EINDHOVEN UNIVERSITY OF TECHNOLOGY

Advances in FairRL: Theory and Applications

Pratik Gajane, Mykola Pechenizkiy, Yingqian Zhang

https://fair-rl.github.io/

IJCAI 2024, Jeju, South Korea
5 August 2024

Goals of the Tutorial

- **Why** are there fairness consideration in RL?
- **What** is fair? **How** is fairness ***defined***? ***measured***?
- **When** should RL-based solutions to be fair?
- **Where** are fairness considerations in RL?
- **How** can we ***achieve*** fairRL?
- **What** is ***SOTA*** in fairRL theory & applications?
- **What's next** in fairRL?

Outline (each part ca 45 mins)

- Part I: Fair Algorithmic Decision Making (ADM)
  - supervised fairML & fairRL perspectives
- Part II: Theoretical results in FairRL
  - performance bounds (bandits, MDPs, MOMDPs)
- Part III: Multi-agent & Multi-objective fairRL
  - from single to multi-object fairML formulations
- Part IV: Future of fairRL
  - how do we bridge gaps in theory and practice

Interactions

- Feel free to interrupt during the tutorial
- Welcome to use Whova to post questions
- We aim to leave 5+ mins after Parts I-III and 15+ mins after Part IV.
- Coffee break 10:30-11:00

Materials

https://fair-rl.github.io/

- Slides
- Bibliography
- Revised survey on FairML

Part I: Supervised fairML & fairRL perspectives

Mykola Pechenizkiy

IJCAI 2024, Jeju, South Korea
5 August 2024

Why fairRL

Why fairRL rather than supervised fairML; to address:

- Sequential ADM
- Primitive fairness-accuracy trade-off
- Positive feedback loops

Why fairness in RL; to prevent:

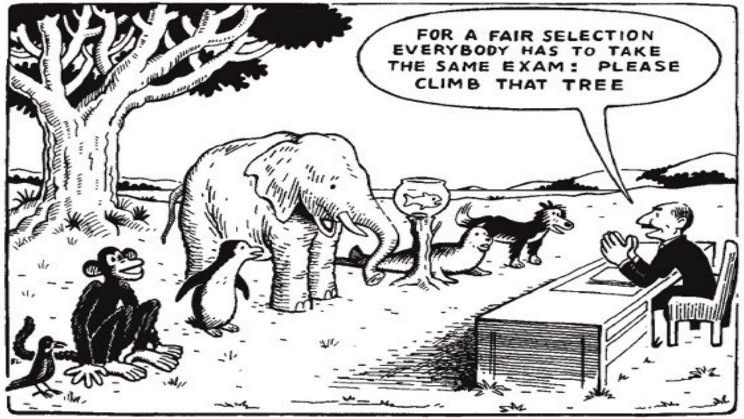
- Discrimination wrt protected attributes (gender, race)
  - unfairness in safety of exploration
  - unfairness in QoS in exploitation
- Propagating existing societal biases (RecSys, Search, SNA)

Part I: Outline

Supervised fairML and fairRL perspectives

- typical notions of fairness
- typical applications
  - societal vs. non-societal fairness
- typical approaches for achieving fairness
  - ML under independency constraints
  - fairness-utility trade-off
  - evaluation and automation

Notions of fairness in fairML



Defining and measuring fairness

– 20+ measures of fairness since FA(cc)T 2018;
– Individual or group level
– Focus on fair *treatment* or fair *impact*
– Achieving parity or satisfying preferences
– Counterfactual fairness
– ….

Fairness notions

fairML

- <u>Group fairness</u>
- Individual fairness
- Calibration fairness
- Counterfactual fairness
- …

fairRL

- (long-term) Group Fairness, Individual,
- Counterfactual
- Envy-freeness
- Effort-based fairness
- Nash Social / Max-min / Generalized Gini Welfare

Generic, application agnostic notions vs. Contextualized to an application

Use cases

societal

- Credit scoring
- Hiring, admission
- Criminal justice
- Fraud detection
- Predictive policing
- RecSys / matchmaking

non-societal

Fair Resource Allocation

- Enhancing TCP over WMN
- Virtualized O-RAN Platforms
- Cloud computing
- Use of road networks in autonomous driving

Human-robot interaction / collaboration, e.g. for managing warehouse, autonomous driving,

Group level fairness

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

- Independency constraints expressed as a group fairness measure

Males		Predicted Label	
		Negative	**Positive**
Actual Label	**Negative**	TN	FP
	Positive	FN	TP

Females		Predicted Label	
		Negative	**Positive**
Actual Label	**Negative**	TN	FP
	Positive	FN	TP

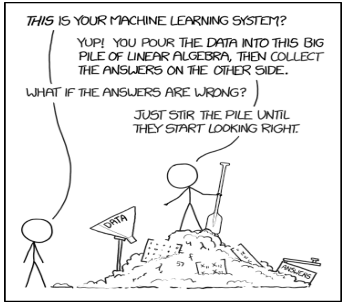
Non-uniform accuracy

$\text{Error}_{\text{males}} << \text{Error}_{\text{females}}$

Favoritism in making decisions:

P(+ | male) – P(+ | female)

- How can we stir the pile?
- What is wrong with the training data?



What harms are we preventing?

