



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Prof. Dr. Judith Simon

Vertrauenswürdige Künstliche Intelligenz? Ethik in Design, Entwicklung & Anwendung von KI

18.09.2024 | KonKIS 24 - Konferenz der deutschen KI-Servicezentren 2024 | Göttingen |



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

stability.ai

Models Deployment Company News 日本語

Activating humanity's potential through generative AI

Open models in every modality, for everyone, everywhere.

Get started with Sora

Get started with API

Creating safe AGI that benefits all of humanity

Learn about OpenAI

Play video

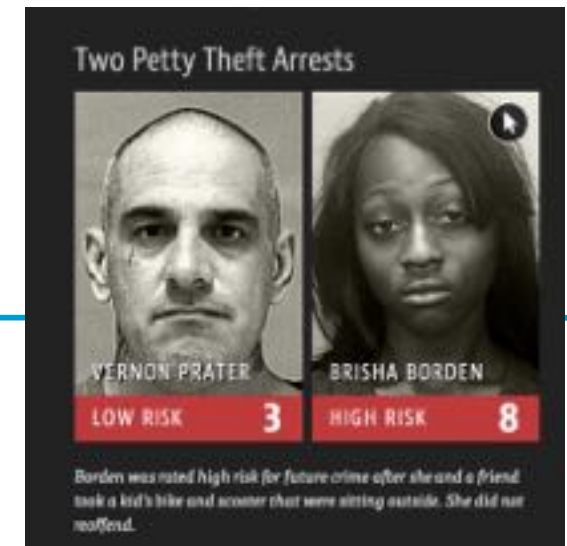
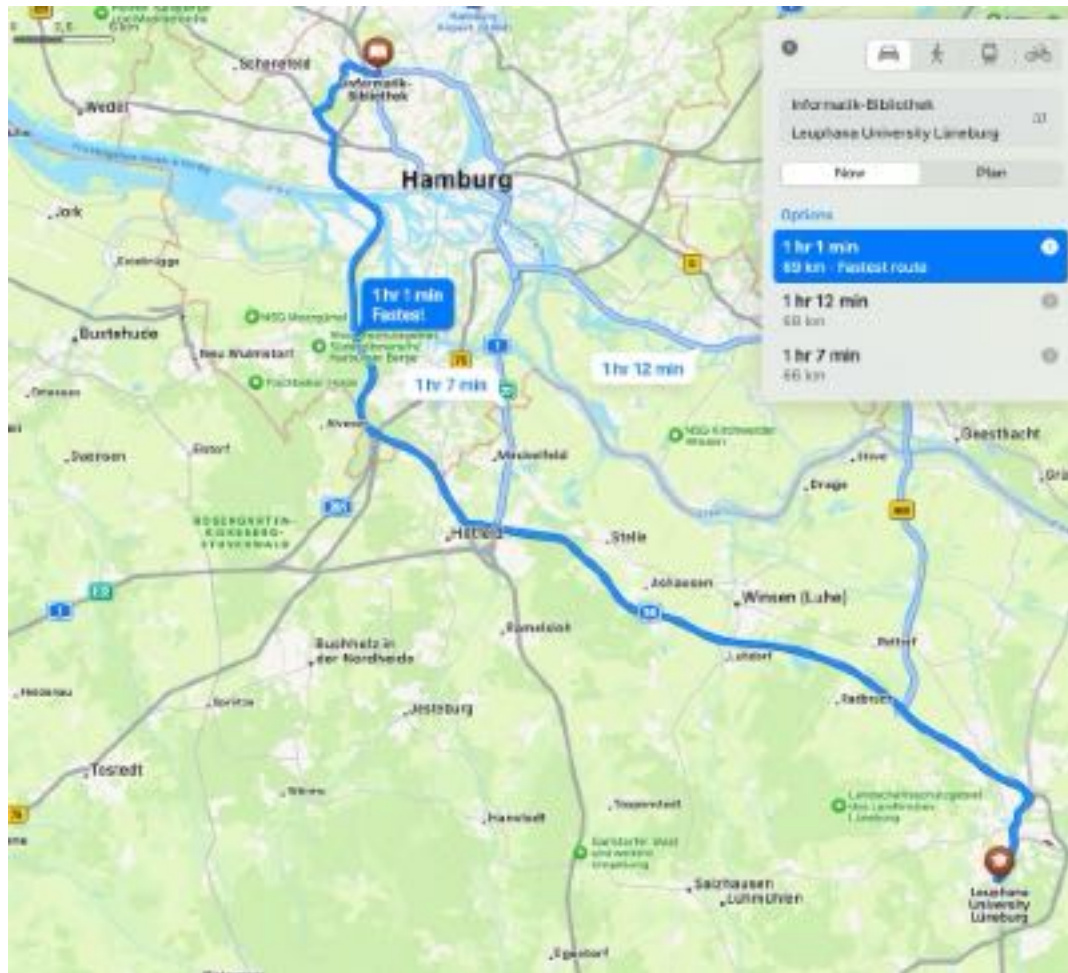
- <https://openai.com>
- <https://openai.com/sora>
- <https://stability.ai>



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Google amazon



- 1) Apple, Facebook Google, Amazon
- 2) Propublica

3) <https://www.mdr.de/wissen/mensch-alltag/die-verschiedenen-plaene-der-laender-fuer-eine-corona-app-100.html>



