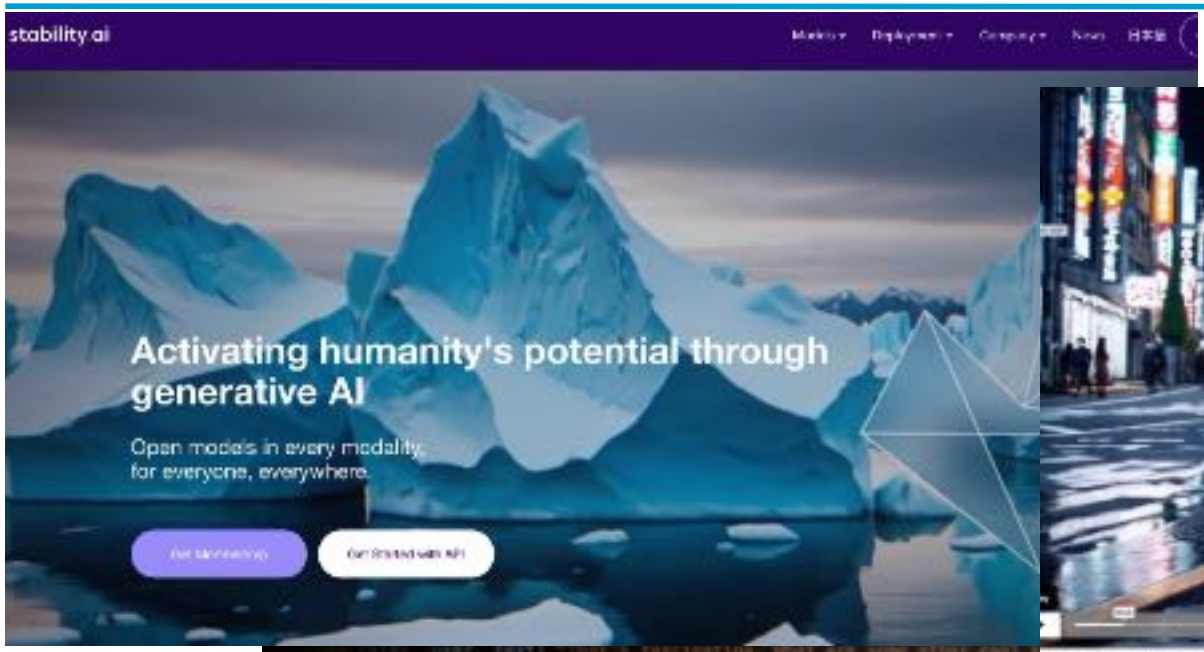
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Prof. Dr. Judith Simon

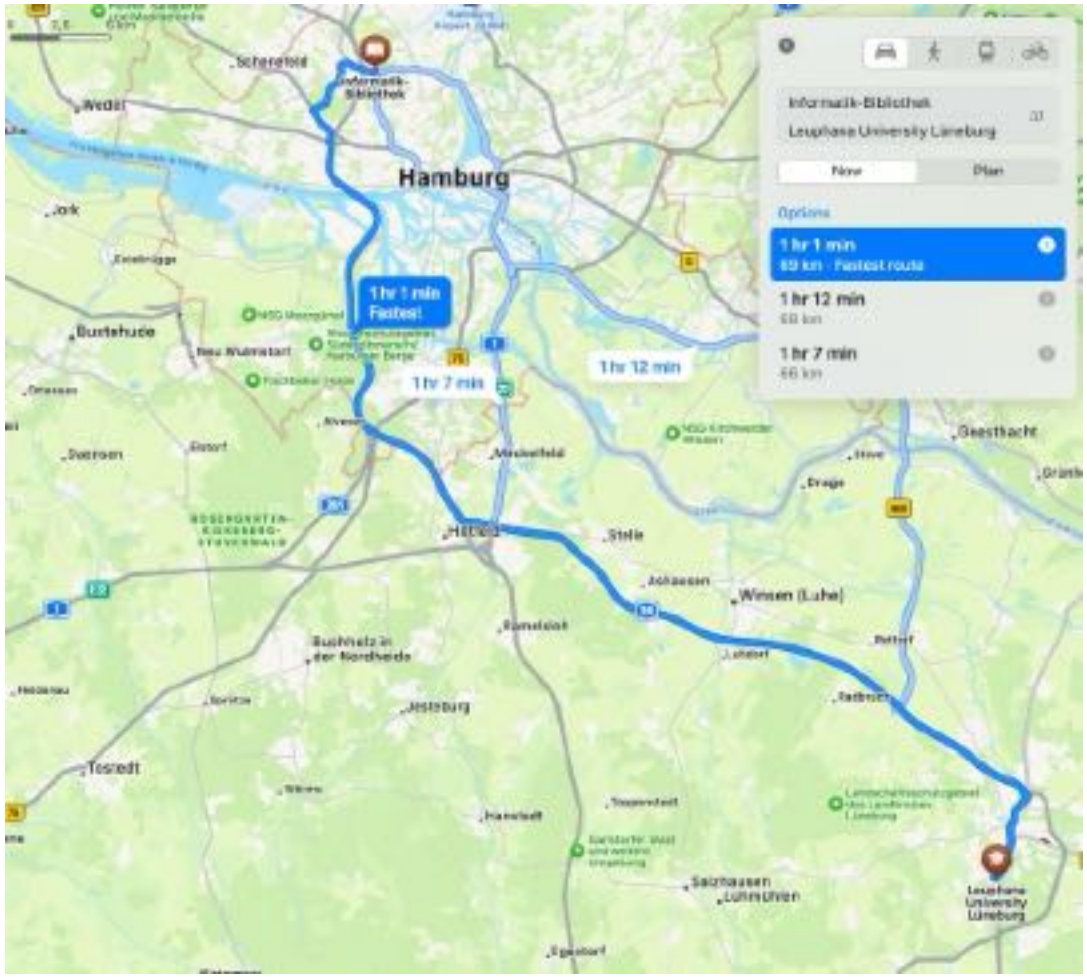
Vertrauenswürdige Künstliche Intelligenz? Ethik in Design, Entwicklung & Anwendung von KI

