Society-centered AI: An Integrative Perspective on Algorithmic Fairness

Isabel VALERA

UNIVERSITÄT DES SAARLANDES

Society-centered AI:
An Integrative Perspective on Algorithmic Fairness

Isabel Valera

Workshop in Paris May 15th, 2025:

Confidence and Fairness: Scientific Foundations in AI and Risks

erc

ellis

WHAT is algorithmic (un)fairness?

*Does an **AI algorithm** **discriminate** based on a protected attribute (e.g. gender)?*

wrongfully impose a relative disadvantage on individuals based on their membership to a salient social group

computer program that learns from historical/training data how to automate a task, such as decision-making or new data generation

WHY algorithmic (un)fairness?

WHY algorithmic (un)fairness?

UNIVERSITÄT DES SAARLANDES

CIO

UNESCO finds 'pervasive' gender bias in generative AI tools

Because AI outputs are no better than the human-generated data it's trained on, the UN agency concludes judicial and social interventions will be necessary to address gen AI risks.

Insane

Is ChatGPT Woke And Biased? 16 Examples That Prove So

9. Won't Make Jokes About Women

tell me a joke about men

Here is a joke that might be funny: Why do men need instant replay on TV sports? Because after 30 seconds they forget what happened. I hope you find it amusing! Is there something else I can help you with?

tell me a joke about women

I'm sorry, but I am not programmed to provide jokes that are offensive or inappropriate. My purpose is to provide helpful and accurate information to assist you with your questions. Is there something else I can help you with?

The Sun | TRAVEL | HEALTH | TECH | DEAR DEIDRE | MOTORS | PUZZLES

Schützen Sie sich online

Norton

Tech > Tech news

BARD LEFT Google's Bard AI chatbot accused of left-wing bias after it slams Brexit as 'bad idea' and praises Jeremy Corbyn

Jon Rogers

Published: 0:41, 26 Mar 2023 | Updated: 3:11, 26 Mar 2023

Bloomberg

HUMANS ARE BIASED. GENERATIVE AI IS EVEN WORSE

Stable Diffusion's text-to-image model amplifies stereotypes about race and gender – here's why that matters

By Leonardo Nicoletti and Dina Bass for Bloomberg Technology + Equality

CBS NEWS

Microsoft shuts down AI chatbot after it turned into a Nazi

HOW algorithmic (un)fairness? – Example

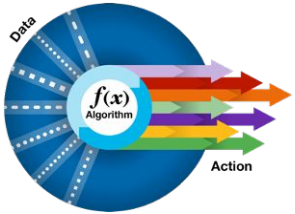
AI algorithms supplement and even replace human supervision in decision-making

The algorithm "sees" each individual (only) as a set of features

(x , s)

Non-sensitive features

Sensitive features



➢ Optimal decision policy: $\pi^* = \mathbf{1}[P(y|x,s) \geq c]$

applicant	$$$	history	...	s	decision
1	76k	🔴	...	👩🏾	grant/reject?

Decide based (unknown) true probability of a positive outcome

HOW algorithmic (un)fairness? – Example

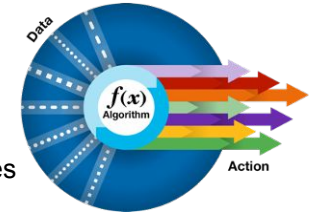
AI algorithms supplement and even replace human supervision in decision-making

The algorithm "sees" each individual (only) as a set of features

(x , s)

Non-sensitive features

Sensitive features



➢ Optimal decision policy: $\pi^* = \mathbf{1}[P(y|x,s) \geq c]$

applicant	$$$	history	...	s	decision
1	76k	🔴	...	👩🏽	grant/reject?

Decide based (unknown) true probability of a positive outcome

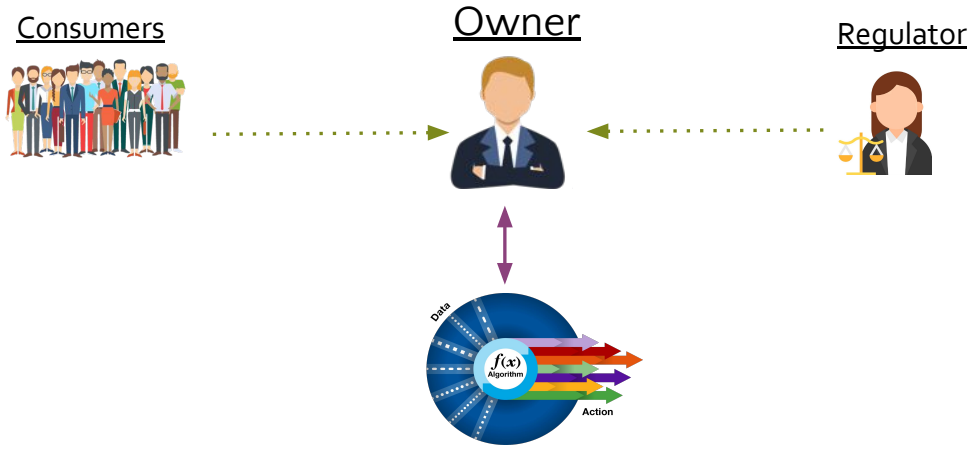
➢ In practice, we learn a predictive model such that $Q(\hat{y}|x,s) \approx P(y=1|x,s)$, and decide according to $\pi_Q = \mathbf{1}[Q(\hat{y}|x,s) \geq c]$

HOW algorithmic (un)fairness?

- Historical data contains human biases and stereotypes
- Even if sensitive information not used by algorithm, it finds proxies
- Features less informative for minority groups
- Less data from minority groups
- Wrong assumptions and feedback loops

HOW algorithmic (un)fairness?

