Diskriminierung durch (lernende) Algorithmen

Meike Zehlike

30. September 2021

Humboldt Universität zu Berlin, Zalando

Agenda

1. Einführung Maschinelles Lernen
2. Bias und Fairness
3. Diskriminierende KI
4. Die Daten sind das Problem
5. Die Algorithmen sind das Problem

Einführung Maschinelles Lernen

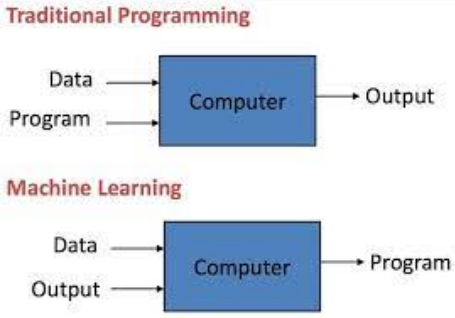


Abbildung 1: Traditionelles Vorgehen vs Machine Learning

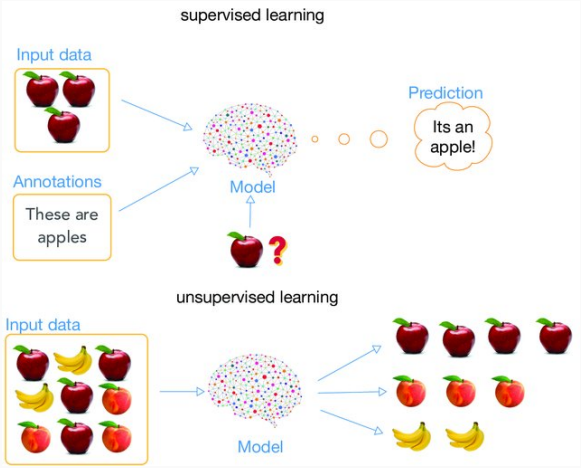


Abbildung 2: Supervised vs Unsupervised Learning

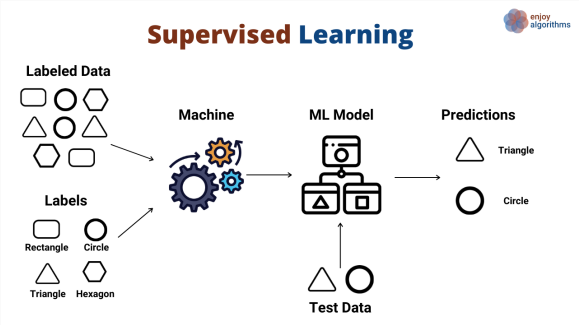


Abbildung 3: Supervised Machine Learning Pipeline

- Um das beste Modell von allen möglichen Modellen zu finden wird ein Optimierungsproblem gelöst:

$$\min \sum_i (\hat{y}_i - y_i)^2$$

- Das beste Modell ist also dasjenige, das die wenigsten Fehler macht

Bias und Fairness

Was ist Bias? [2]

Ein Computer System ist biased, wenn:

Was ist Bias? [2]

Ein Computer System ist biased, wenn:
es bestimmte *Gruppen* oder *Individuen systematisch* und *auf unfaire Art und Weise* zum Vorteil anderer diskriminiert.

Was ist Bias? [2]

Ein Computer System ist biased, wenn:
es bestimmte *Gruppen* oder *Individuen systematisch* und *auf unfaire Art und Weise* zum Vorteil anderer diskriminiert.

Ein System diskriminiert auf unfaire Art, wenn:

Was ist Bias? [2]

Ein Computer System ist biased, wenn:
es bestimmte *Gruppen* oder *Individuen systematisch* und *auf unfaire Art und Weise* zum Vorteil anderer diskriminiert.

Ein System diskriminiert auf unfaire Art, wenn:
es einem *Individuum* oder einer *Gruppe bestimmter Individuen*

Was ist Bias? [2]

Ein Computer System ist biased, wenn:
es bestimmte *Gruppen* oder *Individuen systematisch* und *auf unfaire Art und Weise* zum Vorteil anderer diskriminiert.