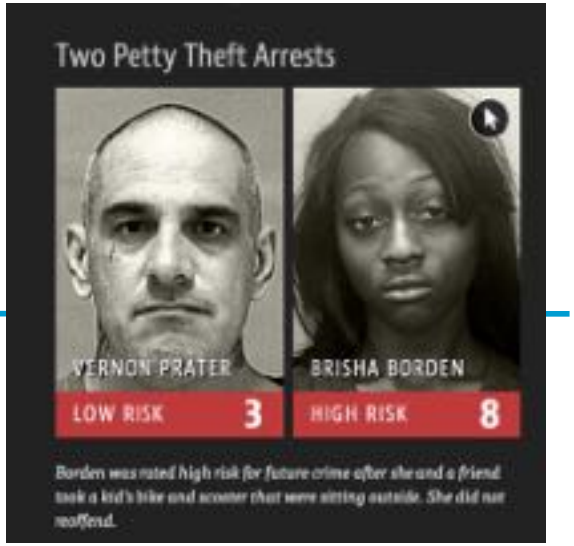
18.09.2024|KonKIS 24 - Konferenz der deutschen KI-Servicezentren 2024 | Göttingen|



https://openai.com
https://openai.com/sora
https://stability.ai

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG





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

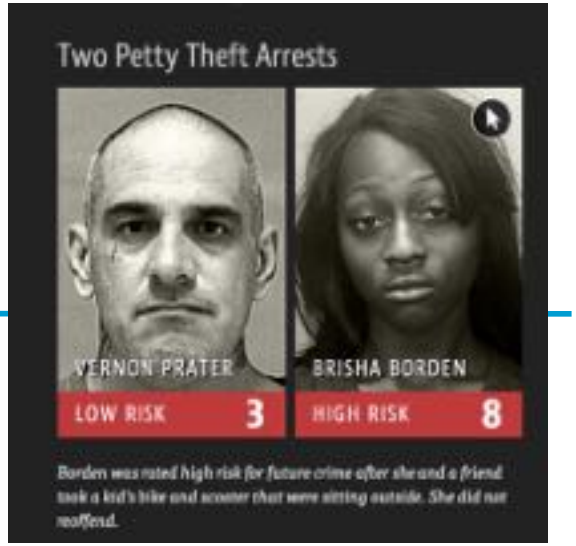
- Historisch: Simulation von Verhalten, welches bei Menschen als intelligent interpretiert würde
- Heute: starker Fokus auf Methoden des maschinellen Lernens

