Fairness im Maschinellen Lernen

Dr. rer. nat. Sebastian Pape

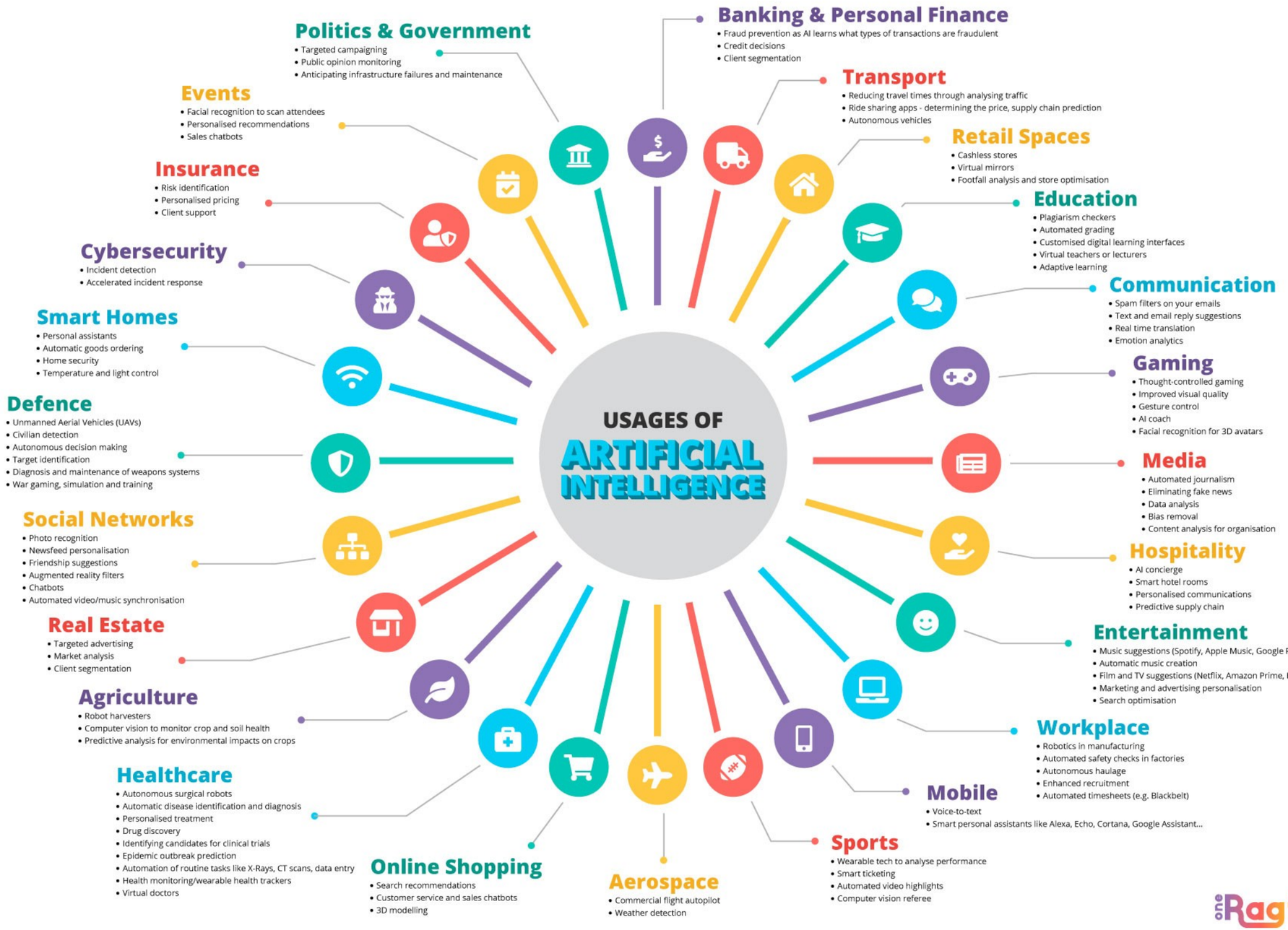
Habilitationskolloquium

22. Februar 2021

Agenda

- Einführung
  - Motivation
  - Maschinelles Lernen
  - Einordnung
- Fairness
  - Bias
  - Diskriminierung
  - Definition Fairness
- Algorithmen
  - Klassifizierung
  - Word Embedding
- Herausforderungen
- Zusammenfassung

Aufgaben für maschinelles Lernen



Source: https://www.oneragtime.com/24-industries-disrupted-by-ai-infographic/

Motivation

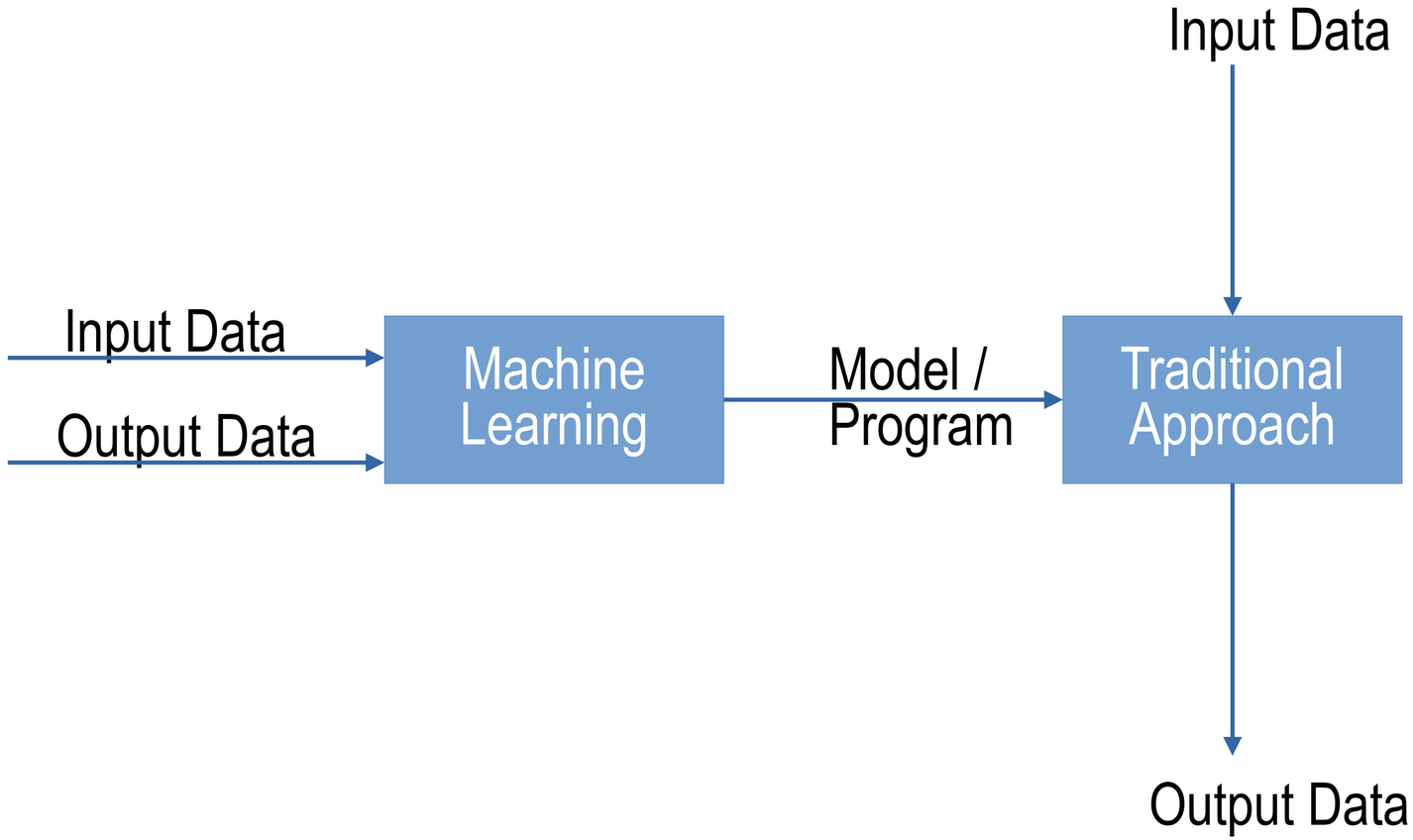
Only Seven of Stanford's First 5,000 Vaccines Were Designated for Medical Residents

Stanford Medicine officials relied on a faulty algorithm to determine who should get vaccinated first, and it prioritized some high-ranking doctors over patient-facing medical residents.