Ein System diskriminiert auf unfaire Art, wenn:
es einem *Individuum* oder einer *Gruppe bestimmter Individuen*
aufgrund *unvernünftiger* oder *unangemessener* Kriterien

Was ist Bias? [2]

Ein Computer System ist biased, wenn:
es bestimmte *Gruppen* oder *Individuen systematisch* und *auf unfaire Art und Weise* zum Vorteil anderer diskriminiert.

Ein System diskriminiert auf unfaire Art, wenn:
es einem *Individuum* oder einer *Gruppe bestimmter Individuen*
aufgrund *unvernünftiger* oder *unangemessener* Kriterien
eine *Chance* oder ein *Gut verweigert*, oder wenn

Was ist Bias? [2]

Ein Computer System ist biased, wenn:
es bestimmte *Gruppen* oder *Individuen systematisch* und *auf unfaire Art und Weise* zum Vorteil anderer diskriminiert.

Ein System diskriminiert auf unfaire Art, wenn:
es einem *Individuum* oder einer *Gruppe bestimmter Individuen*
aufgrund *unvernünftiger* oder *unangemessener* Kriterien
eine *Chance* oder ein *Gut verweigert*, oder wenn
es denselben ein unerwünschtes Ergebnis zuweist.

Was ist Bias? [2]

Bias ist unfaire Diskriminierung, die *systematisch* auftritt.

Was ist Bias? [2]

Bias ist unfaire Diskriminierung, die *systematisch* auftritt.

Bias ist systematische Diskriminierung, die zusammen mit einem *unfairen Ergebnis* auftritt.

Typen von Bias [2]

Typen von Bias [2]

- **Preexisting Bias:** Bias der unabhängig und vor der Schaffung des Systems existiert.

Typen von Bias [2]

- **Preexisting Bias:** Bias der unabhängig und vor der Schaffung des Systems existiert.
- **Technical Bias:** technische Limitierungen und Zwänge (Hardware und Software Limitierungen, Generation von Pseudozufallszahlen, Formalisierung menschlicher Konstrukte)

Typen von Bias [2]

- **Preexisting Bias:** Bias der unabhängig und vor der Schaffung des Systems existiert.
- **Technical Bias:** technische Limitierungen und Zwänge (Hardware und Software Limitierungen, Generation von Pseudozufallszahlen, Formalisierung menschlicher Konstrukte)
- **Emergent Bias:** tritt erst nach der Einführung des Systems auf (Veränderung im Nutzerverhalten, Veränderung sozialer Konzepte)

Chancengleichheit—Equality of Opportunity

Chancengleichheit—Equality of Opportunity

- ein Framework aus der politischen Philosophie

Chancengleichheit—Equality of Opportunity

- ein Framework aus der politischen Philosophie
- vier verschiedene Geschmacksrichtungen [1]:

Chancengleichheit—Equality of Opportunity

- ein Framework aus der politischen Philosophie
- vier verschiedene Geschmacksrichtungen [1]: libertäre EO,

Chancengleichheit—Equality of Opportunity

- ein Framework aus der politischen Philosophie
- vier verschiedene Geschmacksrichtungen [1]: libertäre EO, formelle EO,

Chancengleichheit—Equality of Opportunity

- ein Framework aus der politischen Philosophie
- vier verschiedene Geschmacksrichtungen [1]: libertäre EO, formelle EO, substantielle EO

Chancengleichheit—Equality of Opportunity

- ein Framework aus der politischen Philosophie
- vier verschiedene Geschmacksrichtungen [1]: libertäre EO, formelle EO, substantielle EO (Rawlsian EO, luck-egalitarian EO)

Chancengleichheit—Equality of Opportunity

- ein Framework aus der politischen Philosophie
- vier verschiedene Geschmacksrichtungen [1]: libertäre EO, formelle EO, substantielle EO (Rawlsian EO, luck-egalitarian EO)

Libertäre Chancengleichheit [3]

Libertäre Chancengleichheit [3]

- Eine jede hat die Freiheit zu tun und zu lassen was sie möchte mit all dem, was ihr legitimerweise gehört (Selbst, Geschäft, etc.),

Libertäre Chancengleichheit [3]

- Eine jede hat die Freiheit zu tun und zu lassen was sie möchte mit all dem, was ihr legitimerweise gehört (Selbst, Geschäft, etc.),
  solange dies nicht moralische Rechte anderer Menschen verletzt.

