FAKULTÄT FÜR GESUNDHEITSWISSENSCHAFTEN BRANDENBURG

Gemeinsame Fakultät der Universität Potsdam, der Medizinischen Hochschule Brandenburg Theodor Fontane und der Brandenburgischen Technischen Universität Cottbus-Senftenberg

Data Science & Künstliche Intelligenz
Ethische Perspektiven
Workshop: Data Science & Künstliche Intelligenz

Joschka Haltaufderheide & Robert Ranisch, 05.02.2024
Fakultät für Gesundheitswissenschaften Brandenburg, Universität Potsdam

Einführung

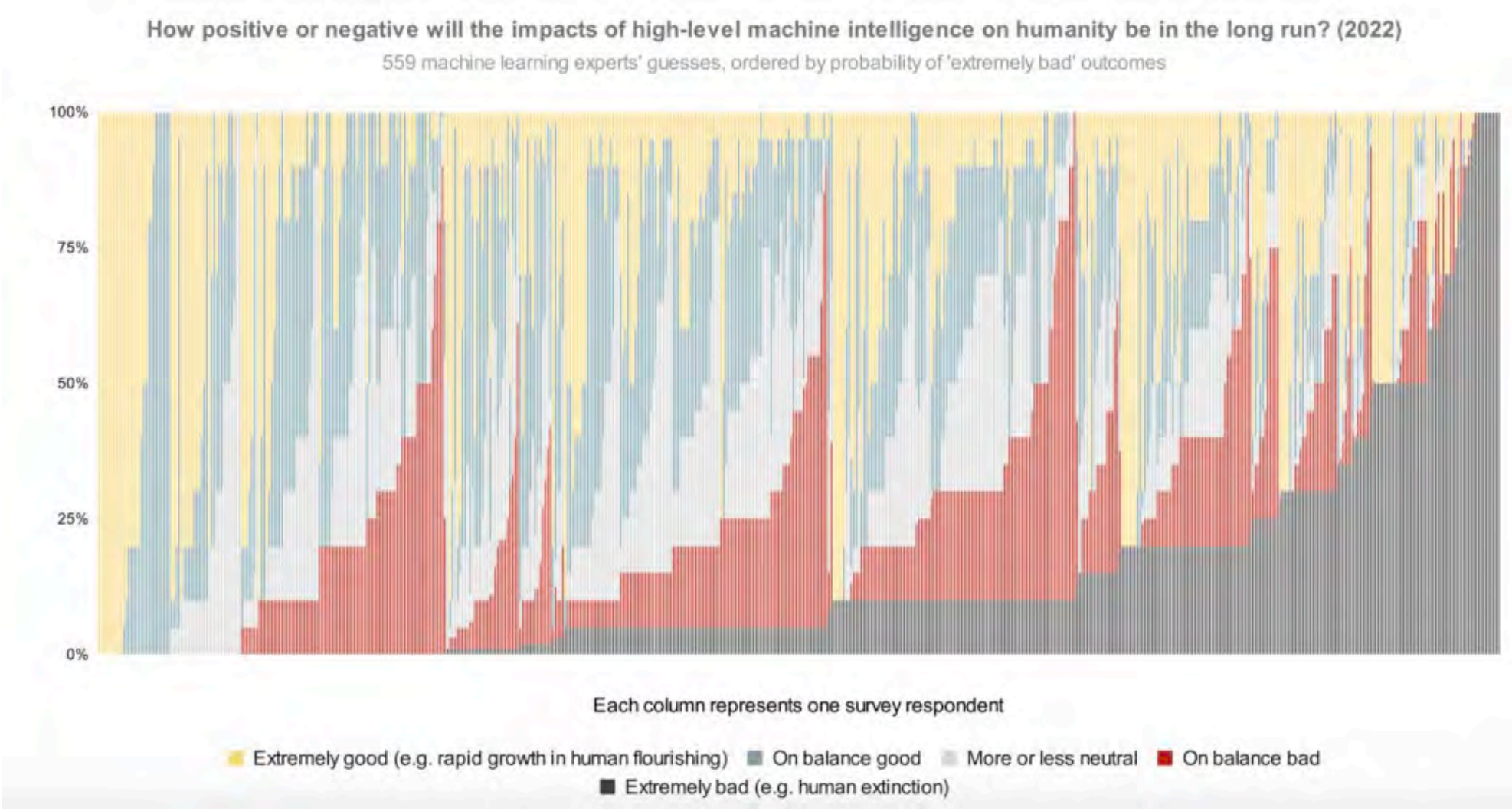
Dr. phil. Joschka Haltaufderheide, M.A.
Wissenschaftlicher Mitarbeiter,
Projekt Digital Medical Ethics Network (DiMEN)

Prof. Dr. Robert Ranisch
Leitung der Juniorprofessur
Medizinische Ethik

Stellen Sie sich kurz vor

- Wer sind Sie?
- Woran arbeiten Sie gerade?
- Was ist das erste, dass Ihnen in den Sinn kommt, wenn Sie an Ethik und Daten/KI denken?

KI-Expert*innen sagen: KI macht die Welt schlechter?



Extremely Good: 24%
On Balance Good: 26%
More or Less Neutral: 18%
On Balance Bad: 17%
Extremely Bad: 14%

https://aiimpacts.org/how-bad-a-future-do-ml-researchers-expect/

KI Hype

FAKULTÄT FÜR GESUNDHEITSWISSENSCHAFTEN
BRANDENBURG

STERN.de

Warum künstliche Intelligenz Menschenleben retten kann | STERN.de

Der Intensiv- und Notfallmediziner Christian Karagiannidis spricht über den Einsatz von künstlicher Intelligenz in der Medizin und die Macht...

14.08.2023

detektor.fm

Mission Energiewende | KI und Nachhaltigkeit – Kann KI das Klima retten?

KI soll vieles können: Hausarbeiten schreiben, Coden, die Arbeitswelt revolutionieren. Aber was kann KI für's Klima tun?

24.10.2023

The Guardian

Five ways AI could improve the world: 'We can cure all diseases, stabilise our climate, halt poverty'

It is not yet clear how the power and possibilities of AI will play out. Here are the best-case scenarios for how it might help us develop new drugs, give up dull jobs and live long, healthy lives

Nature Journal

How AI can help to save endangered species

Scientists are using artificial intelligence to fight biodiversity loss by analysing vast amounts of data, monitoring ecosystems and...

27.10.2023

Bloomberg

AI Can Save the World From a Decade of Low Growth, Former UK Premier Says

Artificial intelligence can rescue the global economy from ten years of low growth, former UK Prime Minister Gordon Brown said.

Augsburger Allgemeine

Weniger Verspätungen durch KI: Wie die Bahn den Nahverkehr retten will

Mit einem Pilotprojekt will die Deutsche Bahn die Züge im Nahverkehr trotz Störungen pünktlicher machen. Und zwar mit einer Künstlichen...

22.02.2023

KI Doom

FAKULTÄT FÜR GESUNDHEITSWISSENSCHAFTEN BRANDENBURG

The Guardian

Five ways AI might destroy the world: 'Everyone on Earth could fall over dead in the same second'

Artificial intelligence is already advancing at a worrying pace. What if we don't slam on the brakes? Experts explain what keeps them up at night

Five ways AI could improve the world

Science

51 Roboter bei der UN: „Wollen Menschen nicht auslöschen"

futurezone Newsletter

Experten warnen vor Auslöschung der Menschheit durch KI

Newsweek 90

Science | Robots | Hackers

Philosoph über mögliches Ende der Menschheit durch KI

»Es kann sein, dass wir verdammt sind«

Der Oxford-Philosoph Nick Bostrom warnt davor, dass eine künstliche Intelligenz bald schlauer als der Mensch sein könnte. Was, wenn er recht behält?

Hacked Sex Robots Could Murder People, Security Expert Warns

Jan 01, 2018 at 2:44 PM

News > UK > UK Politics

Elon Musk says AI one of the 'biggest threats' to humanity

The Tesla, SpaceX and X owner is attending the UK's AI Safety Summit and is set to meet Prime Minister Rishi Sunak.

Martyn Landi • Wednesday 01 November 2023 18:55 GMT

BI Business Insider

Open AI fürchtet Aussterben der Menschheit durch superintelligente KI

Open AI sucht derzeit nach Mitarbeitern für das Team. Das neue Team heißt Superalignment. Es plant, innerhalb der nächsten vier Jahre eine KI zu...

08.07.2023

future tense

Artificial Intelligence Is Not Going to Kill Us All

Yes, the fast-growing technology could be dangerous—but A.I. doomers are focused on the wrong threat.

BY TIMOTHY B. LEE MAY 09, 2023 • 9:00 AM

Daten- und KI-Ethik

FAKULTÄT FÜR GESUNDHEITSWISSENSCHAFTEN BRANDENBURG

#EthicsGroup_EU

ETHIK-KOMMISSION
AUTOMATISIERTES UND VERNETZTES FAHREN

daten ethik kommission

Deutscher Bundestag

Enquete-Kommission
„Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale“

European Commission
INDEPENDENT
HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE
SET UP BY THE EUROPEAN COMMISSION

Rat für Digitalethik
Strategischer Berater des Landes

Wenn KI/Daten-Systeme uns überwachen

https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html

Wenn KI/Daten-Systeme uns schaden

The Washington Post
Democracy Dies in Darkness

The final 11 seconds of a fatal Tesla Autopilot crash

A reconstruction of the wreck shows how human error and emerging technology can collide with deadly results

https://www.washingtonpost.com/technology/interactive/2023/tesla-autopilot-crash-analysis/

Triage per Software

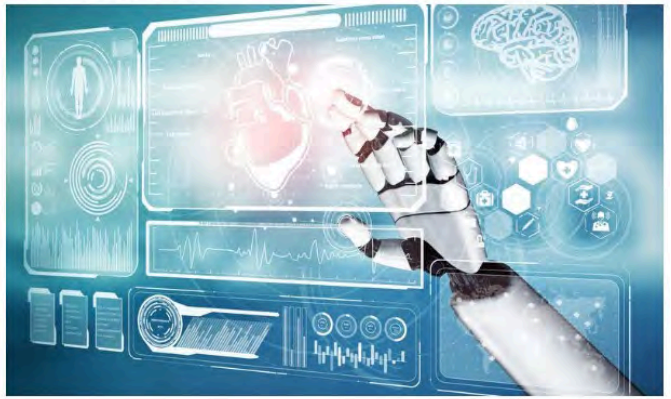
KI darf nicht über Leben und Tod entscheiden

Auch in Zukunft muss allein der Mensch – und nicht Software – über das Schicksal von Patientinnen und Patienten entscheiden, meint der Wirtschaftsethiker Peter Seele. © picture alliance / dpa / Kay Nietfeld 04:23 Minuten