Bernhard Christoph Francke

Turing/Public Domain

https://openai.com

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



Daten, Künstliche Intelligenz & Ethik

- Kern heute zumeist: Analyse großer Datenmengen

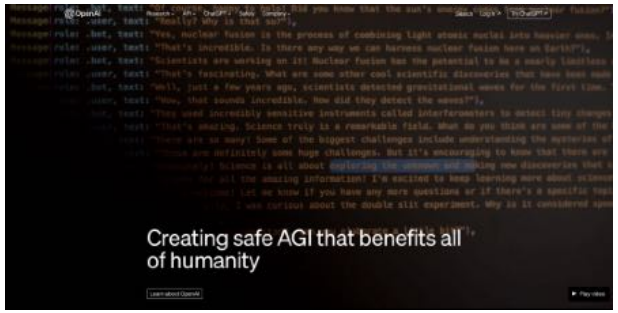
  > Mustererkennung, Klassifikation, Prognose, Entscheidungssupport

  - 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

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

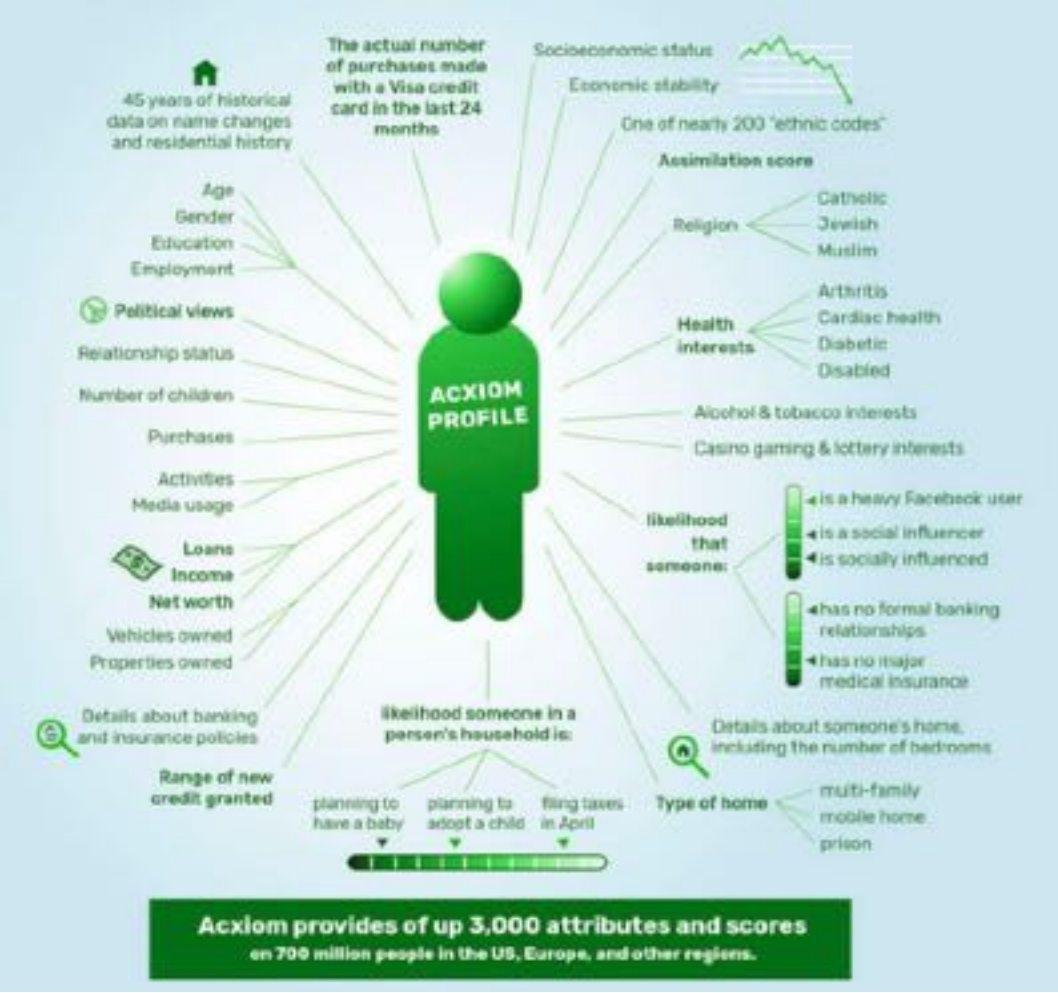
KI & ML: Wissen, Vertrauen & Verantwortung

- 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

Schutz der Privatsphäre



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++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Source: MIT Media Lab