Libertäre Chancengleichheit [3]

- Eine jede hat die Freiheit zu tun und zu lassen was sie möchte mit all dem, was ihr legitimerweise gehört (Selbst, Geschäft, etc.),
  solange dies nicht moralische Rechte anderer Menschen verletzt.
- Im Kontext algorithmischer Entscheidungssysteme (ADM) bedeutet dies totale Freiheit jeden beliebigen Algorithmus zu implementieren.

Libertäre Chancengleichheit [3]

- Eine jede hat die Freiheit zu tun und zu lassen was sie möchte mit all dem, was ihr legitimerweise gehört (Selbst, Geschäft, etc.),
  solange dies nicht moralische Rechte anderer Menschen verletzt.
- Im Kontext algorithmischer Entscheidungssysteme (ADM) bedeutet dies totale Freiheit jeden beliebigen Algorithmus zu implementieren.
- Dieser kann jede verfügbare Information benutzen, auch solche die eigentlich gesetzlich geschützt sind (Geschlecht, Migrationshintergrund), um statistisch akkuratere Entscheidungen zu treffen.

Formelle Chancengleichheit [3]

Formelle Chancengleichheit [3]

- Erstrebenswerte Positionen müssen allen zugänglich sein, die für die Erfüllung der Position *notwendige* Eigenschaften mitbringen.

Formelle Chancengleichheit [3]

- Erstrebenswerte Positionen müssen allen zugänglich sein, die für die Erfüllung der Position *notwendige* Eigenschaften mitbringen.
- Die Positionen werden aufgrund der Qualifikation der Kandidatinnen zugewiesen.

Formelle Chancengleichheit [3]

- Erstrebenswerte Positionen müssen allen zugänglich sein, die für die Erfüllung der Position *notwendige* Eigenschaften mitbringen.
- Die Positionen werden aufgrund der Qualifikation der Kandidatinnen zugewiesen.
- Im Kontext von ADM fordert formelle Chancengleichheit sensible Informationen aus dem System zu entfernen.

Substantielle Chancengleichheit [3]

Substantielle Chancengleichheit [3]

- Erstrebenswerte Positionen müssen allen zugänglich sein, die für die Erfüllung der Position notwendige Eigenschaften mitbringen.

Substantielle Chancengleichheit [3]

- Erstrebenswerte Positionen müssen allen zugänglich sein, die für die Erfüllung der Position notwendige Eigenschaften mitbringen.
- Die Positionen werden aufgrund der Qualifikation der Kandidatinnen zugewiesen.

Substantielle Chancengleichheit [3]

- Erstrebenswerte Positionen müssen allen zugänglich sein, die für die Erfüllung der Position notwendige Eigenschaften mitbringen.
- Die Positionen werden aufgrund der Qualifikation der Kandidatinnen zugewiesen.
- zusätzlich braucht es fairen Zugang zu solchen Qualifikationen

Substantielle Chancengleichheit [3]

- Erstrebenswerte Positionen müssen allen zugänglich sein, die für die Erfüllung der Position notwendige Eigenschaften mitbringen.
- Die Positionen werden aufgrund der Qualifikation der Kandidatinnen zugewiesen.
- zusätzlich braucht es fairen Zugang zu solchen Qualifikationen
- Rawls'sche EO:

Substantielle Chancengleichheit [3]