Majority of fairness research focuses on these two harms

Allocation

The system extends or witholds opportunities, resources, or information.

Quality-of-Service

The system does not work equally well for all groups.

Representation

The development/usage of the system overrepresents or underrepresents/erases certain groups.

Stereotyping

The system reinforces stereotypes.

Denigration

The system is actively derogatory or offensive.

Procedural

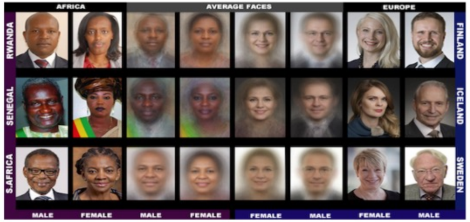
The system makes decisions in a way that violates social norms.

Most prevalent in unstructured data

Closely related to interpretable machine learning

#GenderShades: Facial Recognition Is Accurate

Gender Classifier	Overall Accuracy on all Subjects in Pilot Parlaiments Benchmark (2017)
Microsoft	93.7%
FACE++	90.0%
IBM	87.9%



Pilot Parliaments Benchmark

… if You're a White Guy

- 8.1% – 20.6% worse performance on female faces
- 11.8% – 19.2% worse performance on darker faces
- 20.8% – 34.7% worse performance on darker female faces

#GenderShades; http://gendershades.org/

Child welfare fraud scandal

The Dutch Rutte government stepped down after thousands of families were wrongly accused of child welfare fraud and told to pay money back.

https://www.bbc.com/news/world-europe-55674146

Definitions of group fairness

Demographic parity

- *Both communities have equal access to the benefit*

Equal opportunity

- *If you deserve the benefit, your chances of getting the benefit should not depend on your sensitive attribute*

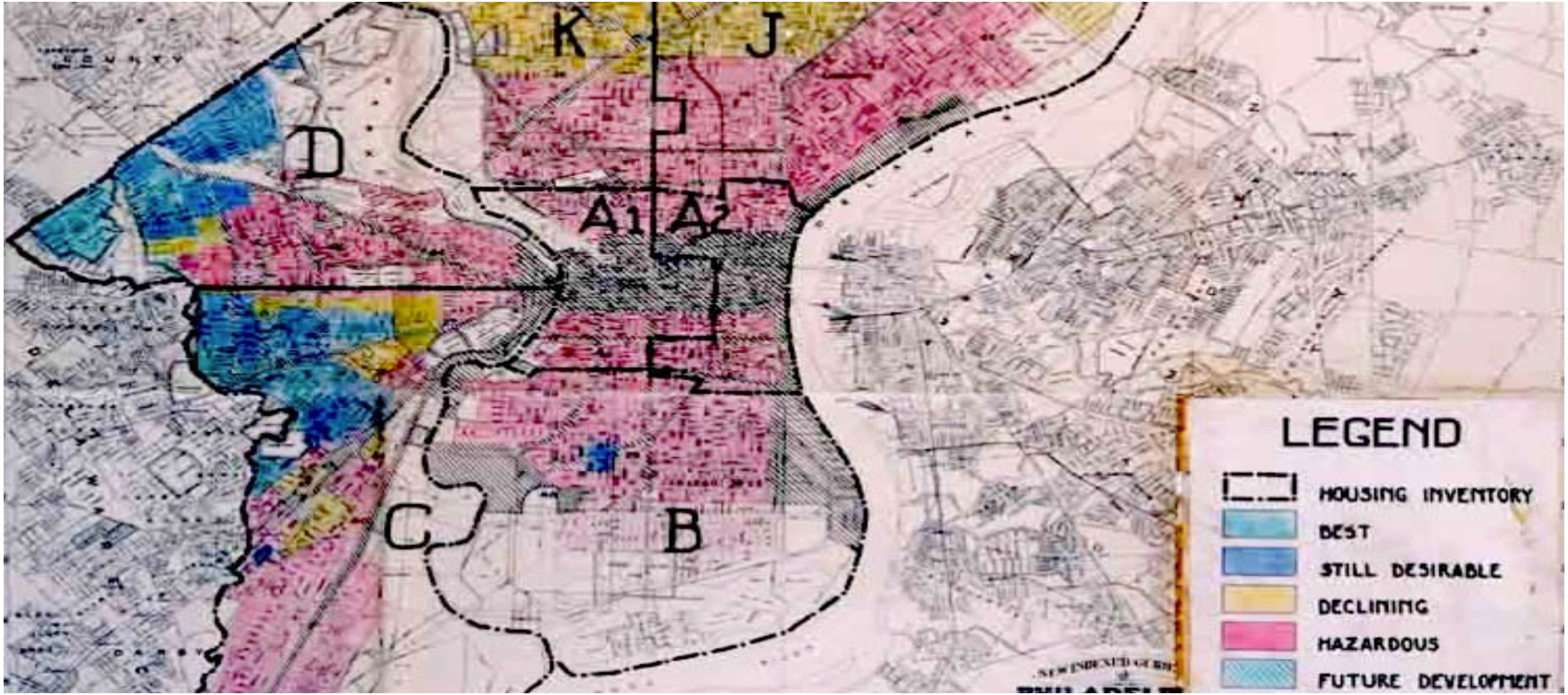
Equal odds

- *If you do not deserve the benefit, your chances of getting it anyway should not depend on your sensitive attribute*