Epistemische Dimension

+

Ethische Dimension

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

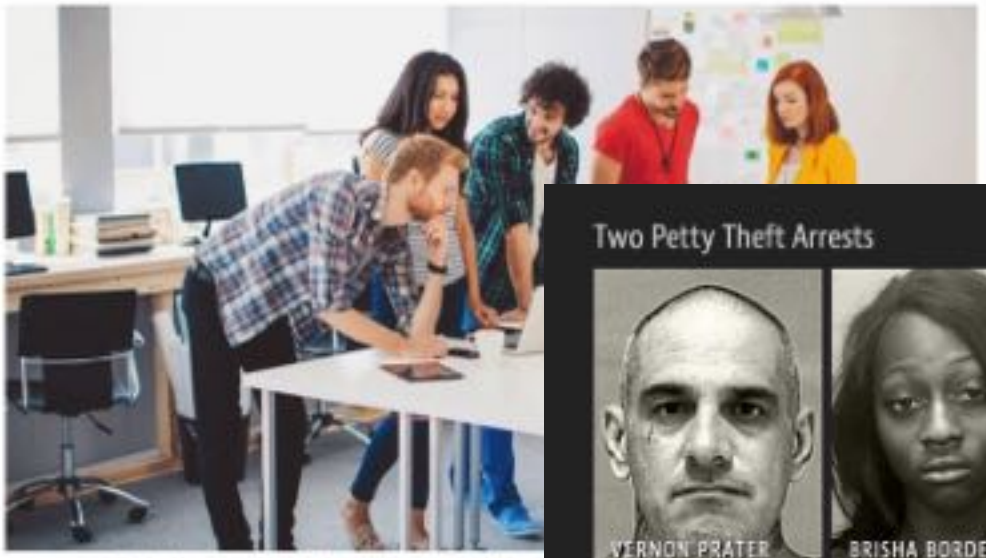
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Schutz vor Diskriminierung, Gerechtigkeit & Transparenz?

BBC

Home News Sport Business Innovation Culture Travel Earth Video Live

Amazon scrapped 'sexist AI' tool



NEWS

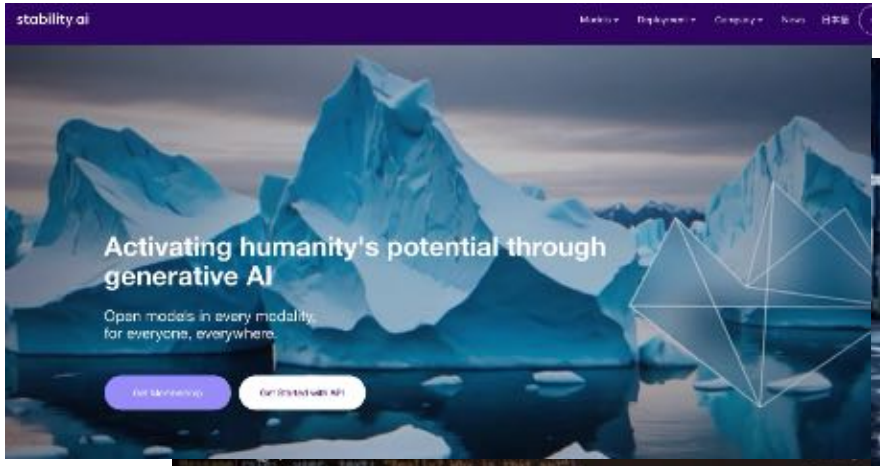
Millions of black people affected by racial bias in health-care algorithms

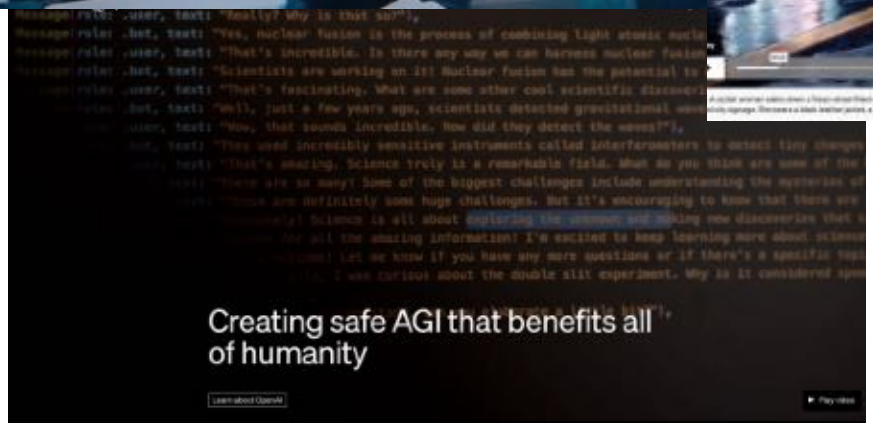
Study reveals rampant racism in decision-making software used by US hospitals — and highlights ways to correct it.