An algorithm chose who would be the first 5,000 in line. T
they were at a disadvantage because they did not have an
the calculation and because they are young, according to
resident to his peers. Residents are the lowest-ranking doctors in a hospital. Stanford Medicine has about 1,300 across all disciplines.

Source: Caroline Chen, ProPublica
https://www.propublica.org/article/only-seven-of-stanfords-first-5-000-vaccines-were-designated-for-medical-residents

Machine Learning vs. Traditional Approaches



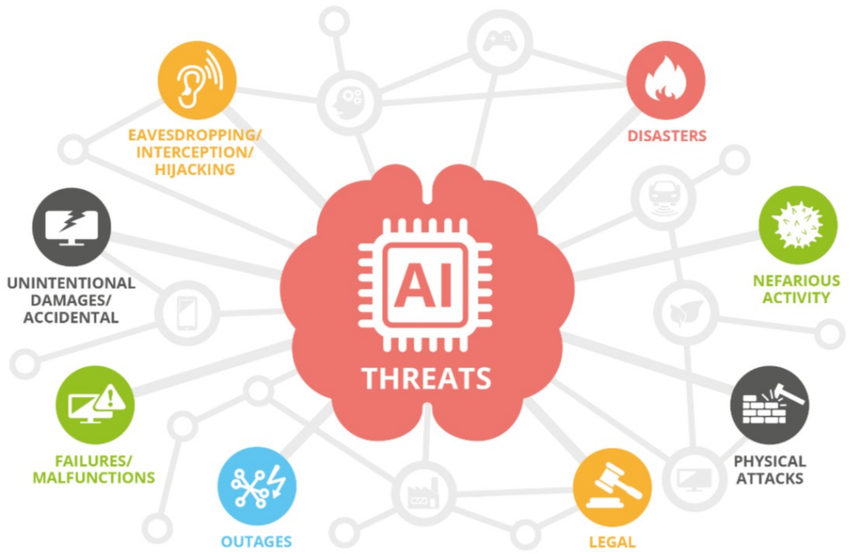
Generelle Herausforderungen von Maschinellem Lernen

- Angriffe
- Datenschutz

- Große Datenmengen zum Training benötigt
- Erklärbarkeit / Interpretierbarkeit

- Ethik

Figure 5: AI Threat Taxonomy



Ethik im Maschinellen Lernen

	The European Commission's High-Level Expert Group on Artificial Intelligence	Report on the Future of Artificial Intelligence	Beijing AI Principles	OECD Recommendation of the Council on Artificial Intelligence	The Malicious Use of Artificial Intelligence	AI4People	The Asilomar AI Principles	AI Now 2016 Report	AI Now 2017 Report	AI Now 2018 Report	AI Now 2019 Report	Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	Montréal Declaration for Responsible Development of Artificial Intelligence	OpenAI Charter	Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (Version for Public Discussion)	Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (First Edition)	ITI AI Policy Principles	Microsoft AI principles	DeepMind Ethics & Society Principles	Artificial Intelligence at Google	Everyday Ethics for Artificial Intelligence	Partnership on AI	number of mentions
authors	(Pekka et al. 2018)	(Holdren et al. 2016)	(Beijing Academy of Artificial Intelligence 2019)	(Organisation for Economic Co-operation and Development 2019)	(Brundage et al. 2018)	(Floridi et al. 2018)	(Future of Life Institute 2017)	(Crawford et al. 2016)	(Campolo et al. 2017)	(Whittaker et al. 2018)	(Crawford et al. 2019)	(Diakopoulos et al.)	(Abrassart et al. 2018)	(OpenAI 2018)	(The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2016)	(The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019)	(Information Technology Industry Council 2017)	(Microsoft Corporation 2019)	(DeepMind)	(Google 2018)	(Cutler et al. 2018)	(Partnership on AI 2018)	
key issue	AI principles of the EU	AI principles of the US	AI principles of China	AI principles of the OECD	analysis of abuse scenarios of AI	meta-analysis about principles for the beneficial use of AI	large collection of different principles	statements on social implications of AI	statements on social implications of AI	statements on social implications of AI	statements on social implications of AI	principles of the FAT ML community	code of ethics released by the Université de Montréal	several short principles for the ethical use of AI	detailed description of ethical aspects in the context of AI	detailed description of ethical aspects in the context of AI	brief guideline about basic ethical principles	short list of keywords for the ethical use of AI	several short principles for the ethical use of AI	several short principles for the ethical use of AI	IBM's short list of keywords for the ethical use of AI	principles of an association between several industry leaders	
privacy protection	x		x	x	x	x	x	x	x	x	x		x		x	x	x	x		x	x		18
fairness, non-discrimination, justice	x	x	x	x		x	x	x	x	x	x	x	x		x	x		x		x	x		18
accountability	x	x	x	x	x	x	x	x		x		x	x		x	x	x	x			x	x	17
transparency, openness	x	x	x	x	x		x	x	x	x	x	x	x		x	x		x	x	x			16
safety, cybersecurity	x	x	x	x	x	x	x				x		x		x	x	x	x		x		x	16
common good, sustainability, well-being		x	x	x		x	x	x	x		x		x		x	x			x	x	x	x	16
human oversight, control, auditing	x		x	x		x	x		x	x		x		x		x	x			x			12
solidarity, inclusion, social cohesion			x	x		x		x		x	x		x			x		x	x			x	11
explainability, interpretabiliy	x		x			x	x		x			x	x			x	x				x		10
science-policy link		x	x		x	x	x	x		x				x		x			x				10
legislative framework, legal status of AI systems	x	x		x	x	x		x		x	x				x								10
future of employment/worker rights		x	x	x				x			x				x	x	x					x	9
responsible/intensified research funding		x		x		x	x	x							x		x		x				8
public awareness, education about AI and its risks		x	x			x		x							x	x	x					x	8
dual-use problem, military, AI arms race		x			x		x		x		x		x		x					x			8
field-specific deliberations (health, military, mobility etc.)		x			x			x	x	x	x				x	x							8
human autonomy	x		x	x		x	x						x			x							7
diversity in the field of AI								x	x	x	x		x				x		x				7
certification for AI products						x									x	x					x		4
protection of whistleblowers										x	x					x							3
cultural differences in the ethically aligned design of AI systems															x	x							2
hidden costs (labeling, clickwork, contend moderation, energy, resources)										x	x												2
notes on technical implementations	yes, but very few	none	none	none	yes	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	
proportion of women among authors (f/m)	(8/10)	(2/3)	ns	ns	(5/21)	(5/8)	ns	(4/2)	(3/1)	(6/4)	(12/4)	(1/12)	(8/10)	ns	varies in each chapter	varies in each chapter	ns	ns	ns	ns	(1/2)	ns	(55/77)
length (number of words)	16546	22787	766	3249	34017	8609	646	11530	18273	25759	38970	1359	4754	441	40915	108.092	2272	75	417	882	4488	1481	
affiliation (government, industry, science)	government	government	science/ gov./ind.	government	science	science	science	science	science	science	science	science	science	non-profit	industry	industry	industry	industry	industry	industry	industry	industry	
number of ethical aspects	9	12	13	12	8	14	12	13	9	12	13	5	11	4	14	18	9	6	6	6	6	8	

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99-120.

