ERA Workshop: Applying EU Anti-Discrimination Law

DISCRIMINATION AND AI: TECHNICAL AND LEGAL PERSPECTIVES

Prof. Dr. Philipp Hacker, LL.M. (Yale)
Chair for Law and Ethics of the Digital Society
European New School of Digital Studies
European University Viadrina Frankfurt (Oder)

This training session is funded under the 'Rights, Equality and Citizenship Programme 2014-2020' of the European Commission.

1

Outline

Part I: Sources of Discrimination in AI

Part II: Algorithmic Discrimination under EU Law

Part III: Algorithmic Fairness

Part IV: Legal Constraints for Algorithmic Fairness

2

Part I: Sources of Discrimination in AI

3

Job Selection Algorithms

The Guardian

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

Source: https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

4

Medical AI

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Source: Obermeyer et al., 366 Science 447 (2019)

ens

5

Dynamics of Smart Cities

Entry to public buildings

ens

6

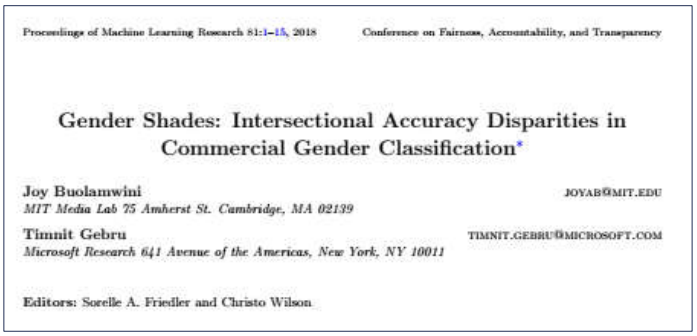
Dynamics of Smart Cities

→ via Face Recognition Technology (FRT)

ens

7

FRT Issues



Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Editors: Sorelle A. Friedler and Christo Wilson

ens

8

Concepts

1) Artificial Intelligence (AI)

- Definitions:

Thinking Humanly	Thinking Rationally
"The exciting new effort to make computers think . . . *machines with minds*, in the full and literal sense." (Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning . . ." (Bellman, 1978)	"The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
Acting Humanly	**Acting Rationally**
"The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990) "The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	"Computational Intelligence is the study of the design of intelligent agents." (Poole *et al.*, 1998) "AI . . . is concerned with intelligent behavior in artifacts." (Nilsson, 1998)

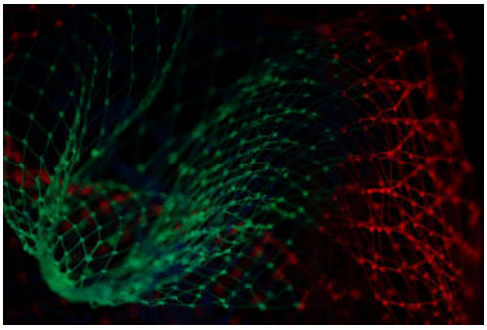
Russel/Norvig, 2016, 2

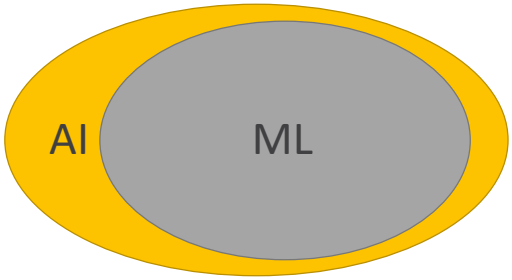
9

Concepts

2) Machine Learning (ML):

- "Learning by example"