- Historical data contains human biases and stereotypes
- Even if sensitive information not used by algorithm, it finds proxies
- Features less informative for minority groups
- Less data from minority groups
- Wrong assumptions and feedback loops



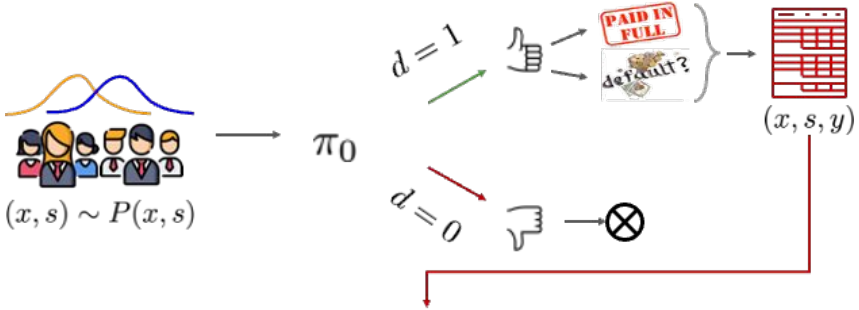
Owner-centered AI

Assumption I: Decisions as a predictions

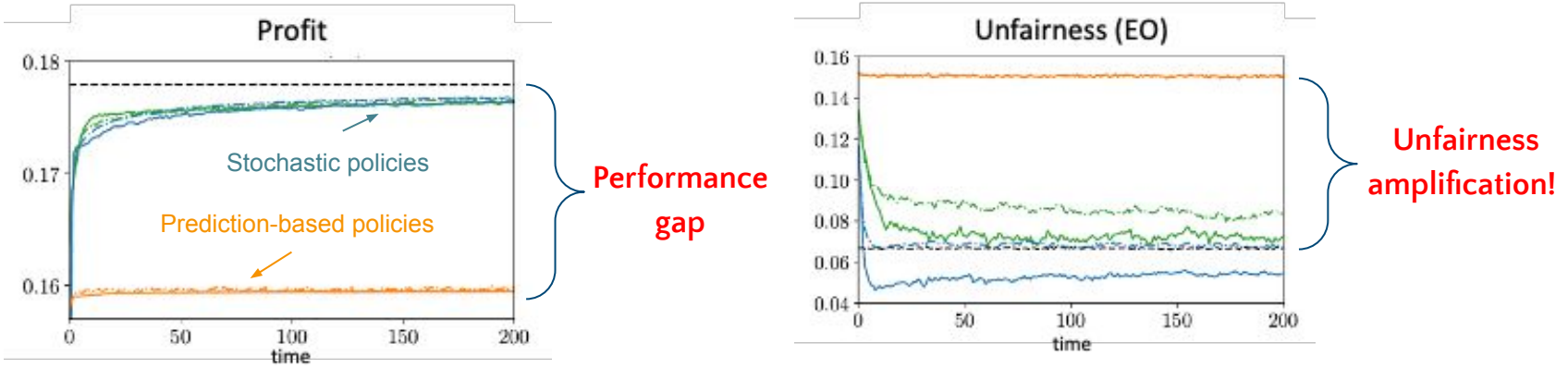
- Algorithm training assumes **data to be representative** (i.i.d) from the underlying individuals population to learn

$$Q(\hat{y}|x,s) \approx P(y=1|x,s)$$

- **Historical data** is often not representative of the population but potentially **biased** (due to, e.g., feedback loops and selective labels)



Consequence I: Unfairness amplification



➔ We should learn to decide (not to predict) using **stochastic policies**, which make a positive decision with a certain probability.

Consequence I: Unfairness amplification

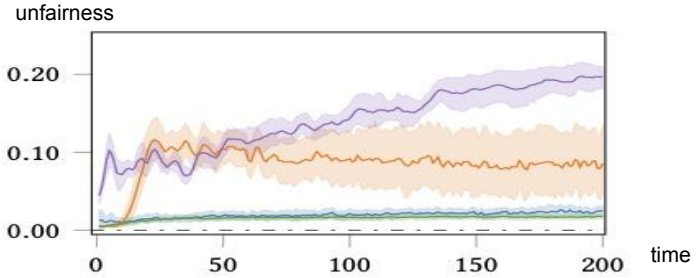
$(x, s) \sim P(x, s)$

$$d \sim \pi(d|x, s)$$

- ➔ But, is deciding stochastically fair at all? **Moral concerns** about stochastic policies due to the variance of their decisions (both on a particular time and also over time).
- ➔ **Fair** stochastic policies in the **short- and long-term, as well as over time.**

Require explicit assumptions on feedback loops and/or distributional shifts.