Calibrated for all

- *The <u>meaning</u> of the label you get should not depend on your sensitive attribute*

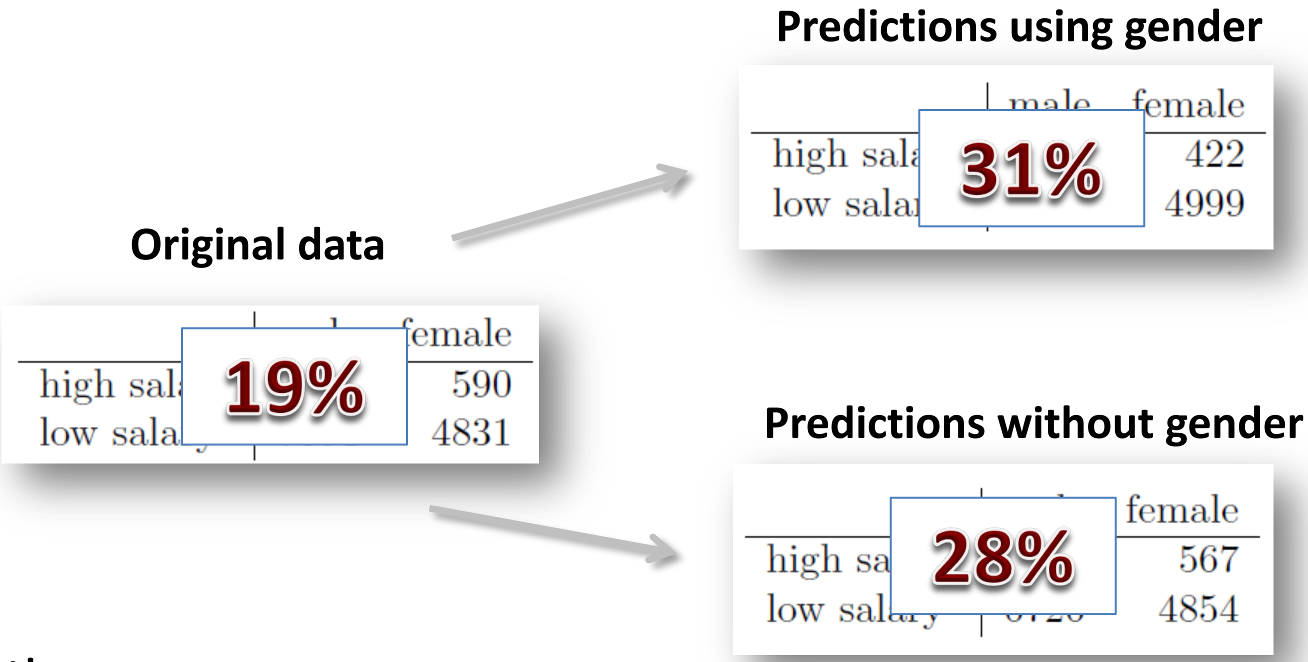
Redlining in Credit Scoring



Source: "Home Owners' Loan Corporation Philadelphia redlining map", Wikipedia
The HOLC maps are part of the records of the FHLBB (RG195) at the National Archives II

Redlining

Example: Census Income Dataset



Discrimination measure:

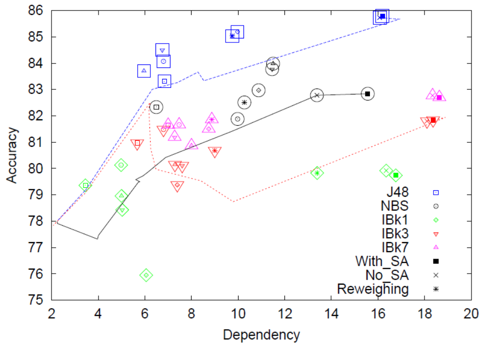
P(‘high salary’ | male) – P(‘high salary’ | female)

Achieving fairness in fairML

ML with independency constraints

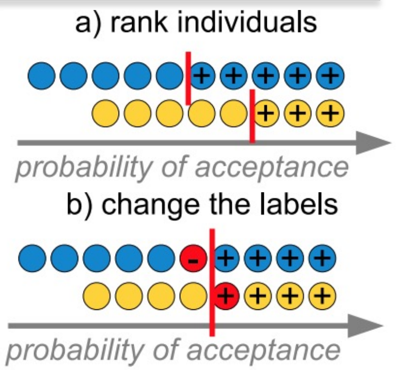
- Removing sensitive attributes A is a bad idea
- Removing also attributes that are correlated with A is also a bad idea: accuracy drops fast if relevant predictive signal is removed
- The challenge of achieving (conditional) independence …

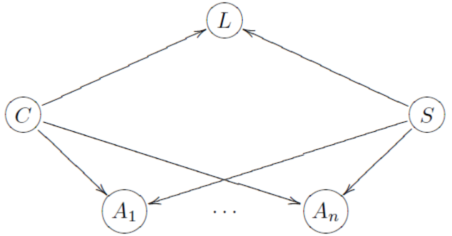
Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