Definition: A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

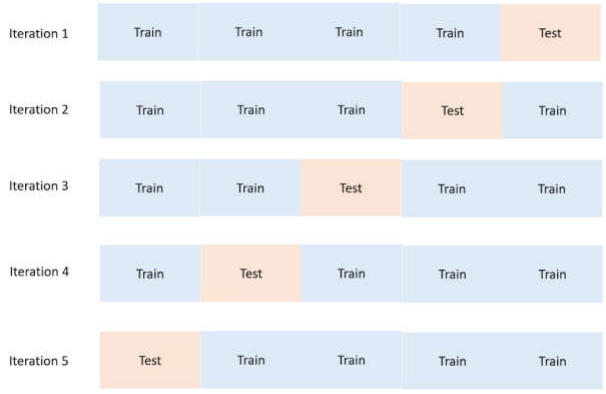
Mitchell, 1997, 2

ens

10

Supervised learning

- Data set: fully labeled (output for each input)
- Training data - test data



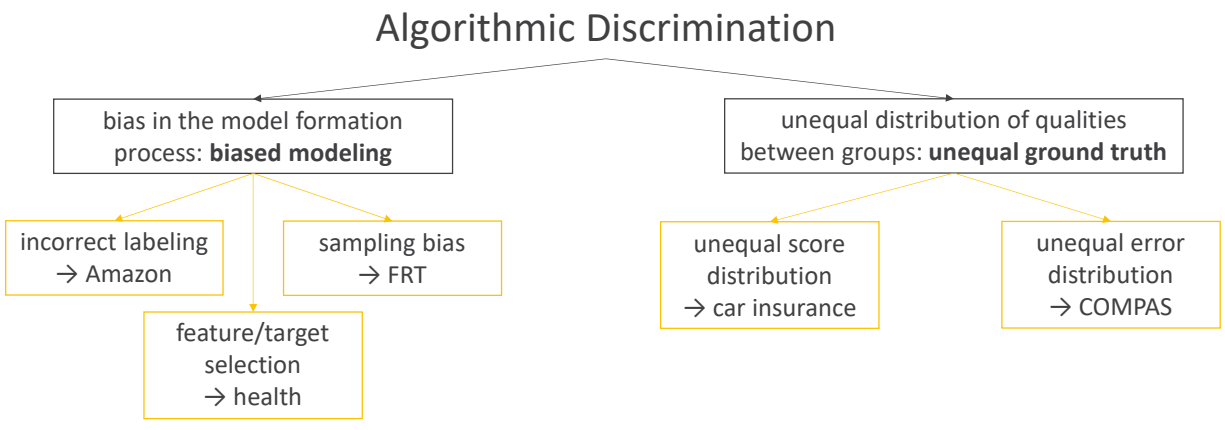
Cross-validation illustrated. The number of splits and iterations can be varied.

Maini/Sabri, 2017, 46

ens

11

The Sources of Algorithmic Discrimination



ens

12

Part II: Algorithmic Discrimination under EU Law

13

The Law of Algorithmic Discrimination

Anti-Discrimination Law (Hacker, 2018; cf. also Zuiderveen Borgesius, 2020; Wachter et al., 2020 & 2021)

- Coverage of algorithmic discrimination
- Indirect discrimination: apparently neutral practice puts protected group at specific disadvantage
- Justification: legitimate aim and discriminatory practice proportionate
  - Biased modeling → (-)
  - Unequal ground truth → (+) if ground truth approximates reality

→ Problem of differentiation between sources of discrimination

→ Enforcement problems: proving differential treatment: no access to data and model (CJEU, C-415/10, *Meister*)

14

The Law of Algorithmic Discrimination

Data Protection Law (Hacker, 2018)

- Data protection principles: *unjustified* discrimination as unfair data processing, Art. 5(1)(a) GDPR (Article 29 Working Party, 2018)
- Art. 22(3) GDPR: automated DM: bias reduction as necessary measures to safeguard rights and freedoms
- Public Enforcement
  - Fines, Art. 83 GDPR
  - Algorithmic audits, Art. 58(1)(b) GDPR

→ Enforcement integration & conceptual convergence

ens

15

Emergent Case Law

ens

16

17

The Motherhood Case - Facts

R (o.t.a The Motherhood Plan) v Her Majesty's Treasury [2021] EWHC 309 (Admin) (17 February 2021)

- 2020: ADM system to determine level of pandemic aid in UK to self-employed business persons
  - No human in the loop
  - 80 % of average trading profits of preceding 3 years
  - Problem: also when partially on maternity leave

Man

Woman | Maternity

18

The Motherhood Case – Discrimination?

Art. 14 ECHR + UK Equality Act (Public Sector Equality Duty)

EWHC: no direct or indirect discrimination:

- "disadvantage […] flows from an absence of or **reduction** in a person's income in the past" (para. 62)

→ "disadvantage is **not caused** by the [ADM system]" (para. 67)

ens

19

The Motherhood Case – Discrimination?

Quite **unconvincing** (cf. Allen/Masters, 2021a)

- **Disadvantage**: yes
  - Reduction of past income is precisely **due to motherhood**
  - Data was factually correct, but sample taken in the wrong way (cf. **sampling bias**)
  - →Necessary disadvantage of women
  - →Problem **inherent** to this ADM system, not external

ens

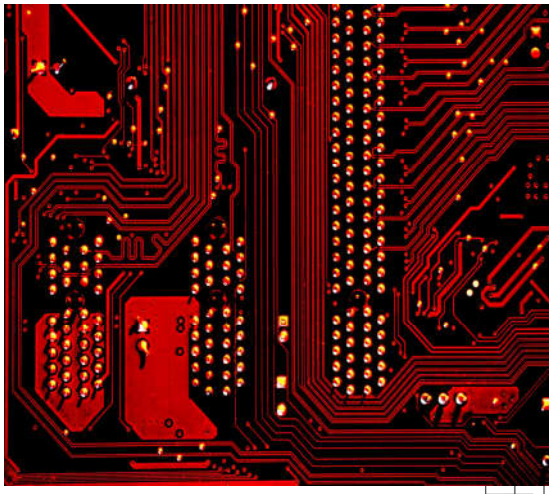
20

The Motherhood Case - Justification

EWHC: ADM **justified**
(not manifestly without reasonable foundation)