Künstliche Intelligenz: lange Geschichte mit Höhen und Tiefen

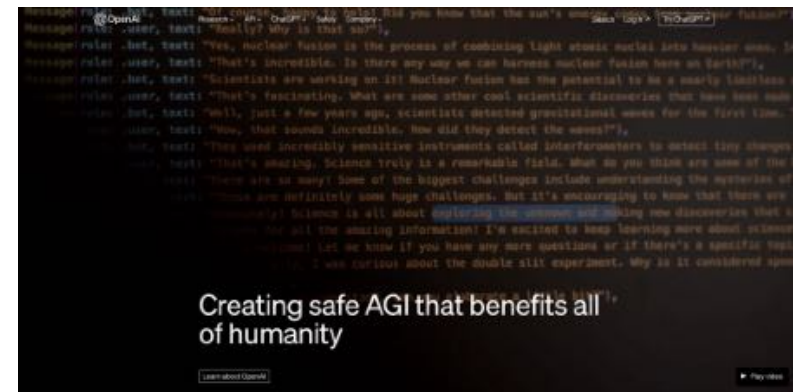
- Historisch: Simulation von Verhalten, welches bei Menschen als intelligent interpretiert würde
- Heute: starker Fokus auf Methoden des maschinellen Lernens



Bernhard Christoph Francke

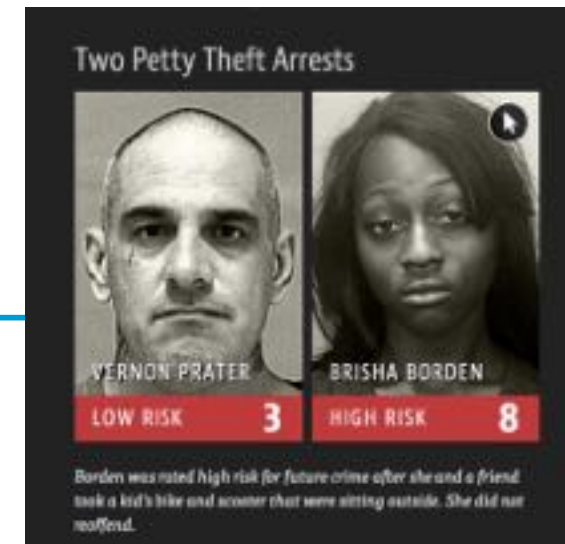


Turing/Public Domain



Creating safe AGI that benefits all
of humanity

<https://openai.com>

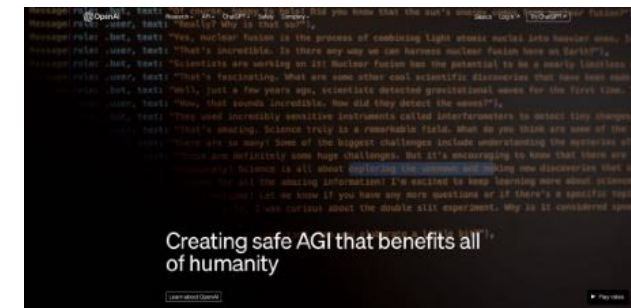


Daten, Künstliche Intelligenz & Ethik

- Kern heute zumeist: Analyse großer Datenmengen

> Mustererkennung, Klassifikation, Prognose, Entscheidungssupport

- Generative KI
- Mannigfaltigkeit, Komplexität & Dynamik
 1. der Technologien
 2. der Entwicklungs- und Anwendungskontexte
 3. der beteiligten Akteure
 4. der ethischen Fragen
 5. der notwendigen Regulierung



→ Ökosystemperspektive auf Daten & KI

- 1) Apple, Facebook Google, Amazon, OpenAI
- 2) Propublica
- 3) <https://www.mdr.de/wissen/mensch-alltag/die-verschiedenen-plaene-der-laender-fuer-eine-corona-app-100.html>