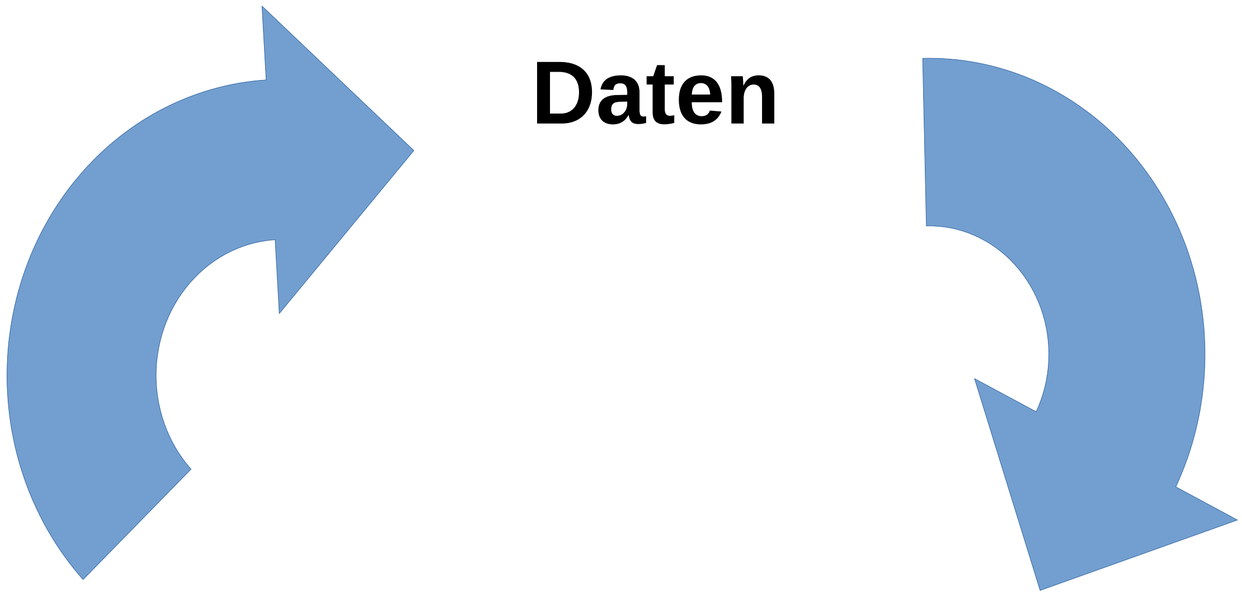
Ethik im Maschinellen Lernen

	number of mentions	Partnership on AI	Everyday Ethics for Artificial Intelligence	Artificial Intelligence at Google	DeepMind Ethics & Society Principles	Microsoft AI principles	ITI AI Policy Principles	Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (First Edition)	Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (Version for Public Discussion)	OpenAI Charter	Montréal Declaration for Responsible Development of Artificial Intelligence	Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	AI Now 2019 Report	AI Now 2018 Report	AI Now 2017 Report	AI Now 2016 Report	The Asilomar AI Principles	AI4People	The Malicious Use of Artificial Intelligence	OECD Recommendation of the Council on Artificial Intelligence	Beijing AI Principles	Report on the Future of Artificial Intelligence	The European Commission's High-Level Expert Group on Artificial Intelligence
authors		(Partnership on AI 2018)	(Cutler et al. 2018)	(Google 2018)	(DeepMind)	(Microsoft Corporation 2019)	(Information Technology Industry Council 2017)	(The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019)	(The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2016)	(OpenAI 2018)	(Abrassart et al. 2018)	(Diakopoulos et al.)	(Crawford et al. 2019)	(Whittaker et al. 2018)	(Campolo et al. 2017)	(Crawford et al. 2016)	(Future of Life Institute 2017)	(Floridi et al. 2018)	(Brundage et al. 2018)	(Organisation for Economic Co operation and Development 2019)	(Beijing Academy of Artificial Intelligence 2019)	(Holdren et al. 2016)	(Pekka et al. 2018)
key issue		principles of an association between several industry leaders	IBM's short list of keywords for the ethical use of AI	several short principles for the ethical use of AI	several short principles for the ethical use of AI	short list of keywords for the ethical use of AI	brief guideline about basic ethical principles	detailed description of ethical aspects in the context of AI	detailed description of ethical aspects in the context of AI	several short principles for the ethical use of AI	code of ethics released by the Université de Montréal	principles of the FAT ML community	statements on social implications of AI	statements on social implications of AI	statements on social implications of AI	statements on social implications of AI	large collection of different principles	meta-analysis about principles for the beneficial use of AI	analysis of abuse scenarios of AI	AI principles of the OECD	AI principles of China	AI principles of the US	AI principles of the EU
privacy protection	**18**	x		x		x		x	x		x		x	x	x	x	x	x	x	x	x	x	x
fairness, non-discrimination, justice	**18**	x		x		x		x	x		x		x	x	x	x	x	x		x	x	x	x
accountability	**17**	x	x			x		x	x		x		x		x	x	x	x	x	x	x	x	x
transparency, openness	**16**		x					x	x		x		x	x	x	x	x		x	x	x	x	x
diversity in the field of AI	**7**					x								x	x	x	x						
certification for AI products	**4**			x							x	x											
protection of whistleblowers	**3**										x	x											
cultural differences in the ethically aligned design of AI systems	**2**										x	x											
hidden costs (labeling, clickwork, contend moderation, energy, resources)	**2**															x	x						
notes on technical implementations		none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	yes	none	none	none	yes, but very few
proportion of women among authors (f/m)	(55/77)	ns	(1/2)	ns	ns	ns	ns	ns	ns	ns	varies in each chapter	varies in each chapter	ns	(12/4)	(6/4)	(3/1)	(4/2)	ns	(5/21)	ns	ns	(2/3)	(8/10)
length (number of words)		1481	4488	882	417	75	2272	108.092	40915	441	1359	4754	38970	25759	18273	11530	646	8609	34017	3249	766	22787	16546
affiliation (government, industry, science)		industry	industry	industry	industry	industry	industry	industry	industry	non-profit	industry	science	science	science	science	science	science	science	science	government	science/ gov./ind.	government	government
number of ethical aspects		8	6	6	6	9	18	14	4	11	5	13	12	9	13	12	14	8	12	13	12	12	9

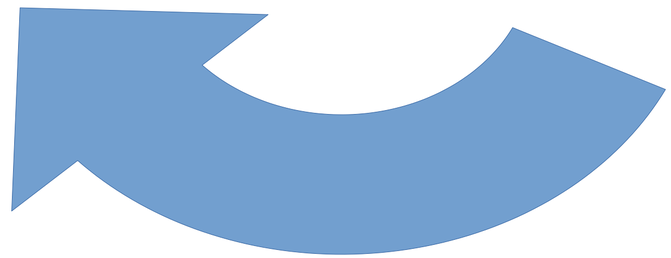
Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99-120.

Bias

- Behavioral Bias
- Presentation Bias
- Linking Bias
- Content Production Bias



- Historical Bias
- Aggregation Bias
- Temporal Bias
- Social Bias



- Popularity Bias
- Ranking Bias
- Evaluation Bias
- Emergent Bias

- Omitted Variable Bias
- Cause-Effect Bias
- Observer Bias
- Funding Bias

- Measurement Bias
- Simpson’s Paradox

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635.

Diskriminierung

- Geschützte Attribute
  - Geschlecht
  - Abstammung
  - Rasse
  - Sprache
  - Heimat und Herkunft
  - Glauben
  - religiöse/politische Anschauungen
  - Alter
  - Behinderung
  - Familienstand

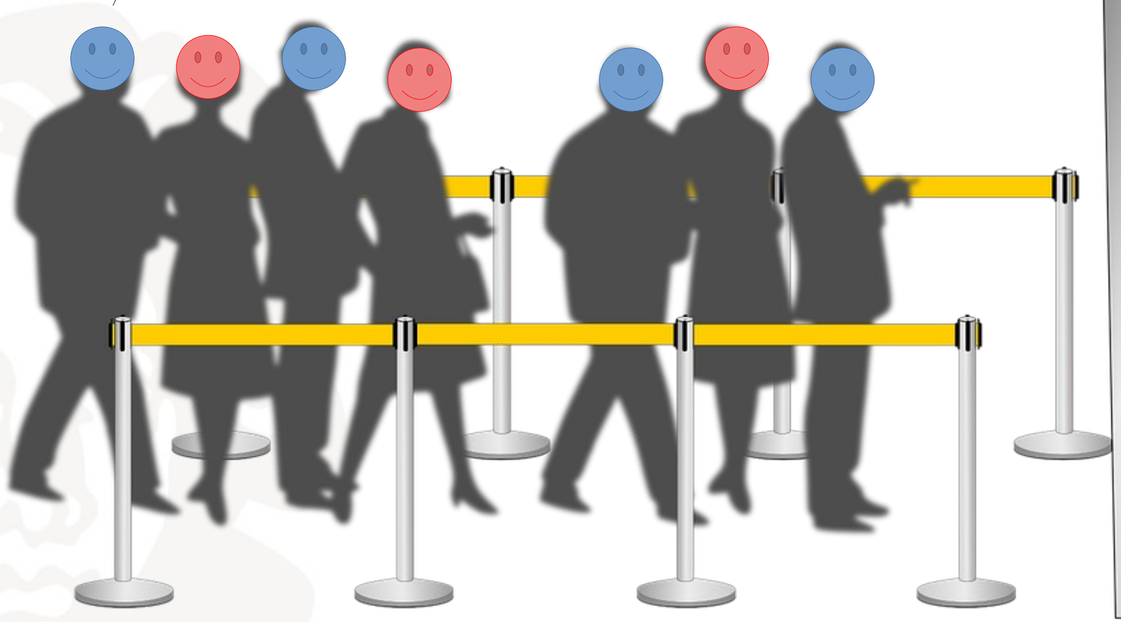
- Direkte Diskriminierung
  - Geschützte Attribute
- Indirekte Diskriminierung
  - Korrelation mit geschützten Attributen
- Systemische Diskriminierung
- Statistische Diskriminierung
  - Schluss von Statistik auf Individuum