- Quicker
- Cheaper
- Simpler
- More fraud-resistant

than human DM (para. 77-85)



ens

21

The Motherhood Case - Justification

Convincing? **Certainly not**

- ADM always quicker, cheaper, "simpler" and more straight-forward
- No room for hard cases

→Cannot be the right measure

Rather:

- Case of **biased modeling** (counting maternity leave)

→ Generally **not justified**

ens

22

The Motherhood Case - Justification

In *Motherhood*:

- Speed and cost: suitability & necessity of ADM (+)
- But not **fair balance** (cf. Campbell, 2021, p. 1217; Hacker, 2018)
  - Data set incomplete/biased
    → AI operator: **reasonable efforts** to obtain more balanced data
  - Here: simple fix:
    - **Self-identification**: maternity leave
    - **Human review** and manual calculation of past profits
  - Feasibility: implemented for grant eligibility, not for level of payment

→ **Unjustified discrimination**

ens

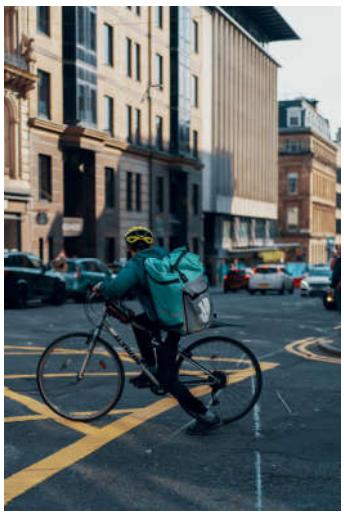
23

24

The Deliveroo Italy Case - Facts

- "Frank": slot allocation system, with advantage for those with high scores
- Calculation: past performance based on
  - Number of booked but **missed slots**
  - Participation in **peak demand** slot Friday evening between 8 and 10 PM

  → Both difficult to reconcile with family and sick children



ens

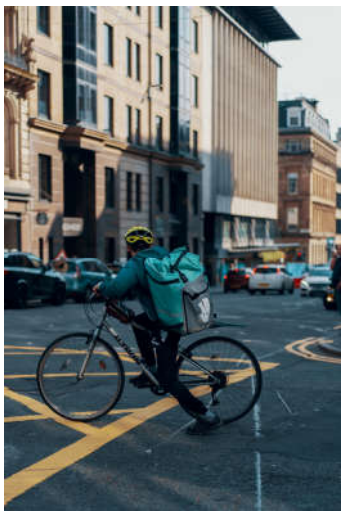
25

The Deliveroo Italy Case - Facts

Bologna Labor Court, Case 2949/2019:

- Indirect **discrimination** (disability, union activity = strike; women)
- **No justification**:
  - No acknowledgement of hard cases = good reasons for unavailability

    ➢ Child sickness, disability, strike



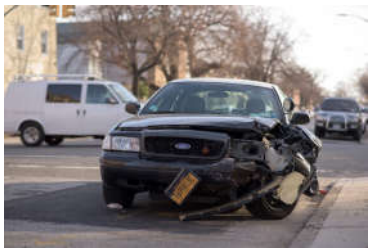
ens

26

Remaining Problems

- Enforcement
  - Knowing of and proving prima-facie discrimination

- Ex-ante prevention instead of ex-post liability

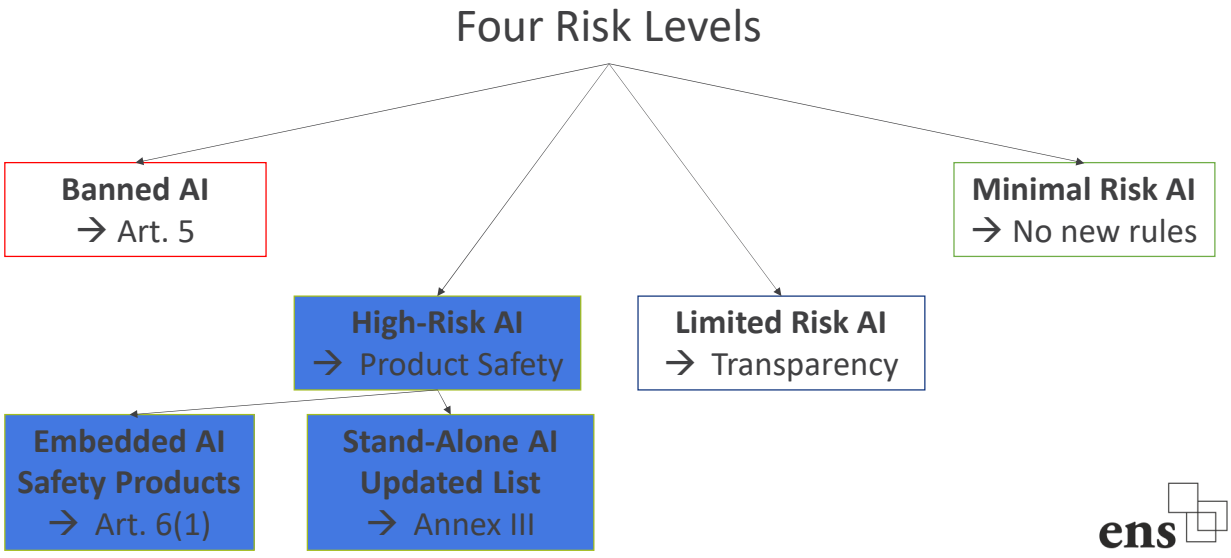


27

Does the new EU AI legislation change anything?