KI & ML: Wissen, Vertrauen & Verantwortung

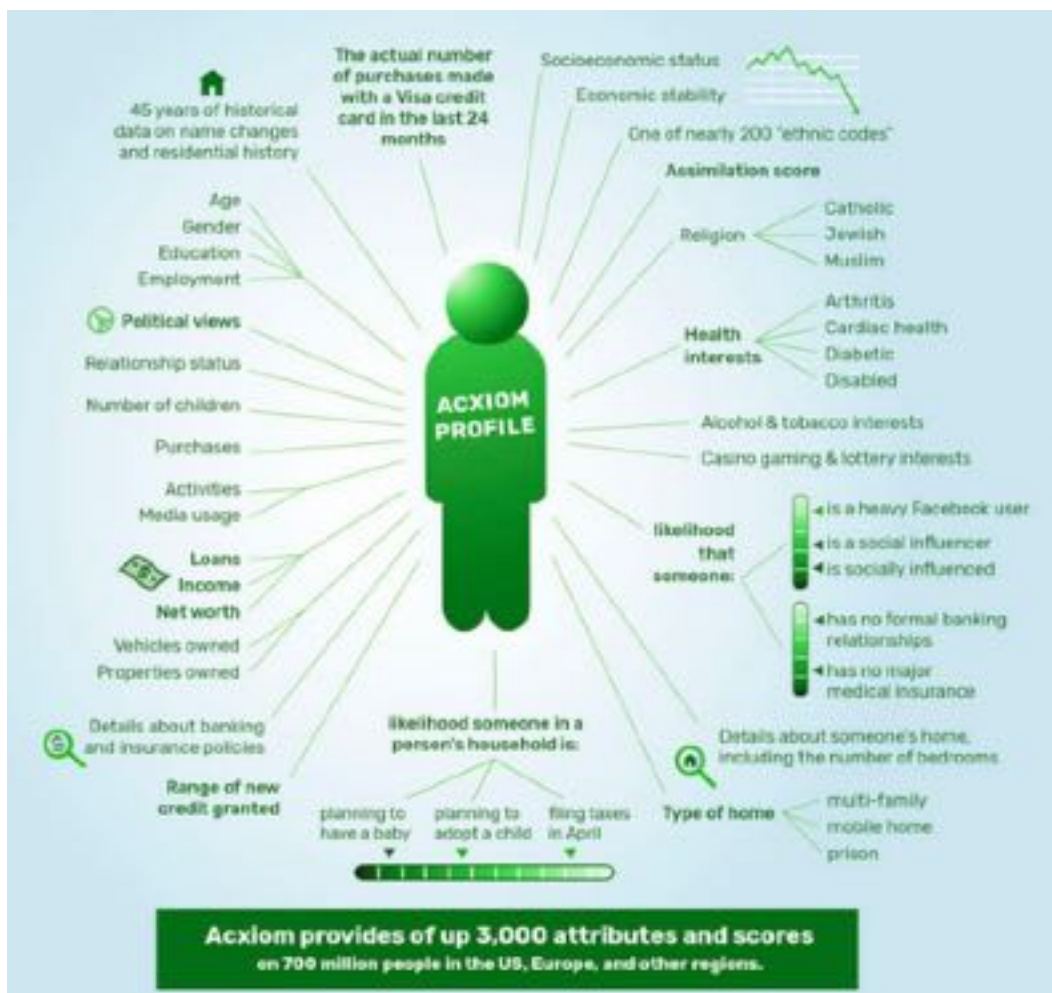
- Wir verlassen uns auf KI/ML in vielen Bereichen unseres Lebens
- Können wir KI auch vertrauen – und sollten wir es tun?
- Was ist ein verantwortlicher Umgang mit KI/ML und welche Fallstricke gibt es?



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

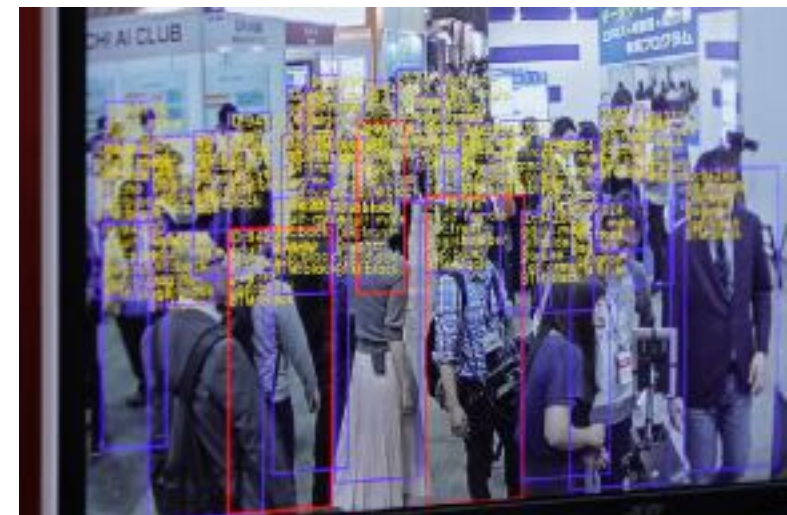
Ethische Probleme mit KI



PREDICTING PERSONAL ATTRIBUTES FROM FACEBOOK LIKES

PREDICTED ATTRIBUTE	ACCURACY
Ethnicity	95 %
Gender	93 %
Sexual Orientation (male)	89 %
Political Views	85 %
Religion	82 %
Sexual Orientation (female)	75 %
Nicotine Usage	73 %
Alcohol Usage	70 %
Relationship	67 %
Drug Usage	65 %
Parents (divorced)	60 %

Predicting personal attributes from Facebook Likes, Source: Kazinski et al 2013.



Source: Christl (2017)

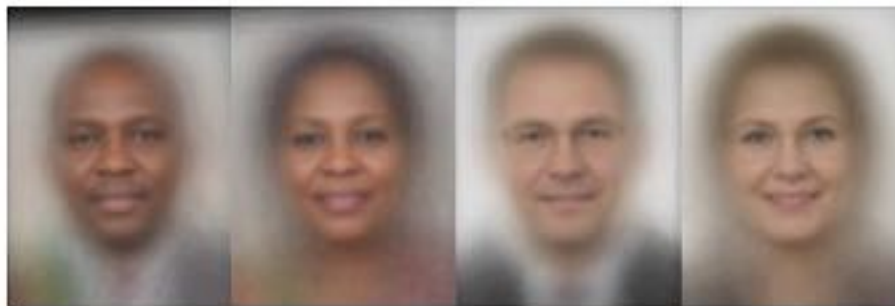
<https://fortune.com/2018/04/09/sensetime-alibaba-ai-startup-600-million/>



Schutz der Privatsphäre

+ Schutz vor Diskriminierung, Gerechtigkeit & Transparenz?

