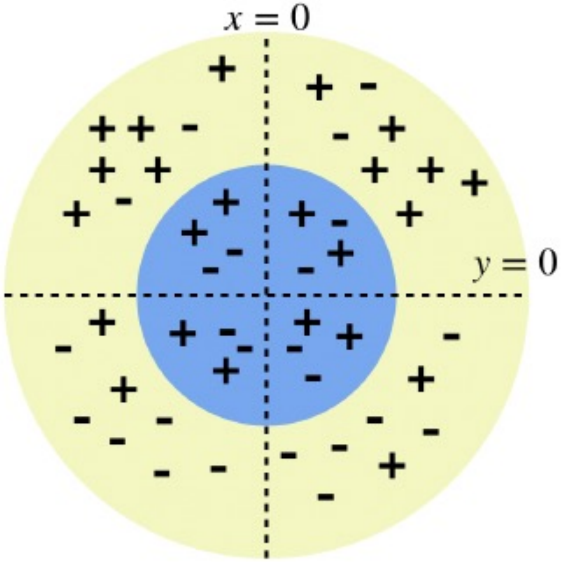
- Calculate $\hat{\mu}_{ij}^{t} = \sum_{s=1}^{t} \frac{X_{ij}^{s}}{n_{ij}^{t}} + \sqrt{\frac{\log(NKt)}{n_{ij}^{t}}}$
- $p^{t} = argmax_{p}\, NSW(p, \hat{\mu}^{t})$

- $\boldsymbol{E[R_T]} = \widetilde{\boldsymbol{O}}\,(\boldsymbol{N}^{1/2}\boldsymbol{K}^{1/2}\boldsymbol{T}^{1/2} + \boldsymbol{NK})$

Classification

[Krishnaswamy, Jiang, Wang, Cheng, Munagala, 2021]

- **Standard Notion of Fairness:** Statistical Parity or Equalized odds



Can *every* group of individuals be treated at least as well as it can be classified in itself?

Classification

[Krishnaswamy, Jiang, Wang, Cheng, Munagala, 2021]

- **Utility of an individual:** $u_i(f) = \mathbb{I}[f(x_i) = y_i]$
- **Utility of a group:** $u_S(f) = \frac{1}{|S|} \sum_{i \in S} u_i(f)$
- **Optimal Classifier for a group:** $f_S^* = argmax_{f \in F} u_S(f)$
- **Best-effort Guarantees**
  - Return f such that $u_S(f) \geq \alpha \cdot u_S(f_S^*)$, with $\alpha \leq 1$, for each $S \subseteq N$
- **Observation:** No imperfect classifier f provides any reasonable guarantee to best-effort
  - Let $S = \{i \in N: f(x_i) \neq y_i\}$ and $u_s(f_S^*) = 1$
- **Randomized Classifiers:** Let D_f be a distribution over F
  - $u_i(D_f) = \mathbb{E}_{f \sim D_f}[u_i(f)]$
  - $u_S(D_f) = \frac{1}{|S|} \sum_{i \in S} \mathbb{E}_{f \sim D_f}[u_i(f)]$

Classification

[Krishnaswamy, Jiang, Wang, Cheng, Munagala, 2021]