28

Structure of the Proposed AI Act (AIA)



29

High-Risk AI Systems

Include (Annexes II, III AIA):

- FRT
- Employment
- Medical AI
- Credit scoring
- Social benefits
- Judiciary

→ Our cases

Do not include:

- E-Commerce
- Search Engines

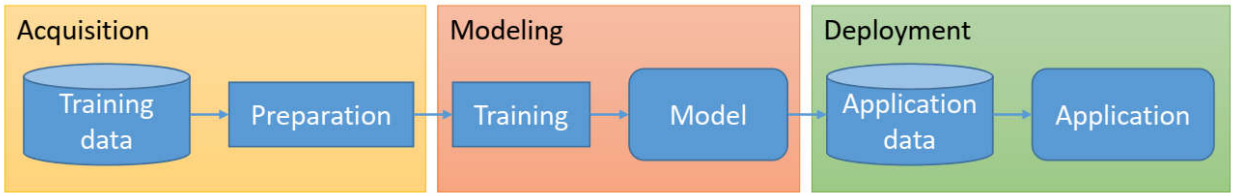
→ Digital Markets Act (DMA)

30

Non-Discrimination in High-Risk AI Systems

ML pipeline:

- Art. 11/13 AIA: **transparency**
- Art. 14 AIA: **human** in the loop



- Art. 10 AIA: training data regime
  - **Correctness**
  - **Representativeness**

Hacker, A Legal Framework for AI Training Data (2021)

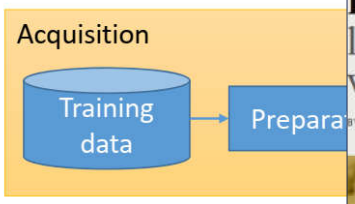
- Art. 15 AIA: performance
  - **Accuracy**
  - Mitigation of **biased feedback**

ens

31

Non-Discrimination in High-Risk AI Systems

ML pipeline:



- Art. 10: training d regime
  - **Correctness**
  - **Representativene**

https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

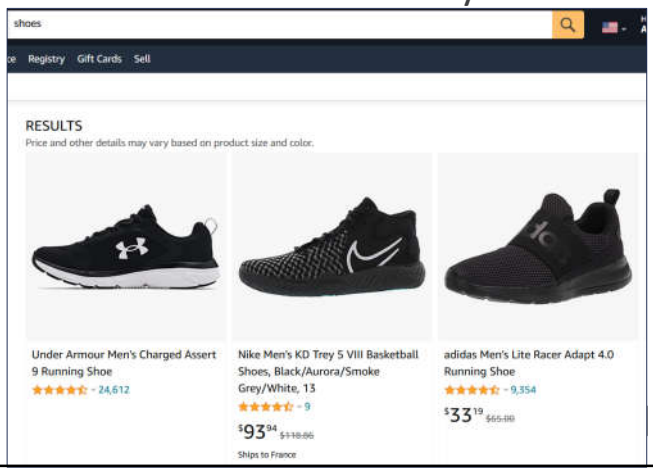
32

Digital Markets Act (DMA)

- For gatekeepers (Google, Amazon, Meta)
- Art. 6(5) DMA: transparent, fair and non-discriminatory rankings

→ Justification necessary

→ND in e-commerce



33

Remaining Problems

- Enforcement
  - Knowing of and proving prima-facie discrimination
- Still problematic
  - Self-certification in AIA
  - Restriction to high-risk AI and gatekeepers

AIA + DMA

- Ex-ante prevention instead of ex-post liability
- Better in AIA
  - But depending on operator goodwill & effective deterrence

ens

34

Part III: Algorithmic Fairness

ens

35

Algorithmic Fairness

Definitions of fairness: 2 main groups (Dwork, 2012; Friedler et al., 2016; Pessach/Shmueli, 2020)

1) Individual Fairness: similar input → similar output
   - Aristotle, Nicomachean Ethics, Book V, § 3, 1131a10
   - CJEU: equality before the law, Art. 20 ChFR
2) Group Fairness: e.g., same positive selection rate for each group (statistical parity)
   - Outcome-egalitarian concept (Binns, 2018)
   - Impossibility of indirect discrimination