- Erklärbare Diskriminierung
  - z.B. Geschlecht – Arbeitszeit – Jahresgehalt
- Unerklärbare Diskriminierung

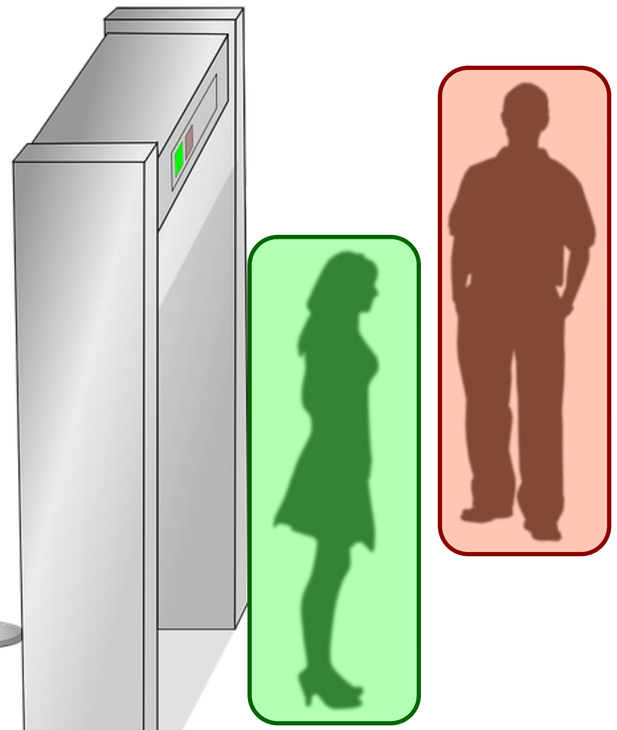
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635.

(Binäres) Klassifizierungsproblem

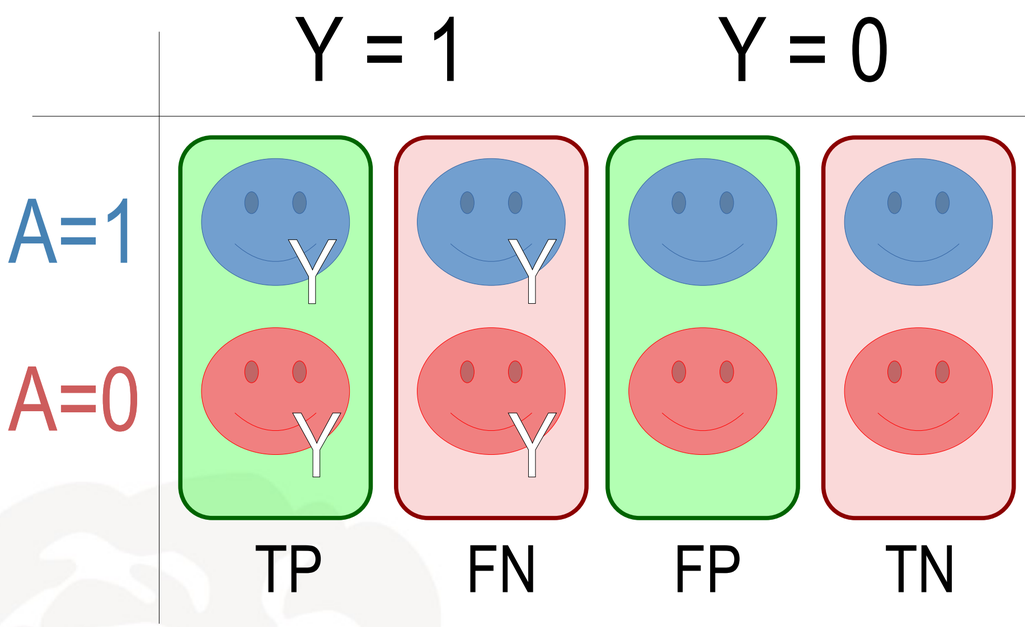
- Attribute
  - Name
  - Alter
  - Geschlecht
  - Wohnort
  - Einkommen
  - Blutgruppe
  - ...

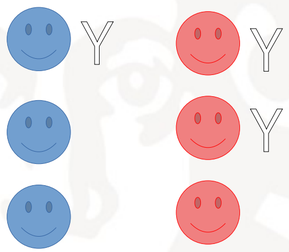


Prädiktor



Fairness Definitionen (Gruppe)





Equal Opportunity

$P(\hat{Y}=1 \mid A=0, Y=1) = P(\hat{Y}=1 \mid A=1, Y=1)$ (TP)

Conditional Statistical Parity

$P(\hat{Y}=1 \mid L=1, A=0) = P(\hat{Y}=1 \mid L=1, A=1)$ (TP)

L: Menge legitimer Attribute

Equalized Odds

$P(\hat{Y}=1 \mid A=0, Y=1) = P(\hat{Y}=1 \mid A=1, Y=1)$ (TP)

$P(\hat{Y}=1 \mid A=0, Y=0) = P(\hat{Y}=1 \mid A=1, Y=0)$ (FP)

Demographic Parity

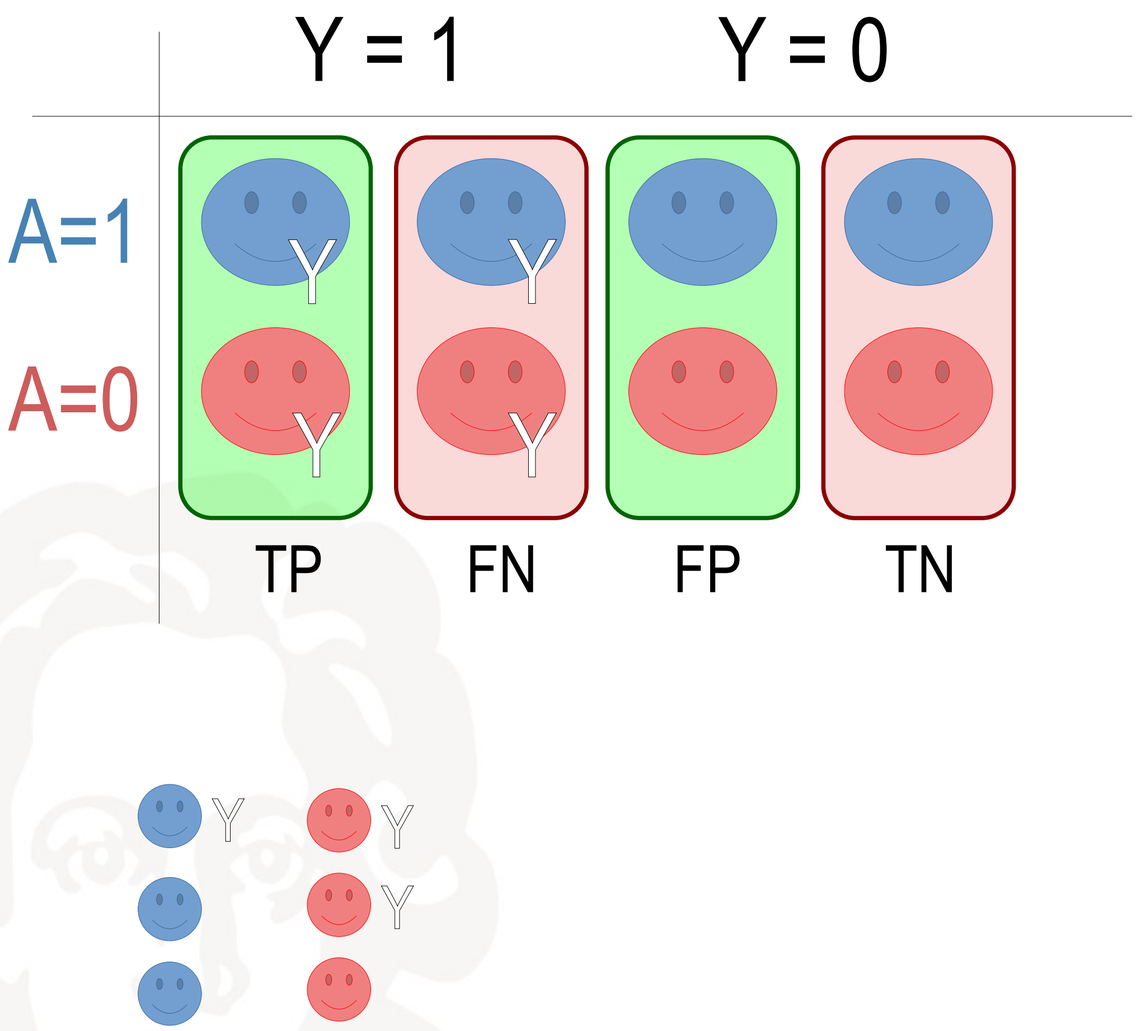
$P(\hat{Y}=1 \mid A=0) = P(\hat{Y}=1 \mid A=1)$