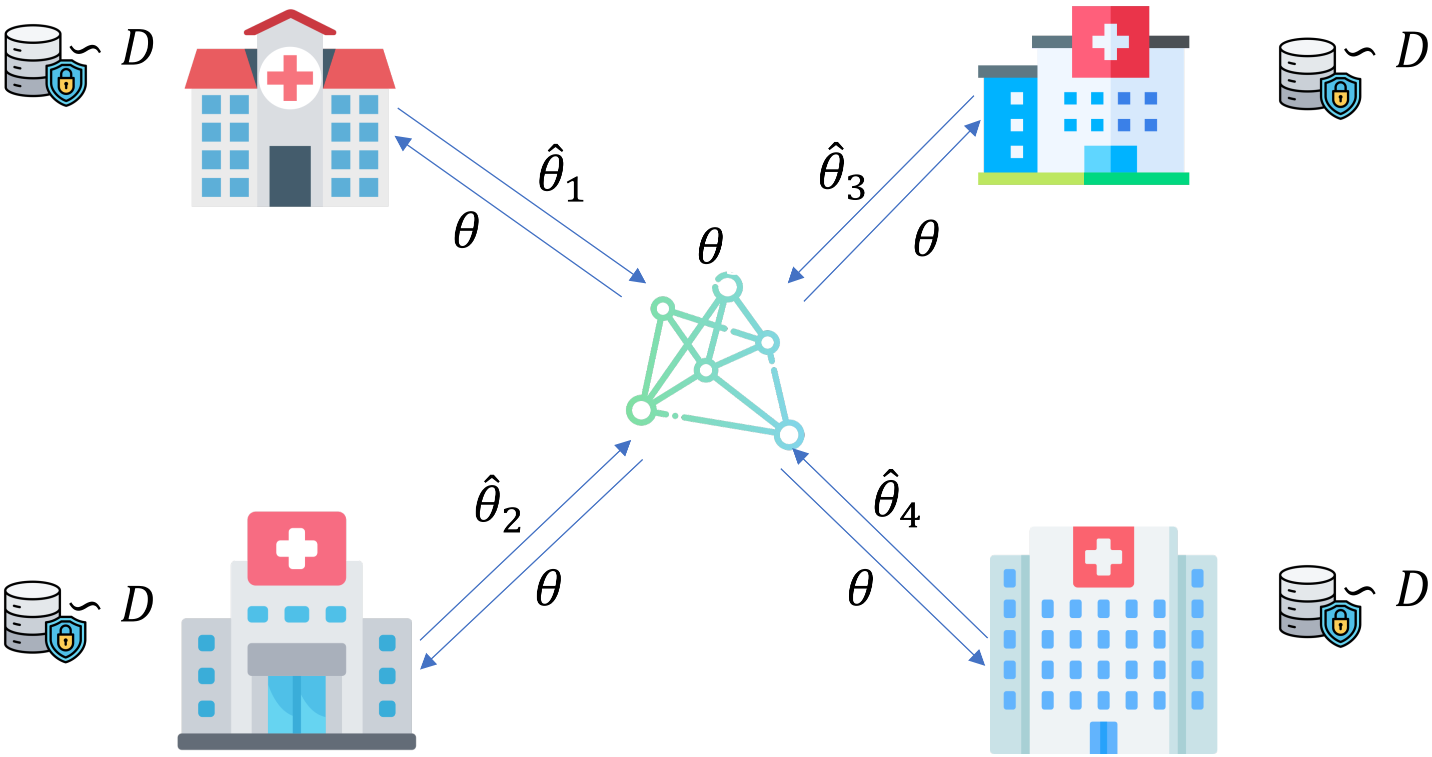
- **Theorem:** There is an instance in which there is no distribution D_f over classifiers

  such that for all $S \subseteq N$ with $u_s(f_S^*) = 1$, $u_s(D_f) > \frac{|S|}{|N|}$
- $\boldsymbol{D_f^{NSW} = argmax_{D_f \in \Delta(F)} \prod_{i \in N} u_i(D_f)}$
- **Theorem:**

  1. For every group $S \subseteq N$ that admits a perfect classifier, $u_S(D_f^{NSW}) \geq \frac{|S|}{|N|}$
  2. For every group $S \subseteq N$, $u_S(D_f^{NSW}) \geq \frac{|S|}{|N|}[u_S(f_S^*)]^2$