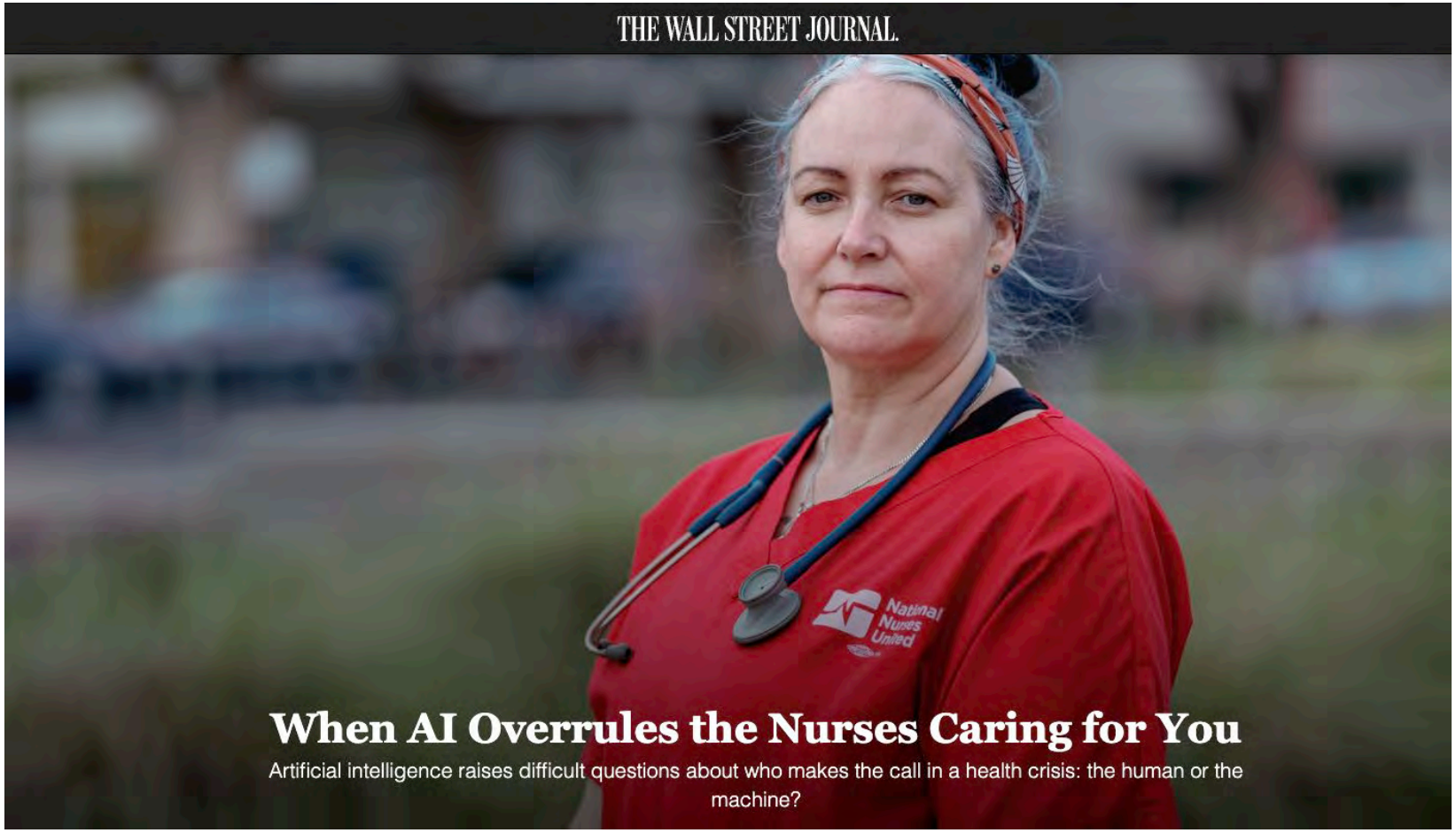
Künstliche Intelligenz in der Medizin

Wenn der Computer den Arzt ersetzt: Darf der „Todes-Algorithmus" über Leben entscheiden?

Montag, 18.10.2021 | 21:07



Wenn KI/Daten-Systeme uns dominieren



https://www.wsj.com/articles/ai-medical-diagnosis-nurses-f881b0fe

Wenn KI/Daten-Systeme uns täuschen

Wenn KI/Daten-Systeme uns manipulieren

SOCIETY

Facebook Manipulates Our Moods For Science And Commerce: A Roundup

JUNE 30, 2014 · 12:31 PM ET

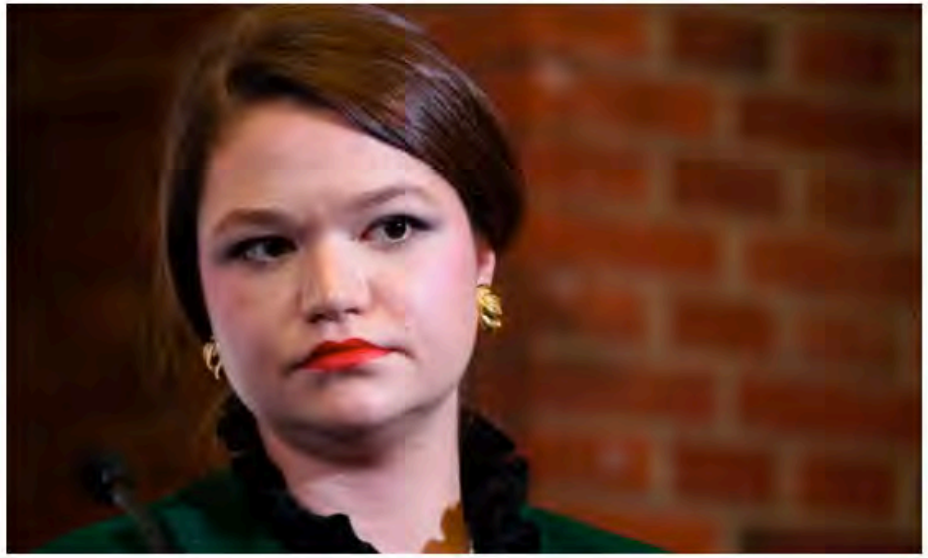
Elise Hu

Facebook researchers manipulated newsfeeds of nearly 700,000 users to study "emotional contagion."

iStockPhoto

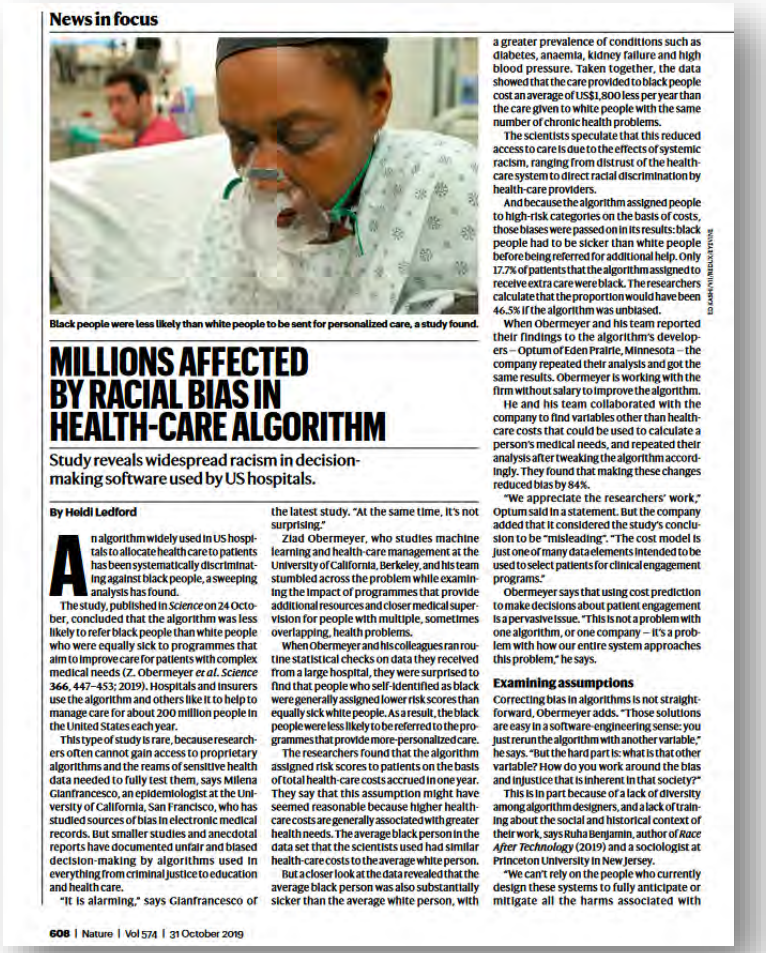
Fresh Cambridge Analytica leak 'shows global manipulation is out of control'

Company's work in 68 countries laid bare with release of more than 100,000 documents



Wenn KI/Daten-Systeme uns diskriminieren

https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html



News in focus

Black people were less likely than white people to be sent for personalized care, a study found.

MILLIONS AFFECTED BY RACIAL BIAS IN HEALTH-CARE ALGORITHM

Study reveals widespread racism in decision-making software used by US hospitals.

By Heidi Ledford

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.

The study, published in *Science* on 24 October, concluded that the algorithm was less likely to refer black people than white people who were equally sick to programmes that aim to improve care for patients with complex medical needs (Z. Obermeyer *et al. Science* **366**, 447–453; 2019). Hospitals and insurers use the algorithm and others like it to help to manage care for about 200 million people in the United States each year.

This type of study is rare, because researchers often cannot gain access to proprietary algorithms and the reams of sensitive health data needed to fully test them, says Milena Gianfrancesco, an epidemiologist at the University of California, San Francisco, who has studied sources of bias in electronic medical records. But smaller studies and anecdotal reports have documented unfair and biased decision-making by algorithms used in everything from criminal justice to education and health care.

"It is alarming," says Gianfrancesco of the latest study. "At the same time, it's not surprising."

Ziad Obermeyer, who studies machine learning and health-care management at the University of California, Berkeley, and his team stumbled across the problem while examining the impact of programmes that provide additional resources and closer medical supervision for people with multiple, sometimes overlapping, health problems.

When Obermeyer and his colleagues ran routine statistical checks on data they received from a large hospital, they were surprised to find that people who self-identified as black were generally assigned lower risk scores than equally sick white people. As a result, the black people were less likely to be referred to the programmes that provide more-personalized care.

The researchers found that the algorithm assigned risk scores to patients on the basis of total health-care costs accrued in one year. They say that this assumption might have seemed reasonable because higher health-care costs are generally associated with greater health needs. The average black person in the data set that the scientists used had similar health-care costs to the average white person.

But a closer look at the data revealed that the average black person was also substantially sicker than the average white person, with a greater prevalence of conditions such as diabetes, anaemia, kidney failure and high blood pressure. Taken together, the data showed that the care provided to black people cost an average of US$1,800 less per year than the care given to white people with the same number of chronic health problems.

The scientists speculate that this reduced access to care is due to the effects of systemic racism, ranging from distrust of the health-care system to direct racial discrimination by health-care providers.

And because the algorithm assigned people to high-risk categories on the basis of costs, those biases were passed on in its results: black people had to be sicker than white people before being referred for additional help. Only 17.7% of patients that the algorithm assigned to receive extra care were black. The researchers calculate that the proportion would have been 46.5% if the algorithm was unbiased.

When Obermeyer and his team reported their findings to the algorithm's developers – Optum of Eden Prairie, Minnesota – the company repeated their analysis and got the same results. Obermeyer is working with the firm without salary to improve the algorithm.

He and his team collaborated with the company to find variables other than health-care costs that could be used to calculate a person's medical needs, and repeated their analysis after tweaking the algorithm accordingly. They found that making these changes reduced bias by 84%.

"We appreciate the researchers' work," Optum said in a statement. But the company added that it considered the study's conclusion to be "misleading". "The cost model is just one of many data elements intended to be used to select patients for clinical engagement programs."

Obermeyer says that using cost prediction to make decisions about patient engagement is a pervasive issue. "This is not a problem with one algorithm, or one company – it's a problem with how our entire system approaches this problem," he says.

Examining assumptions

Correcting bias in algorithms is not straightforward, Obermeyer adds. "Those solutions are easy in a software-engineering sense: you just rerun the algorithm with another variable," he says. "But the hard part is: what is that other variable? How do you work around the bias and injustice that is inherent in that society?"

This is in part because of a lack of diversity among algorithm designers, and a lack of training about the social and historical context of their work, says Ruha Benjamin, author of *Race After Technology* (2019) and a sociologist at Princeton University in New Jersey.

"We can't rely on the people who currently design these systems to fully anticipate or mitigate all the harms associated with

608 | Nature | Vol 574 | 31 October 2019

Leitfragen

1. Was bedeutet „Ethik“?
   a. Handwerkszeug: Normen, Werte, Handlungen
   b. Ethisch argumentieren: Deontologie und Teleologie
   c. Beispiel: Trolley Problem

2. Was hat Ethik mit Daten/KI zu tun?
   a. Ethische Fragen an Technologien
   b. Was ist ein Datum?
   c. Verzerrungen in Daten

3. Was hat das mit mir zu tun?
   a. Forschungsverantwortung

4. Und was mach ich jetzt?
   a. Fallbeispiel Forschungsethik
   b. Fallbeispiel Forschung und Daten

Grundbegriffe I

Begriff		Definition
Moral		Das **Normensystem**, das in einer Gesellschaft oder **menschlichen Gemeinschaft** als **gültig** angesehen wird
Ethik		Das **methodische Nachdenken** oder die philosophische Reflexion über Moral
KI-Ethik Daten-Ethik		Ethik im **Anwendungsbereich** spezifischer Technologischer Systeme

Fallbeispiel Selbstfahrende Autos

WIRED LONG READS BUSINESS CULTURE GEAR SCIENCE SECURITY VIDEO

JAMIE SUSSKIND IDEAS 12.12.2018 06:00 AM

Should we allow self-driving cars to help people kill themselves?

