C³W Chaos Computer Club Wien

sicherheitskultur.at

Problematische Algorithmen greifen in unser Leben ein – was sind die Probleme und mögliche Lösungsansätze

Philipp Schaumann

https://sicherheitskultur.at/Manipulation.htm

Wo werden Algorithmen (sinnvoll ?) eingesetzt ?

- Algorithmen übernehmen die Kontrolle, wo Menschen zu langsam sind (z.B. bei Waffen)
- Algorithmen sind im Einsatz wo dadurch Arbeitsplätze eingespart werden können
- Algorithmen sind im Einsatz wo dadurch zusätzliche „Erkenntnisse“ gewonnen werden können (z.B. bei Bewertungen)

Algorithmen treffen Entscheidungen über das Schicksal von Menschen

Deep Learning Algorithmen machen (systembedingte) Fehler

- **Die obere Zeile zeigt jeweils das Originalschild, die Zeile drunter zeigt wie das Schild verändert wurde und der Text gibt dann an, wie das System das Schild interpretiert hat.**

https://arxiv.org/abs/1802.06430

Algorithmen steuern die Welt
Seit 2013: Vorhersage von Kriminalität (Predictive Policing)

Computer analysieren Orte im Hinblick auf frühere Straftaten (oder Ereignisse), werten manchmal Überwachungskameras aus und schicken Polizisten dorthin, wo mit höchster Wahrscheinlichkeit ein Verbrechen passieren wird.

Dort, wo Polizisten Passanten überprüfen, dort finden sie auch auffälliges – Selbstbestätigung der Vorhersage

Algorithmen steuern die Welt – Seit 2013: Vorhersage von Rückfallswahrscheinlichkeit

Computer analysieren die Wahrscheinlichkeit, dass ein Gefängnis-Insasse mit einer bestimmten Geschichte und einem bestimmten sozialen Umfeld rückfällig wird.
Sie entscheiden über (vorzeitige) Entlassung und schlagen auch das Strafmaß vor.

http://www.mosac.mo.gov/page.jsp?id=45498

Studies and Tests: Algorithms do have bias

2016: Studien auf Grund von Präsident Obama:

- White House report on Big Data . . . cautions against re-encoding bias and discrimination into algorithmic systems.
  https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-and-civil-rights

- Software used to predict future criminals is biased against blacks
  https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

- What Algorithmic Injustice Looks Like in Real Life
  https://www.propublica.org/article/what-algorithmic-injustice-looks-like-in-real-life

2020 Studien zu „Algorithmus Correctional Offender Management Profiling for Alternative Sanctions“ (Compas) und LSI-R (Level of Service Inventory-Revised):

- Vorhersage-Qualität zwischen 60 und 70%
- Einsatz in EU nicht OK, aber …. Kasse Hamburg zum Erkennen Betrugsverdacht, Zoll zur für prüfwürdige Steuererklärungen
- ABER: Datensätze vorteilhaft für das Training von Richtern

Details http://www.heise.de/-4661585

Diskriminierung auch auf Grund von (harmlosen) Gruppenmerkmalen

Typischerweise soll die Diskriminierung über Merkmale wie

- Rasse / ethnische Herkunft
- Geschlecht
- Sexuelle Orientierung
- Alter
- Religion

vermieden werden

Der gleiche Effekt kann aber auch bei anderen Gruppenmerkmalen passieren, z.B.

- typische Wohngebiete (Gräzel)
- typische Sprachstil (z.B. für Farbige in den USA)
- typische Musikauswahl
- . . .

Amazon – Recruiting thru Deep Learning stopped

Der Algorithmus diskriminierte gegenüber Frauen, weil er auf der bisherigen Einstellungspraxis beruhte.

https://www.heise.de/newsticker/meldung/Amazon-KI-zur-Bewerbungspruefung-benachteiligte-Frauen-4189356.html

Bundestagsstudie: Robo-Recruiting birgt hohe Diskriminierungsgefahr

https://www.heise.de/news/Bundestagsstudie-Robo-Recruiting-birgt-hohe-Diskriminierungsgefahr-4875132.html

Thema Bewerbung für einen Job

Einstellungsgespräche waren immer subjektiv, aber . . .

Die Vorurteile des Einstellenden werden ersetzt durch die mathematische Willkür eines unbekannten Algorithmus mit unbekannter Parametrisierung und fragwürdigen Input Daten.

Die Inputs des Algorithmus sind z.B.

- Bewerbungsunterlagen (und andere Texte) der jetzigen Angestellten durch Deep Learning analysiert
- Spuren der BewerberIn im Web (oder die fehlenden Spuren)
- „Friends“ in Social Networks, deren Postings, Wohnorte, Likes,
- Konsumverhalten der BewerberIn,
- …..

Einstellung auf Grund von Persönlichkeitsanalysen

Tweets verraten (angeblich) mit 70% Wahrscheinlichkeit, ob jemand eine psychische Erkrankung hat. Alle diese Kandidaten werden nicht eingeladen.

Genauso können verwendet werden können aber auch die Telefon-Meta-Daten
(Zahl eingehender versus ausgehender Anrufe, Gesprächsdauer für jede Kategorie, Uhrzeit der Gespräche, etc.)