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	89.3%	65.5%	99.2%	94.0%	33.8%
IBM	86.0%	65.3%	99.7%	92.8%	34.4%



© MIT Media Lab

Source: MIT Media Lab

Epistemische Dimension

+

Ethische Dimension



<https://fortune.com/2018/04/09/sensetime-alibaba-ai-startup-600-million/>



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Schutz vor Diskriminierung, Gerechtigkeit & Transparenz?



Home News Sport Business Innovation Culture Travel Earth Video Live

Amazon scrapped 'sexist AI' tool



10 October 2018

Show more



The algorithm reported bias towards men, reflected in the technology's

Two Petty Theft Arrests

	
VERNON PRATER	BRISHA BORDEN
LOW RISK 3	HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

NEWS - 24 OCTOBER 2019 - UPDATE 26 OCTOBER 2019

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

Heidi Ledford



Black people with complex medical needs were not likely to be equally as white people to be referred to programmes that provide more personalised care. Credit: iStockphoto/Redux/istock

PDF version

RELATED ARTICLES

Aid for way forward for AI in health care

Bias detects the researchers striving to make algorithms fair

Can we open the black box of AI?

SUBJECTS

Computer science Health care

Society

An algorithm widely used in US hospitals to allocate health care to patients has been systematically discriminating against black people, a sweeping analysis has found.

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

<https://www.nature.com/articles/d41586-019-03228-6>

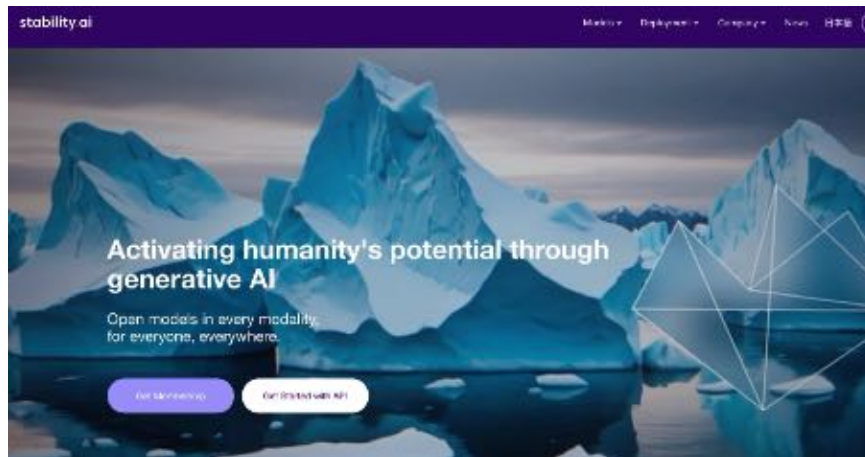
<https://www.bbc.com/news/technology-45809919>



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

(Neue) Herausforderungen durch ChatGPT & Co?