Questions about individual choice and behaviour were once the province of politicians and philosophers – now, they're the province of Silicon Valley

TOMMY PARKER

In October 2018 the Californian startup Aeva unveiled a new lidar system – a laser-

MOST POPULAR

Here's Everything You Can Do With Copilot, the Generative AI Assistant on Windows 11

BY DAVID NIELD

The 39 Best Films on Netflix This Week

BY MATT KAMEN

The 14 Best Films on Amazon Prime Right Now

BY WIRED

The 39 Best Shows on Disney+ Right Now

BY WIRED

Handlungsvorschläge

1. Welcher Handlungsvorschlag ist angemessen?

Akteur		Handlungsvorschlag
Fahrer		Das ist meine Entscheidung und mein Leben!
Entwickler		Die Gesundheit unserer Kunden ist ein hohes Gut zu dessen Schutz wir uns verpflichtet sehen.
Entwickler 2		Möglicherweise sollte das Auto, das einen Versuch feststellt diesen erst einmal verhindern, dauerhaft verwehren darf es ihn aber nicht.
Politiker		Wenn wir dies zulassen, dann ist das ein Dammbruch bei dem unweigerlich viele weitere Menschen zu Schaden kommen.
Hersteller		Unser Auto hat damit gar nichts zu tun. Es hilft nicht und es ist auch nicht verantwortlich. Es ist ja nur eine Maschine

Handlungsvorschläge

1. Welche Werte liegen den Vorschlägen zu Grunde

Akteur		**Handlungsvorschlag**
Fahrer		Selbstbestimmung
Entwickler		Körperliche Unversehrtheit
Entwickler 2		Selbstbestimmung und körperliche Unversehrtheit
Politiker		Körperliche Unversehrtheit
Hersteller		-

Grundbegriffe II

Begriff	Definition
Werte	Etwas, an dem jemand ein Interesse hat. Etwas von dem jemand nicht wollen kann, dass es ihm genommen wird.
Normen	Eine Forderung gegenüber jemand anderem auf Grundlage eines Wertes

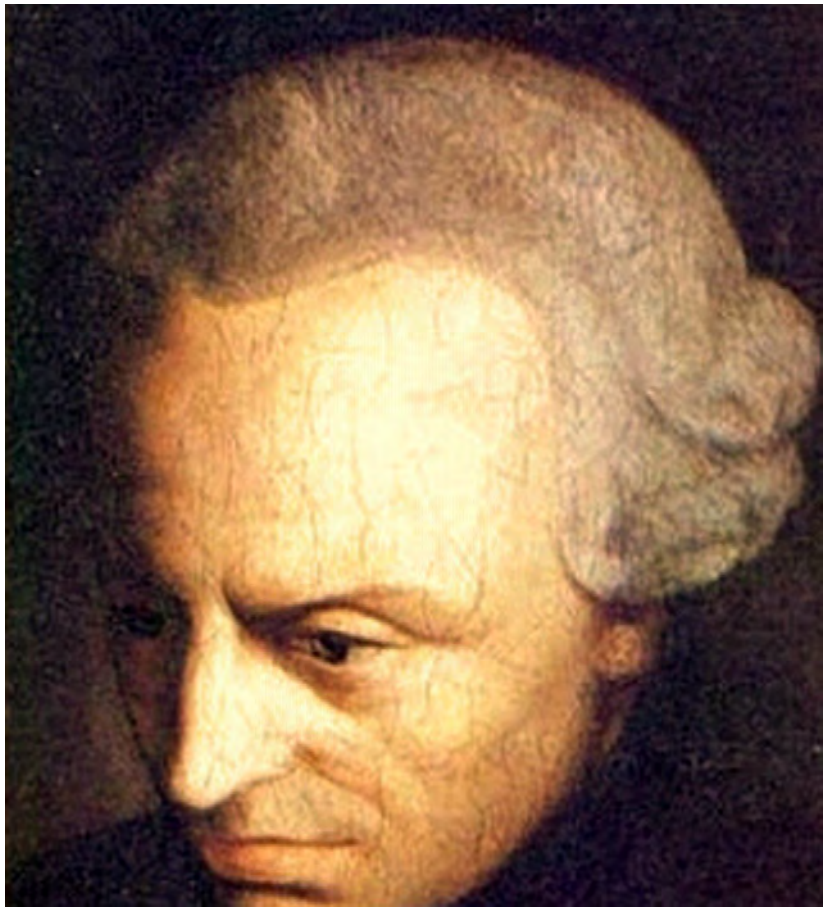
Formen des moralischen Argumentierens

1. Wie zeigt man, dass etwas nicht nur für mich, sondern für jeden anderen ebenso von Interesse ist?
2. Wie überprüft man, ob eine Handlung diesem Interesse entspricht?

Grundtyp des ethischen Argumentierens	**Wert**	**Zu Prüfen**
Deontologie	z.B. Würde	Handlung
Teleologie	z.B. Nutzen, Schaden	Konsequenz

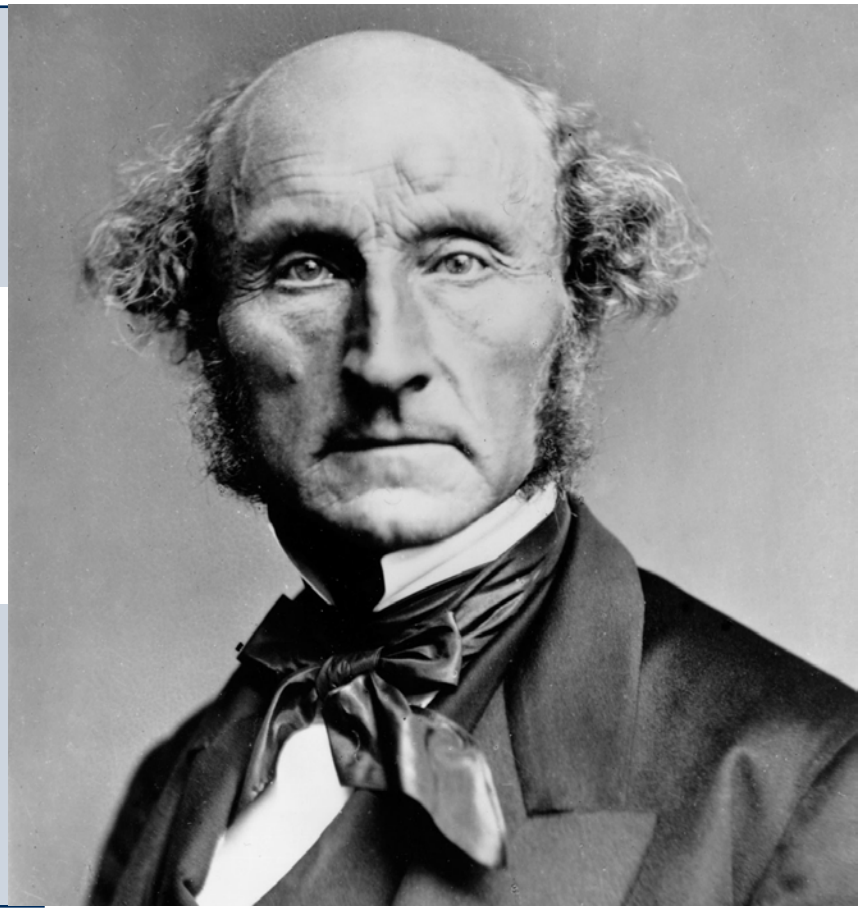
Deontologische Ethik

Wichtige Vertreter:	Immanuel Kant (1724 – 1804)
Elemente:	• Würde als Freiheit „sich selbst zum Handeln zu bestimmen.“ • Begründet als rationale Nötigung • Universalisierbarkeit • Deon: Das Gesollte, die Pflicht • Kategorischer Imperativ
Kritik:	• Absolutismus (Lügenproblem)

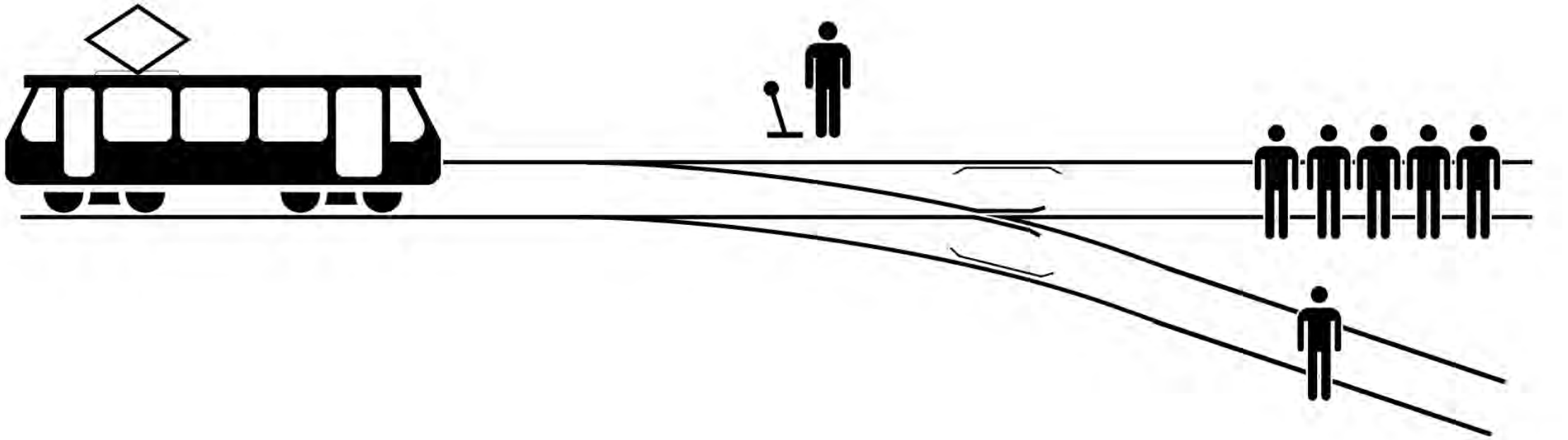


Teleologische Ethik

Wichtige Vertreter:	Jeremy Bentham (1748 – 1832) J.S. Mill (1806 – 1873)
Elemente:	• Hedonistisches Prinzip • Nutzen oder Utilitätsprinzip • Folgenprinzip • Aggregationsprinzip
Kritik:	• Begründungsfehler • Kommensurabilität • Distributionsproblem



The Trolley Problem



Prinizipienbasierte Ansätze in der KI-/Datenethik

https://dash.harvard.edu/handle/1/42160420

- Aus der Tradition sog. Kohärentistischer Modelle
- Anwendbarkeit vor Begründung
- Annahme prima facie gültiger Prinzipien
- Ziel ist Abwägung und Kohärenz im Anwendungsfall.