- Erstrebenswerte Positionen müssen allen zugänglich sein, die für die Erfüllung der Position notwendige Eigenschaften mitbringen.
- Die Positionen werden aufgrund der Qualifikation der Kandidatinnen zugewiesen.
- zusätzlich braucht es fairen Zugang zu solchen Qualifikationen
- **Rawls'sche EO:**
  - Jene mit demselben Talent und Bestreben müssen die gleichen Aussichten haben die sozialen Positionen zu erlangen,

Substantielle Chancengleichheit [3]

- Erstrebenswerte Positionen müssen allen zugänglich sein, die für die Erfüllung der Position notwendige Eigenschaften mitbringen.
- Die Positionen werden aufgrund der Qualifikation der Kandidatinnen zugewiesen.
- zusätzlich braucht es fairen Zugang zu solchen Qualifikationen
- **Rawls'sche EO:**
  - Jene mit demselben Talent und Bestreben müssen die gleichen Aussichten haben die sozialen Positionen zu erlangen, unabhängig von zufälligen Faktoren wie sozio-ökonomischer Status der Eltern.

Substantielle Chancengleichheit [3]

- Erstrebenswerte Positionen müssen allen zugänglich sein, die für die Erfüllung der Position notwendige Eigenschaften mitbringen.
- Die Positionen werden aufgrund der Qualifikation der Kandidatinnen zugewiesen.
- zusätzlich braucht es fairen Zugang zu solchen Qualifikationen
- **Rawls'sche EO:**
  - Jene mit demselben Talent und Bestreben müssen die gleichen Aussichten haben die sozialen Positionen zu erlangen, unabhängig von zufälligen Faktoren wie sozio-ökonomischer Status der Eltern.
  - Nimmt an, dass Talent und Bestreben sich über alle Individuen hinweg vergleichen lässt.

Substantielle Chancengleichheit [3]

- Erstrebenswerte Positionen müssen allen zugänglich sein, die für die Erfüllung der Position notwendige Eigenschaften mitbringen.
- Die Positionen werden aufgrund der Qualifikation der Kandidatinnen zugewiesen.
- zusätzlich braucht es fairen Zugang zu solchen Qualifikationen
- **Rawls'sche EO:**
  - Jene mit demselben Talent und Bestreben müssen die gleichen Aussichten haben die sozialen Positionen zu erlangen, unabhängig von zufälligen Faktoren wie sozio-ökonomischer Status der Eltern.
  - Nimmt an, dass Talent und Bestreben sich über alle Individuen hinweg vergleichen lässt.
- **Luck-egalitarian EO:**

Substantielle Chancengleichheit [3]

- Erstrebenswerte Positionen müssen allen zugänglich sein, die für die Erfüllung der Position notwendige Eigenschaften mitbringen.
- Die Positionen werden aufgrund der Qualifikation der Kandidatinnen zugewiesen.
- zusätzlich braucht es fairen Zugang zu solchen Qualifikationen
- Rawls'sche EO:
  - Jene mit demselben Talent und Bestreben müssen die gleichen Aussichten haben die sozialen Positionen zu erlangen, unabhängig von zufälligen Faktoren wie sozio-ökonomischer Status der Eltern.
  - Nimmt an, dass Talent und Bestreben sich über alle Individuen hinweg vergleichen lässt.
- Luck-egalitarian EO:
  - Bietet eine relative Sichtweise auf Talent und Bestreben und erlaubt, dass diese vom sozio-ökonomischen Hintergrund beeinflusst werden.

Substantielle Chancengleichheit [3]

- Erstrebenswerte Positionen müssen allen zugänglich sein, die für die Erfüllung der Position notwendige Eigenschaften mitbringen.
- Die Positionen werden aufgrund der Qualifikation der Kandidatinnen zugewiesen.
- zusätzlich braucht es fairen Zugang zu solchen Qualifikationen
- Rawls'sche EO:
  - Jene mit demselben Talent und Bestreben müssen die gleichen Aussichten haben die sozialen Positionen zu erlangen, unabhängig von zufälligen Faktoren wie sozio-ökonomischer Status der Eltern.
  - Nimmt an, dass Talent und Bestreben sich über alle Individuen hinweg vergleichen lässt.
- Luck-egalitarian EO:
  - Bietet eine relative Sichtweise auf Talent und Bestreben und erlaubt, dass diese vom sozio-ökonomischen Hintergrund beeinflusst werden.
  - Nimmt an, dass Talent und Bestreben nur innerhalb einer Gruppe verglichen werden darf.