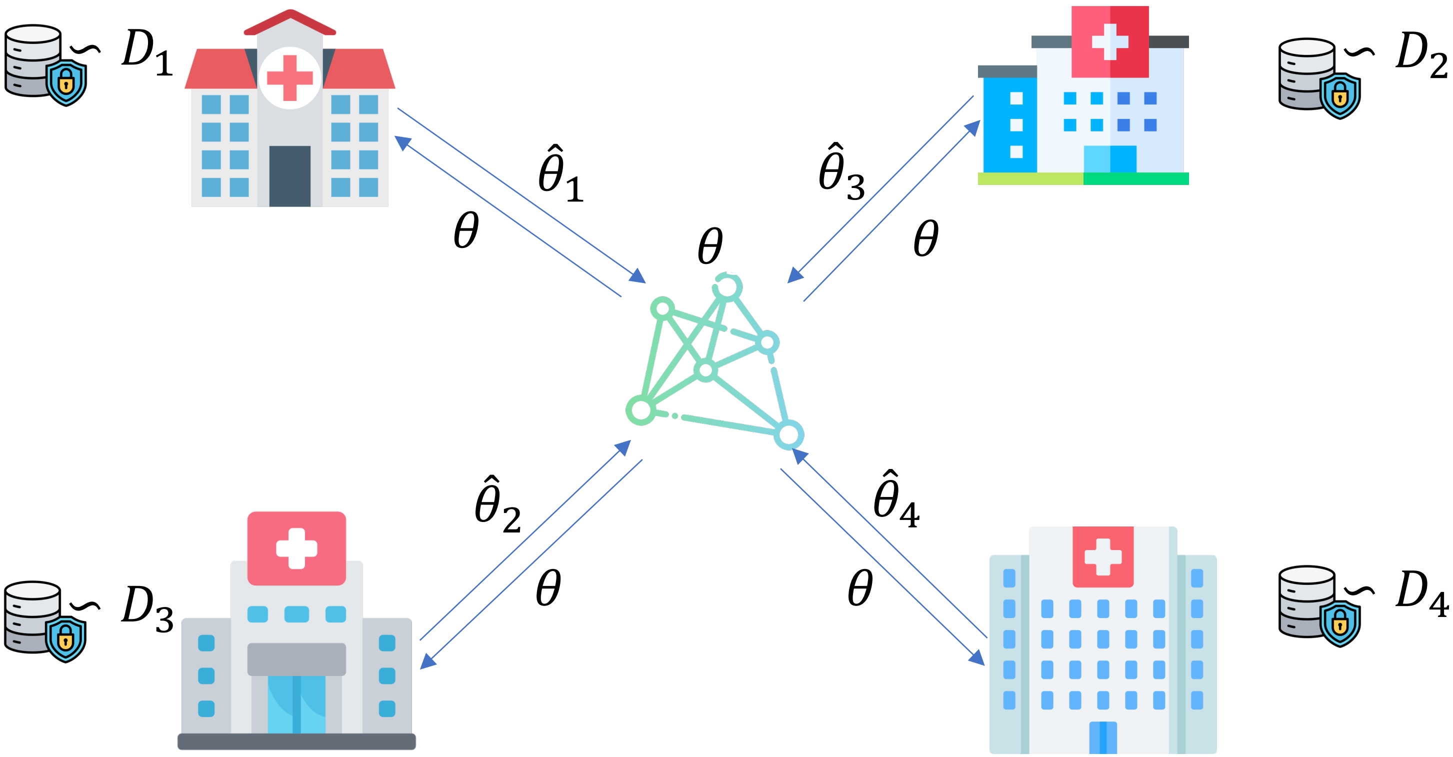
Core in ML

Federated Learning



- Goal: Choose $f_\theta\colon \mathbb{R}^d \to \mathbb{R}$ from $F = \{f_\theta\colon \theta \in P \subseteq \mathbb{R}^d\}$

Federated Learning



- Goal: Choose $f_\theta\colon \mathbb{R}^d \to \mathbb{R}$ from $F = \{f_\theta\colon \theta \in P \subseteq \mathbb{R}^d\}$

Federated Learning

[Chaudhury, Li, Kang, Li, Mehta, 2022]

- **Utility of each agent:**
  - $u_i(\theta) = M - \mathbb{E}_{(x,y)\sim D_i}\left[\ell_i(f_\theta(x), y)\right]$
- **Goal:** Choose θ that is fair for all agents
- **Core:** A parameter vector $\theta \in P$ is in the core if for all $\theta' \in P$ and $S \subseteq N$, it holds

  $u_i(\theta) \geq \frac{|S|}{|N|} u_i(\theta')$ for all $i \in S$, with at lost one strict inequality
- **Pareto Optimality:** A parameter vector $\theta \in P$ is Pareto Optimal if there exists no $\theta' \in P$ such that $u_i(\theta') \geq u_i(\theta)$ for all $i \in N$, with at lost one strict inequality
- **Proportionality:** A parameter vector $\theta \in P$ is proportionally fair if for all $\theta' \in P$, it holds

  $u_i(\theta) \geq \frac{u_i(\theta')}{|N|}$ for all $i \in N$

Federated Learning

[Chaudhury, Li, Kang, Li, Mehta, 2022]

- **Theorem:** When the agents' utilities are continuous and the set of maximizers of any conical combination of the agents' utilities is convex, a parameter vector $\theta \in P$ in the core always exists