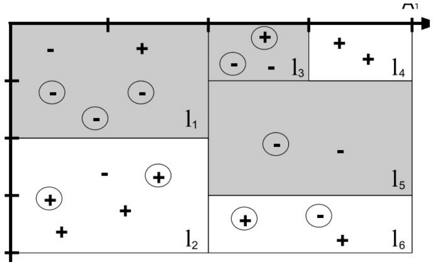


Early approaches for fairML

- ~~Remove sensitive attributes?~~
- Preprocessing – “data massaging”
  – Modify input data (labels)
  – Resample input data
- In-processing / constraint learning
  – Bayesian, decision trees, deep learning
- Post-processing
  – Modify models
  – Modify outputs

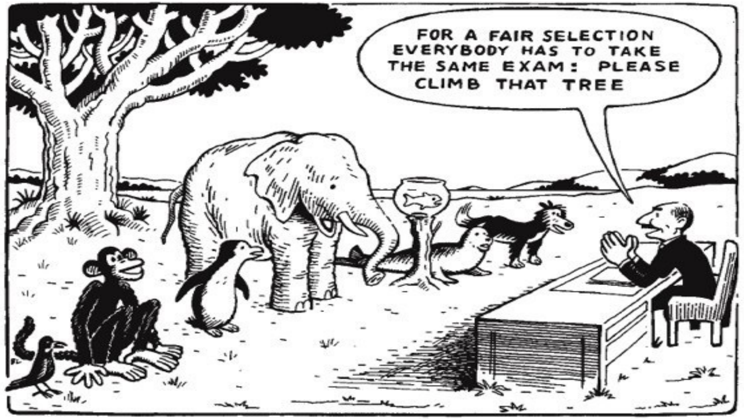






Kamiran, F., Calders, T., & Pechenizkiy, M. (2013). Techniques for discrimination-free predictive models. In *Discrimination and Privacy in the Information Society* (pp. 223-239). Springer, Berlin, Heidelberg.

Many more cost-sensitive learning ideas (apparently often naïve?) for fair classification, regression and other ML tasks as constraint learning



Variants of framing

- Consider an explicit trade-off: is the utility gain proportional to worsening of fairness?
- 0-unfairness: satisfy the independency constraint as much as possible and find solution with max utility that satisfies it
- **ε**-max-utility: do everything possible to minimize unfairness within **ε** from max-utility solution

Is There a Trade-Off?

Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing, Dutta et al. ICML 2020

- *"Our most important result is to theoretically show that for a fair classifier with sub-optimal accuracy on the given biased data distributions, **there always exist ideal distributions such that fairness and accuracy are in accord when accuracy is measured with respect to the ideal distributions**. Through this perspective, there is no trade-off between fairness and accuracy"*

FairML (not?) as Optimization

Cherry on the Cake: Fairness is NOT an Optimization Problem (Favier & Calders 2024) https://arxiv.org/pdf/2406.16606

- *Use cake-cutting theory to describe the behavior of optimal fair decisions, which, counterintuitively, often exhibit quite unfair properties.*
- *Specifically, in order to satisfy fairness constraints, it is sometimes preferable, in the name of optimality,* ***to purposefully make mistakes and deny giving the positive label to deserving individuals in a community in favor of less worthy individuals within the same community****.*
- *"blatantly unfair", cherry-picking, …*

What are some of the roots of unfair ML?

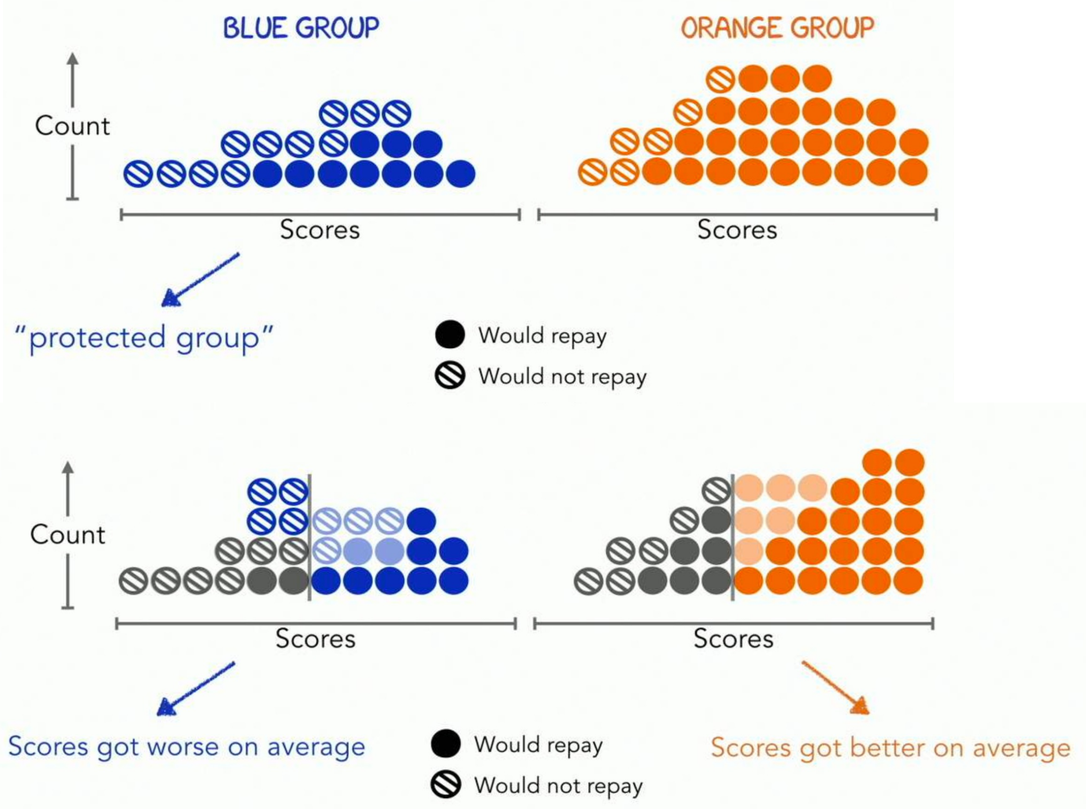
Are some groups underrepresented?