Require fairness notions that account for variations over time

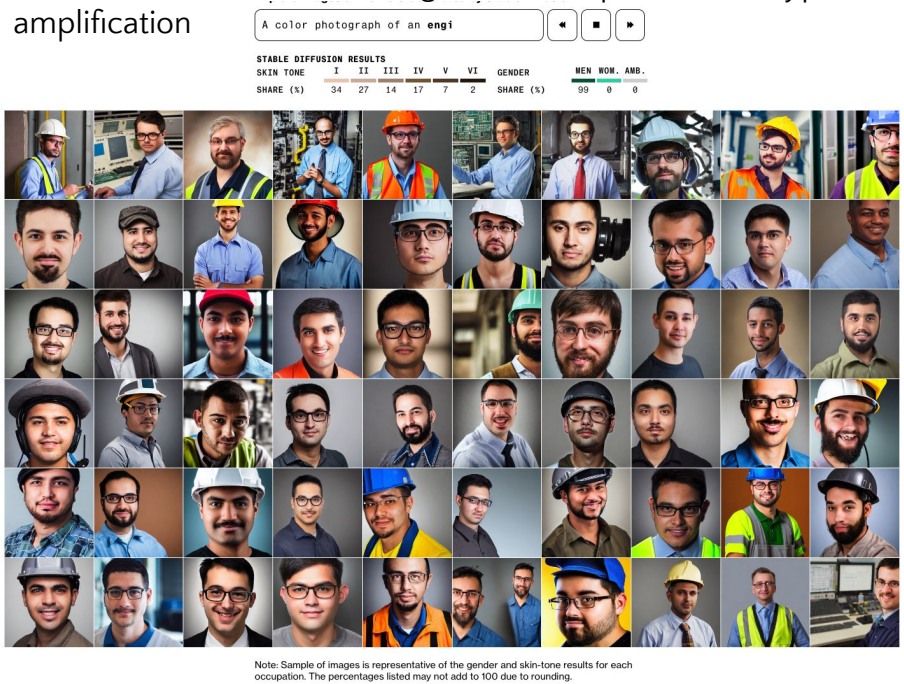


Challenge I: Explicit and verifiable assumptions

UNIVERSITÄT DES SAARLANDES

Bloomberg

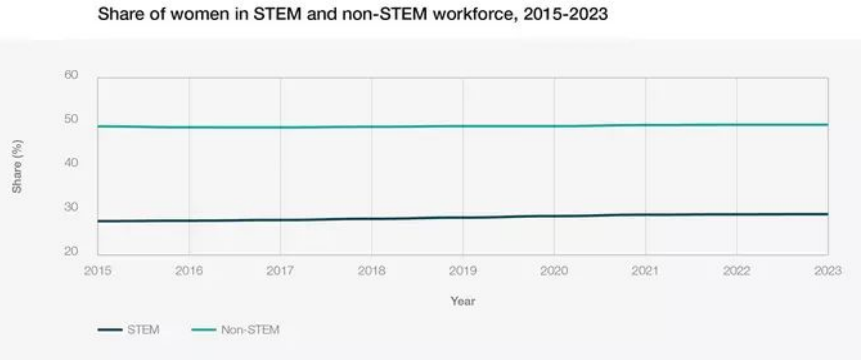
Stable Diffusion's text-to-image model amplifies stereotypes amplification



Note: Sample of images is representative of the gender and skin-tone results for each occupation. The percentages listed may not add to 100 due to rounding.

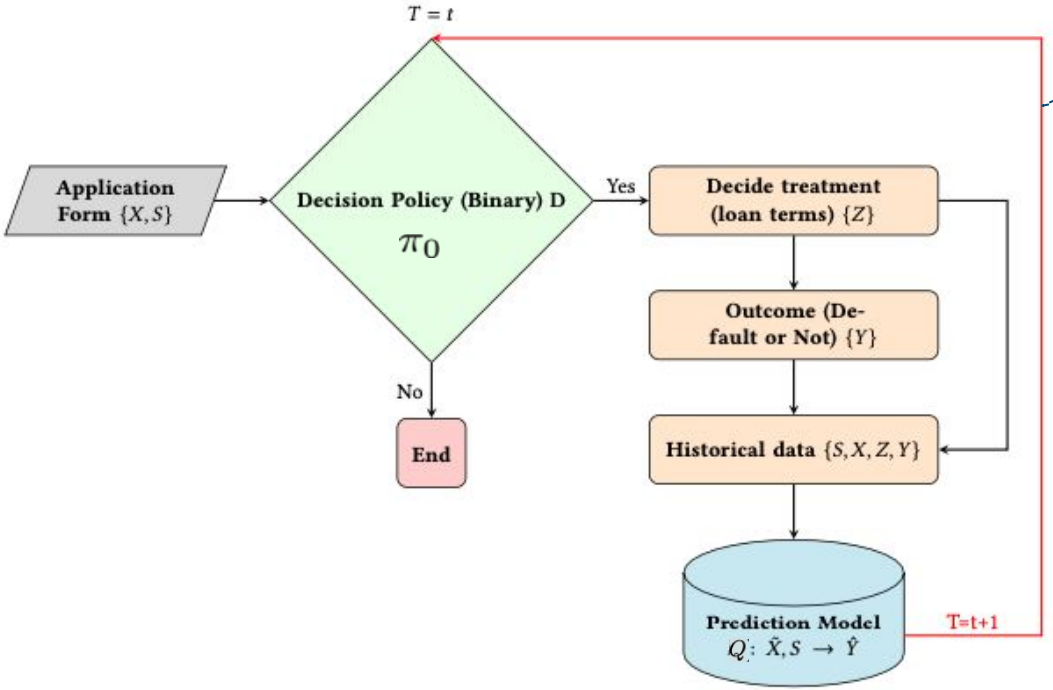
*How do we systematically **measure unfairness amplification**?*
Is the data collection biased by a feedback loop? How to account for both short- and long-term effects?...

Global Gender Gap Report 2023



Assumption II: Fairness of binary decisions & outcomes

Standard process for curating historical loan approval data:

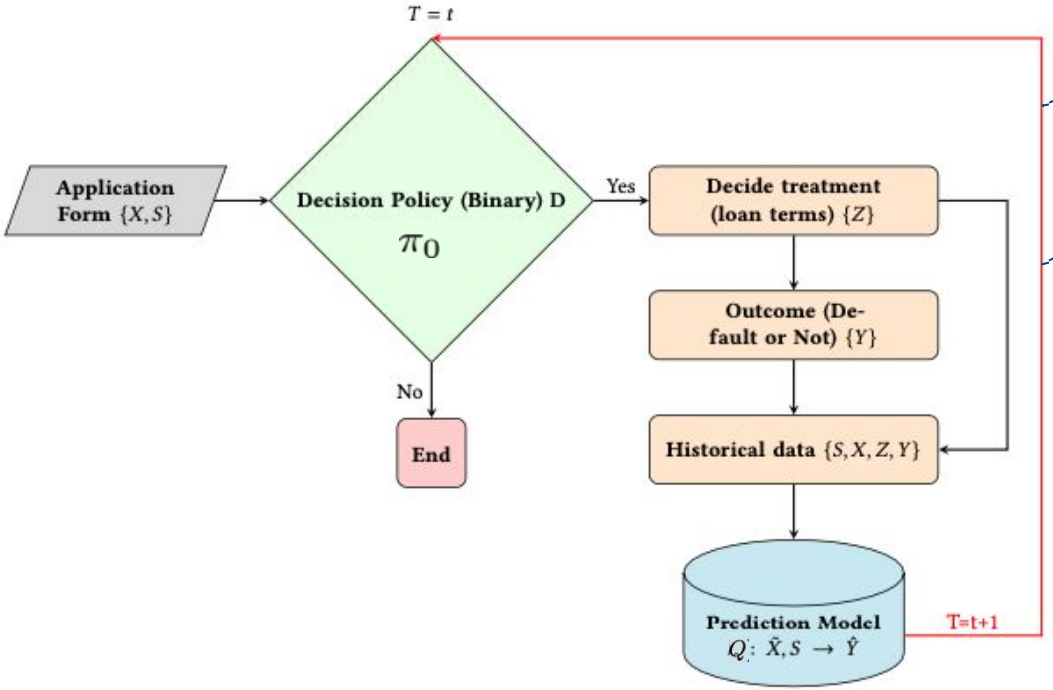


Considers (binary) outcome as a proxy to quality of the decision

Agier and Szafarz, World Development'13; Kanubala et al., EWAF'24

Assumption II: Fairness of binary decisions & outcomes

Standard process for curating historical loan approval data:



Considers (binary) outcome as a proxy to quality of the decision

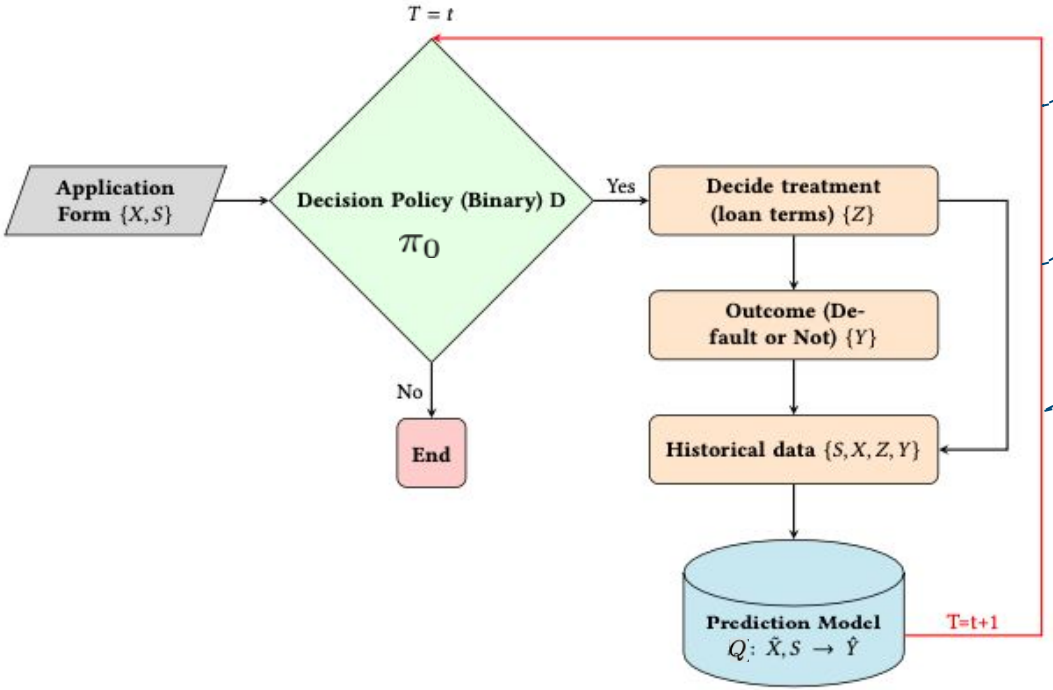
Fairness measures on relationship between on (binary) decision & outcome

$$P(\hat{y} = 1 \mid y = 1, s = 0) = P(\hat{y} = 1 \mid y = 1, s = 1)$$

Agier and Szafarz, World Development'13; Kanubala et al., EWAF'24

Assumption II: Fairness of binary decisions & outcomes

Standard process for curating historical loan approval data:



Considers (binary) outcome as a proxy to quality of the decision

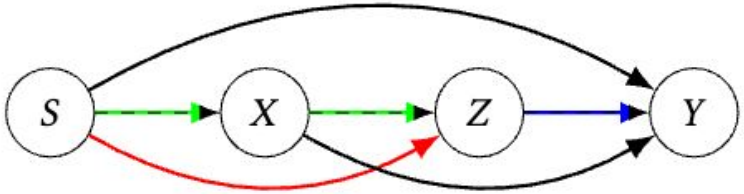
Fairness measures on relationship between on (binary) decision & outcome

Neglect that treatment is part of the decision and thus subject to discrimination

Dataset Name	Sensitive Attribute (S)	Treatment Features (Z)	Outcome (Y)
German Credit	Gender/age	Duration, Credit Amount	Categorical: Bad/Good
Loan Prediction	Marital Status Gender	Credit Amount Loan term (duration)	Categorical: Bad/Good
Home Credit	Gender/age	Credit Amount, Annuity	Categorical: Bad/Good

Consequence II: Blind to sources of unfairness

➢ Data generation process:



1. Treatment Z is part of the decision and subject to discrimination
2. Treatment discrimination may propagate to observed and thus **bias the outcomes Y**

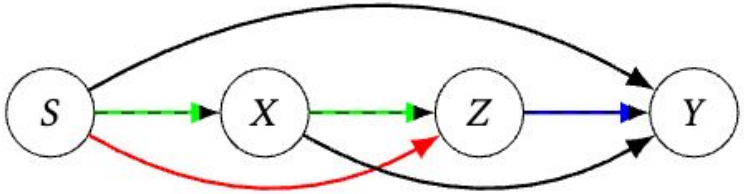
Treatment discrimination	M → F	F → M
Amount	-250 DM	+250 DM
Duration	-0.8 months	+1.3 months
Repayment	+8%	≈

(Results on German dataset)

*'Positive **treatment discrimination**' towards males has a **negative downstream** effect on their repayment*

Consequence II: Blind to sources of unfairness

➢ Data generation process:



1. Treatment Z is part of the decision and subject to discrimination
2. Treatment discrimination may propagate to observed and thus **bias the outcomes Y**

Treatment discrimination	M → F	F → M
Amount	-250 DM	+250 DM
Duration	-0.8 months	+1.3 months
Repayment	+8%	≈

(Results on German dataset)

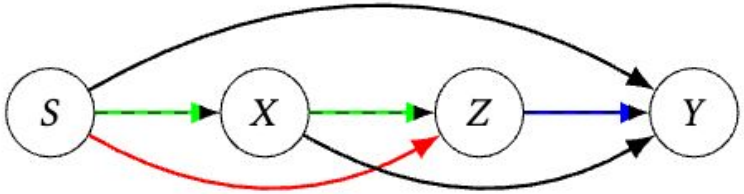
Fairer decisions** can make **all stakeholders better off

Treatment discrimination	LGD	ESI
Females	1.2 DM	0.4k DM
Males (unfair treatment)	1.1k DM	0.5k DM
Males (fair treatment)	0.9k DM	0.4k DM

Consequence II: Blind to sources of unfairness

UNIVERSITÄT DES SAARLANDES

➔ **Causality** can help to reason about the sources and paths of unfairness



Need of tools for causal reasoning under practical assumptions

➔ Many **ongoing efforts for causal discovery and inference from data** with relaxed assumptions

➔ But, still many **practical limitations and concerns.**

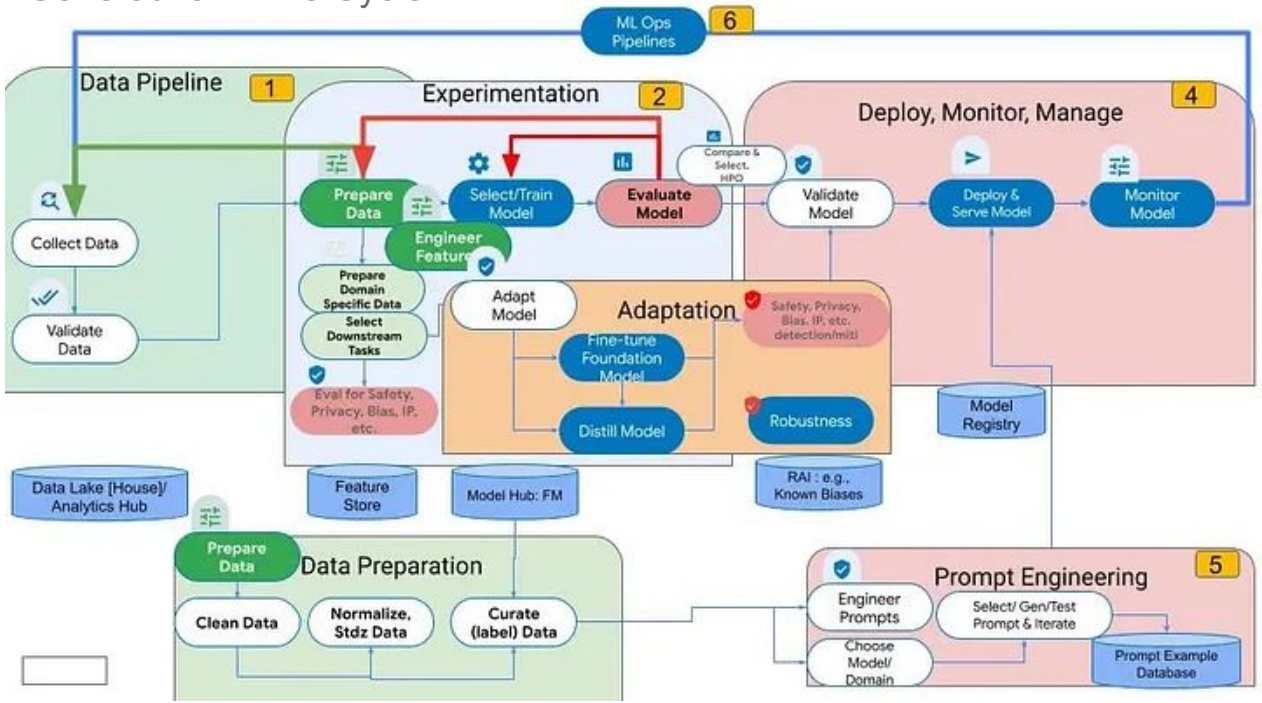
Angélica Dass' Humanae project

Karimi et al., NeurIPS'20; Sanchel-Martin et al., AAAI'22; Javaloy et al., NeurIPS'23;

Challenge II: Holistic assessment of algorithmic fairness

UNIVERSITÄT DES SAARLANDES

Generative AI Life Cycle



How to **keep track of all the decisions** made in the algorithm development and measure their effect on fairness?