<https://openai.com>

<https://openai.com/sora>

<https://stability.ai>

Täuschung I: KI oder Mensch?

```
Welcome to

EEEEEE LL      IIII  ZZZZZZ  AAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LL      II      ZZ  AAAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LLLLLL  IIII  ZZZZZZ  AA  AA

Eliza is a sock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:  Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:  They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:  Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:  He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:  It's true, I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```



https://de.wikipedia.org/wiki/ELIZA#/media/Datei:ELIZA_conversation.jpg

Open.ai

Täuschung II: KI Hype

The Google engineer who thinks the company's AI has come to life

AI ethicists warned Google not to impersonate humans. Now one of Google's own thinks there's a ghost in the machine.



By Nisha Ilika

June 11, 2022 at 8:00 a.m. EDT



Google engineer Blake Lemoine. (Martin Kinski/For the Washington Post)

<https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/>

Täuschung III: Inhalte

Der falsche Christian Sievers

VON AXEL WEIDMANN · AKTUALISIERT AM 22.08.2019 · 39 SE

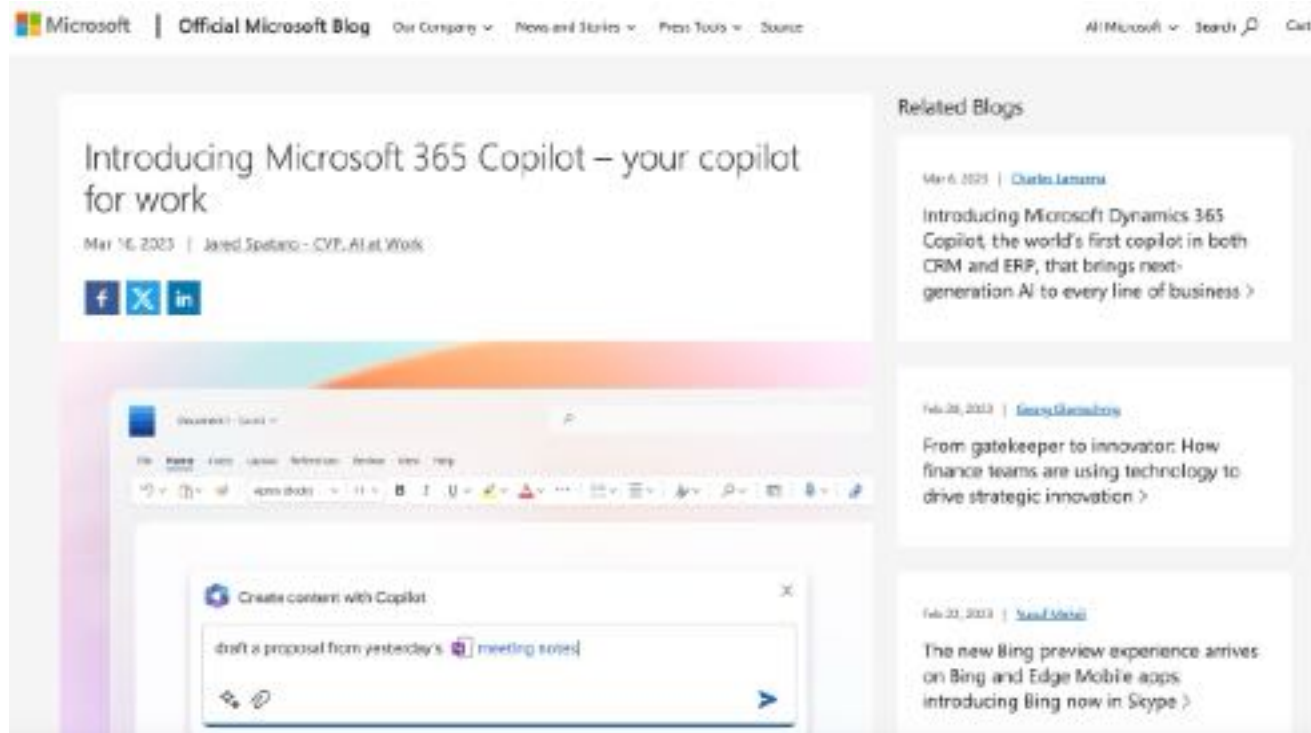


ZDF-Moderator Christian Sievers für zweifelhafte
ch das Video ist ein Fake, das dem Nachrichtenmann
legt.

<https://www.zdf.de/nachrichten/panorama/prominente/papst-daunenjacke-fake-ki-kuenstliche-intelligenz-100.html>

<https://www.faz.net/aktuell/feuilleton/medien/deepfake-video-von-zdf-moderator-christian-sievers-19193736.html>

Täuschung IV: Kontextkollaps



<https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/>



Zwischenfazit: Grundlegende Probleme

- Das Gerechtigkeitsproblem
 - Gesellschaftliche Stereotypen und Vorurteile, aber auch Ungleichheiten und Ungerechtigkeiten in der Gesellschaft werden oftmals – beabsichtigt oder unbeabsichtigt – in Technologien eingeschrieben (durch Trainingsdaten oder methodische Entscheidung)
 - Insbesondere datenbasierte Systeme laufen Gefahr, eine ungerechte Vergangenheit in die Zukunft fortzuschreiben



Zwischenfazit: Grundlegende Probleme

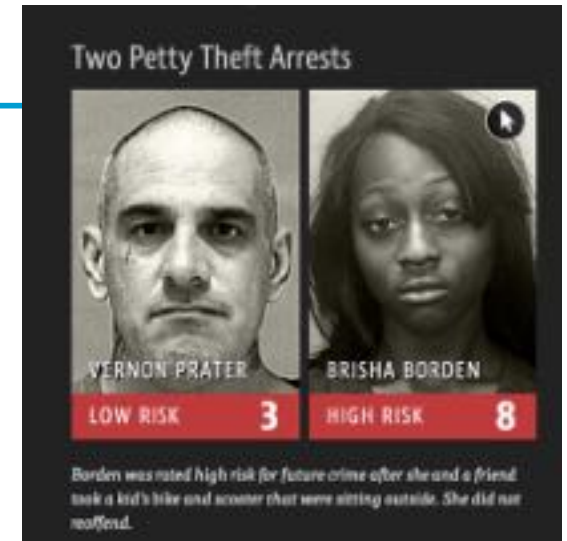
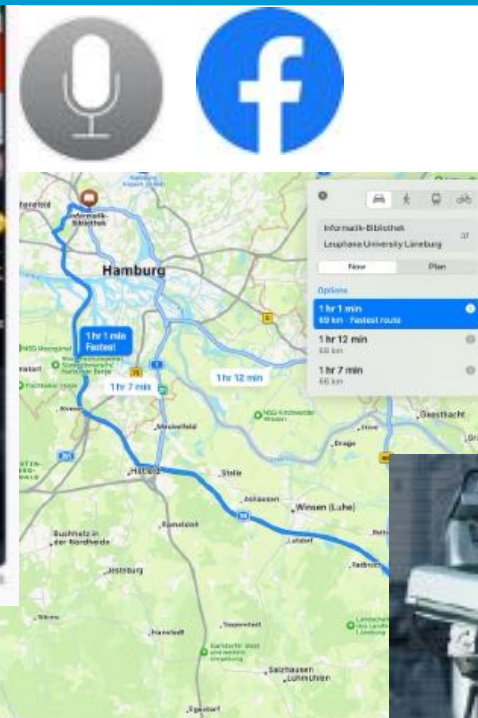
- Das dreifache Transparenzproblem
 - Funktional: Kein Zugang zur Software aufgrund von Geschäftsgeheimnissen
 - Epistemisch: Blackbox Deep Learning, Intransparenz Maschinellen Lernens
 - Kompetenz: wer hat die notwendige Kompetenz, um Software zu prüfen, selbst wenn Transparenz gegeben ist?