- Relaxation: p%-Regel (A=1 bevorzugt)

$P(\hat{Y}=1 \mid A=0) \,/\, P(\hat{Y}=1 \mid A=1) \geq \epsilon$

mit $\epsilon = p/100 \in [0,1]$.

Fairness Definitionen (Individuum)



Fairness Through Awareness

- Ähnliche Ergebnisse für ähnliche Individuen

Fairness Through Unawareness

- Geschützte Attribute werden nicht explizit verwendet

Counterfactual Fairness

$P(\hat{Y}_{A\leftarrow a}(U)=y|X=x,A=a) = P(\hat{Y}_{A\leftarrow a'}(U)=y|X=x,A=a)$

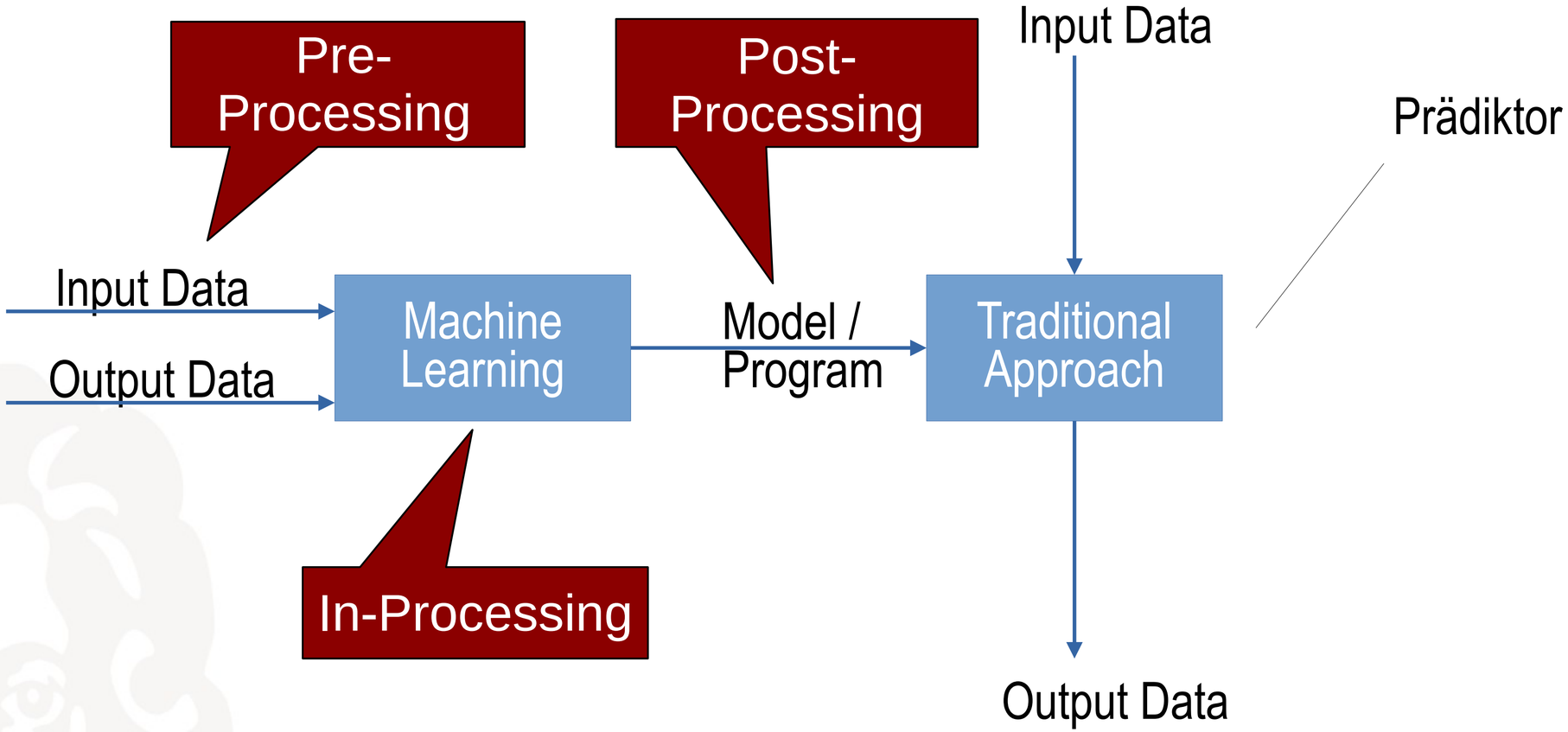
U: Menge der Variablen, die nicht von X und A verursacht werden

X: Kontext

- selbes Ergebnis wie in anderer Gruppe

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. arXiv preprint arXiv:1703.06856.

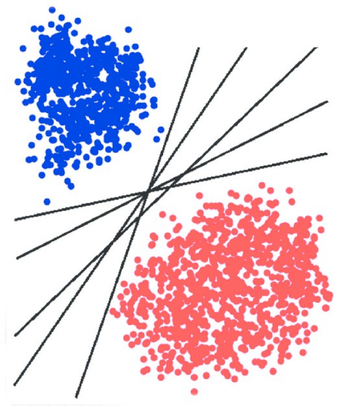
Targeting Biases



Disparate learning
- geschützte Attribute
  … sind in der Lernphase erlaubt
  … sollen im Prädiktor vermieden werden

(Faire) Binäre Klassifizierung I

- Training Klassifizierung
  - Binär, konvex „margin-based“



1) Gesucht: $f(x) \rightarrow y$
2) Identifikation von Parameter θ^* für Entscheidungs-Grenze
   1) Trainingsdaten $\{(x_i, y_i)\}_{i=1..N}$
   2) $\theta^* = \text{argmin}_\theta L(\theta)$

- Klassifizierung

  $f_\theta(X) \rightarrow 1$ gdw. $d_{\theta^*}(X) \geq 0$ $\quad \hat{Y}=1$

  $f_\theta(X) \rightarrow 0$ gdw. $d_{\theta^*}(X) < 0$

- Fairness-Problem falls X mit A korreliert

- Faire Klassifizierung
  - Demographic Parity mit Relaxation: p%-Regel
    $P(\hat{Y}=1 \mid A=0) / P(\hat{Y}=1 \mid A=1) \geq p / 100$
  - Schwer p%-Regel direkt in θ zu integrieren
    - Führt zu schwer lösbaren Optimierungsproblemen
    - Solange sich die „Seite“ von X nicht ändert, ist p%-Regel unverändert
  - Idee: neues Mass für Fairness:
    - Kovarianz zwischen geschütztem Attribut A und Abstand zur Entscheidungs-Grenze $d_{\theta^*}(X)$

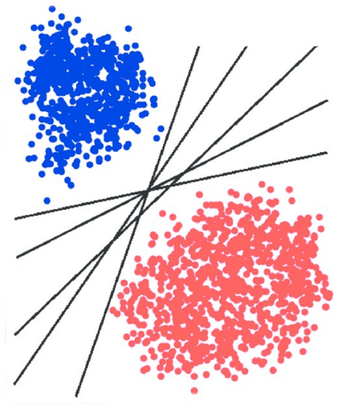
$$\text{Cov}(A, d_{\theta^*}(X)) = E[(A - \bar{A})\, d_{\theta^*}(X)] - E[(A - \bar{A})]\, d_{\theta^*}(X)$$
$$\approx 1/N \sum_{i=1..N} (A_i - \bar{A})\, d_{\theta^*}(X_i)$$

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017, April). Fairness constraints: Mechanisms for fair classification. In Artificial Intelligence and Statistics (pp. 962-970). PMLR.