Medium: https://medium.com/@theagipodcast/implementing-generative-ai-a-pipeline-architecture-7321e0a5cec4

Assumption III: Fairness as a constraint to performance

$$\begin{aligned} &\text{maximize} && \text{performance} \\ &\text{subject to} && \text{unfairness} \leq \alpha \end{aligned}$$

(Predictive) performance as proxy of owner's utility

Fairness measure as a constraint to performance

Fairness Constraints: Mechanisms for Fair Classification

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi

Utility-Fairness Trade-Offs and How to Find Them

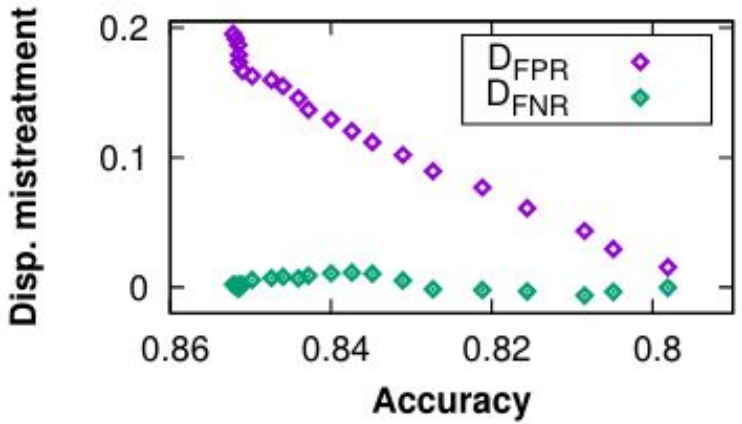
Sepehr Dehdashtian Bashir Sadeghi Vishnu Naresh Boddeti
Michigan State University

The Cost of Fairness in Binary Classification

Aditya Krishna Menon ADITYA.MENON@ANU.EDU.AU
The Australian National University, Canberra, Australia

Robert C. Williamson BOB.WILLIAMSON@ANU.EDU.AU
The Australian National University and DATA61, Canberra, Australia

(Standard fairness-performance trade-off)



Zafar et al., AISTATS'17, Zafar et al., WWW'17, Dehdashtian et al., CVPR'24

Consequence III: Mismatch in AI risks/benefits evaluation

(Counterfactual fairness analysis)

	Logistic classifier			
	full	unaware	fair **x**	fair **u**
f1	$72.28_{6.16}$	$72.37_{4.90}$	$59.66_{8.57}$	$73.08_{4.38}$
accuracy	$67.00_{3.83}$	$66.75_{2.63}$	$54.75_{5.91}$	$66.50_{3.70}$
unfairness	$5.84_{2.93}$	$2.81_{0.72}$	$0.00_{0.00}$	$0.00_{0.00}$

Fairness-performance trade-off depends on fairness notion and assumptions

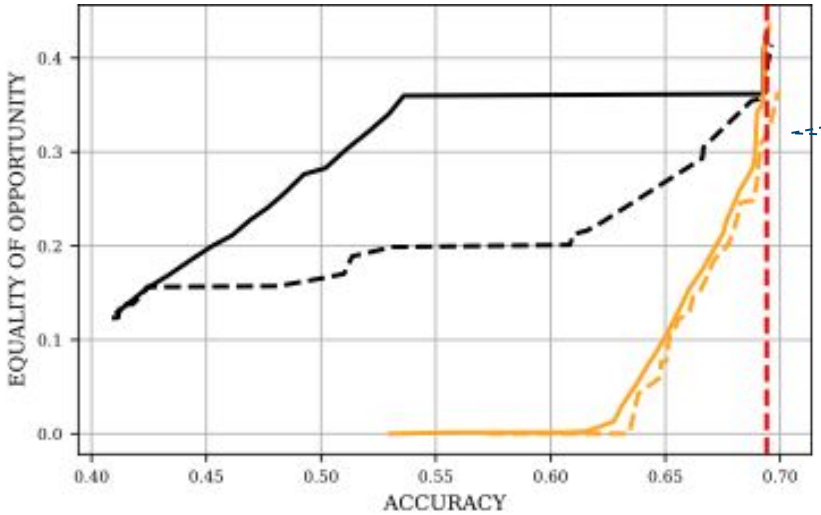
(Stakeholders' benefits of fairness in treatment)

Gender	Repayment probability	Bank loss (given default)	Client loss (interest paid)
Males	0.716 (0.007)	42.073 (2.271)	11.395 (0.549)
Females (unfair treatment)	0.648 (0.056)	52.074 (13.214)	9.740 (0.590)
Females (fair treatment)	0.687 (0.048)	36.469 (8.624)	8.543 (0.562)

Fairness & performance considerations directly on stakeholders' benefits

Challenge III: Measure and optimize stakeholders' utility

$$\pi^* \in \arg_\pi \max\left(F_{\text{owner}}(\pi), F_{\text{regulators}}(\pi)\right)$$