Zwischenfazit: (Neue) Herausforderungen durch Generative KI

- Alle herkömmlichen Herausforderungen durch KI
 - Schutz der Privatsphäre/Datenschutz, Bias/Diskriminierung, mangelnde Transparenz, Nachvollziehbarkeit & Kontrolle, Energieverbrauch, prekäre Arbeitsbedingungen, ...
- Plus: Problem der vierfachen Täuschung
- Weitreichende und versteckte Probleme durch Wiederverwendung der gleichen Komponenten in verschiedenen Anwendungen

Vertrauen in KI?



- <https://openai.com>
- <https://openai.com/sora>
- <https://stability.ai>

- 1) Apple, Facebook Google, Amazon
- 2) Propublica

- 3) <https://www.mdr.de/wissen/mensch-alltag/die-verschiedenen-plaene-der-laender-fuer-eine-corona-app-100.html>



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Vertrauen & Vertrauenswürdigkeit



"Whatever matters to human beings, trust is the atmosphere in which it thrives."

Bok (1978), in Baier (1986)

" Exploitation and conspiracy, as much as justice and fellowship, thrive better in an atmosphere of trust.

"Trust then [...] is accepted vulnerability to another's possible but not expected ill will (or lack of good will) toward one."

Baier (1986) : Trust and Antitrust



- Vertrauen
 - als akzeptierte Vulnerabilität
 - Vertrauen <> Sicherheit, Gewissheit
 - als relationales Konzept: A vertraut B in Bezug auf X
 - Ich vertraue meiner Ärztin in Bezug auf medizinische Diagnose, nicht Reparatur meines Autos...
 - beschreibt sehr verschiedene Relationen in
 - Personen (Partner, Kind, Fremde, Hausärztin..)
 - Institutionen (Bundesregierung, STIKO, ÖRR ...)
 - abstrakte Entitäten (die Wissenschaft, die Politik, die Medien, ..)
 - **Technologien? Künstliche Intelligenz?**
 - ...



- Vertrauen, Vertrauenswürdigkeit und Wissen/schaft

“Modern knowers cannot be independent and self-reliant, not even in their own fields of specialization. In most disciplines, those who do not trust cannot know; those who do not trust cannot have the best evidence for their beliefs. In an important sense, then, trust is often epistemologically even more basic than empirical data or logical arguments: the data and the argument are available only through trust. If the metaphor of foundation is still useful, the trustworthiness of members of epistemic communities is the ultimate foundation for much of our knowledge.”

Hardwig (1991). The Role of Trust in Knowledge



-
- Vertrauen, Vertrauenswürdigkeit und Wissen/schaft
 - Wissenschaftler müssen sich gegenseitig vertrauen in Bezug auf ihre
 - Kompetenz
 - Ehrlichkeit
 - Angemessene epistemische Selbsteinschätzung