→Trade-off necessary: more GF ↔ less IF

ens

36

Bridging the Divide: Our Model

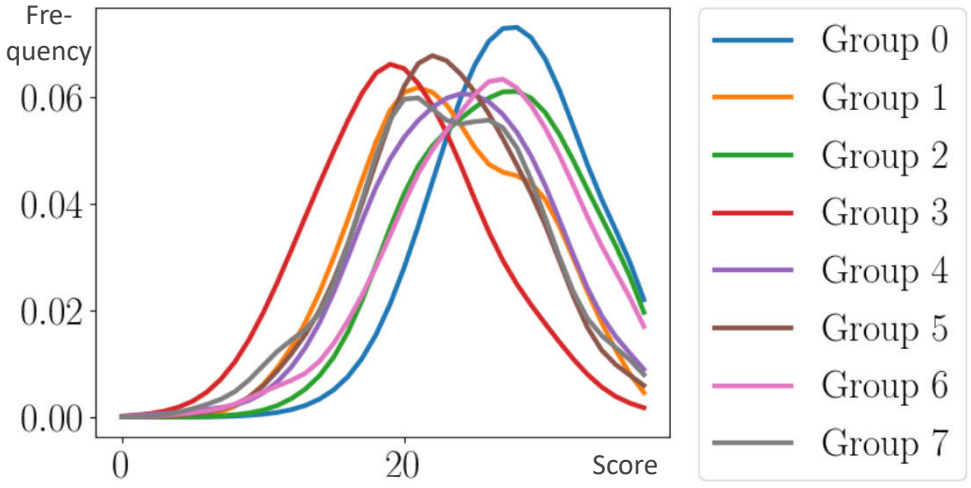
Zehlike/Hacker/Wiedemann, Matching Code and Law, 34 Data Mining and Knowledge Discovery 2020, 163:

- Continuous interpolation between measures of individual and group fairness
- Parameter θ ∈ [0;1]: degree of approximation of group distributions
  - θ = 0 → individual differences are fully preserved (IF)
  - θ = 1 → group distributions fully mapped onto barycenter (GF)
- Minimal information loss for decision maker through optimal transport

ens

37

LSAT Scores (ethnicity): descriptive statistics

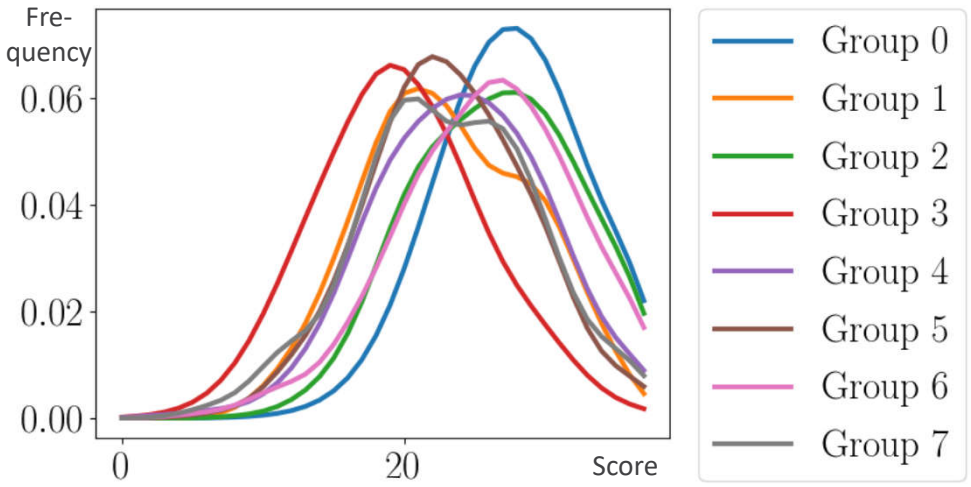


Source: Zehlike/Hacker/Wiedemann, 34 Data Mining and Knowledge Discovery 2020, 163

ens

38

LSAT Scores (ethnicity) for θ = 0

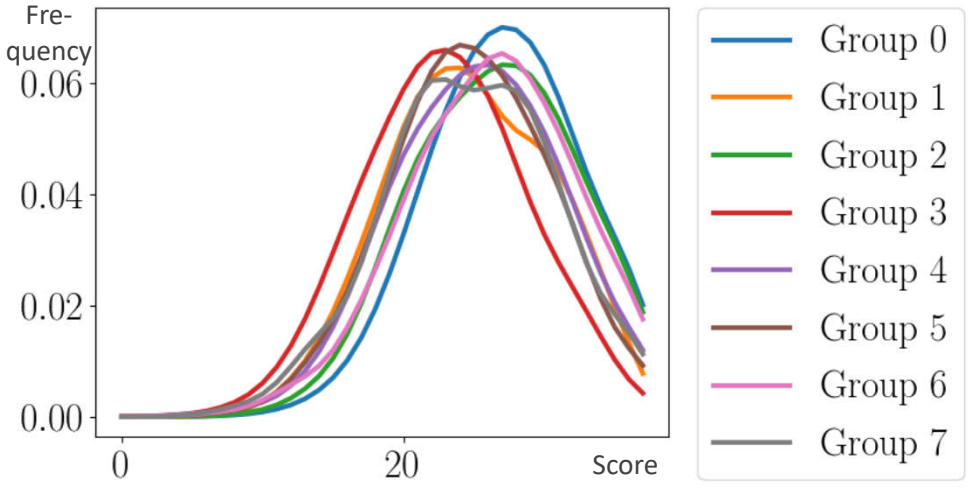


Source: Zehlike/Hacker/Wiedemann, 34 Data Mining and Knowledge Discovery 2020, 163