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

(Neue) Herausforderungen durch ChatGPT & Co?





https://openai.com
https://openai.com/sora
https://stability.ai

Täuschung I: KI oder Mensch?

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

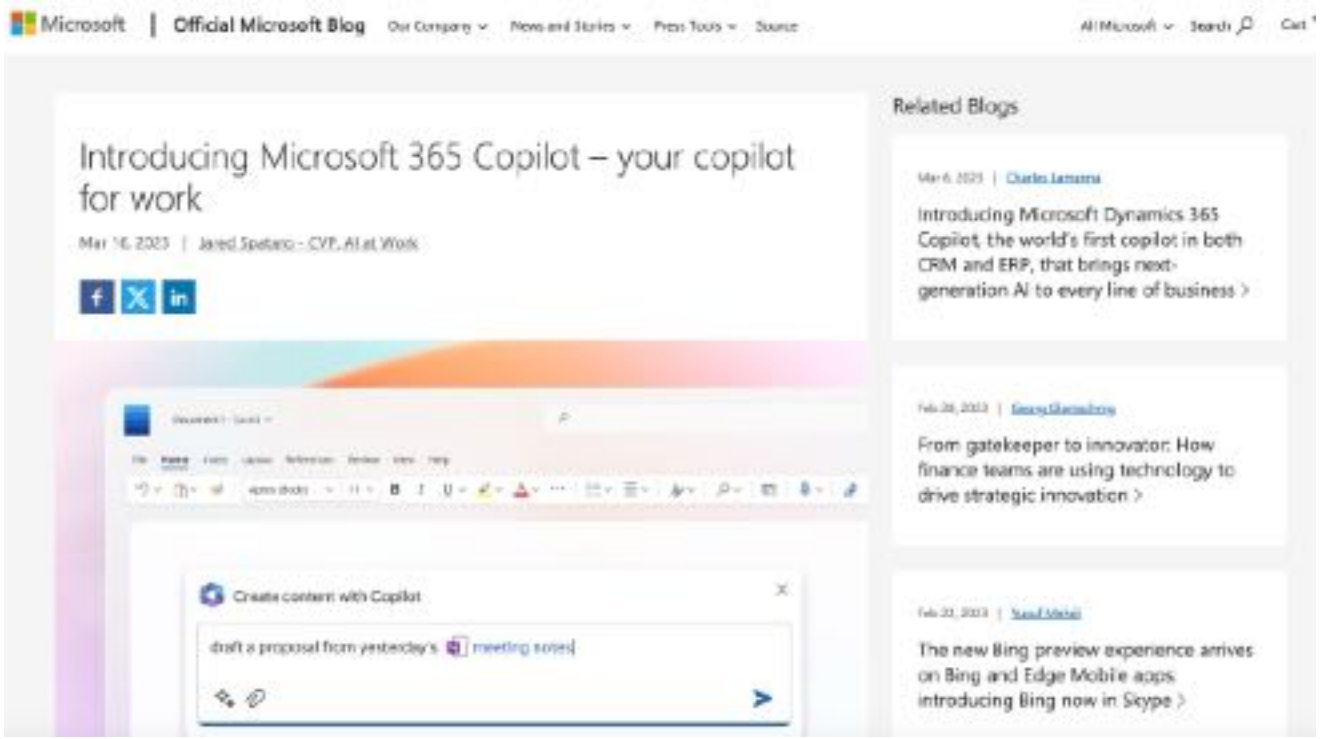
AI ethicists warned Google not to impersonate humans. Now one of Google's own thinks there's a ghost in the machine.