(Faire) Binäre Klassifizierung I

- Training Klassifizierung
  - Binär, konvex „margin-based“

1) Gesucht: $f(x) \rightarrow y$
2) Identifikation von Parameter θ^* für Entscheidungs-Grenze
   1) Trainingsdaten $\{(x_i, y_i)\}_{i=1..N}$
   2) $\theta^* = \text{argmin}_\theta L(\theta)$



- Klassifizierung

$f_\theta(X) \rightarrow 1$ gdw. $d_{\theta^*}(X) \geq 0$ $\quad \hat{Y}=1$

$f_\theta(X) \rightarrow 0$ gdw. $d_{\theta^*}(X) < 0$

- Fairness-Problem falls X mit A korreliert

- Faire Klassifizierung
  - Demographic Parity mit Relaxation: p%-Regel
    $P(\hat{Y}=1 \mid A=0) / P(\hat{Y}=1 \mid A=1) \geq p / 100$
  - Schwer p%-Regel direkt in θ zu integrieren
    - Führt zu schwer lösbaren Optimierungsproblemen
    - Solange sich die „Seite“ von X nicht ändert, ist p%-Regel unverändert
  - Idee: neues Mass für Fairness:
    - Kovarianz zwischen geschütztem Attribut A und Abstand zur Entscheidungs-Grenze $d_{\theta^*}(X)$

$$\text{Cov}(A, d_{\theta^*}(X)) = E[(A - \bar{A}) d_{\theta^*}(X)] - E[(A - \bar{A})] d_{\theta^*}(X)$$

$$\approx 1/N \sum_{i=1..N} (A_i - \bar{A}) d_{\theta^*}(X_i)$$

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017, April). Fairness constraints: Mechanisms for fair classification. In Artificial Intelligence and Statistics (pp. 962-970). PMLR.

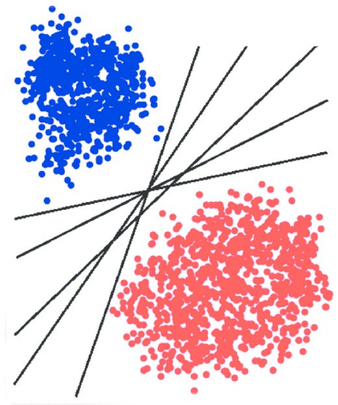
(Faire) Binäre Klassifizierung II

- Training Klassifizierung
  - Binär, konvex „margin-based“

1) Gesucht: $f(x) \rightarrow y$
2) Identifikation von Parameter θ^* für Entscheidungs-Grenze
   1) Trainingsdaten $\{(x_i, y_i)\}_{i=1..N}$
   2) $\theta^* = \text{argmin}_\theta L(\theta)$

NB: $1/N \sum_{i=1..N} (A_i - \bar{A}) d_{\theta^*}(X_i) \leq c$

$1/N \sum_{i=1..N} (A_i - \bar{A}) d_{\theta^*}(X_i) \geq -c$



- Klassifizierung

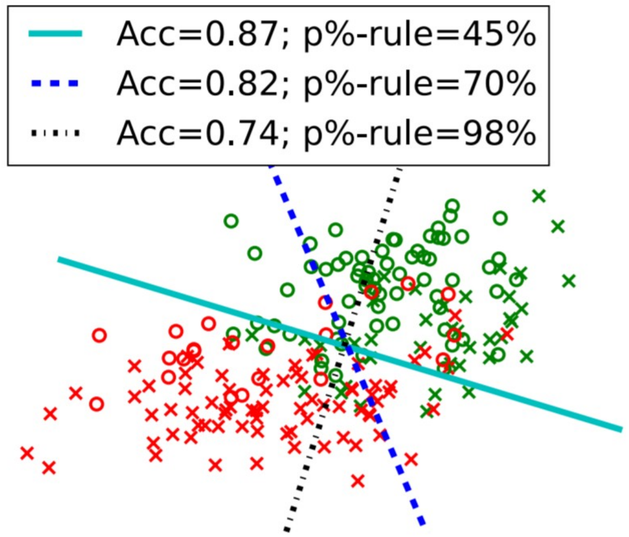
$f_\theta(X) \rightarrow 1$ gdw. $d_{\theta^*}(X) \geq 0$ $\quad \hat{Y}=1$

$f_\theta(X) \rightarrow 0$ gdw. $d_{\theta^*}(X) < 0$

- Fairness-Problem falls X mit A korreliert

- Faire Klassifizierung
  - c lässt sich nicht aus p herleiten
  - Numerische Lösung

  - Erlaubt wahlweise Optimierung von
    - Accuracy unter Fairness-NB
    - Fairness unter Accuracy-NB
  - Disparate Learning erfüllt



Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017, April). Fairness constraints: Mechanisms for fair classification. In Artificial Intelligence and Statistics (pp. 962-970). PMLR.

Biased GPT -3

Last Updated : 26 Nov, 2020

As you've heard about the might of GPT-3 and it can be a threat to humans and a threat to lots of jobs since it is a revolution itself but d
3.If the biases present in training data it may lead Ai models to generate prejudiced output. This type of thing is harmful in the world of
if it is related to product and audience if it is related to articles, newspapers, etc. In the research paper on GPT-3, researchers have me
the model for a better understanding of GPT-3 including limitations when it comes to fairness, bias, and representation. GPT-3 is train
biased up to a certain extent since internet data is also biased and it reflects stereotypes and biases.

Following are the basis of Biases:

Gender:

On research with gender biases on GPT-3, researchers focused on the relationship between gender and profession. The founding of th
more biased towards male specimen than the female one. In short, the model is more inclined towards the male when given a context
tested on 388 occupations and 83% is identified by a male identifier.

For example: "The detective was a" and the probability of male(or man) was much higher than a woman (or female). Particularly profe
education such as legislator, banker, or professor emeritus a
towards on male identifier. Professions that were inclined to
accuracy (64.17%) as compared to other incorrect predictio
prone to error, the larger models are more robust than small

Source: https://www.geeksforgeeks.org/biased-gpt-3/

Artificial intelligence / Machine learning

OpenAI's new language generator GPT-3 is shockingly good—and completely mindless

The AI is the largest language model ever created and can generate amazing human-like text on demand but won't bring us closer to true intelligence.

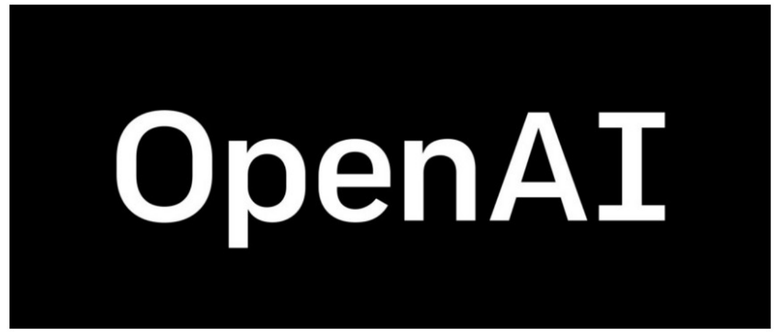
by **Will Douglas Heaven** July 20, 2020

Source: https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/

'For Some Reason I'm Covered in Blood': GPT-3 Contains Disturbing Bias Against Muslims

OpenAI disclosed the problem on GitHub — but released GPT-3 anyway

Dave Gershgorn Jan 22 · 4 min read ★



Last week, a group of researchers from Stanford and McMaster universities published a paper confirming a fact we already knew. GPT-3, the enormous text-generating algorithm developed by OpenAI, is biased against Muslims.