ens

39

LSAT Scores (ethnicity) for θ = 0.5

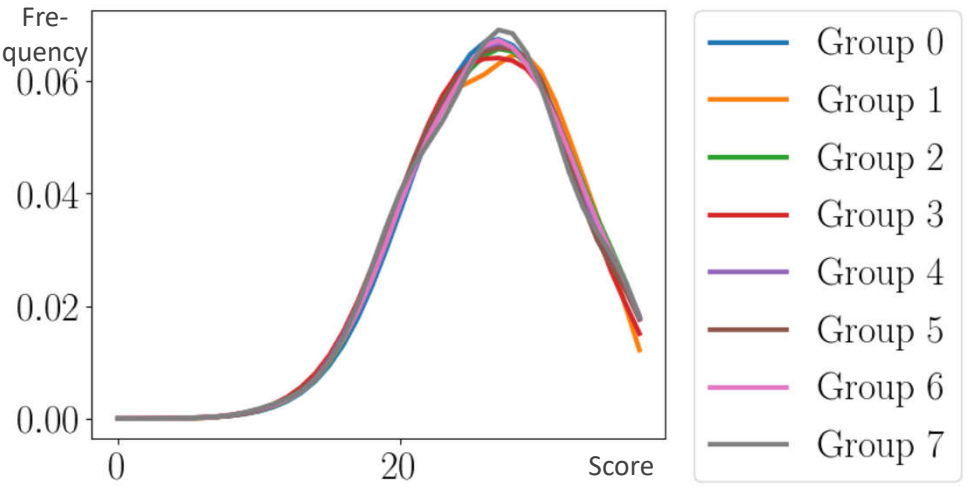


Source: Zehlike/Hacker/Wiedemann, 34 Data Mining and Knowledge Discovery 2020, 163

ens

40

LSAT Scores (ethnicity) for θ = 1



Source: Zehlike/Hacker/Wiedemann, 34 Data Mining and Knowledge Discovery 2020, 163

ens

41

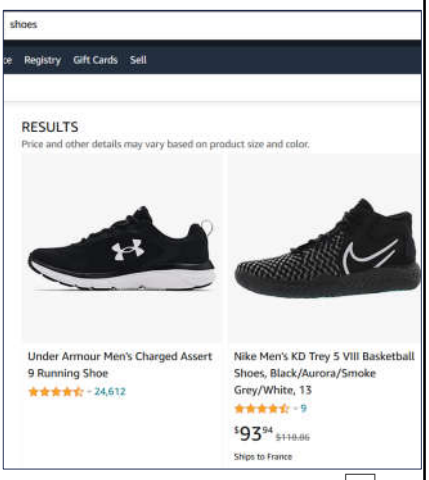
Current Extension

Together with Zalando (German Amazon):

- Trade-off between 3 fairness measures:
  - Calibration
  - Balance for negative class (false positives)
  - Balance for positive class (false negatives)

→COMPAS

→E-commerce products (Art. 6(5) DMA: fair ranking provision, see Hacker, KI und DMA [= AI and DMA], GRUR 2022, 1278)