- **Theorem:** When the agents' utilities are concave, then the parameter vector $\theta \in P$ that maximizes the NSW is in the core

maximize $\prod_{i \in N} u_i(\theta)$

subject to $\theta \in P$

maximize $\sum_{i \in N} \log(u_i(\theta))$

subject to $\theta \in P$

Core in AI

Peer Review Model

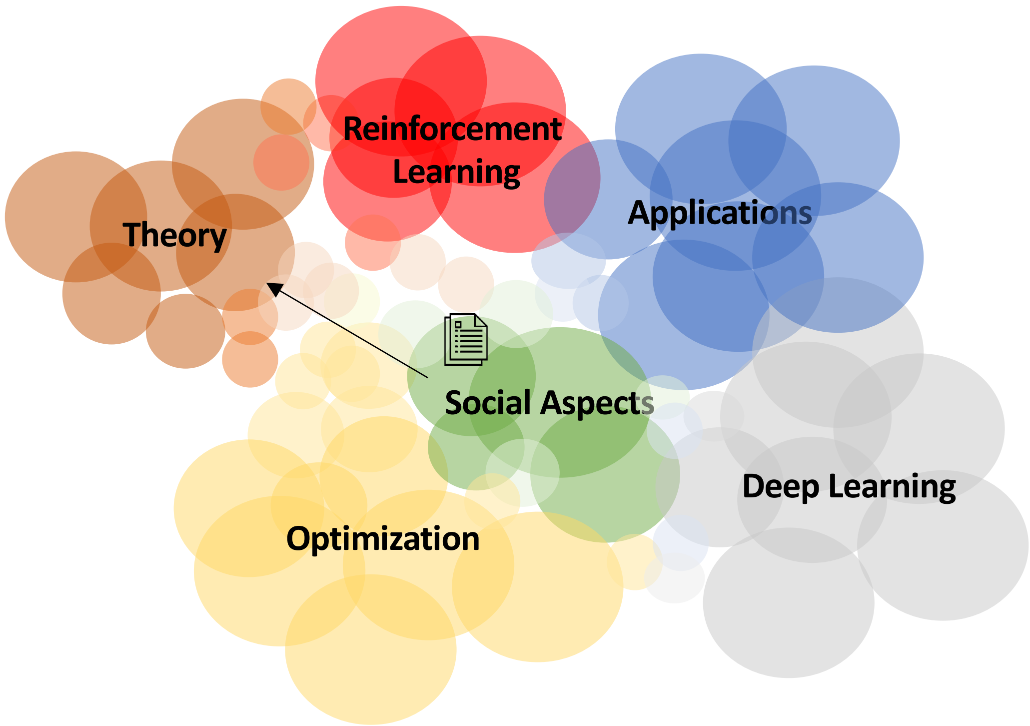
[Aziz, M, Shah, 2023]

- A set $N = [n]$ of authors that serve as reviewers
- Each author i submits a set of papers P_i

An assignment of $\cup_{i \in N} P_i$ over N is *valid* if:

- No agent is assigned to review her own papers
- Each paper is assigned to k_p reviewers
- Each reviewer is assigned to review up to k_a papers