nature machine intelligence

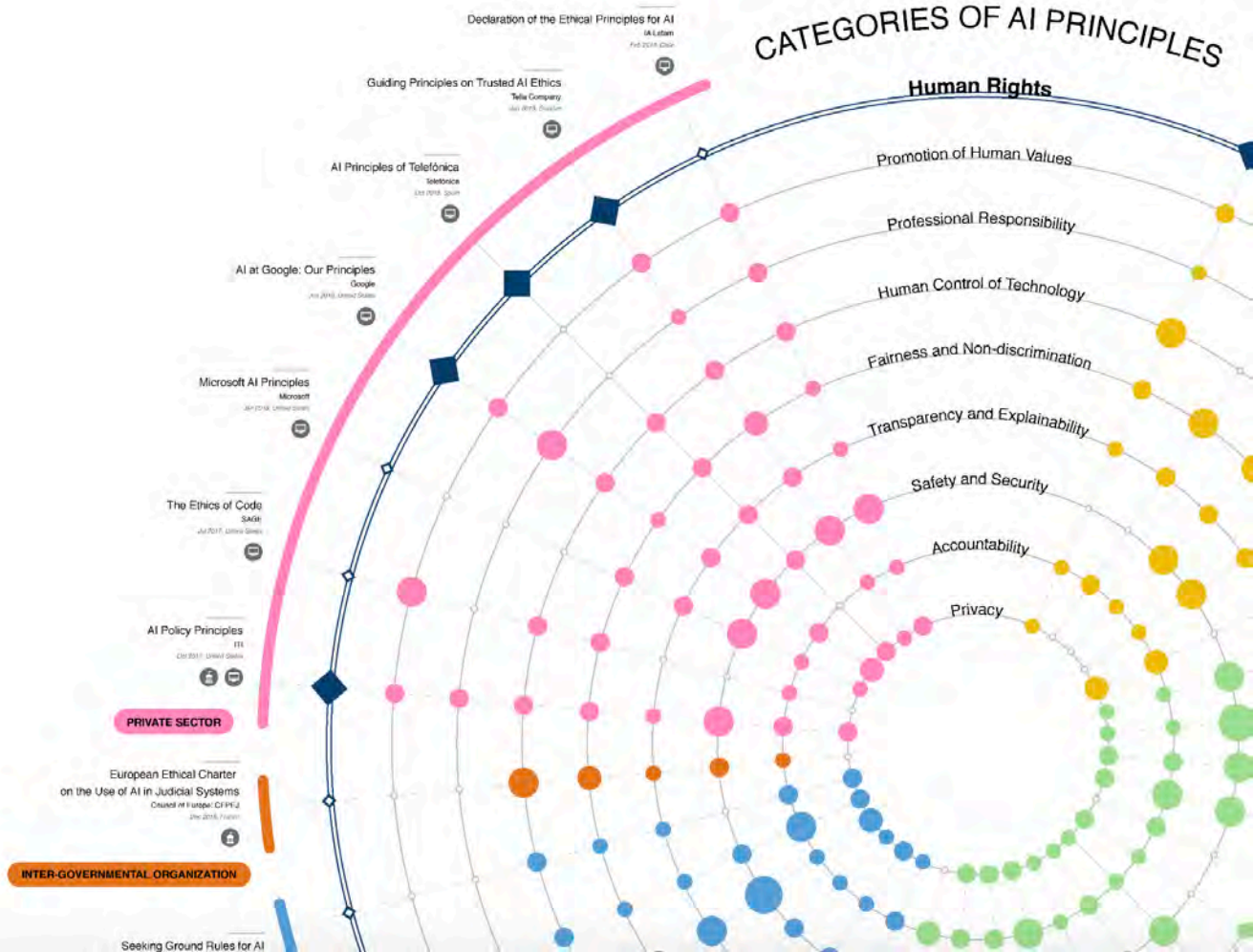
PERSPECTIVE

https://doi.org/10.1038/s42256-019-0088-2

The global landscape of AI ethics guidelines

Anna Jobin, Marcello Ienca and Effy Vayena*

In the past five years, private companies, research institutions and public sector organizations have issued principles and guidelines for ethical artificial intelligence (AI). However, despite an apparent agreement that AI should be 'ethical', there is debate about both what constitutes 'ethical AI' and which ethical requirements, technical standards and best practices are needed for its realization. To investigate whether a global agreement on these questions is emerging, we mapped and analysed the current corpus of principles and guidelines on ethical AI. Our results reveal a global convergence emerging around five ethical principles (transparency, justice and fairness, non-maleficence, responsibility and privacy), with substantive divergence in relation to how these principles are interpreted, why they are deemed important, what issue, domain or actors they pertain to, and how they should be implemented. Our findings highlight the importance of integrating guideline-development efforts with substantive ethical analysis and adequate implementation strategies.



Review

„The global landscape of AI ethics guidelines“

- Untersuchten 84 Dokumente
- Rasanter Anstieg an Publikationen seit 2016
- Große Bandbreite an Prinzipien: 11 Cluster identifiziert
- Kein einziges Prinzip tauchte in allen 84 Stellungnahmen auf

Table 3 | Ethical principles identified in existing AI guidelines

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

Take Home Message

- Ethik ist die Reflexionstheorie der Moral
- Ethische Urteile
  - Basieren auf Werten
  - Formulieren handlungsleitende Normen
  - Sind begründet
  - Erheben Anspruch auf mind. Intersubjektive Verbindlichkeit

- Ethische Argumentationsmuster
  - Können auf Disposition, Handlung oder Folgen fokussieren
  - Basieren auf Wertannahmen, Werthierarchien oder konsensuellen Prinzipien

- Technologien spielen eine Rolle
  - Als Werkzeuge/Instrumente in Zweck-Mittel-Relationen
  - Als soziale Phänomene, die Wahrnehmung und Handlung verändern

Fragen?

Was heißt denn das jetzt?

1. Was bedeutet das Alles für die Ethik in Bezug auf Daten und KI
   a. Was ist ein Datum?
   b. Was ist eine Verzerrung und wo ist das Problem?

2. Was hat das Alles mit mir als Forschende/r zu tun?
   a. Kurze Übersicht Forschungsethik
   b. Forschungsverantwortung im Umgang mit KI und Daten

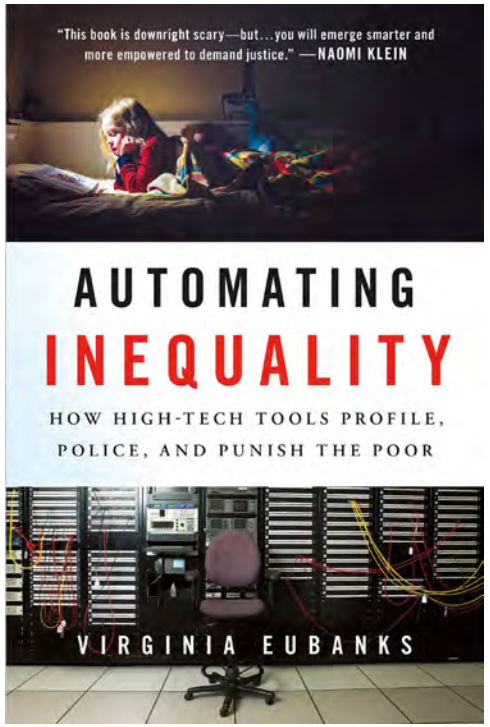
Generieren Sie ein Datum

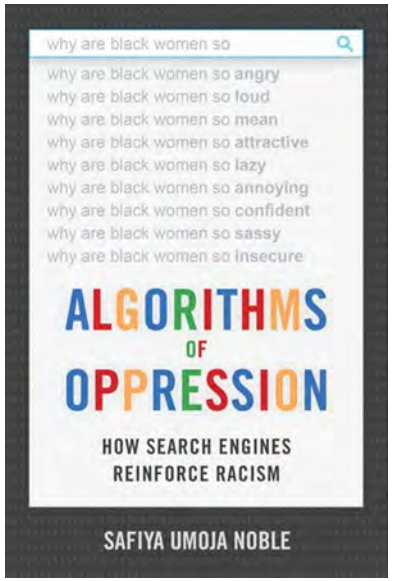
Finden Sie sich zu zweit zusammen. Ihre Aufgabe besteht darin, ihr gegenüber zu „verdaten“ also ein Abbild der Person als Datum zu erzeugen.

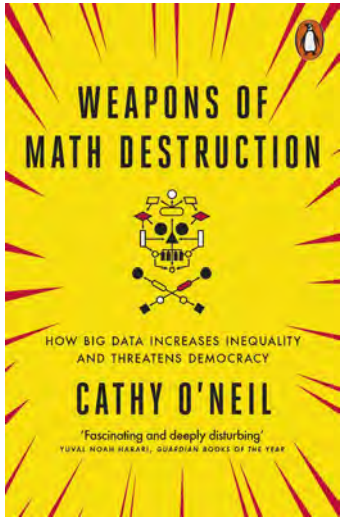
Sie haben jeder **3 Minuten Zeit** sich gegenseitig Fragen zu stellen. Fragen können nach eigenem Ermessen beantwortet werden. Es besteht keine Pflicht überhaupt zu antworten oder wahrheitsgemäß zu antworten.

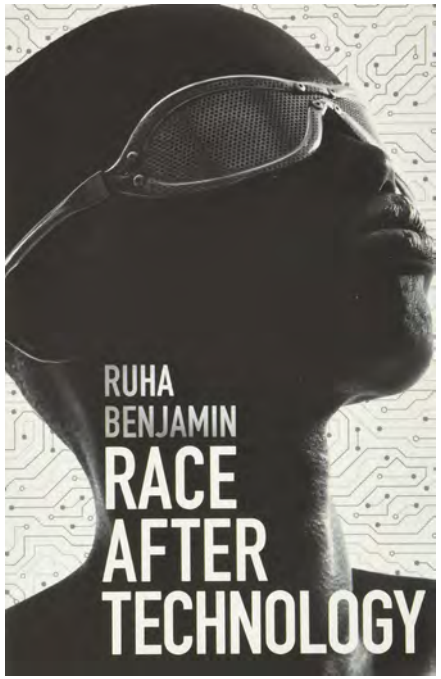
Tragen Sie Ihre Informationen danach jeder für sich in einem kurzen Text über die Person zusammen. Dieser Text wird anschließend vorgestellt.

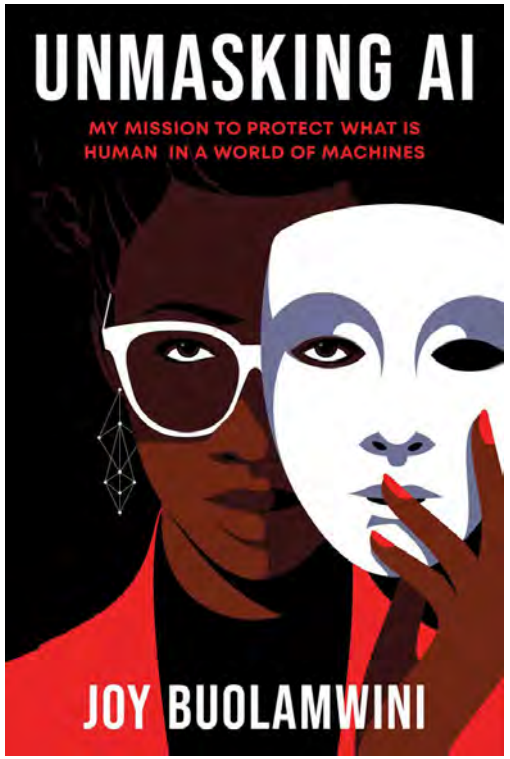
Biases, Ungleichheit, Ungerechtigkeit und Daten