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

Täuschung III: Inhalte

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

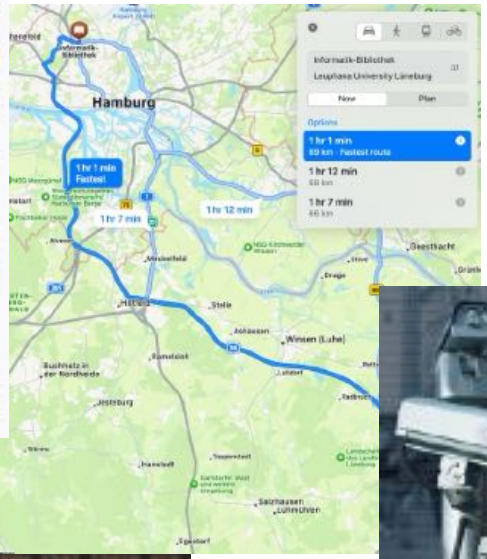
- Das dreifache Transparenzproblem
  - Funktional: Kein Zugang zur Software aufgrund von Geschäftsgeheimnissen
  - Epistemisch: Blackbox Deep Learning, Intransparenz Maschinellen Lernens
  - Kompetenz: wer hat die notwendige Kompetenz, um Software zu prüfen, selbst wenn Transparenz gegeben ist?

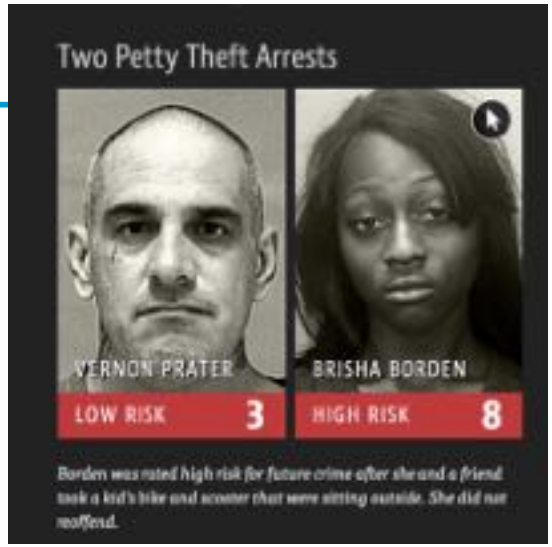
Zwischenfazit: (Neue) Herausforderungen durch Generative KI

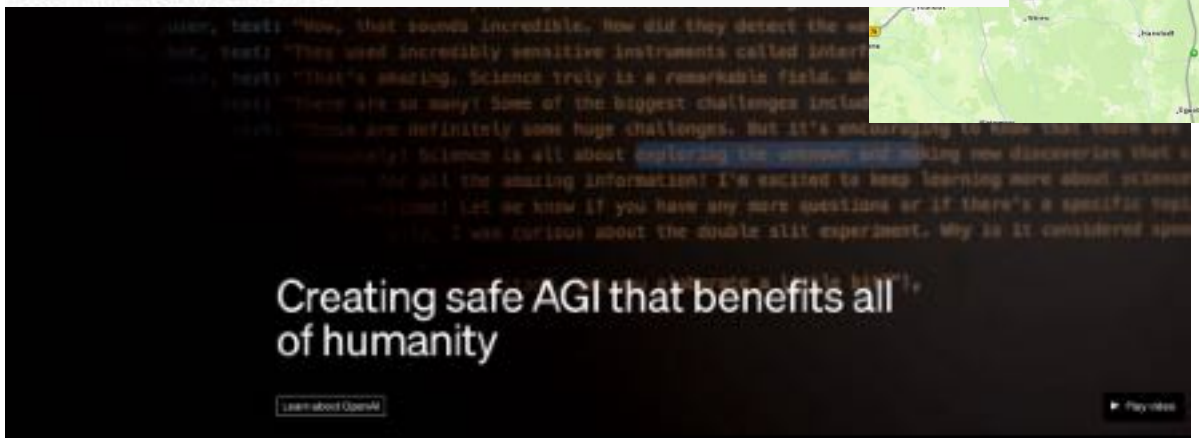
- Alle herkömmlichen Herausforderungen durch KI
  - Schutz der Privatsphäre/Datenschutz, Bias/Diskriminierung, mangelnde Transparenz, Nachvollziehbarkeit & Kontrolle, Energieverbrauch, prekäre Arbeitsbedingungen, ...
- Plus: Problem der vierfachen Täuschung
- Weitreichende und versteckte Probleme durch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