Diskriminierende KI

Beispiele

- Apples Credit Scoring Algorithmus bewertet Frauen schlechter

Beispiele

- Apples Credit Scoring Algorithmus bewertet Frauen schlechter
- Google Ad Service suggeriert Schwarze seien kriminell

Beispiele

- Apples Credit Scoring Algorithmus bewertet Frauen schlechter
- Google Ad Service suggeriert Schwarze seien kriminell
- Google Bilderkennung erkennt auf Bildern mit schwarzen Menschen Gorillas

Beispiele

- Apples Credit Scoring Algorithmus bewertet Frauen schlechter
- Google Ad Service suggeriert Schwarze seien kriminell
- Google Bilderkennung erkennt auf Bildern mit schwarzen Menschen Gorillas
- Siri und Alexa reagieren bei sexueller Belästigung unangemessen

Beispiele

- Apples Credit Scoring Algorithmus bewertet Frauen schlechter
- Google Ad Service suggeriert Schwarze seien kriminell
- Google Bilderkennung erkennt auf Bildern mit schwarzen Menschen Gorillas
- Siri und Alexa reagieren bei sexueller Belästigung unangemessen
- Tay (Windows Chatbot) wird zum Rassisten

Beispiele

- Apples Credit Scoring Algorithmus bewertet Frauen schlechter
- Google Ad Service suggeriert Schwarze seien kriminell
- Google Bilderkennung erkennt auf Bildern mit schwarzen Menschen Gorillas
- Siri und Alexa reagieren bei sexueller Belästigung unangemessen
- Tay (Windows Chatbot) wird zum Rassisten
- Amazon's Hiring Tool sortiert Lebensläufe von Frauen aus

Beispiele

- Apples Credit Scoring Algorithmus bewertet Frauen schlechter
- Google Ad Service suggeriert Schwarze seien kriminell
- Google Bilderkennung erkennt auf Bildern mit schwarzen Menschen Gorillas
- Siri und Alexa reagieren bei sexueller Belästigung unangemessen
- Tay (Windows Chatbot) wird zum Rassisten
- Amazon's Hiring Tool sortiert Lebensläufe von Frauen aus
- Google Ads zeigt Werbung für gutbezahlte Jobs überwiegend Männern

Beispiele

- Apples Credit Scoring Algorithmus bewertet Frauen schlechter
- Google Ad Service suggeriert Schwarze seien kriminell
- Google Bilderkennung erkennt auf Bildern mit schwarzen Menschen Gorillas
- Siri und Alexa reagieren bei sexueller Belästigung unangemessen
- Tay (Windows Chatbot) wird zum Rassisten
- Amazon's Hiring Tool sortiert Lebensläufe von Frauen aus
- Google Ads zeigt Werbung für gutbezahlte Jobs überwiegend Männern
- Gesichtserkennung zum Entsperren von iPhones funktioniert für schwarze Menschen wesentlich schlechter als für Männer

Beispiele

- Apples Credit Scoring Algorithmus bewertet Frauen schlechter
- Google Ad Service suggeriert Schwarze seien kriminell
- Google Bilderkennung erkennt auf Bildern mit schwarzen Menschen Gorillas
- Siri und Alexa reagieren bei sexueller Belästigung unangemessen
- Tay (Windows Chatbot) wird zum Rassisten
- Amazon's Hiring Tool sortiert Lebensläufe von Frauen aus
- Google Ads zeigt Werbung für gutbezahlte Jobs überwiegend Männern
- Gesichtserkennung zum Entsperren von iPhones funktioniert für schwarze Menschen wesentlich schlechter als für Männer
- Ein Algorithmus zur Bettenbelegung auf einer Intensivstation diskriminiert gegen Schwarze