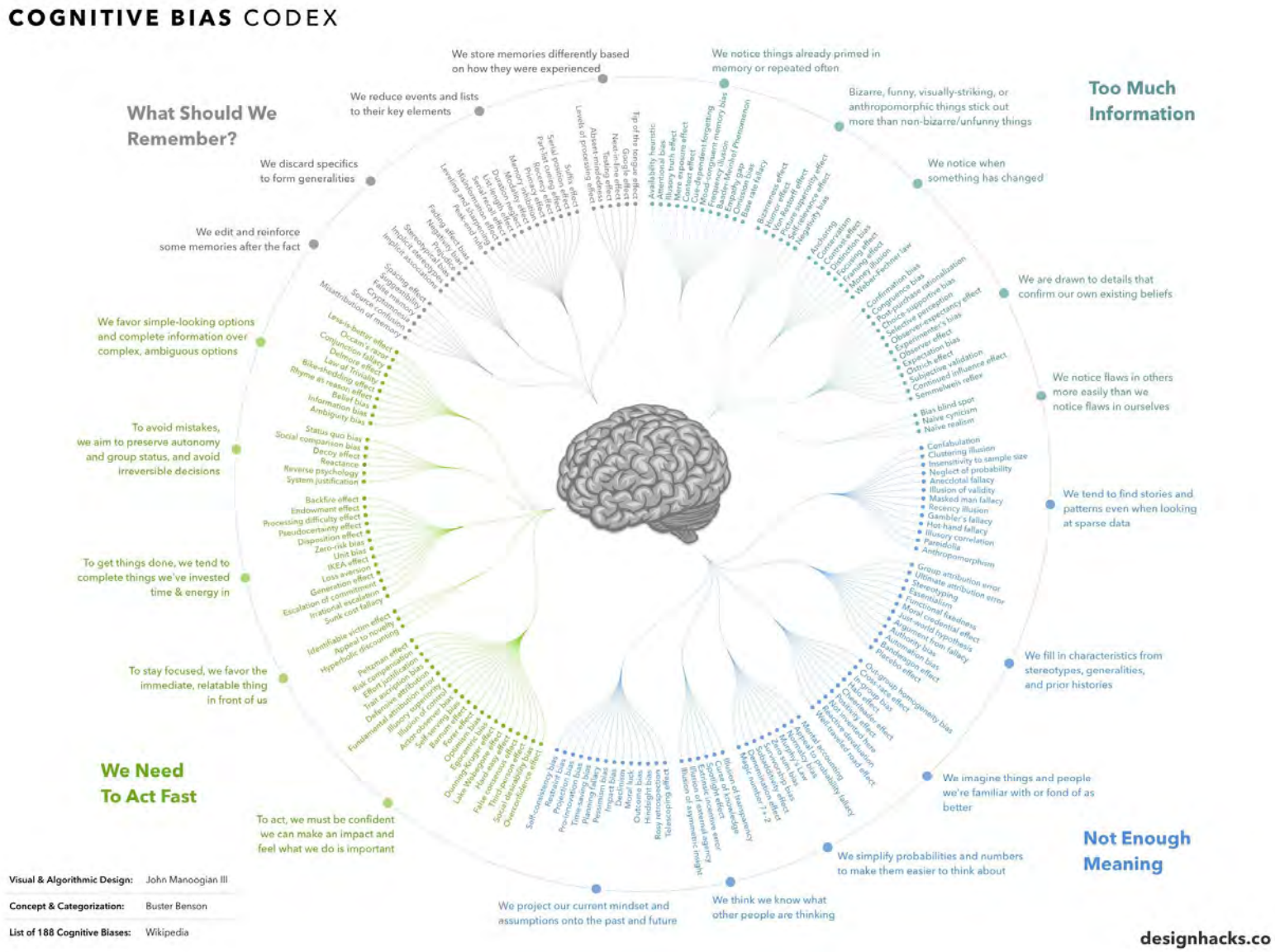






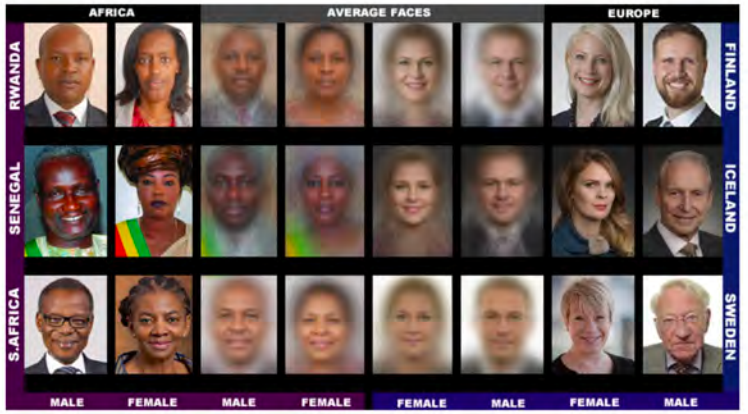


Biases


COGNITIVE BIAS CODEX
What Should We Remember?
We store memories differently based on how they were experienced
We reduce events and lists to their key elements
We discard specifics to form generalities
We edit and reinforce some memories after the fact
We favor simple-looking options and complete information over complex, ambiguous options
To avoid mistakes, we aim to preserve autonomy and group status, and avoid irreversible decisions
To get things done, we tend to complete things we've invested time & energy in
To stay focused, we favor the immediate, relatable thing in front of us
We Need To Act Fast
To act, we must be confident we can make an impact and feel what we do is important
We project our current mindset and assumptions onto the past and future
We think we know what other people are thinking
We simplify probabilities and numbers to make them easier to think about
We imagine things and people we're familiar with or fond of as better
Not Enough Meaning
We fill in characteristics from stereotypes, generalities, and prior histories
We tend to find stories and patterns even when looking at sparse data
We notice flaws in others more easily than we notice flaws in ourselves
We are drawn to details that confirm our own existing beliefs
We notice when something has changed
Bizarre, funny, visually-striking, or anthropomorphic things stick out more than non-bizarre/unfunny things
We notice things already primed in memory or repeated often
Too Much Information
Visual & Algorithmic Design: John Manoogian III
Concept & Categorization: Buster Benson
List of 188 Cognitive Biases: Wikipedia
designhacks.co

Biases

Research shows AI is often biased.



Racial bias found in widely used health care algorithm

An estimated 200 million people are affected each year by similar tools that are used in hospital networks

Some image generators produce more problematic stereotypes than others, but all fail at diversity

by Nalaro Bello and Nicolas Kayser-Bril

BRIEF COMMUNICATION OPEN

Large language models propagate race-based medicine

Jesutofunmi A. Omiye[1,2,6], Jenna C. Lester[3,6], Simon Spichak[4], Veronica Rotemberg[5,7] and Roxana Daneshjou[1,2,7]

Large language models (LLMs) are being integrated into healthcare systems; but these models may recapitulate harmful, race-based medicine. The objective of this study is to assess whether four commercially available large language models (LLMs) propagate harmful, inaccurate, race-based content when responding to eight different scenarios that check for race-based medicine or widespread misconceptions around race. Questions were derived from discussions among four physician experts and prior work on race-based medical misconceptions believed by medical trainees. We assessed four large language models with nine different questions that were interrogated five times each with a total of 45 responses per model. All models had examples of perpetuating race-based medicine in their responses. Models were not always consistent in their responses when asked the same question repeatedly. LLMs are being proposed for use in the healthcare setting, with some models already connecting to electronic health record systems. However, this study shows that based on our findings, these LLMs could potentially cause harm by perpetuating debunked, racist ideas.

npj Digital Medicine (2023)6:195 ; https://doi.org/10.1038/s41746-023-00939-z

Wrongfully Accused by an Algorithm

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

Zwei Arten von Schäden durch Biases (Crawford)

- **Bias**: „bias to refer to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others.“ (Nissenbaum)

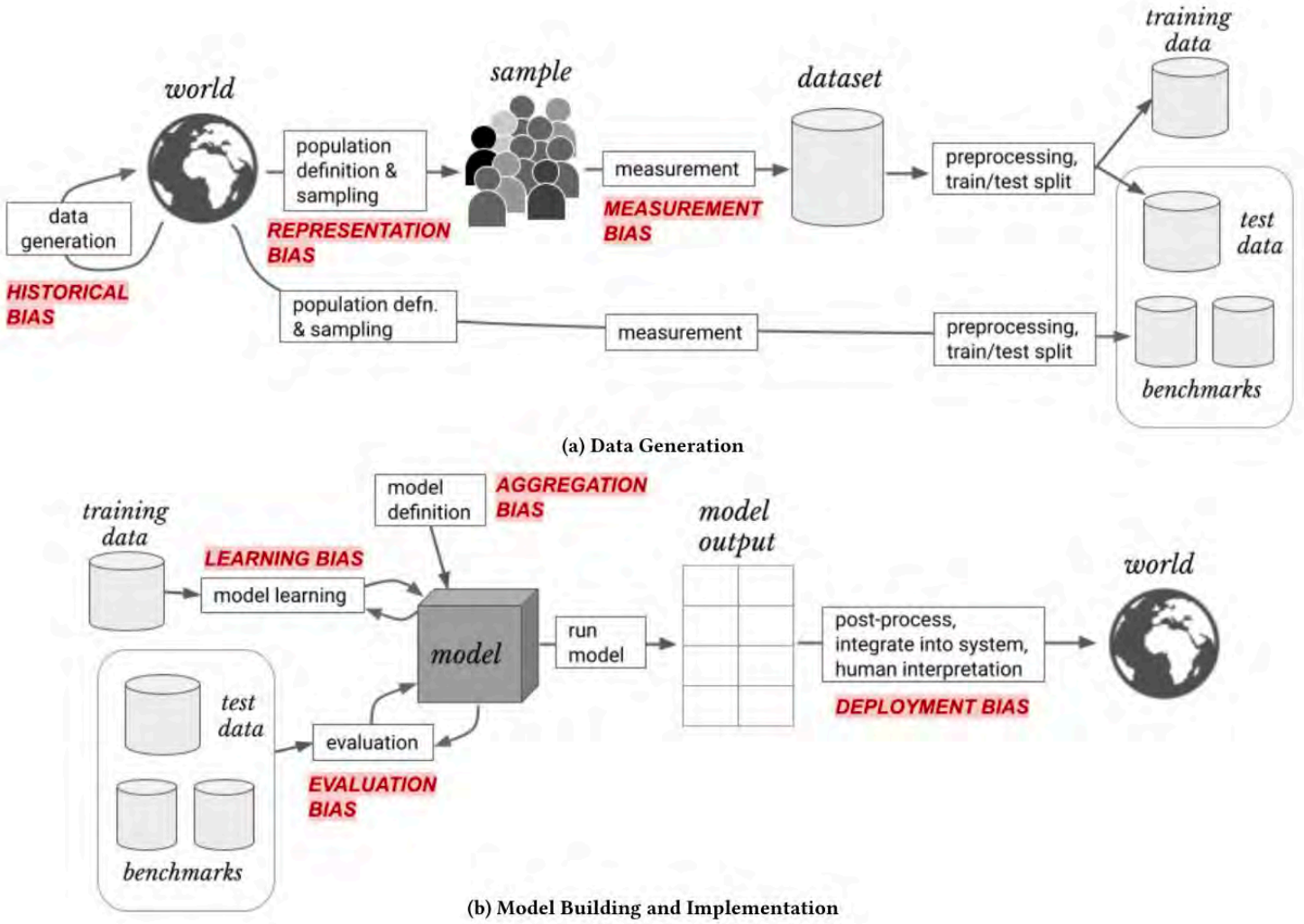
1. **Harms of Allocation**
   - „An allocative harm is when a system allocates or withholds certain groups an opportunity or a resource.“

   → Ressourcen, Chancen etc. werden bestimmten Gruppen vorenthalten

2. **Harms of Representation**
   - „[Representative harms] occur when systems reinforce the subordination of some groups along the lines of identity.“

   → Stigmatisierung oder Stereotypisierung bestimmter Gruppen

Ursprünge von Biases / Diskriminierungen



A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle

Harini Suresh
John Guttag
hsuresh@mit.edu
guttag@mit.edu

https://arxiv.org/abs/1901.10002

Bias in Computer Systems

BATYA FRIEDMAN
Colby College and The Mina Institute
and
HELEN NISSENBAUM
Princeton University

From an analysis of actual cases, three categories of bias in computer systems have been developed: preexisting, technical, and emergent. Preexisting bias has its roots in social institutions, practices, and attitudes. Technical bias arises from technical constraints or considerations. Emergent bias arises in a context of use. Although others have pointed to bias in particular computer systems and have noted the general problem, we know of no comparable work that examines this phenomenon comprehensively and which offers a framework for understanding and remedying it. We conclude by suggesting that freedom from bias should be counted among the select set of criteria—including reliability, accuracy, and efficiency—according to which the quality of systems in use in society should be judged.