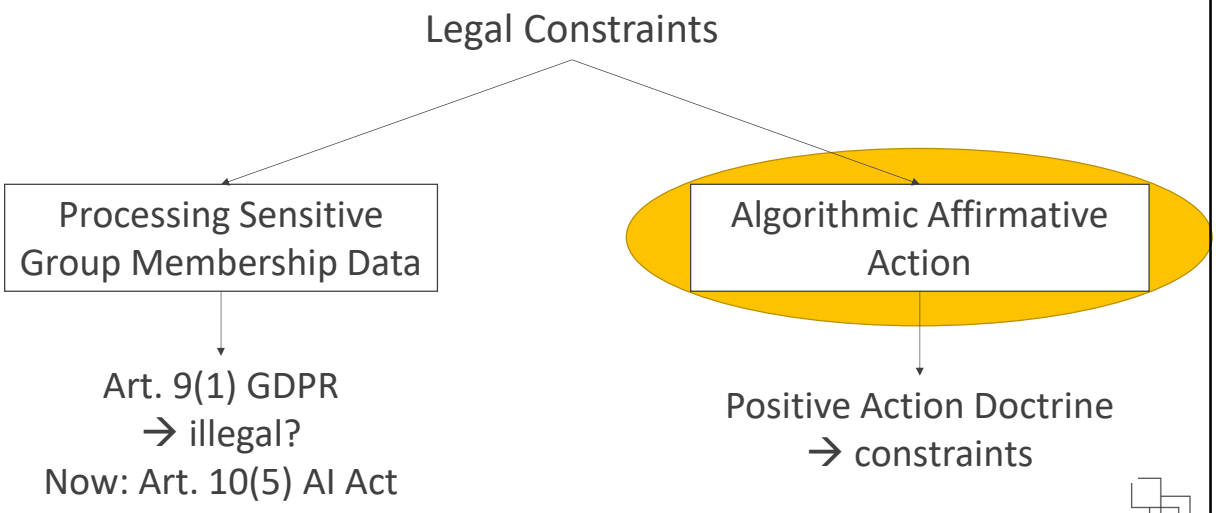
ens

42

Part IV: Legal Constraints for Algorithmic Fairness

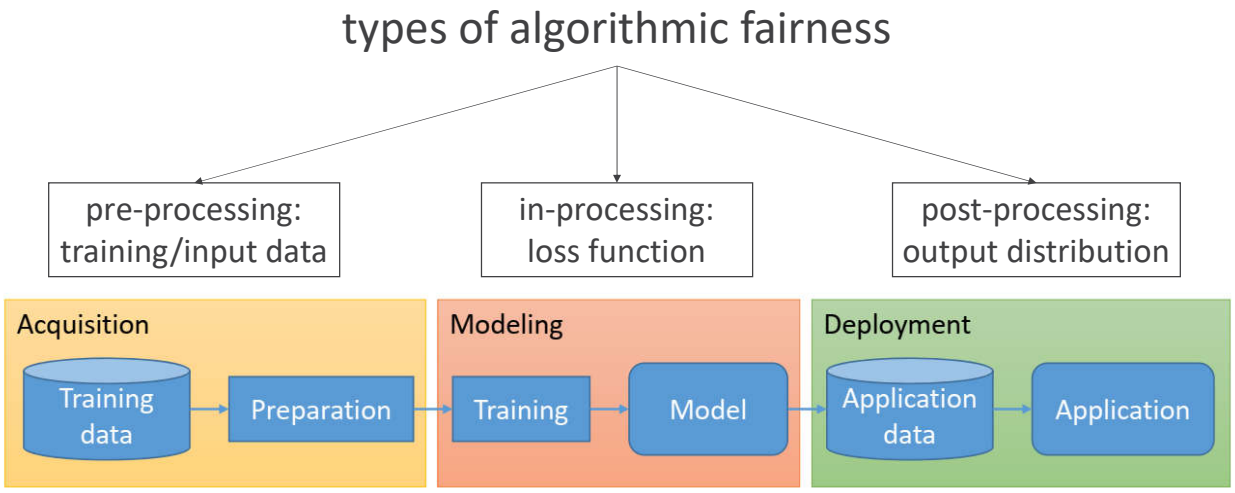
43

The Legality of Algorithmic Fairness



44

Algorithmic Fairness Procedures



45

Algorithmic Affirmative Action

CJEU guidelines:

1) **During** selection phase (**results**):
Marschall: restrictive criteria: only on the basis of all available information of the specific case
→ human in the loop, no automatic re-ranking

? → post-processing approaches

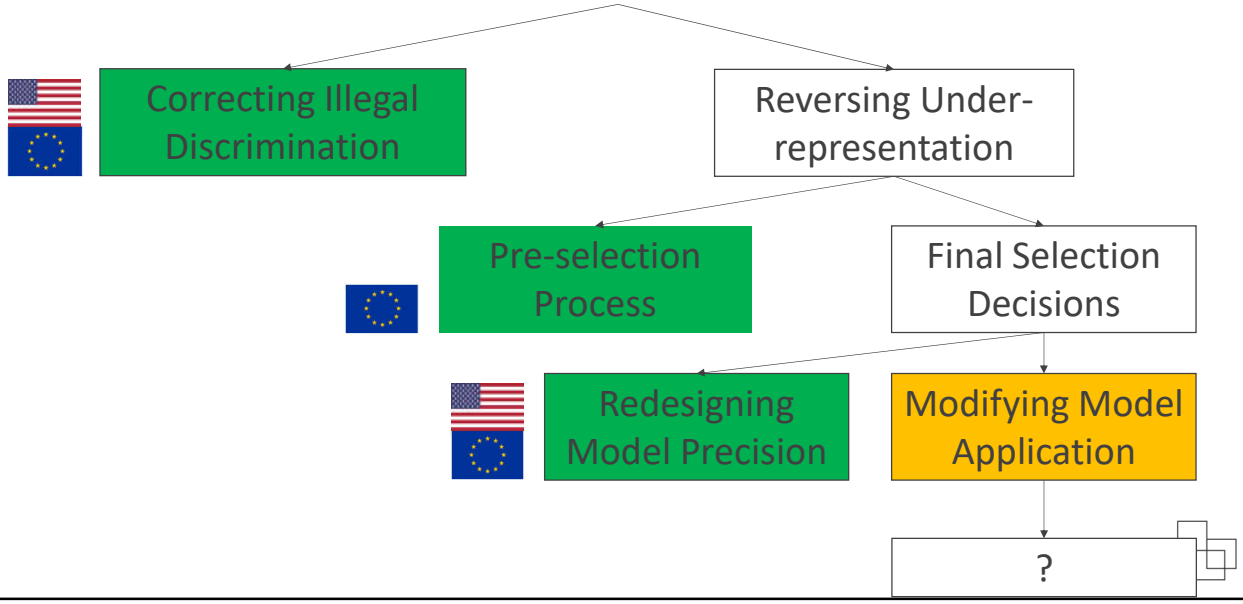
2) **Before** selection phase (**opportunity**):
Badeck: lenient criteria: even quota possible