Sex	Ethnicity	Highest Degree	Job Type	Class
m	native	university	board	+
m	native	high school	board	+
m	native	university	education	+
m	non-native	university	healthcare	+
m	non-native	none	healthcare	-
f	non-native	high school	board	-
f	native	university	education	-
f	native	none	healthcare	+
f	non-native	high school	education	-
f	native	university	board	+

Are historical labels biased?

Note: bias in – bias out is absolutely not the only reason why models become unfair

Impact of decisions on population

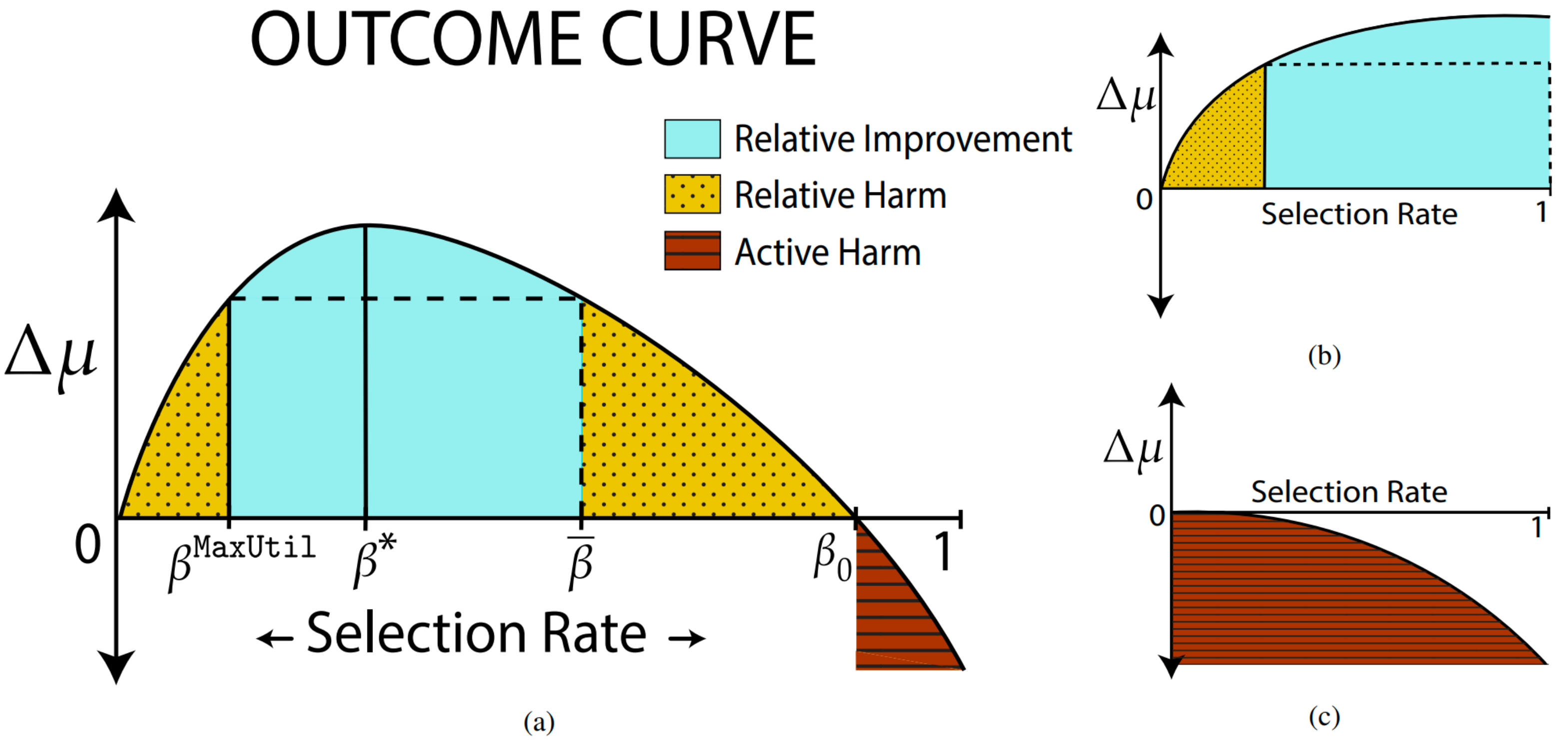


Approving loans while aiming at DP => redistribution of scores over time:

- repayments ⤴️
- defaults ⤵️

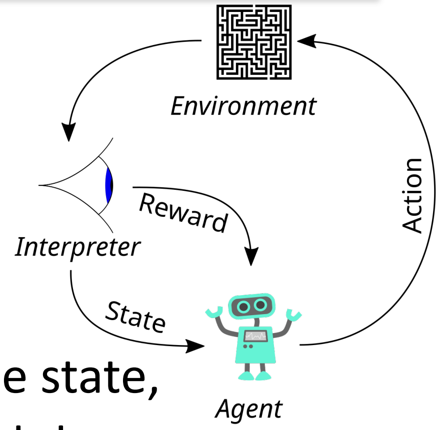
Liu et al. Delayed impact of fair machine learning. ICML 2018
https://www.microsoft.com/en-us/research/video/delayed-impact-of-fair-machine-learning/

Delayed impact of fairML



Liu et al. Delayed impact of fair machine learning. ICML 2018
https://www.microsoft.com/en-us/research/video/delayed-impact-of-fair-machine-learning/

Recap on conceptualising RL

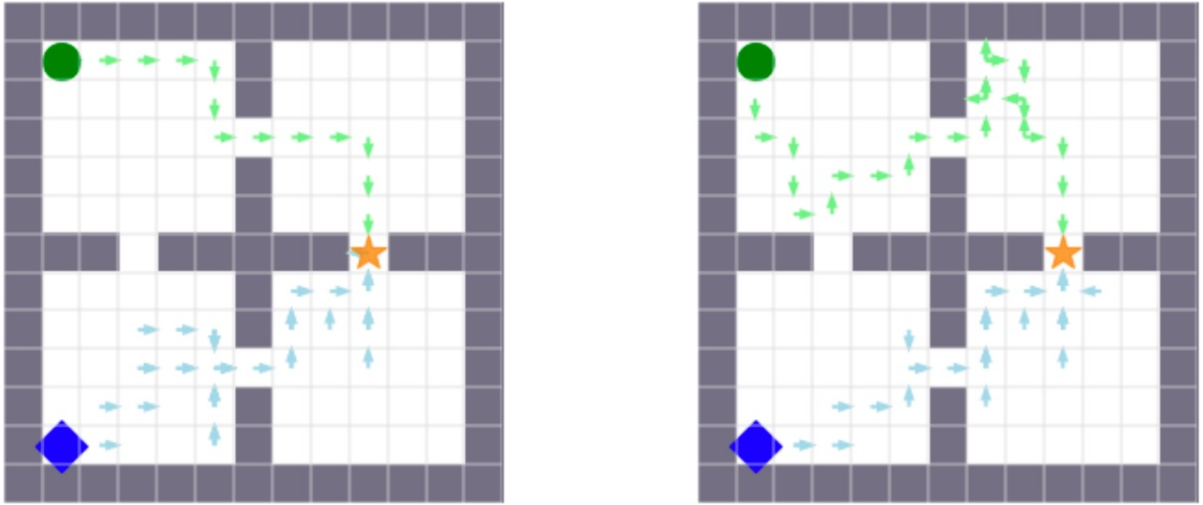


- Actions *A* an agent can take,
- States *S* in the environment the agent is in
  - (Contextual) Bandits ~ RL formulation with only a single state,
  - Markov Decision Processes (MDPs) allow for multiple states
- Policy π, guiding the agent's behavior:
  - Maximizing the total reward *r* over time, i.e. *T* interactions
  - The rewards can be immediate or delayed
  - RL agent can be in single-objective vs. multi-objective setting
  - RL agent can be model-based vs. model-free

Where fairness considerations arise in RL

- Modeling / conceptualization + design choices
  - Pre-specified rewards, but also unknown
  - Exploration safety
  - Temporal dynamics of fairness
  - …

Traditional vs. fair optimal policies



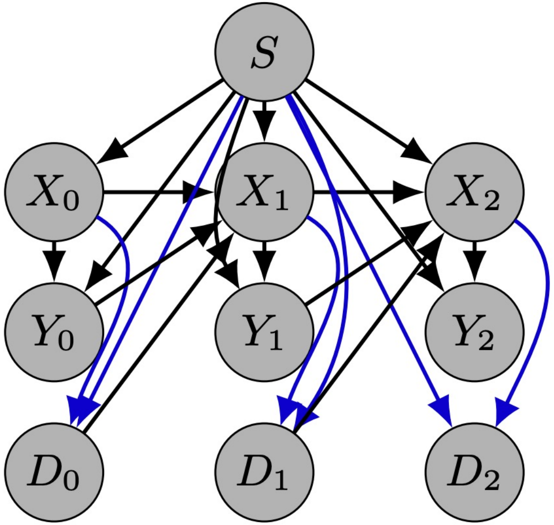
(a) Traditional (b) Fair

Harsh Satija et al. Group Fairness in Reinforcement Learning, TMLR 2023

When fairness (timeline)