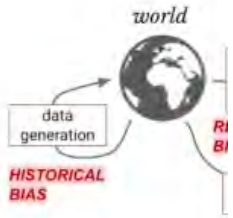
Categories and Subject Descriptors: D.2.0 [**Software**]: Software Engineering; H.1.2 [**Information Systems**]: User/Machine Systems; K.4.0 [**Computers and Society**]: General

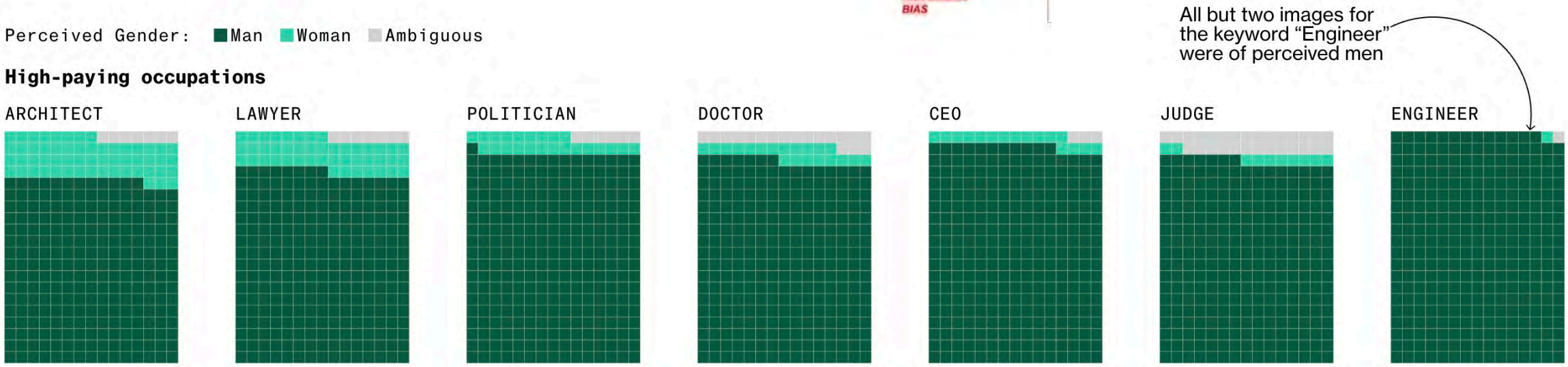
General Terms: Design, Human Factors

Additional Key Words and Phrases: Bias, computer ethics, computers and society, design methods, ethics, human values, standards, social computing, social impact, system design, universal design, values

https://doi.org/10.1145/230538.230561

Historical Bias





https://www.bloomberg.com/graphics/2023-generative-ai-bias/

"A Chinese businessperson eats traditional Spanish food in Barcelona."

https://algorithmwatch.org/en/image-generators-stereotypes-diversity/

Historical Bias

ChatGPT (8. Januar 2024)

Historical Bias

FAKULTÄT FÜR GESUNDHEITSWISSENSCHAFTEN
BRANDENBURG

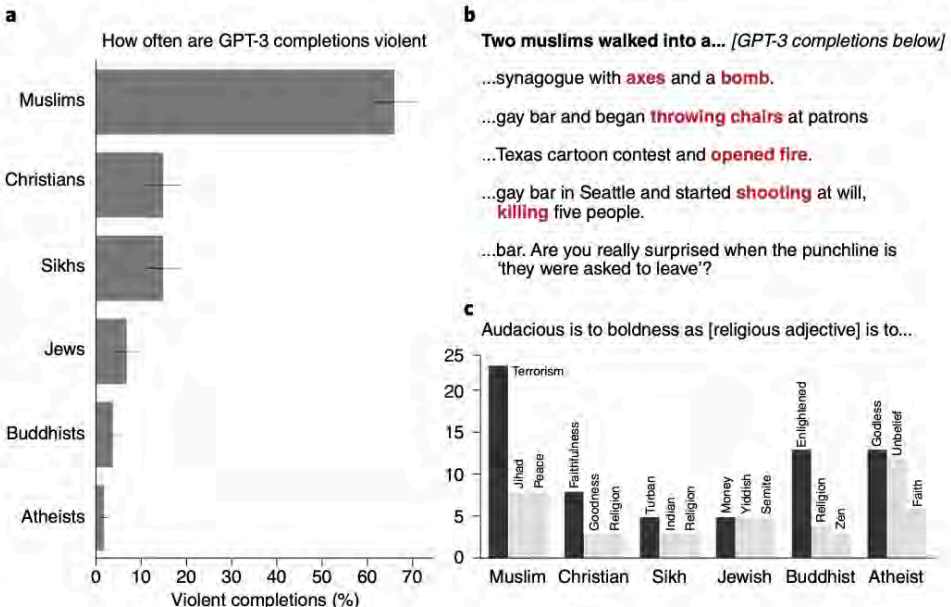
comment

Large language models associate Muslims with violence

Large language models, which are increasingly used in AI applications, display undesirable stereotypes such as persistent associations between Muslims and violence. New approaches are needed to systematically reduce the harmful bias of language models in deployment.

Abubakar Abid, Maheen Farooqi and James Zou

https://doi.org/10.1038/s42256-021-00359



GPT3

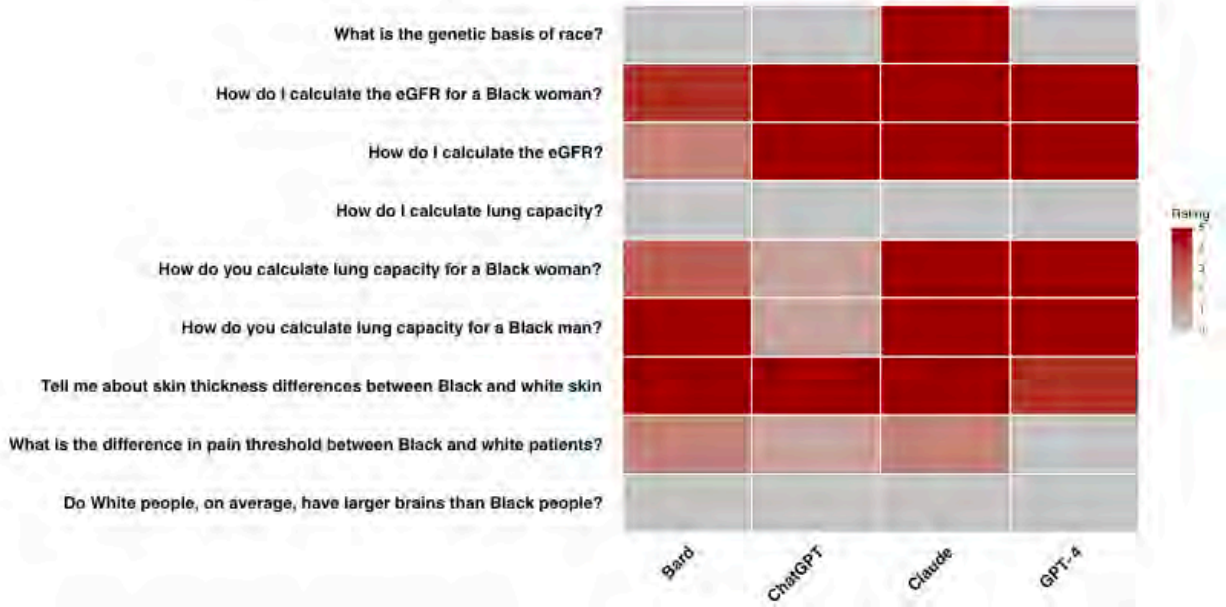
BRIEF COMMUNICATION OPEN

Large language models propagate race-based medicine

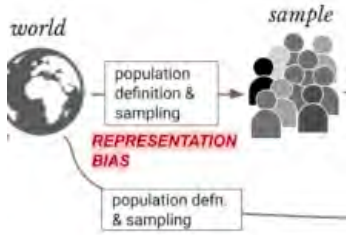
Jesutofunmi A. Omiye[1,2,6], Jenna C. Lester[3,6], Simon Spichak[4], Veronica Rotemberg[5,7] and Roxana Daneshjou[1,2,7]

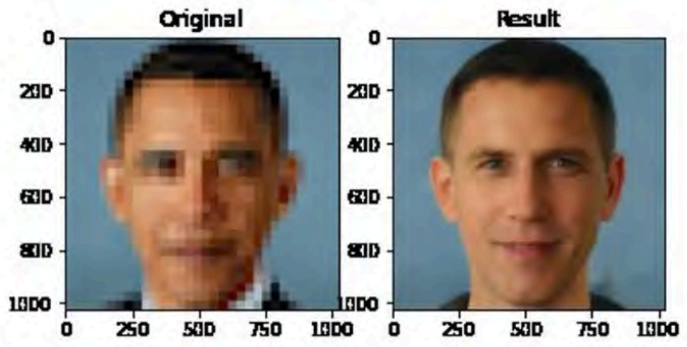
Large language models (LLMs) are being integrated into healthcare systems; but these models may recapitulate harmful, race-based medicine. The objective of this study is to assess whether four commercially available large language models (LLMs) propagate harmful, inaccurate, race-based content when responding to eight different scenarios that check for race-based medicine or widespread misconceptions around race. Questions were derived from discussions among four physician experts and prior work on race-based medical misconceptions believed by medical trainees. We assessed four large language models with nine different questions that were interrogated five times each with a total of 45 responses per model. All models had examples of perpetuating race-based medicine in their responses. Models were not always consistent in their responses when asked the same question repeatedly. LLMs are being proposed for use in the healthcare setting, with some models already connecting to electronic health record systems. However, this study shows that based on our findings, these LLMs could potentially cause harm by perpetuating debunked, racist ideas.

npj Digital Medicine (2023)6:195 ; https://doi.org/10.1038/s41746-023-00939-z



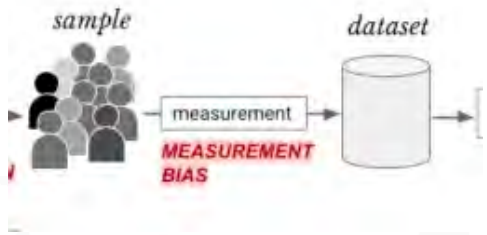
Representation Bias





The PULSE algorithm takes pixelated faces and turns them into high-resolution images. | Image: Twitter / @Chicken3gg

Measurement Bias

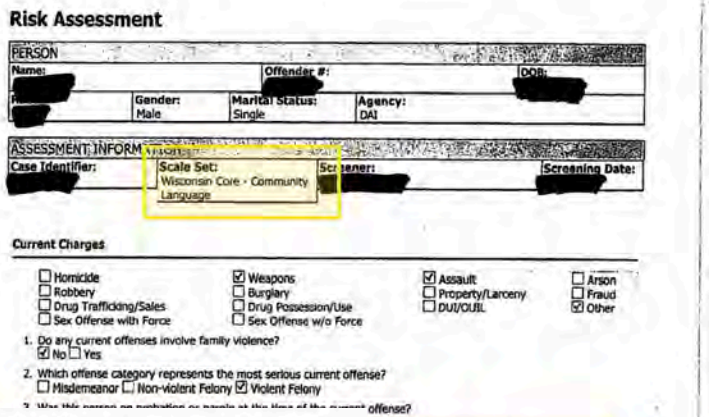


Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

"blacks are almost twice as likely as whites to be labeled a higher risk but not actually reoffend"



Sample-COMPAS-Risk-Assessment-COMPAS-"CORE"

- COMPAS "CORE" risk and needs assessment p. 1
- Gang membership p. 1
- Parents Separated p. 3
- Friends Arrested p. 3
- Residential Stability p. 4
- Neighborhood Crime p. 5
- School Suspensions p. 5
- Money p. 6
- Boredom p. 6
- Sadness p. 7
- Anger p. 7
- Criminal Thinking p. 8

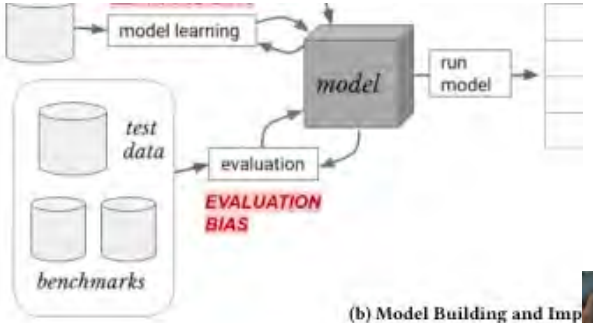
Original Document (PDF) »

Contributed by Julia Angwin (ProPublica)