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

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

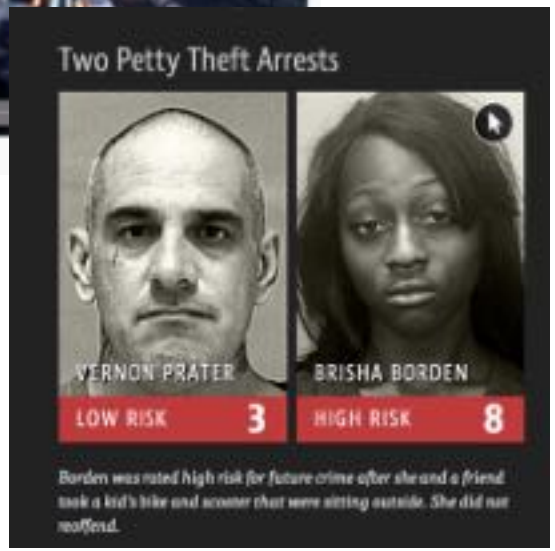
Hardwig (1991). The Role of Trust in Knowledge

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Vertrauen in KI – Vertrauenswürdige KI?

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Vertrauen in KI – Vertrauenswürdige KI?



https://openai.com/sora
Propublica.com
https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

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

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

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Vertrauen in KI – Vertrauenswürdige KI?