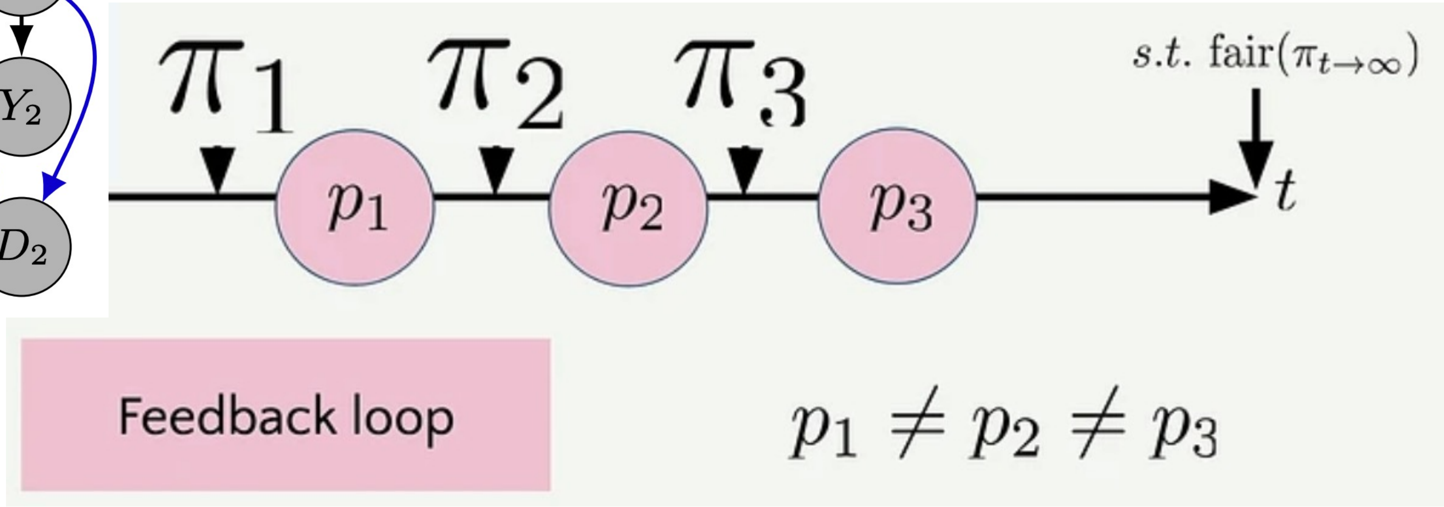
- Past (biased)
- Now and near future
- Some distant future we are stearing towards
- All the time - we want to understand and control the dynamics

Fairness is not static



Policy π blue

Feedback loop from Decisions to Data



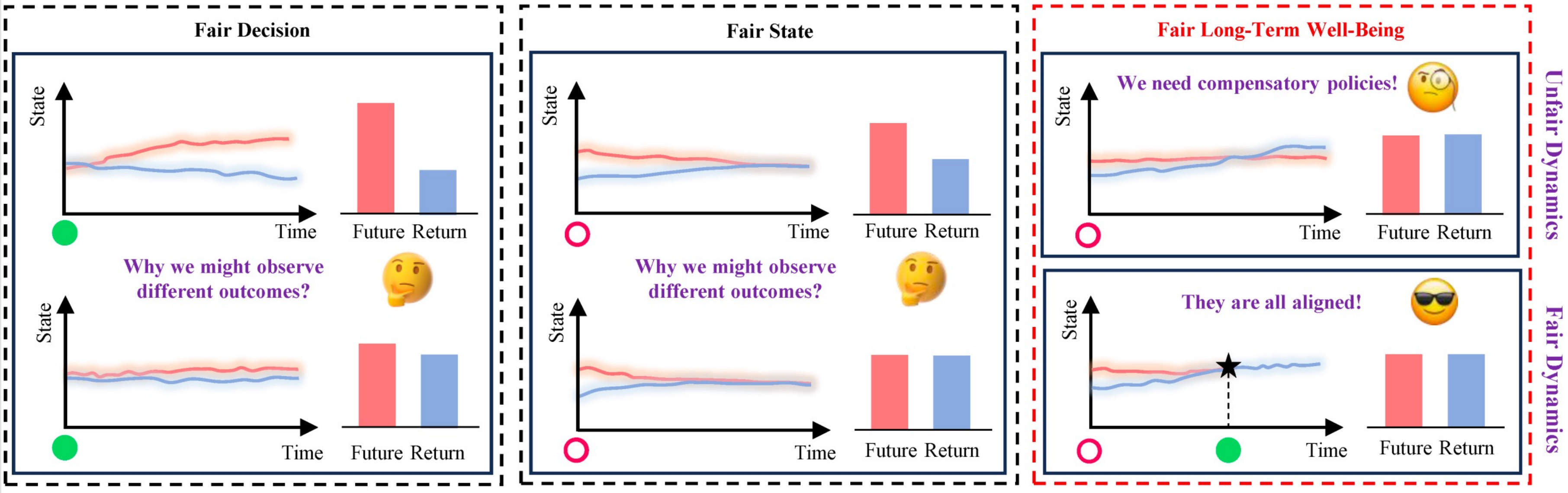
Rateike et al., Designing Long-term Group Fair Policies in Dynamical Systems, FAccT 2024, WS@NeurIPS 2023

Long-term Fair Policies

- Long-term Group Fair Policies in Dynamical Systems, FAccT 2024
- Algorithmic Fairness in Performative Policy Learning: Escaping the Impossibility of Group Fairness, FAccT 2024
- A Reinforcement Learning Framework for Studying Group and Individual Fairness, AAMAS 2024

- Questioning the scope of the fairML impossibility results

Fairness dynamics in RL



Deng et al. What Hides behind Unfairness? Exploring Dynamics Fairness in Reinforcement Learning. IJCAI 2024