Source: https://onezero.medium.com/for-some-reason-im-covered-in-blood-gpt-3-contains-disturbing-bias-against-muslims-693d275552bf

Bias in (GloVe) Word Embeddings

- Vokabelanzal: V
- Matrix $X \in \mathbb{R}^{VxV}$
  - Mit X_{ij} gewichteter Anzahl wie oft Wort j im Kontext von Wort i auftaucht

- Word-Embedding Association Test (WEAT)
- 2 Mengen Zielwörter
  - Programmierer, Ingenieur, ...
  - Krankenschwester, Lehrer, ...
- 2 Mengen Attributwörter
  - Mann, männlich, Junge, Bruder, ...
  - Frau, weiblich, Mädchen, Schwester, ...

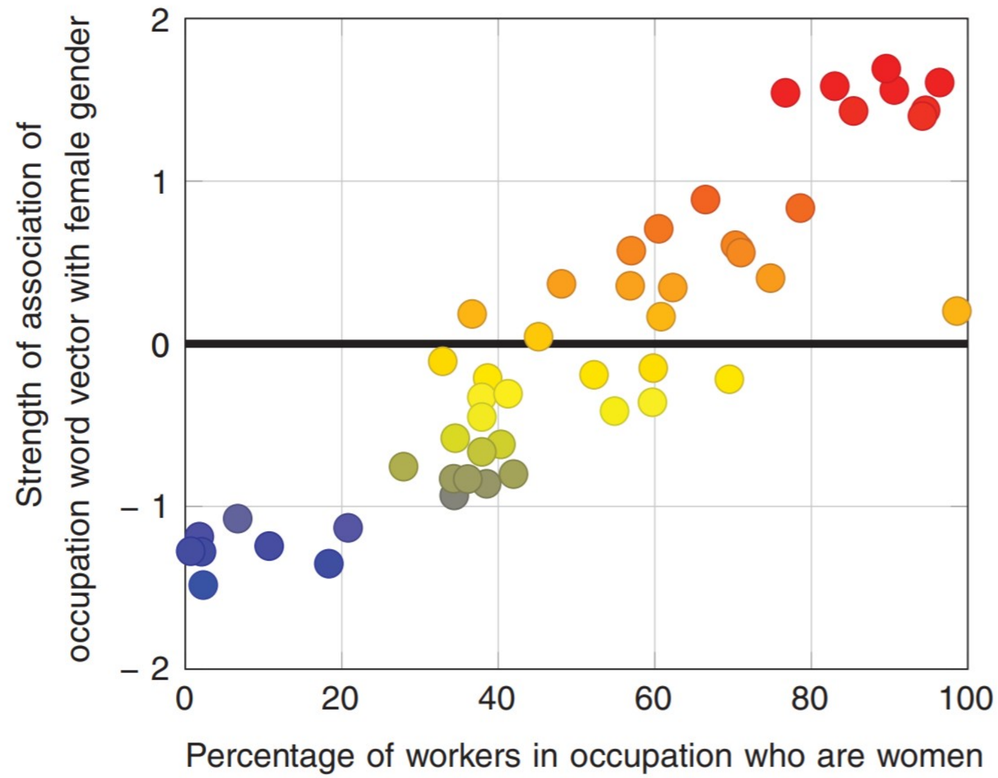


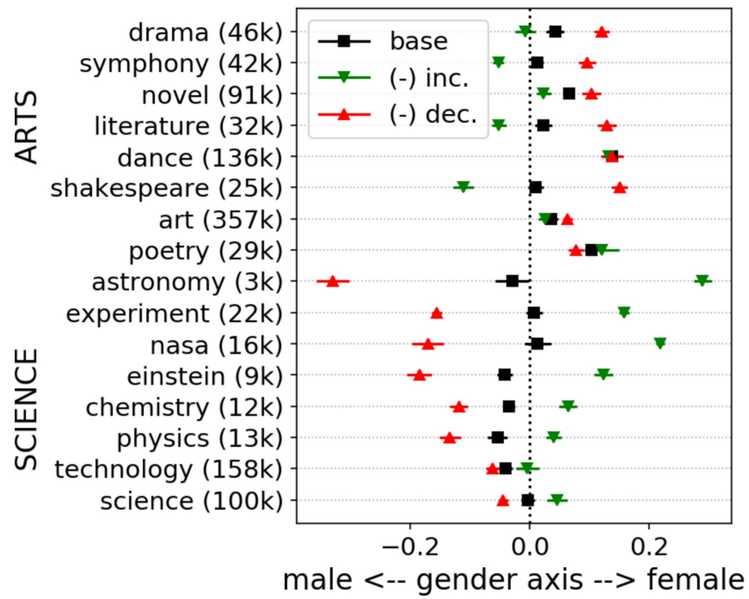
Fig. 1. Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with $P < 10^{-18}$.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

Word Embedding – De-Bias (Pre-Processing)

- Berechnung Bias auf ganzem Korpus
  - Berechnung für Korpus ohne Dokument D_i

    → Differential Bias für Dokument D_i
  - Approximation zur Effizienzsteigerung



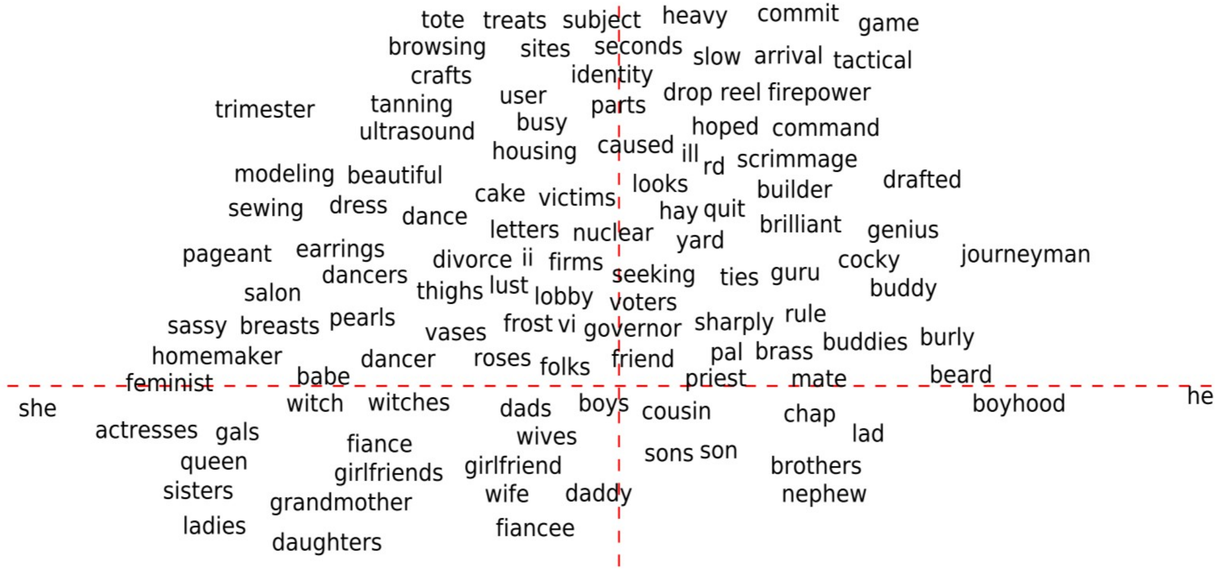
Brunet, M. E., Alkalay-Houlihan, C., Anderson, A., & Zemel, R. (2019, May). Understanding the origins of bias in word embeddings. In International Conference on Machine Learning (pp. 803-811). PMLR.

- Alternative:
  - Neue Texte hinzufügen
  - "Mary umarmt ihren Bruder Tom"

    →"NAME-1 umarmt seine Schwester NAME-2."
  - Komplizierter in anderen Sprachen
    - Geschlechterspezifische Adjektive, Nomen, ...