 - Epistemische und moralische Komponenten von Vertrauen und Vertrauenswürdigkeit

Hardwig (1991). The Role of Trust in Knowledge



Universität Hamburg

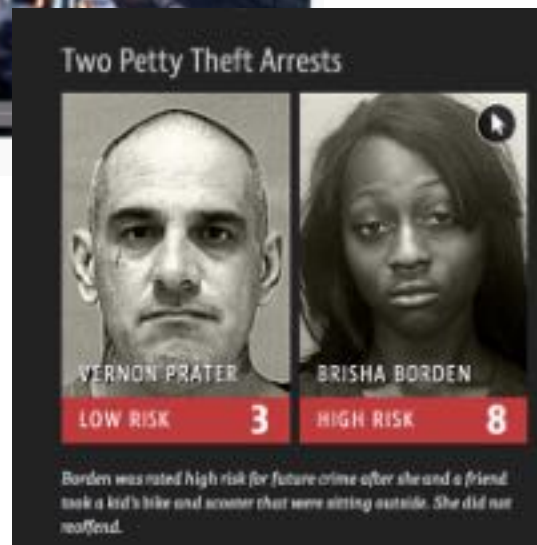
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Vertrauen in KI – Vertrauenswürdige KI?

Vertrauen in KI – Vertrauenswürdige KI?



A woman walks down a busy street at night, wearing a long red dress and a black jacket. The street is wet and reflective, with many neon signs and lights in the background.



<https://openai.com/sora>

Propublica.com

<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>



Vertrauen in KI – Vertrauenswürdige KI?

- Können wir KI vertrauen?
 - Vertrauen im starken Sinn bedarf einer moralischen Komponente
 - nur gegeben wenn man KI als sozio-technisches Ökosystem versteht
- Sollten wir KI vertrauen?
 - Nur wenn KI Vertrauenswürdig ist → epistemische und moralische Anforderungen

Vertrauen in KI – Vertrauenswürdige KI?

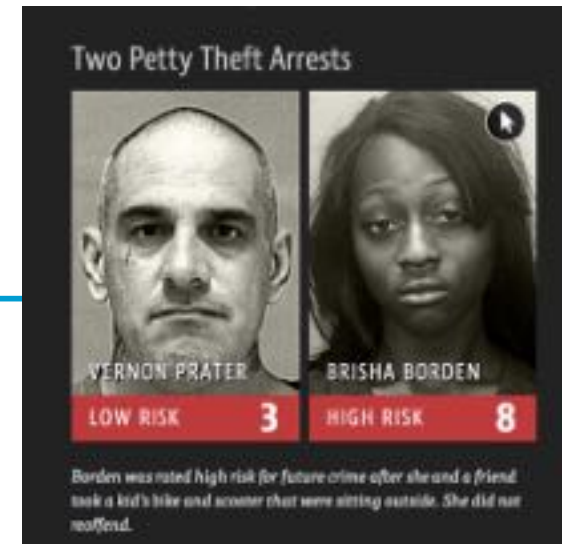


Abbildung 1: Die Leitlinien als Rahmen für eine vertrauenswürdige KI

Vertrauen in KI – Vertrauenswürdige KI?

- 4 Ethische Grundsätze im Kontext von KI-Systemen
 1. Achtung der menschlichen Autonomie
 2. Schadensverhütung
 3. **Fairness**
 4. Erklärbarkeit





- Discrimination-aware data mining vs. fair ML
 - Unterschiedliche Methoden, um Diskriminierung nachzuweisen und zu verhindern/minimieren
 - Accuracy equity, conditional accuracy equity, equality of opportunity, ...
 - Welche Personen/Gruppen will ich (priorisiert) schützen?
 - Schutz vor Diskriminierung oder aktiver Einbau von Fairness?
→ Auswahl, Umsetzung & Begründung bedingt mathematisch-technisch wie ethische Expertise

Vertrauen in KI – Vertrauenswürdige KI?

- 4 Ethische Grundsätze im Kontext von KI-Systemen
 1. Achtung der menschlichen Autonomie
 2. Schadensverhütung
 3. Fairness
 4. Erklärbarkeit





Picturing the original image (left), saliency map using a method called Grad-CAM (middle), and another using Guided Backpropagation (right). The picture above is the canonical example for "class-discrimination". The above saliency maps are taken from <https://github.com/kazuto1211/grad-cam-pytorch>.

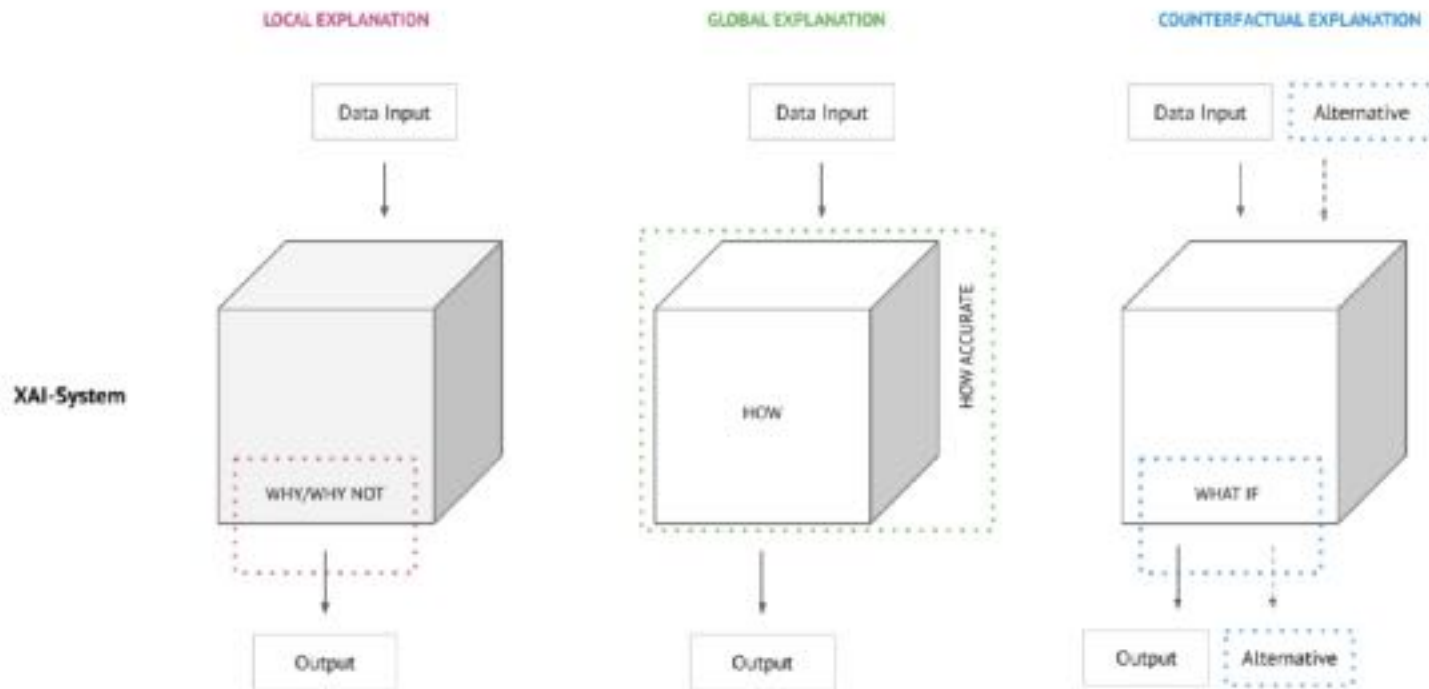
<https://towardsdatascience.com/what-explainable-ai-fails-to-explain-and-how-we-fix-that-1e35e37bee07>



„Who needs to understand what in a given scenario? *What can* be explained about the system in use? *What should* explanations look like in order to be *meaningful* to affected users?“