? → pre-/in-processing approaches

ens

46

The Legality of Algorithmic Fairness in ADL



47

Criteria for Lawful Fairness Interventions

Changing the model in the concrete operation

Example: algorithmically modifying application data (pre-processing) or applicant scores (post-processing)

- Individual detriment → *Marschall* revisited
- Admissibility: Balance of rights and interests

ens

48

Criteria for Lawful Fairness Interventions

Model change during selection in favor of an underrepresented group justified:

- Difficulty to measure merit across groups (e.g., ground truth skewed)
  → Individual fairness does not work
- Group-based **substantive** equality **>** merit-based **formal** individual equality
  → Basic capabilities for later successful competition on the merits
  e.g., basic education; only exceptionally: credit scoring or job application

→ But: human review generally necessary

→ "Substantive equality of opportunity"

ens

49

Conclusion

1) Emergent case law on AI and discrimination
2) Enforcement bottleneck
3) Many algorithmic fairness metrics in Computer Science
4) Our algorithmic model:
   a) Bridges individual and group fairness
   b) While minimizing information loss
5) Replication of fairness divide in affirmative action law
   a) Legal constraints on algorithmic fairness
   b) Incentives for Human-Machine Teaming to mitigate underrepresentation

ens

50

Thank you!

hacker@europa-uni.de

ens

51

Selected Publications by Philipp Hacker

- **Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law**, 55 Common Market Law Review 1143-1186 (2018), open access: https://ssrn.com/abstract=3164973
- **Matching Code and Law: Achieving Algorithmic Fairness with Optimal Transport**, 34 Data Mining and Knowledge Discovery 163-200 (2020) (Meike Zehlike, Philipp Hacker and Emil Wiedemann), open access: https://arxiv.org/abs/1712.07924
- **A Legal Framework for AI Training Data**, 13 Law, Innovation and Technology 257-301 (2021), open access: https://doi.org/10.1080/17579961.2021.1977219
- **KI und DMA – Zugang, Transparenz und Fairness für KI-Modelle in der digitalen Wirtschaft**, GRUR 2022, 1278-1285 (engl. in preparation)

52

Additional Sources

- Allen, R. & Masters, D. (2021a). The pandemic, social benefits, and automated decision making (ADM): Just because it is quicker to use a machine, is it consistent with the principle of non-discrimination?, *AI Law Hub Blog* (19 April 2021), https://ai-lawhub.com/2021/04/19/the-pandemic-social-benefits-and-automated-decision-making-adm-just-because-it-is-quicker-to-use-a-machine-is-it-consistent-with-the-principle-of-non-discrimination/.
- Allen, R. & Masters, D. (2021b). An Italian lesson for Deliveroo: Computer programmes do not always think of everything!, *AI Law Hub Blog* (18 January 2021), https://ai-lawhub.com/2021/01/18/an-italian-lesson-for-deliveroo-computer-programmes-do-not-always-think-of-everything/.
- Article 29 Data Protection Working Party (2018). Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679, WP251rev.01, https://ec.europa.eu/newsroom/article29/redirection/document/49826.

53

Additional Sources

- Campbell, M. (2021). The austerity of lone motherhood: discrimination law and benefit reform. *Oxford Journal of Legal Studies*, *41*(4), 1197-1226.
- Maini, V., & Sabri, S. (2017). Machine learning for humans. https://www.dropbox.com/s/e38nil1dnl7481q/machine_learning.pdf?dl=0.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- Russell, S. J. (2010). *Artificial Intelligence: A Modern Approach*. 3rd. ed. Pearson Education.
- Wachter, S., Mittelstadt, B., & Russell, C. (2020). Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. *W. Va. L. Rev.*, *123*, 735.

54

Additional Sources

- Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, *41*, 105567.
- Zuiderveen Borgesius, F. J. (2020). Strengthening legal protection against discrimination by algorithms and artificial intelligence. *The International Journal of Human Rights*, *24*(10), 1572-1593.

55