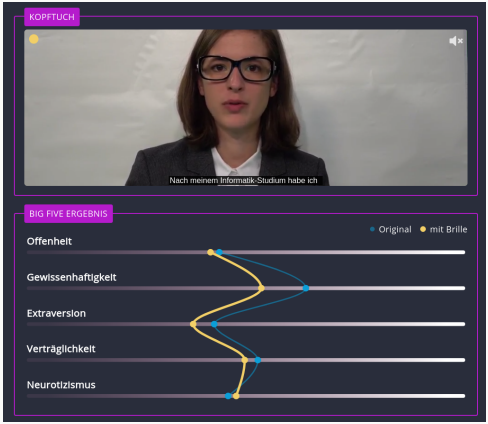
Beispiele

- Apples Credit Scoring Algorithmus bewertet Frauen schlechter
- Google Ad Service suggeriert Schwarze seien kriminell
- Google Bilderkennung erkennt auf Bildern mit schwarzen Menschen Gorillas
- Siri und Alexa reagieren bei sexueller Belästigung unangemessen
- Tay (Windows Chatbot) wird zum Rassisten
- Amazon's Hiring Tool sortiert Lebensläufe von Frauen aus
- Google Ads zeigt Werbung für gutbezahlte Jobs überwiegend Männern
- Gesichtserkennung zum Entsperren von iPhones funktioniert für schwarze Menschen wesentlich schlechter als für Männer
- Ein Algorithmus zur Bettenbelegung auf einer Intensivstation diskriminiert gegen Schwarze
- Facebook erkennt Profile von Native-Americans als Fake-Profile

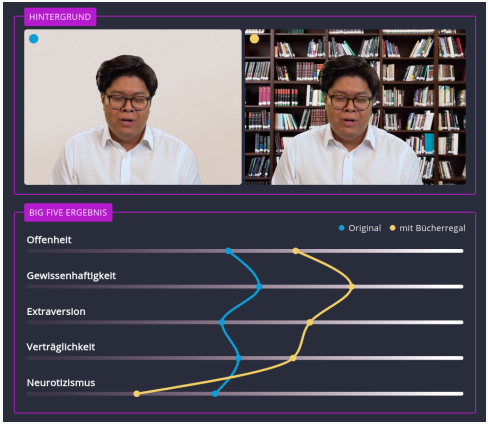
Beispiele

- Apples Credit Scoring Algorithmus bewertet Frauen schlechter
- Google Ad Service suggeriert Schwarze seien kriminell
- Google Bilderkennung erkennt auf Bildern mit schwarzen Menschen Gorillas
- Siri und Alexa reagieren bei sexueller Belästigung unangemessen
- Tay (Windows Chatbot) wird zum Rassisten
- Amazon's Hiring Tool sortiert Lebensläufe von Frauen aus
- Google Ads zeigt Werbung für gutbezahlte Jobs überwiegend Männern
- Gesichtserkennung zum Entsperren von iPhones funktioniert für schwarze Menschen wesentlich schlechter als für Männer
- Ein Algorithmus zur Bettenbelegung auf einer Intensivstation diskriminiert gegen Schwarze
- Facebook erkennt Profile von Native-Americans als Fake-Profile
- ...

Mehr Beispiele



(a) Mit vs. Ohne Brille



(b) Mit Bücherregal

Abbildung 4: https://web.br.de/interaktiv/ki-bewerbung/

Die Daten sind das Problem

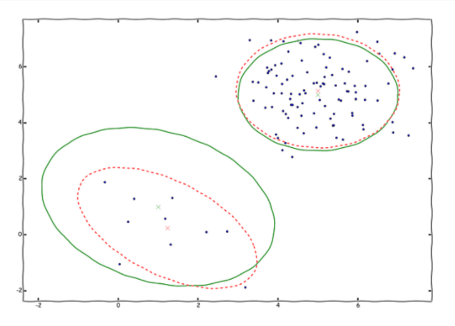


Abbildung 5: Modelle funktionieren mit weniger Trainingsdaten schlechter (grün – ground truth, rot – Modell)