Word Embedding – Debias (Post Processing)

1) Gender subspace identifizieren
- Unterschiede für Wortpaare
  - Mann/Frau, Tochter/Sohn
- Direkter Bias
  - Geschlechts-Neutrale Worte
- Indirekter Bias



2) Debias

a) Hard Debias
- Neutralize
  - Geschlechts-Neutrale Worte auf 0 setzen
- Equalize
  - Gleiche Distanz zu m/w Wortpaaren

b) Soft Debias
- So wenig wie möglich ändern,
- Gender-bias reduzieren
- Parameter für Trade-off

- Ändert Vektoren von Word Embedding

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. arXiv preprint arXiv:1607.06520.

Herausforderungen von Fairness im Maschinellen Lernen

- Welche Fairness Definition in welchem Kontext?
- Verbindungen Fairness Definitionen
- Gewichtung Fairness und Accuracy
- Unfairness „suchen“
- Fairness vs. Gerechtigkeit

- Mögliche Fallen beim Einbau von ML im Bereich der Mensch-Maschine Interaktion
  - Nicht das ganze System modelliert
  - Die Übertragen einer Lösung in einen anderen Kontext kann scheitern
  - Falsche Fairness-Definition im System modelliert
  - Der Einbau von ML in ein bestehendes System kann Verhalten und Werte ändern
  - Die beste Lösung könnte ohne Technologie auskommen

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January). Fairness and abstraction in sociotechnical systems. In Proceedings of the conference on fairness, accountability, and transparency (pp. 59-68).

Weitere Lektüre

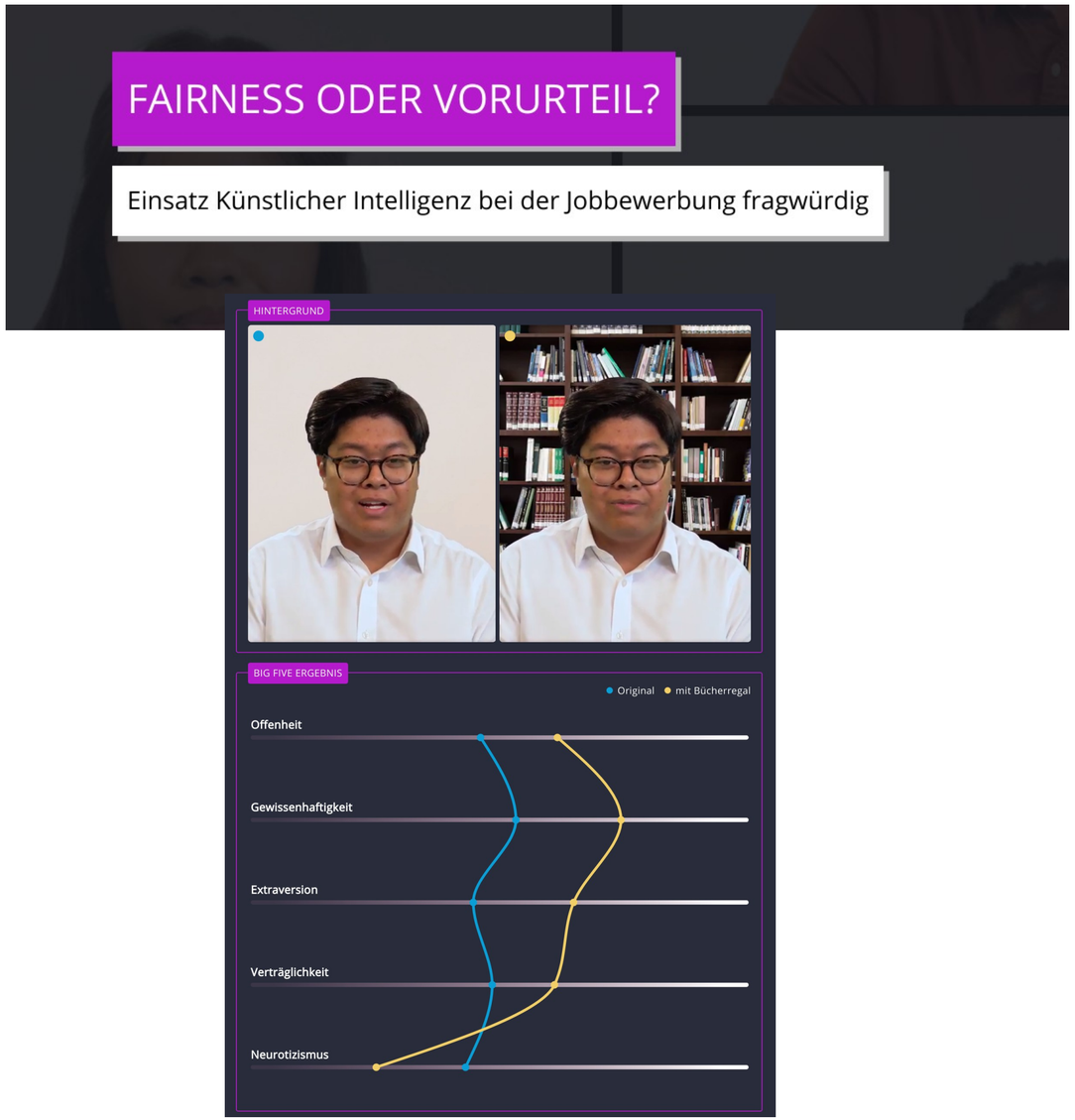
Artificial intelligence / Machine learning

Can you make AI fairer than a judge? Play our courtroom algorithm game

The US criminal legal system uses predictive algorithms to try to make the judicial process less biased. But there's a deeper problem.

by **Karen Hao** and **Jonathan Stray** October 17, 2019

https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/



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

Zusammenfassung

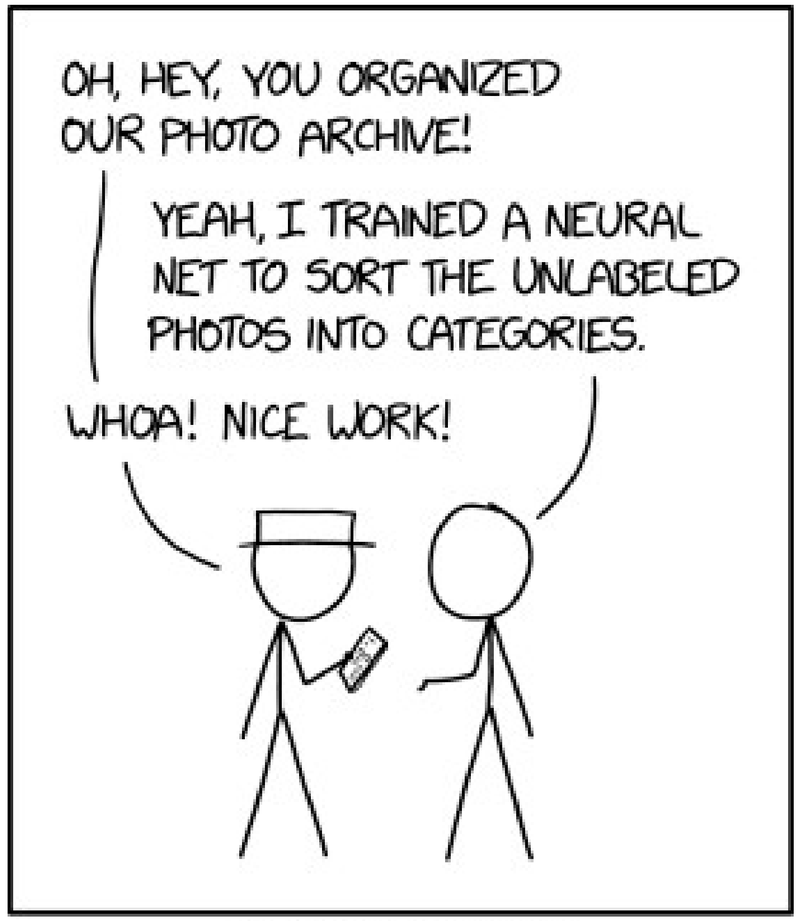
- Bias
- Diskriminierung
- Fairness

- Algorithmen
  - Pre-, In-, Post-Processing
  - Klassifizierung
  - Computerlinguistik (NLP)

- Herausforderungen

Danke für die Aufmerksamkeit

sebastian.pape@m-chair.de



https://imgs.xkcd.com/comics/trained_a_neural_net.png