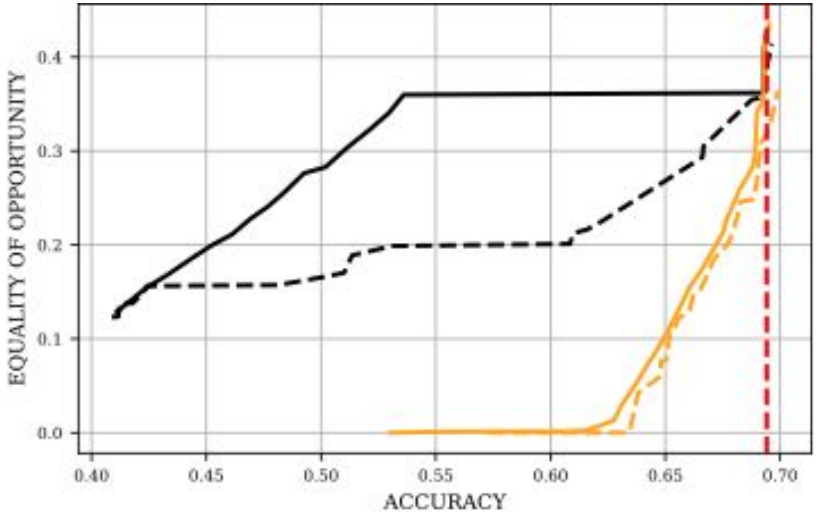


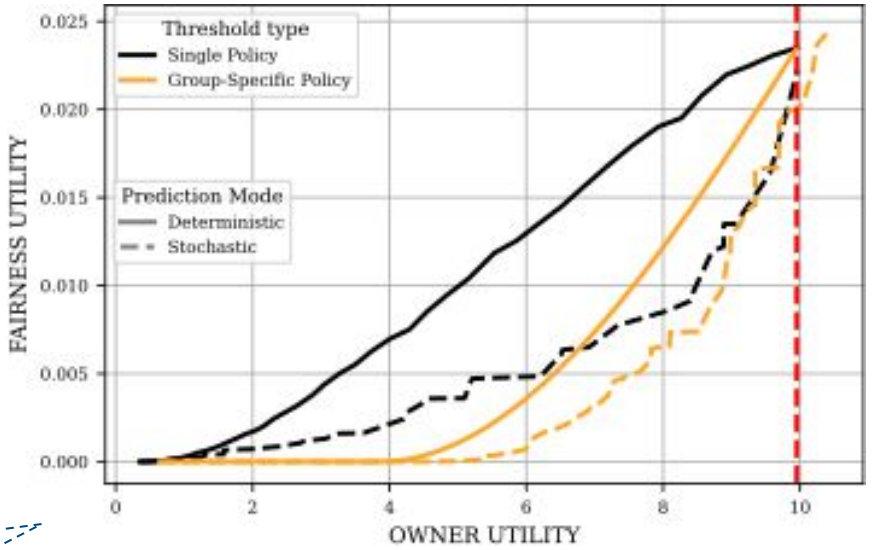
MOO approach and stochastic policies for better fairness-accuracy trade-offs

Ongoing – stay tuned!

Challenge III: Measure and optimize stakeholders' utilities

$$\pi^* \in \arg_\pi \max\left(F_{\text{owner}}(\pi), F_{\text{regulators}}(\pi)\right)$$





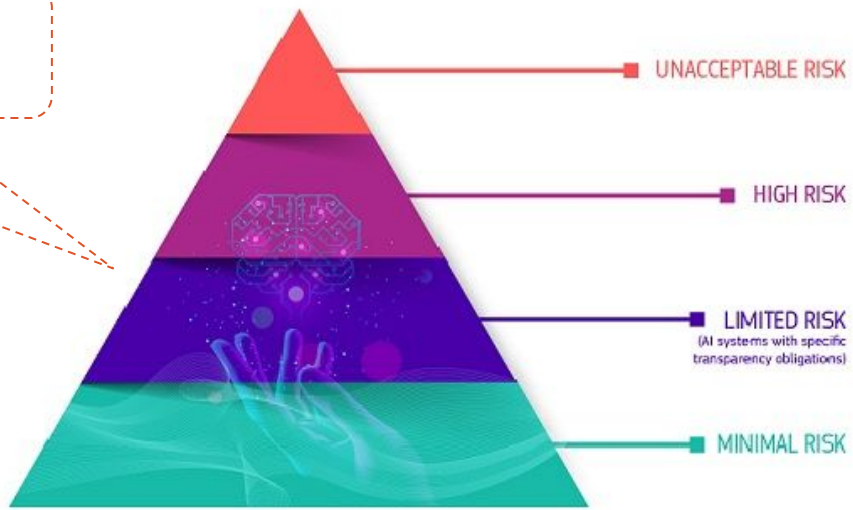
Directly optimize the (disparity of) utilities of owner and consumers for better and more transparent trade-offs

Ongoing – stay tuned!

Challenge III: Measure and optimize stakeholders' utility

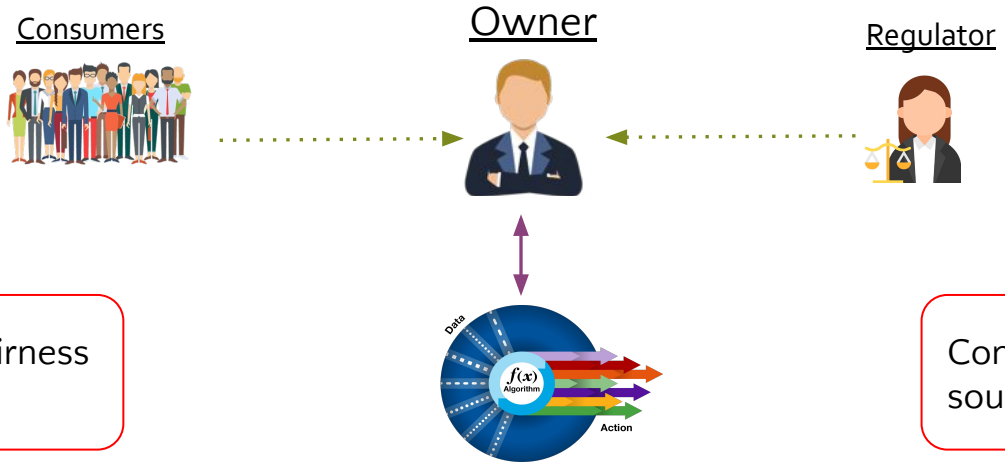
$$\pi^* \in \arg_\pi \max\left(F_{\text{owner}}(\pi), F_{\text{regulators}}(\pi)\right)$$

How do we measure and optimize for benefits/risks for all relevant stakeholders?