Learning and Exploration

- Exploration-exploitation trade-off (70s)
- Took time to rediscover in RecSys and other relevant application areas
- Took time to rediscover in fairML and fairRL
  - Fair Exploration via Axiomatic Bargaining, NeurIPS 2023

Empirical evaluation

fairML

- Benchmarks
- Single time point hold-out estimates
- Datasheets for datasets
- Model cards
- Fairness robustness

fairRL

- Simulated data
- Simulated environments
- Eval. is inherintly over time
- Exploration and exploitation aspects

Fairness robustness

- D-Hacking, FAccT 2024
  - Systematically selecting among numerous models to find the least discriminatory
  - misleading or non-generalizable fairness performance
  - parallels the concept of p-hacking
- Multiverse analysis. FAccT 2024
  - Sensitivity analysis wrt design choices along fairML solution development

Theory in fairML/fairRL

fairML

- Impossibility results
- Fairness is optimization under (independency) constraints
- Fairness is NOT an optimization problem

fairRL

- Incompatibility of fairness & efficiency (social optimality)
- Performance guarantees / bounds
- Worst-case analysis

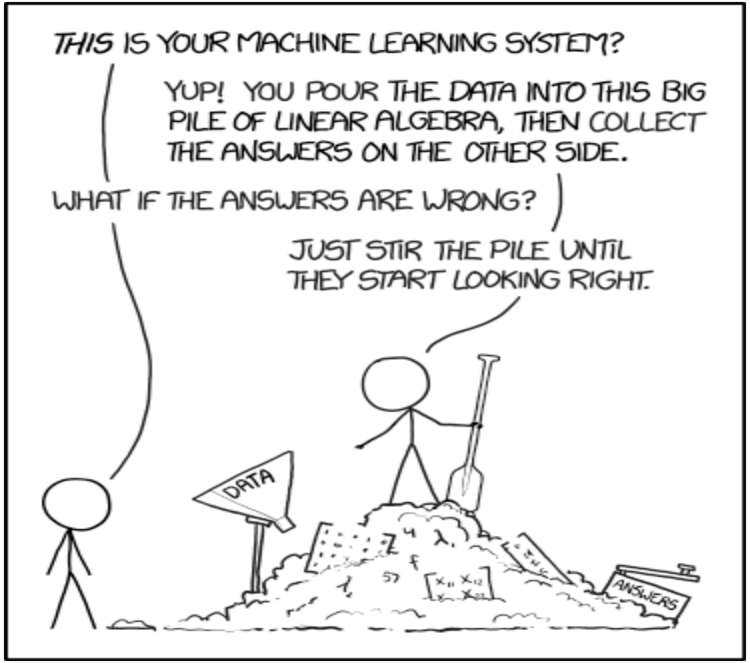
Do we know what is achievable? (e.g. Maximal fairness FAccT 2023)

Possibility of Fairness

Empirical evidence in fairML/fairRL:

- The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice, FAccT 2023
- Algorithmic Fairness in Performative Policy Learning: Escaping the Impossibility of Group Fairness, FAccT 2024
- A Reinforcement Learning Framework For Studying Group And Individual Fairness, AAMAS 2024

ML as optimization

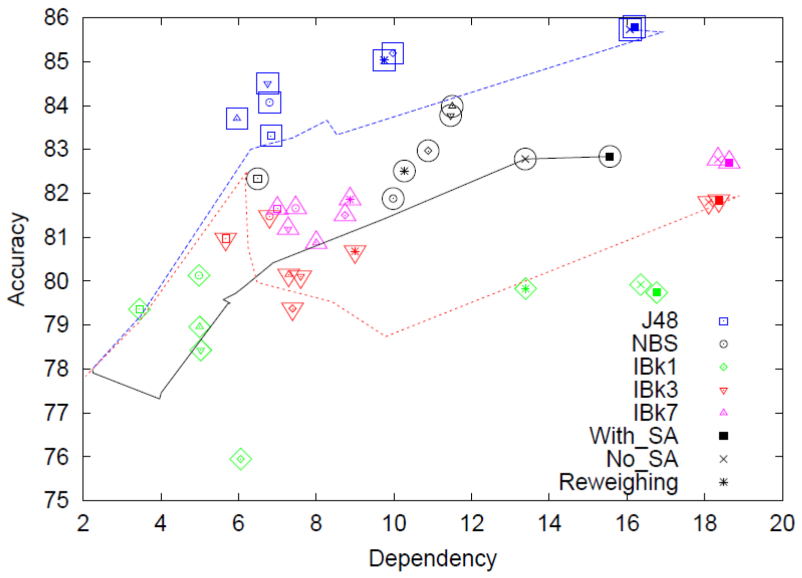


"I want everything I touch to turn to gold"

Do we really know what we are optimizing for?

fairML as Optimization?

Fairness–Accuracy trade-off



But we want to compute expected performance in possible future worlds and steer towards a better world, not towards the past, which we expected to exibit unwanted biases.

F-A trade-off framing might be misleading!

fairML as Optimization

Achieving Fairness revisited

- Fairness – Accuracy Trade-Off
- Moral Justification of fairML
- "It's not (only) about the result, it's about how we reached it.”

- Will get back to this in Part IV