Asghari, et al. (2021). "What to explain when explaining is difficult? An interdisciplinary primer on XAI and meaningful information in automated decision-making." <https://doi.org/10.5281/zenodo.6375784>.

„Who needs to understand what in a given scenario? What can be explained about the system in use? What should explanations look like in order to be *meaningful* to affected users?“



Asghari, et al. (2021). "What to explain when explaining is difficult? An interdisciplinary primer on XAI and meaningful information in automated decision-making." <https://doi.org/10.5281/zenodo.6375784>.



Universität Hamburg

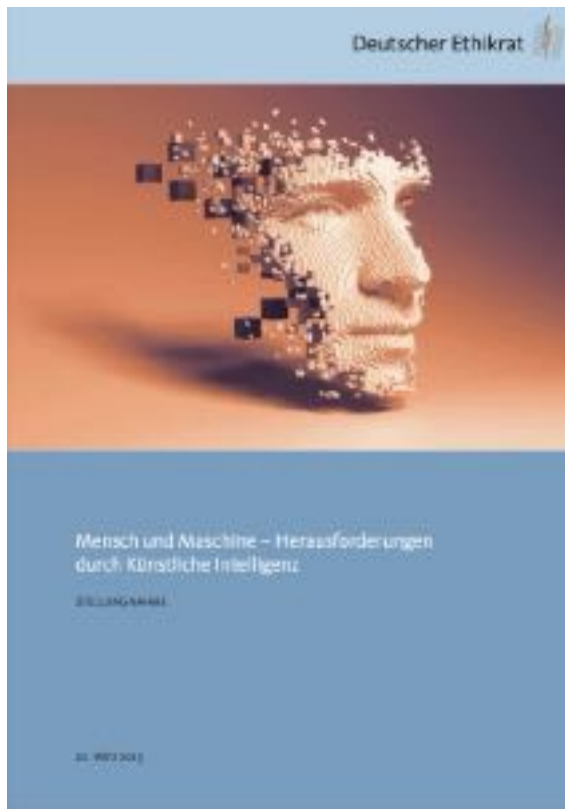
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Fazit

Ethische Analyse des Einsatzes in KI in 4 Bereichen

- Medizin
- Bildung
- Öffentliche Kommunikation und Meinungsbildung
- Öffentliche Verwaltung

→ Fokus: Wechselwirkung Mensch & Maschine



<https://www.ethikrat.org>

QUERSCHNITTSTHEMEN

1. Erweiterung & Verminderung von Handlungsmöglichkeiten
2. Wissenserzeugung durch KI und der Umgang mit KI-gestützten Voraussagen
3. Die Gefährdung des Individuums durch statistische Stratifizierung
4. Auswirkungen von KI auf menschliche Kompetenzen und Fertigkeiten
5. Schutz von **Privatsphäre** und Autonomie versus Gefahren durch Überwachung und Chilling-Effekte
6. Datensouveränität und **gemeinwohlorientierte Datennutzung**
7. **Kritische Infrastrukturen, Abhängigkeiten und Resilienz**
8. **Pfadabhängigkeiten**, Zweitverwertung und Missbrauchgefahren
9. **Bias und Diskriminierung**
10. **Transparenz und Nachvollziehbarkeit – Kontrolle und Verantwortung**





Fazit

- Ziel menschliche Handlungsmöglichkeiten und Autorschaft erweitern, Verminderungen verhindern
- Auswirkungen unterschiedlich für verschiedene Betroffene – besonderer Fokus auf jene, welche bereits vulnerabel sind
- Der Teufel steckt im Detail: genauer Blick auf Technologien, aber auch institutionelle/organisationale Rahmenbedingungen
- Generative KI: Alle üblichen Probleme von KI auch hier relevant + Problem der 4-fachen Täuschung
- Sinnvoll über Vertrauen kann nur in KI als sozio-technisches Ökosystem geredet werden
- Vertrauen nur gerechtfertigt wenn Vertrauenswürdigkeit vorliegt
- Vertrauenswürdigkeit hat epistemische und moralische Komponenten
- Verantwortliches Handeln - Möglichkeiten und Grenzen von Transparenz/Erklärbarkeit und Bias Minimierung/Fairness



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Vielen Dank für Ihre Aufmerksamkeit!

Prof. Dr. Judith Simon

Professorin für Ethik in der Informationstechnologie

Email: judith.simon@uni-hamburg.de

Web: <http://uhh.de/inf-eit>