NeurIPS

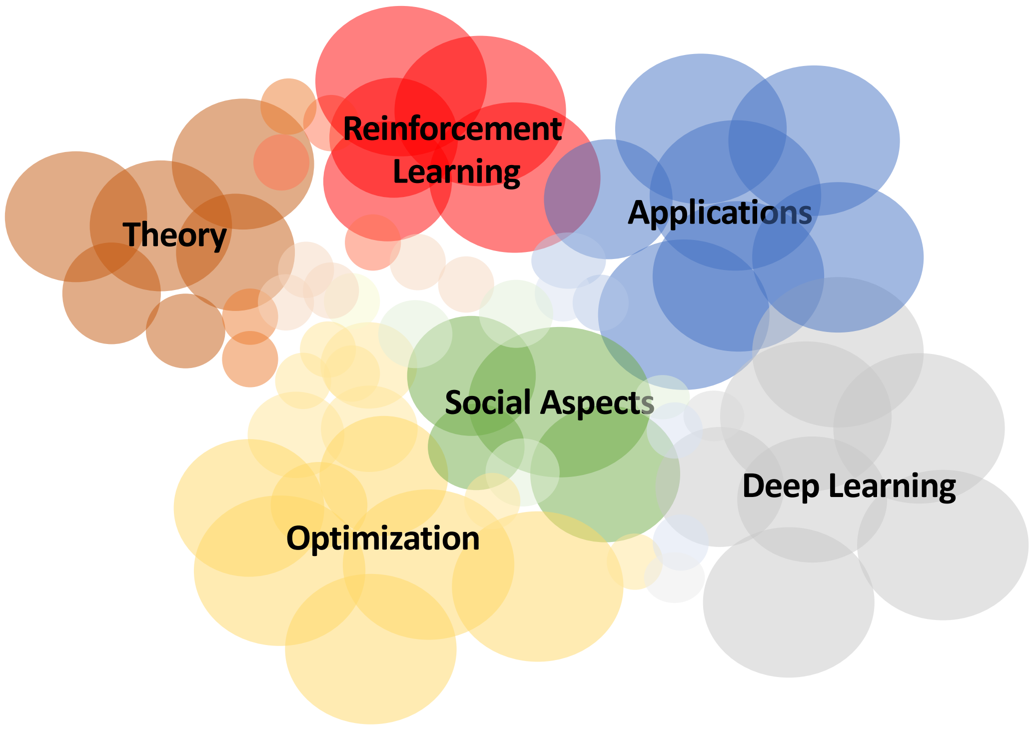


FAccT

COLT

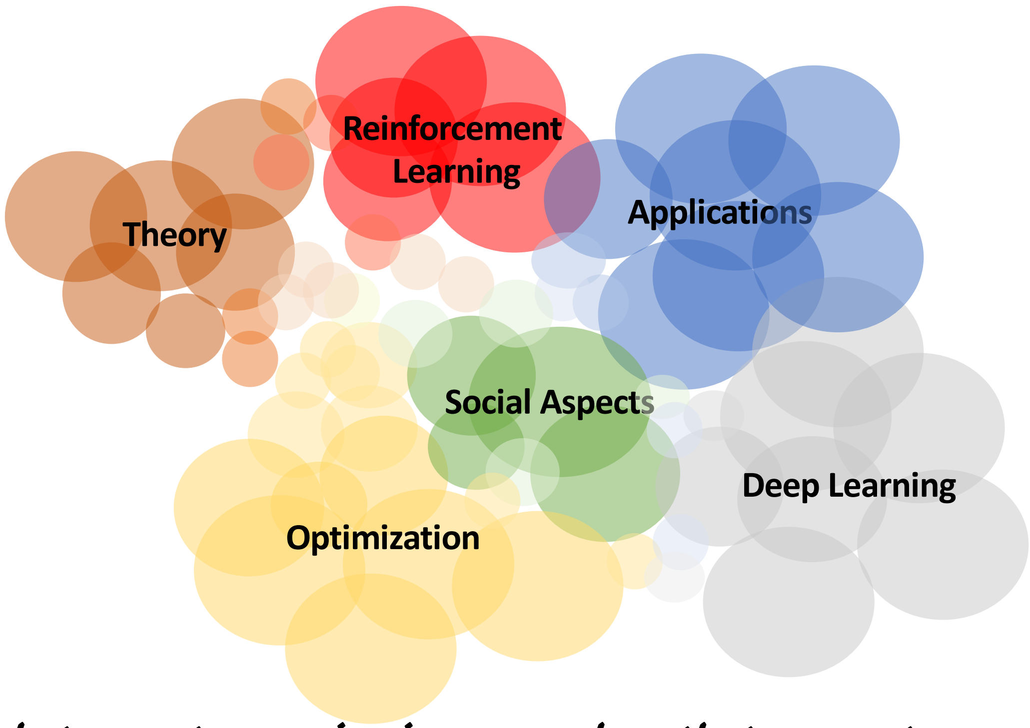
ALT

NeurIPS



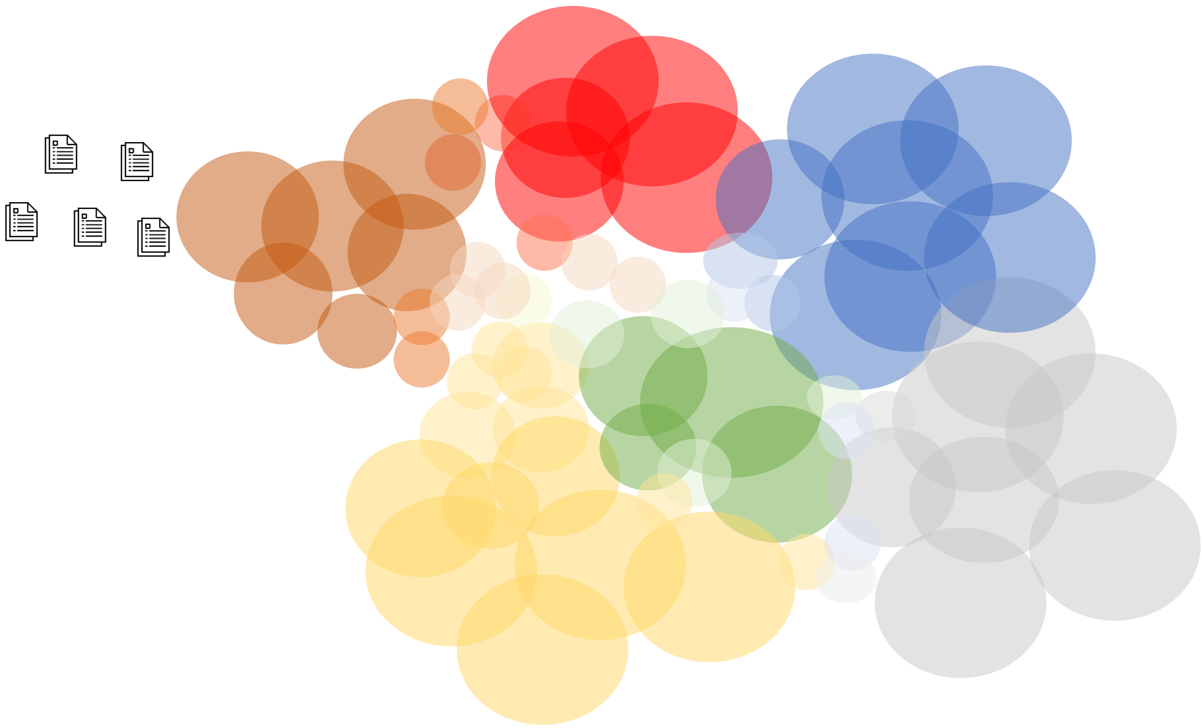
Reinforcement Learning
Theory
Applications
Social Aspects
Deep Learning
Optimization

NeurIPS



Is it possible to create a reviewing procedure that prevents any subcommunity from benefiting by withdrawing from a large conference?

Core as A Notion of Fairness



Peer Review Model

[Aziz, M, Shah, 2023]

An assignment R is in the core if there is no $N' \subseteq N$, $P_i' \subseteq P_i$ for $i \in N'$ and a valid assignment R' of $\cup_{i \in N'} P_i'$ over N' such that

$$\forall\, i \in N', R' \succ_i R$$