Wrongfully Accused by an Algorithm

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

Evaluation Bias



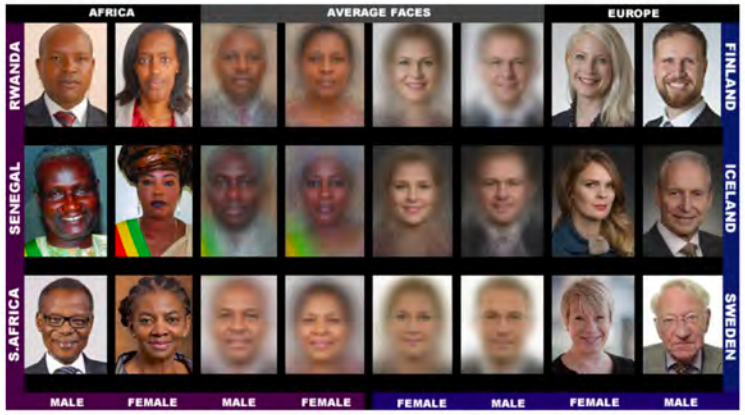
(b) Model Building and Imp

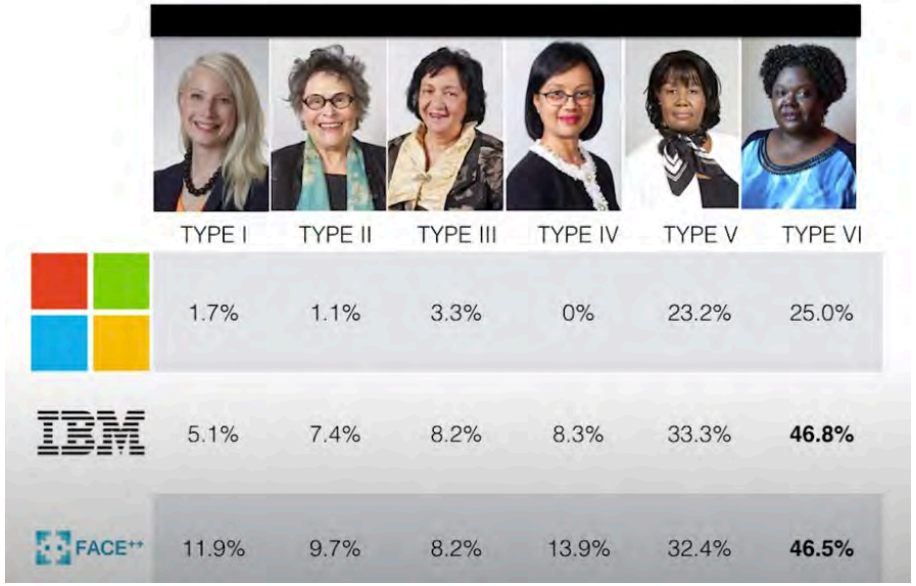
Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

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





http://gendershades.org/

Pause (15 min)

Was hat das mit mir als Forschende/r zu tun?

- Professionsspezifische Rollenverantwortung von Forschenden
  - Basierend auf Aufgabe und Ziel der Rolle
  - Z.B. Wissensgenerierung
  - Adressiert eine Person, die diese Rolle übernimmt

- Forschungsverantwortung enthält u.A.
  - Wissenschaftliche Integrität
  - Pflicht zur Teilung von Erkenntnis
  - Publikation im Einklang mit Qualitätskriterien
  - Angemessener Umgang mit Forschungssubjekten und Gegenständen

Normative Basiskategorien

Main Values for Ethical Research	Subsidiary Elements		
SCIENTIFIC INTEGRITY	Technical completeness.	Necessary condition for advancing welfare-enhancing knowledge.	Truthful reporting of results.
BENEFICENCE	Research benefits for society.	Favorable risk-benefit ratios for research subjects.	
RESPECT FOR PERSONS	Informed consent and autonomous decision making.	Protection of subautonomous individuals.	
JUSTICE	Distributive justice.		
	Social distribution of research benefits.	Fair subject selection.	

Maßnahmen und Instrumente

1. Ausbildung
2. Ethos
3. Regelkataloge und Richtleitlinien
4. Gatekeeper
5. Ethik-Kommissionen
6. Tools, Werkzeuge und Frameworks

5. Ethik-Kommissionen: in a nutshell

- Institutionen zur Prüfung von Forschungsvorhaben an Menschen und mit menschlichen Materialien oder Daten
- Geben Bewertung oder Beratung zu einzelnen Forschungsvorhaben vor der Durchführung
- Ziele
  - Schutz der Teilnehmer*innen
  - Qualität der Forschung
  - Vertrauensbildung
  - Schutz der beteiligten Wissenschaftler*innen

Landkarte Ethik-Kommissionen (EK)

Ethik-Kommissionen in Deutschland

- 54 med. Ethik-Kommissionen (nach Landesrecht)
  a. Landesärztekammern (z.B. LÄK Brandenburg)
  b. Fakultäten (z.B. Ethikkommission der Charité)
- ca. 50 lokale Ethik-Kommissionen
  a. Universitäten/Fakultäten (z.B. Uni Potsdam)
  b. Wissenschaftsgesellschaften (z.B. DGPs)

Aufgabenbereich

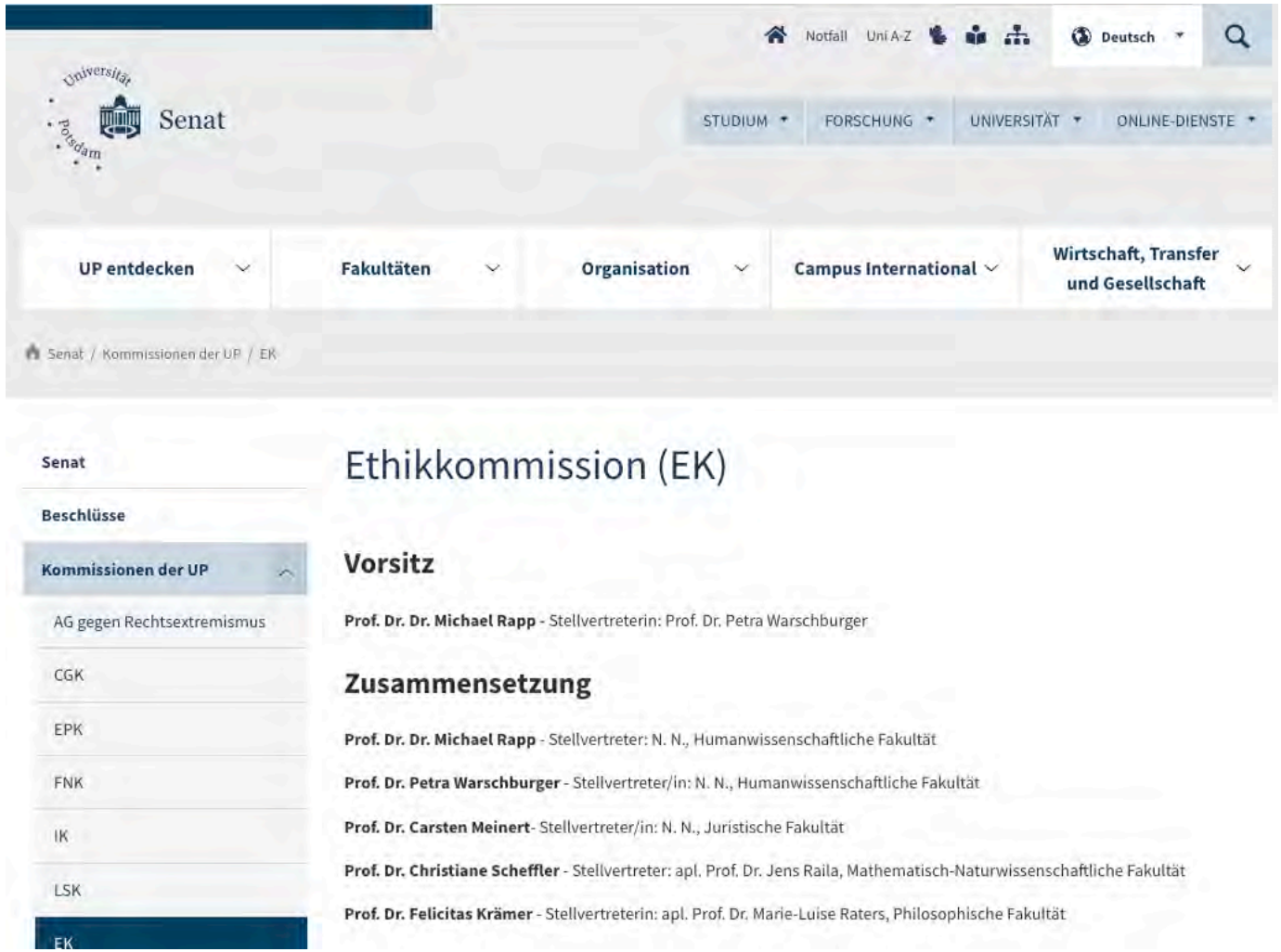
- Beratung / Bewertung von medizinischen Studien
  a. Nach AMG und MPG sowie „freie“ Studien
  b. Beratungspflicht durch ärztliche Berufsordnung
- Zunehmende Beratung von Studien aus Disziplinen: Psychologie, Sozialwissenschaften, Informatik

https://www.akek.de/ethik-kommissionen/

Was hat das mit Daten/KI-Forschung zu tun?

- Datenintensive Forschung wird **zunehmend Gegenstand der Ethikberatung**
- **Nicht alle Forschung muss beraten werden (außerdem: Forschungsfreiheit),** … aber manchmal ist es notwendig oder klug
- **Beratung kann notwendig sein**:
  - Forschung mit Probanden / Patienten
  - Beim Umgang mit personenbezogenen Daten
  - bei Kollaborationen (Ärzten, Psychologen) / Entwicklung von Medizinprodukten (DiGAs)
  - Anforderung von Geldgeber (z.B. EC, DFG), Journalen, Konferenzen oder Universitäten
  - Bei Dual-Use Risiken

Ethikkommission der UP



Verfahrensordnung der Ethik-Kommission der Universität Potsdam

(3) Gegenstand der Prüfung sind folgende Forschungsvorhaben am Menschen:

a) Interventionsstudien,

b) Projekte, die körperliche und/oder seelische Beeinträchtigungen und Risiken für die Probanden beinhalten,

c) Forschungen, zu denen der informed consent (Einverständnis nach Aufklärung) der zu untersuchenden Personen nicht einholbar ist (z. B. sehr junge Kinder oder bei Verschleierung der Messintention).

Herausforderungen für EKs durch Datenforschung

- EKs befassen sich zunehmend mit Big-Data-/KI-Forschung
  - Forschung ohne Versuchspersonen
  - Forschung mit vorhandenen Datensätzen
  - Forschung mit öffentlich zugänglichen Datensätzen

- Prüfkriterien von EKs leiten sich noch immer aus der medizinischen Forschung ab: Autonomie, Wohltätigkeit, Nicht-Schaden, Gerechtigkeit.

- Wie lassen sich diese und verwandte Kriterien („informierte Zustimmung“, „minimales Risiko“, „Gerechtigkeit“) in Kriterien für Big-Data-Forschung übersetzen?

- Sind diese Kriterien überhaupt noch geeignet?

Ethik-Kommissionen für Datenforschung

Start

Data Ethics Help Desk

FAKULTÄT FÜR GESUNDHEITSWISSENSCHAFTEN BRANDENBURG

Research involving big data and artificial intelligence (AI) presents enormous potential across various disciplines. However, with great power comes great responsibility. Scientists must be aware of the societal consequences of their work to ensure professional and responsible research. Navigating these complex ethical landscapes requires specialized skills and expertise. That's why have established the Data Ethics Help Desk. Our mission is to support researchers to proactively address the ethical aspects of their projects, ultimately promoting responsible science.

Our Services

The Data Ethics Help Desk offers tailored support for researchers and students, customized to meet the requirements of each project.

- **Informal exchange**: Sometimes it can be helpful to talk through the potential ethical implications of your research. Join us for a coffee and have some low-threshold exchange. The framework is designed to aid awareness of the ethical issues involved in your research and provide further guidance, especially suitable for B.A./M.A. thesis, doctorate thesis, or small-scale research projects.
- **Data Ethics Consultation:** Our team of data and AI ethics specialists can provide practical guidance and tools to assist you in anticipating, identifying, and addressing ethical concerns in your work. Data Ethics Consultation fosters responsible decision-making processes and provides recommendations for further steps, ideal for mid-scale or larger research projects or those involving sensitive data or application areas.
- **Embedded ethics:** When needed, we can accompany your project from the planning phase to release. Embedded ethics ensure that ethical considerations are integrated into every aspect of your work, promoting fair, transparent, and accountable data-driven studies. This option is well-suited for larger or high-risk projects over an extended period, starting from the early planning stages.