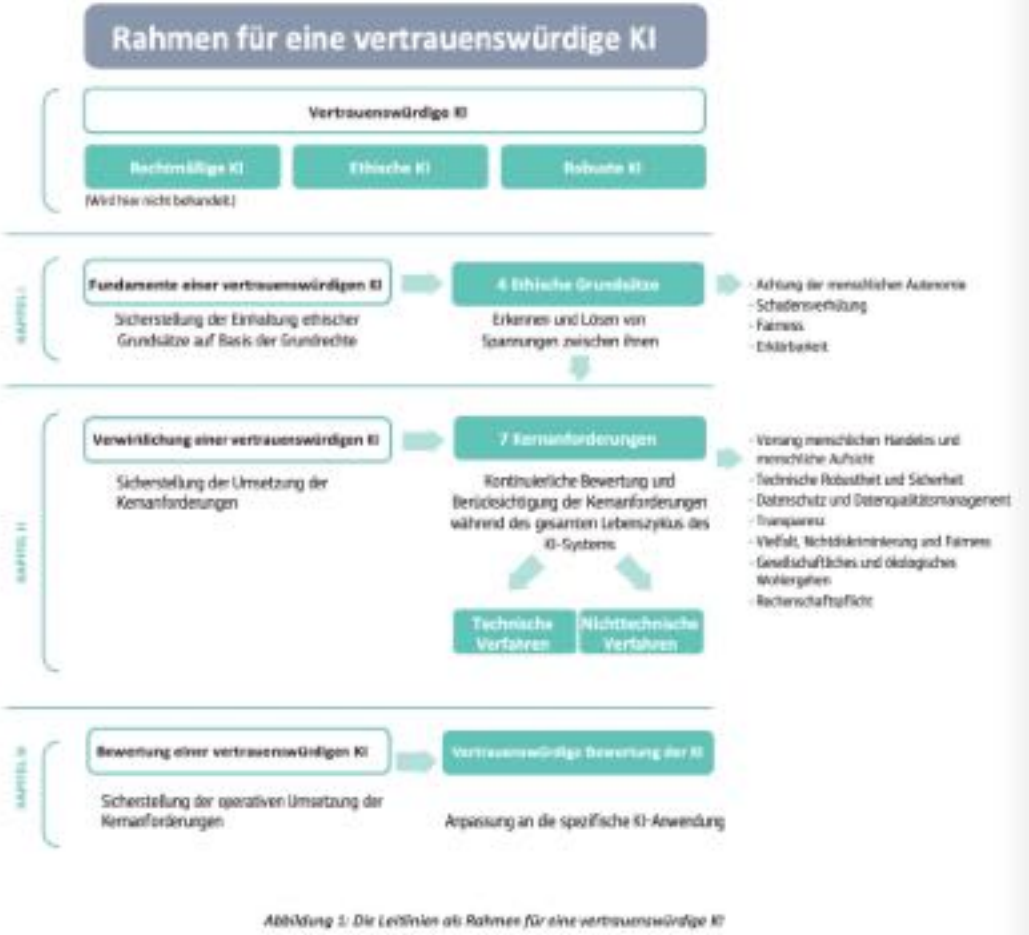


Abbildung 1: Die Leitlinien als Rahmen für eine vertrauenswürdige KI

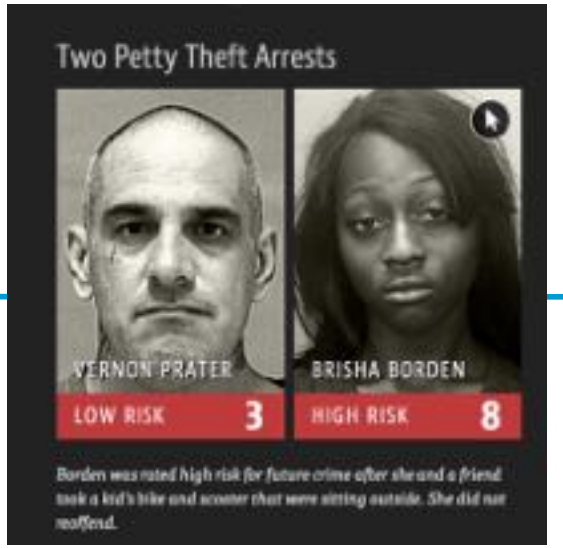
https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Vertrauen in KI – Vertrauenswürdige KI?

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

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



- 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

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Vertrauen in KI – Vertrauenswürdige KI?

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

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

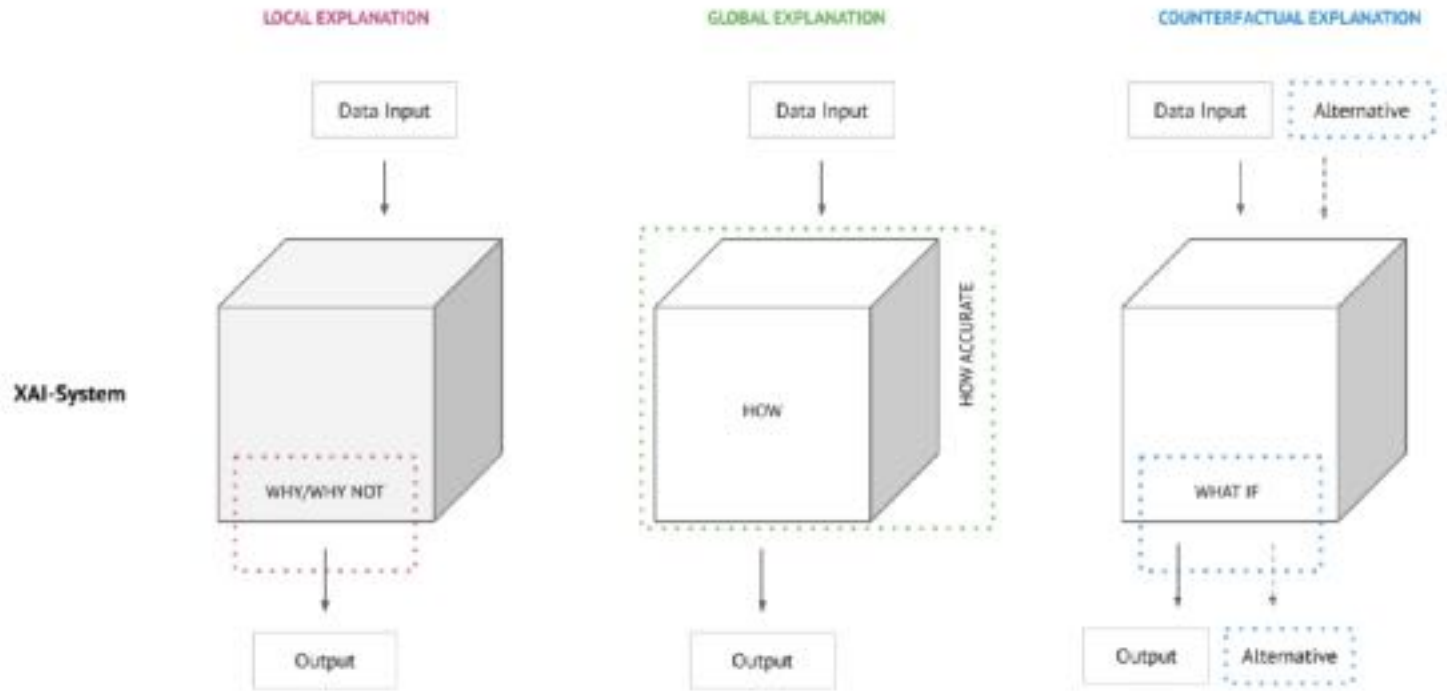
Picturing the original image (left), saliency map using a method called Grad CAM (middle), and another using Guided Backpropagation (right). The picture above is the canonical example for "class-discrimination". The above saliency maps are taken from https://github.com/kazuto1011/grad-cam-pytorch.

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.

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