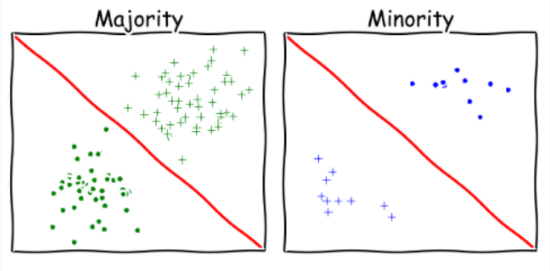


Abbildung 6: Innerhalb der Population kann es große statistische Unterschiede geben

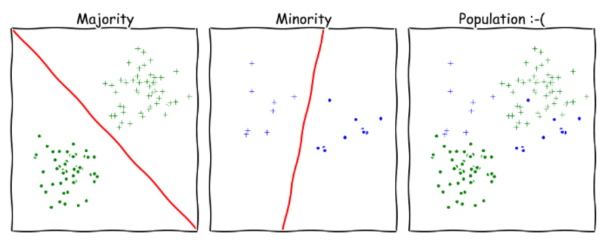


Abbildung 7: Was nun?

Sind Trainingsdaten Neutral?

(a) Was ist hier zu sehen?

Sind Trainingsdaten Neutral?

(a) Was ist hier zu sehen?

(b) Was ist hier zu sehen?

Sind Trainingsdaten Neutral?

(a) Was ist hier zu sehen?

(b) Was ist hier zu sehen?

Abbildung 8: Beschreibung und Parameter werden als Norm und Abweichung von der Norm dargestellt

Sind Trainingsdaten Neutral?

Beispiel Gender Pay Gap:

- Frauen verdienen 80€/h, Männer verdienen 100€/h

Sind Trainingsdaten Neutral?

Beispiel Gender Pay Gap:

- Frauen verdienen 80€/h, Männer verdienen 100€/h
- “Noch immer verdienen Frauen 20% weniger als Männer”

Sind Trainingsdaten Neutral?

Beispiel Gender Pay Gap:

- Frauen verdienen 80€/h, Männer verdienen 100€/h
- “Noch immer verdienen Frauen 20% weniger als Männer”
- Aus Sicht der Frauen: “Noch immer verdienen Frauen 25% weniger als Männer”

Was sind Sensible Daten?

Was ist die Kategorie? Eigenschaft oder Wahl?

Was sind Sensible Daten?

Was ist die Kategorie? Eigenschaft oder Wahl?

Konstruierte Kategorien? Biologisch oder sozial konstruiert?

Was sind Sensible Daten?

Was ist die Kategorie? Eigenschaft oder Wahl?

Konstruierte Kategorien? Biologisch oder sozial konstruiert?

Modularität Modelle sind Systeme mit reduzierter Komplexität

Was sind Sensible Daten?

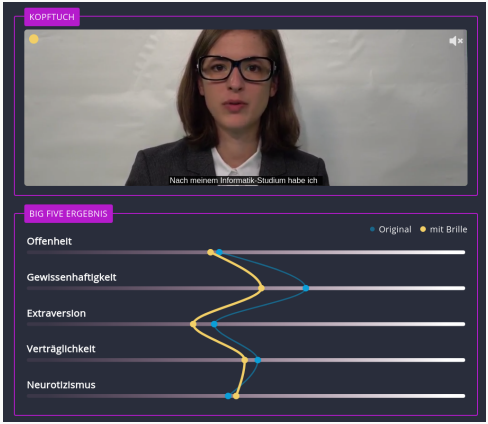
Was ist die Kategorie?	Eigenschaft oder Wahl?
Konstruierte Kategorien?	Biologisch oder sozial konstruiert?
Modularität	Modelle sind Systeme mit reduzierter Komplexität
Ontologische Instabilität	Bedeutung eines Begriffs verändert sich über die Zeit

Was sind Sensible Daten?