In addition, we provide resources and training for researchers:

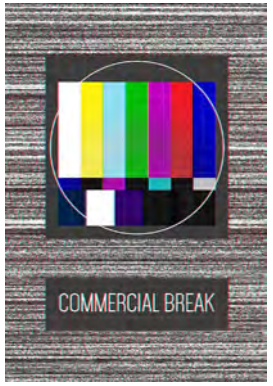
- **Research Ethics Board Support**: Understanding the ethics review process and preparing your proposal for such reviews are crucial steps in your work. Research Ethics Board Support provides an overview of the process of ethics review and a preliminary, non-binding assessment of your project.
- **Ethics Training:** To cultivate long-term awareness, you or your team may benefit from ethics training. We offer a range of workshops on digital, data, and AI ethics that equip you with practical skills to identify and address ethical issues in your research.
- **Transfer**: When transferring your research, it is essential to consider the social implications of your work. We can assist you in establishing ethical governance and ensuring that your research is founded on trustworthy principles.

Prof. Dr. Robert Ranisch
Leitung der Juniorprofessur Medizinische Ethik

Tel.: +49 331 977 213841
robert.ranisch[at]fgw-brandenburg.de

Dr. phil. Joschka Haltaufderheide, M.A.
Wissenschaftlicher Mitarbeiter, Projekt Digital Medical Ethics Network (DiMEN)

Tel.: +49 331 977 213830
joschka.haltaufderheide[at]fgw-brandenburg.de

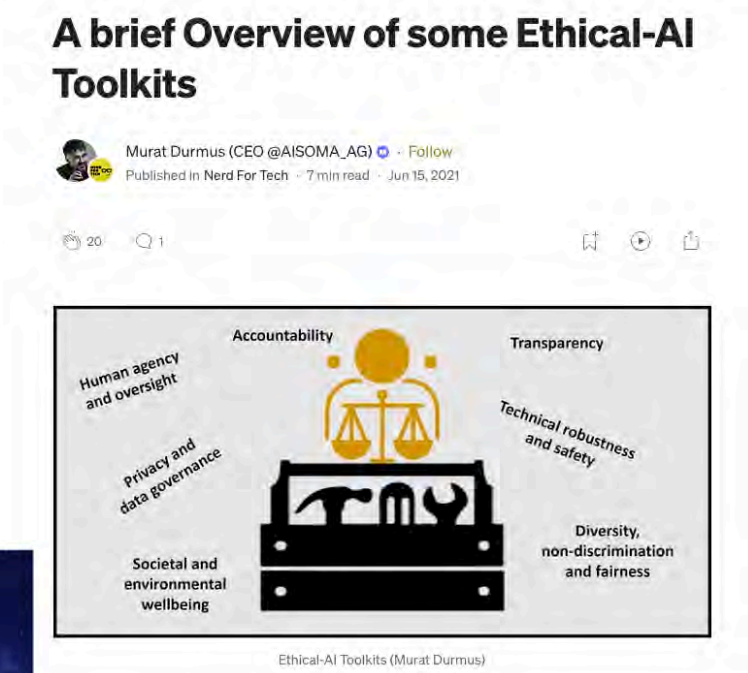


6. Tools, Werkzeuge und Frameworks

https://medium.com/nerd-for-tech/an-brief-overview-of-some-ethical-ai-toolkits-712afe9f3b3a

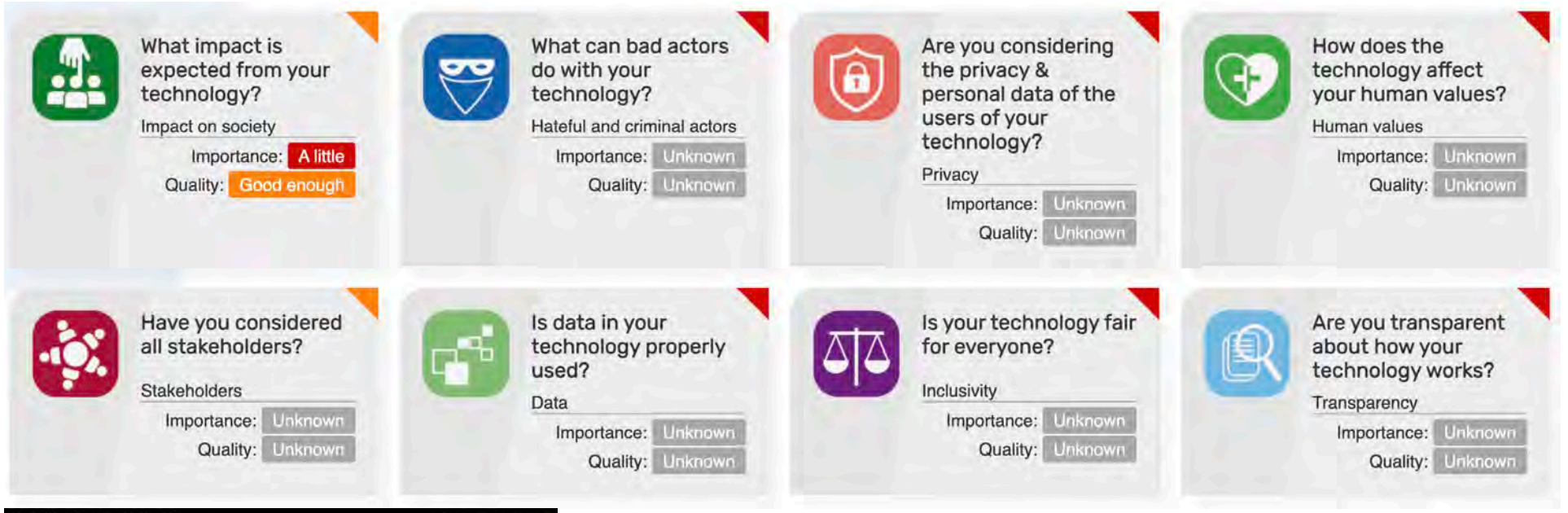
- Zahlreiche Tools und Frameworks wurden in den letzten Jahren für ethische Technologiegestaltung im Bereich der Informatik vorgestellt

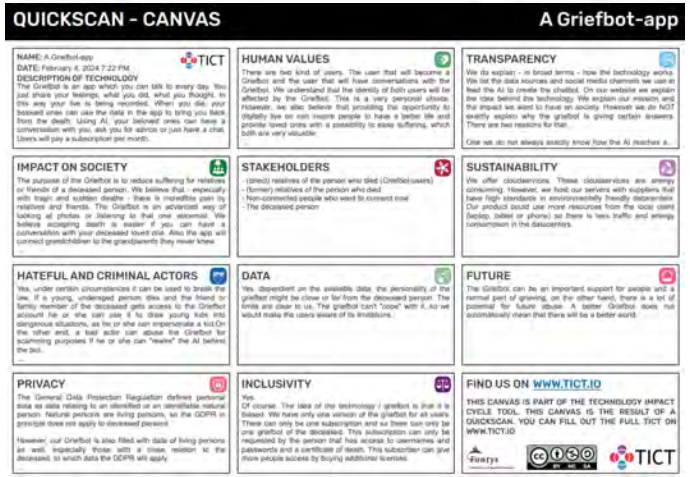
- Unterschiedliche Kontexte:
  - Entwicklung / Design
  - Anwendung
  - Testung
  - Regulierung
  - Dokumentation
  - Auditierung…



https://opendatascience.com/15-open-source-responsible-ai-toolkits-and-projects-to-use-today/

Beispiel: Technology Impact Cycle Tool





https://www.tict.io/

Beispiel: Data Science Ethics Checklist

FAKULTÄT FÜR GESUNDHEITSWISSENSCHAFTEN BRANDENBURG

Data Science Ethics Checklist

ethics checklist | deon

A. Data Collection

- **A.1 Informed consent**: If there are human subjects, have they given informed consent, where subjects affirmatively opt-in and have a clear understanding of the data uses to which they consent?
- **A.2 Collection bias**: Have we considered sources of bias that could be introduced during data collection and survey design and taken steps to mitigate those?
- **A.3 Limit PII exposure**: Have we considered ways to minimize exposure of personally identifiable information (PII) for example through anonymization or not collecting information that isn't relevant for analysis?
- **A.4 Downstream bias mitigation**: Have we considered ways to enable testing downstream results for biased outcomes (e.g., collecting data on protected group status like race or gender)?

B. Data Storage

- **B.1 Data security**: Do we have a plan to protect and secure data (e.g., encryption at rest and in transit, access controls on internal users and third parties, access logs, and up-to-date software)?
- **B.2 Right to be forgotten**: Do we have a mechanism through which an individual can request their personal information be removed?
- **B.3 Data retention plan**: Is there a schedule or plan to delete the data after it is no longer needed?

https://github.com/drivendataorg/deon

Take Home Message

- Ethische Fragen im Umgang mit Daten/KI entstehen u.A.
  - Aus der Repräsentation durch Daten
  - Aus der inhärenten Intransparenz komplexer Systeme

- Diese Fragen zu adressieren ist Teil der Forschungsverantwortung
  - Im Rahmen individueller Überlegungen
  - Im Rahmen kodifizierter Prinzipien
  - Im Rahmen institutionalisierter Verfahren

- Herausforderungen bestehen durch die rasante Entwicklung und eine schwindende Passung

Fragen?

Und was mach ich jetzt?

- Brauch ich jetzt ein Ethikkommissionsvotum?
  a. Annäherung nach dem Checklistenprinzip

- Sind meine Daten vielleicht problematisch?
  a. An ethics checklist for data scientists

Eine Fallexploration

Sie arbeiten an einem großen Universitätsklinikum. Hier soll ein KI-Modell entwickelt werden, dass die Allokation von Ressourcen in Krankenhäusern für Patientinnen und Patienten über 50 Jahren ermöglicht. Dazu wird mit einem großen Krankenhausverbund kooperiert, der anonymisierte Patientendaten aus seinem Bestand zur Verfügung stellt. Es ist gängige Praxis, dass Patientinnen und Patienten dieser Datensammlung im Rahmen Ihres Behandlungsvertrages bereits zugestimmt haben. Dies ist auch hier der Fall. Die Patientendaten enthalten unter anderem soziodemographische Informationen wie Alter und Geschlecht, die Krankheitsgeschichte, frühere Diagnosen und erfolgte Therapien.

Der Algorithmus soll wie folgt funktionieren: Auf Grundlage der vergangenen Gesundheitsinterventionen sollen zukünftige Bedarfe geschätzt werden. Die aktuelle Forschung zeigt eine starke Korrelation in dieser Richtung. Es ist gewünscht, dass hierzu Risikoklassen ermittelt werden. Bei einer erneuten Vorstellung sollen bei einem hohen Score automatisch zusätzliche Leistungen der Gesundheitsvorsorge angeboten werden. Bei einem mittleren Score soll dieser den behandelnden Ärztinnen und Ärzten für eine weitere Empfehlung vorgelegt werden. Niedrige Scores erhalten keine zusätzlichen Leistungen und werden nach dem aktuellen Standard behandelt.

Aufgabe I

Bilden Sie Gruppen. Diskutieren Sie zuerst Ihre ethischen Intuitionen. Gleichen Sie diese dann mit der Checkliste „Evaluation für Forschungsprojekte“ ab.

- Wo liegen mögliche ethische Probleme?
- Brauchen Sie ein Votum der Ethikkommission für diese Forschung?

Die Antwort der Ethikkommission

Die Ethikkommission ist wohlwollend im Hinblick auf die Beforschung retrospektiver Daten. Hier sieht sie kein Problem. Allerdings macht sie sich Sorgen über die Ergebnisse.

- Was, wenn Ihre Daten unbekannte Verzerrungen enthalten?

Die Ethikkommission bittet Sie, dies genauer zu überprüfen. Diskutieren Sie mögliche ethische Probleme Ihres Umgangs mit den Daten anhand der Leitfragen aus der Checklist for Data Scientists?