Theorem: There exists an efficient algorithm, called ***CoBRA***, that finds an assignment in the core.

Experiments with Rea Data

- TPMS (Toronto Paper Matching System)
- PR4A (Peer Review for All)

Dataset	Algo	USW	ESW	α-Core		CV-Pr
				#unb-α	α^*	
CVPR 2017	CoBRA	1.225 ± 0.021	0.000 ± 0.000	0%	1.00+0.00	0%
	TPMS	1.497 ± 0.019	0.000 ± 0.000	89%	3.134 ± 0.306	100%
	PR4A	1.416 ± 0.019	0.120 ± 0.032	51%	1.700 ± 0.078	100%
CVPR 2018	CoBRA	0.224 ± 0.004	0.004 ± 0.001	0%	1.000 ± 0.000	0%
	TPMS	0.286 ± 0.005	0.043 ± 0.004	0%	1.271 ± 0.038	100%
	PR4A	0.282 ± 0.005	0.099 ± 0.001	0%	1.139 ± 0.011	100%
ICLR 2018	CoBRA	0.166 ± 0.001	0.028 ± 0.001	0%	1.000 ± 0.000	0%
	TPMS	0.184 ± 0.001	0.048 ± 0.002	0%	1.048 ± 0.008	90%
	PR4A	0.179 ± 0.001	0.082 ± 0.001	0%	1.087 ± 0.009	100%