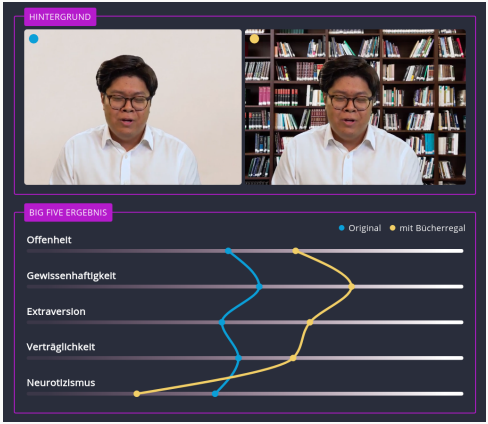
Was ist die Kategorie?	Eigenschaft oder Wahl?
Konstruierte Kategorien?	Biologisch oder sozial konstruiert?
Modularität	Modelle sind Systeme mit reduzierter Komplexität
Ontologische Instabilität	Bedeutung eines Begriffs verändert sich über die Zeit
	Zuschreibungen führen zu veränderten Verhaltensweisen

Die Algorithmen sind das Problem

Welches Problem wird überhaupt gelöst?



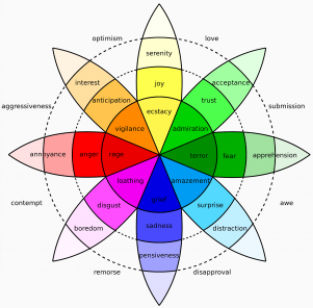
(a) Mit vs. Ohne Brille



(b) Mit Bücherregal

Abbildung 9: https://web.br.de/interaktiv/ki-bewerbung/

Absurde Vereinfachung komplexer Konzepte



(a) Emotionsrad nach Plutchik

(b) Beispiel für Trainingsdaten

Ethische Werte in Optimierungsproblemen

Utilitaristische Ethik:

Utilitaristische Ethik: $\min \sum_i (\hat{y}_i - y_i)^2$

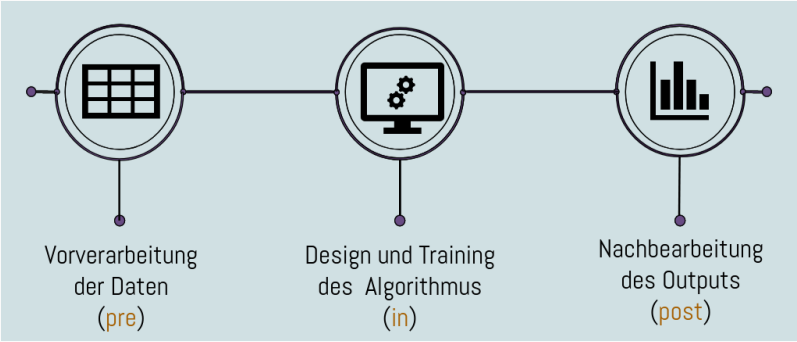
Utilitaristische Ethik: $\min \sum_i (\hat{y}_i - y_i)^2$

Rawls'sche Fairness:

Utilitaristische Ethik: $\min \sum_i (\hat{y}_i - y_i)^2$

Rawls'sche Fairness: $\min \max_i (\hat{y}_i - y_i)^2$

Wie Baut man einen Fairen Algorithmus?



Wie übersetzt man Gerechtigkeit in Mathematik?

$$\min \sum_i (\hat{y}_i - y_i)^2 \quad \text{subject to} \quad P(\hat{y}_i \neq y_i | \text{protected} = 1)$$

$$= P(\hat{y}_i \neq y_i | \text{protected} = 0)$$

Fragen?

References I

R. Arneson.
Equality of opportunity.
In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2015 edition, 2015.

B. Friedman and H. Nissenbaum.
Bias in computer systems.
ACM Transactions on Information Systems, 14(3):330–347, 1996.

H. Heidari, M. Loi, K. P. Gummadi, and A. Krause.
A moral framework for understanding fair ml through economic models of equality of opportunity.
In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 181–190. ACM, 2019.