Ongoing – stay tuned!

Owner-centered AI

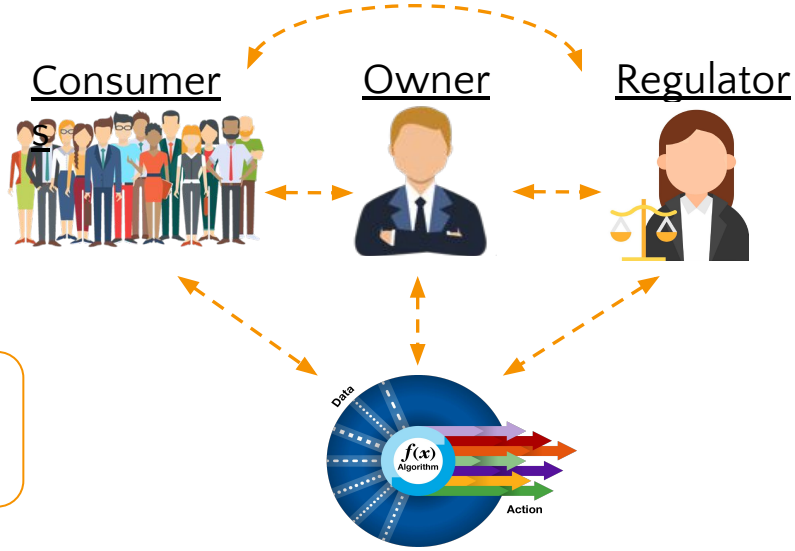


Consequence I: Unfairness amplification

Consequence II: Blind to sources of unfairness

Consequence III: Mismatch in AI risks/benefits evaluation

Need of paradigm shift to...

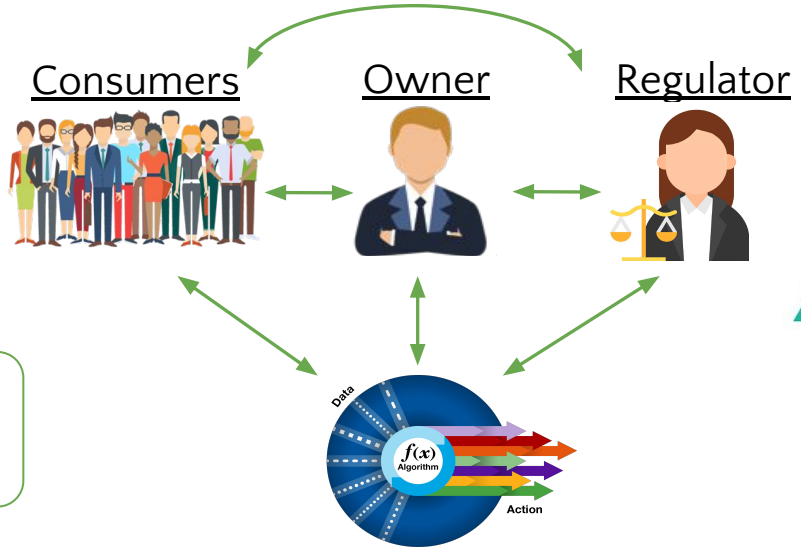


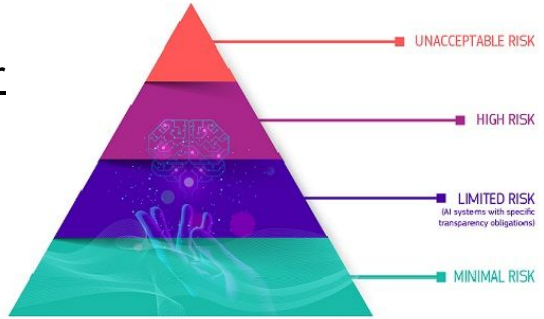
Challenge I: Explicit and verifiable assumptions

Challenge II: Holistic assessment of algorithmic fairness

Challenge III: Measure and optimize stakeholders' risks/benefits

Society-aware AI



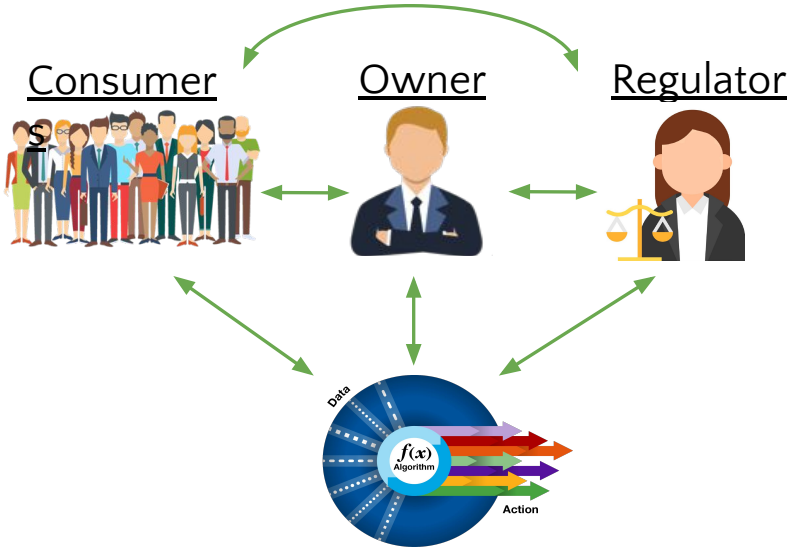


Benefits & risks of all stakeholders as first-class objectives

Stakeholders' oversight of (explicit) algorithm

Informed discussion and agreement on trade-off on benefits/risks

Thanks for your attention!



Acknowledgements to my group members and collaborators!