Einstellung auf Grund von Key-Words

Nicht wirklich besser:
Simple Algorithmen, die die Häufigkeit von bestimmten (recht willkürlichen) Schlüsselwörter prüfen. (Beispiele: innovative, motivated, dynamic, organized, reliable, honest, creative, experience, helped, supervised, confidence, consistent, …)

https://www.cvplaza.com/cv-basics/cv-power-words/
https://www.jobsite.co.uk/worklife/how-to-use-keywords-cv-7317/
https://www.cv-library.co.uk/career-advice/cv/how-use-keywords-cv/
https://www.callcentrehelper.com/the-top-25-words-to-use-on-your-cv-10032.htm
https://www.reed.co.uk/career-advice/what-words-should-i-use-on-my-cv/

Einstellung auf Grund von Persönlichkeitsanalysen

Es wurde eine hohe Korrelation zwischen Spitzenleistungen beim Programmieren und einer bestimmten Variante vom Manga-Comics gefunden.

Nur wenn die IP-Adresse der Bewerberin in den Logs dieser Website aufscheint, gibt es einen Job.

Der Standard, 8.11.2014

Algorithmische Modellierung bestimmt unsere virtuelle Repräsentanz

- die Bank bei der Kreditvergabe,
- der Mobilfunkanbieter – Handyvertrag vs. Wertkarte
- das Dating App beim Vorschlag von „Matches“
- das Call-Center wenn mensch dort anruft und über einen besseren Tarif reden möchte
- bei der Bewertung eines möglichen Terrorismus-Risikos (?)
- . . . und da geht noch viel mehr, siehe China

Die virtuelle Repräsentanz wird wichtiger als die reale Person.

Algorithmen steuern die Welt

2013: Vorhersage von Terrorismus

Computer analysieren Ihr Verhalten im Internet, die Vernetzungen und die Kommunikation und sagen vorher, mit welcher Wahrscheinlichkeit Sie zum “Trouble Maker” werden.

Zukünftige „Trouble Maker“ kommen auf die No-Fly Liste oder werden bei jedem Flug separat verhört.

Die Gedanken hören auf, frei zu sein.

Beispiel Arbeitsamt Österreich (AMS)

- Arbeitslose werden künftig (2019) in drei Gruppen A, B, C eingeteilt und zwar in jene mit hohen, mittleren und niedrigen Chancen am Arbeitsmarkt.
- Wer mit 66-prozentiger Wahrscheinlichkeit innerhalb von sieben Monaten wieder einen Job haben wird, soll ab 2019 als Person mit hoher Arbeitsmarktchance gelten, Gruppe A.
- Wer weniger als 25 Prozent Chance hat innerhalb von zwei Jahren einen Job zu bekommen, gilt dann als Kunde mit niedrigen Chancen und kommt in Gruppen B oder bei ganz schlecht, in C.
- Förderungen, z.B. Schulungen, gibt es hauptsächlich für Gruppe B.

https://derstandard.at/2000089720308/Leseanleitung-zum-AMS-Algorithmus

Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective
https://www.frontiersin.org/articles/10.3389/fdata.2020.00005/full

Beispiel Arbeitsamt Österreich (AMS) (2)

- Grundsätzliche Probleme der „Stabilisierung“ von Benachteiligungen durch Rückkopplungseffekte (wer in die Gruppe „schlecht vermittelbar“ eingestuft wird bekommt weniger Hilfen)
- Variable wie ‚Frau‘ wird generell angewendet, ohne Rücksicht auf den spezifischen Arbeitsmarkt, das gleiche gilt für ‚Migrationshintergrund‘
- Nur 3 Altersgruppen, Probleme an den Grenzen
- Noch weniger transparent als diese Variablen sind die spezifischen Gewichtungen im Algorithmus die darüber entscheiden, wie schwer jeder Faktor in die Entscheidung eingreift
- Unterschiedliche Auswertungen je nach engen lokalen Anforderungen, bezogen nur auf Wohnort, nicht Wunscharbeitsort. D.h. Regionen die von 1 Branche oder Arbeitsgeber dominiert sind führen zu ganz anderen Ergebnissen
- Alle Branchen werden in nur 2 sehr grobe Klassen eingeteilt: Service oder Produktion.
- Mögliche Diskriminierungen von groben Gruppen wird zu: ‚. . . the system captures the “harsh reality” of the labor market by making realistic predictions for job seekers belonging to disadvantaged groups‘
- Viele wichtige Parameter wie Eigeninitiative, Social Skills, Berufserfahrung, Sprachkenntnisse in Fremdsprachen, etc. gehen in die Berechnungen nicht ein

Ausführliche Studie zu den “Problemen” des Algorithmus und zu seinem Einsatz:
https://www.frontiersin.org/articles/10.3389/fdata.2020.00005/full

Beispiel Arbeitsamt Österreich (AMS) (3)

Die Umsetzung im Arbeitsalltag

- Die Trefferraten liegen angeblich um 90%. Ursprünglich wurde der Algorithmus als 'first opinion' klassifiziert, nach Kritik als 'Second Opinion', was 'overruled' werden kann, was aber begründet werden muss (d.h. Mehrarbeit und geringere Effektivität der Bearbeiterin.
- Es ist fraglich, ob diese Mehrarbeit leistbar ist, weil das Ziel des Algorithmus die Erhöhung der Effizienz der Beratung ist. zu viele hochstufungen reduziert die "Performance" der Beraterin und der AMS-Stelle.
- Letztendlich ist die primäre Zielsetzung „Effektivität“ und nicht „Fairness“. Und die Kernparameter bestimmt die Politik, nämlich wie viel Geld für die Unterstützung von Arbeitslosen verfügbar gemacht wird

Ausführliche Studie zu den “Problemen” des Algorithmus und zu seinem Einsatz:
https://www.frontiersin.org/articles/10.3389/fdata.2020.00005/full

“MathWashing” Algorithmen für eine weiße Weste

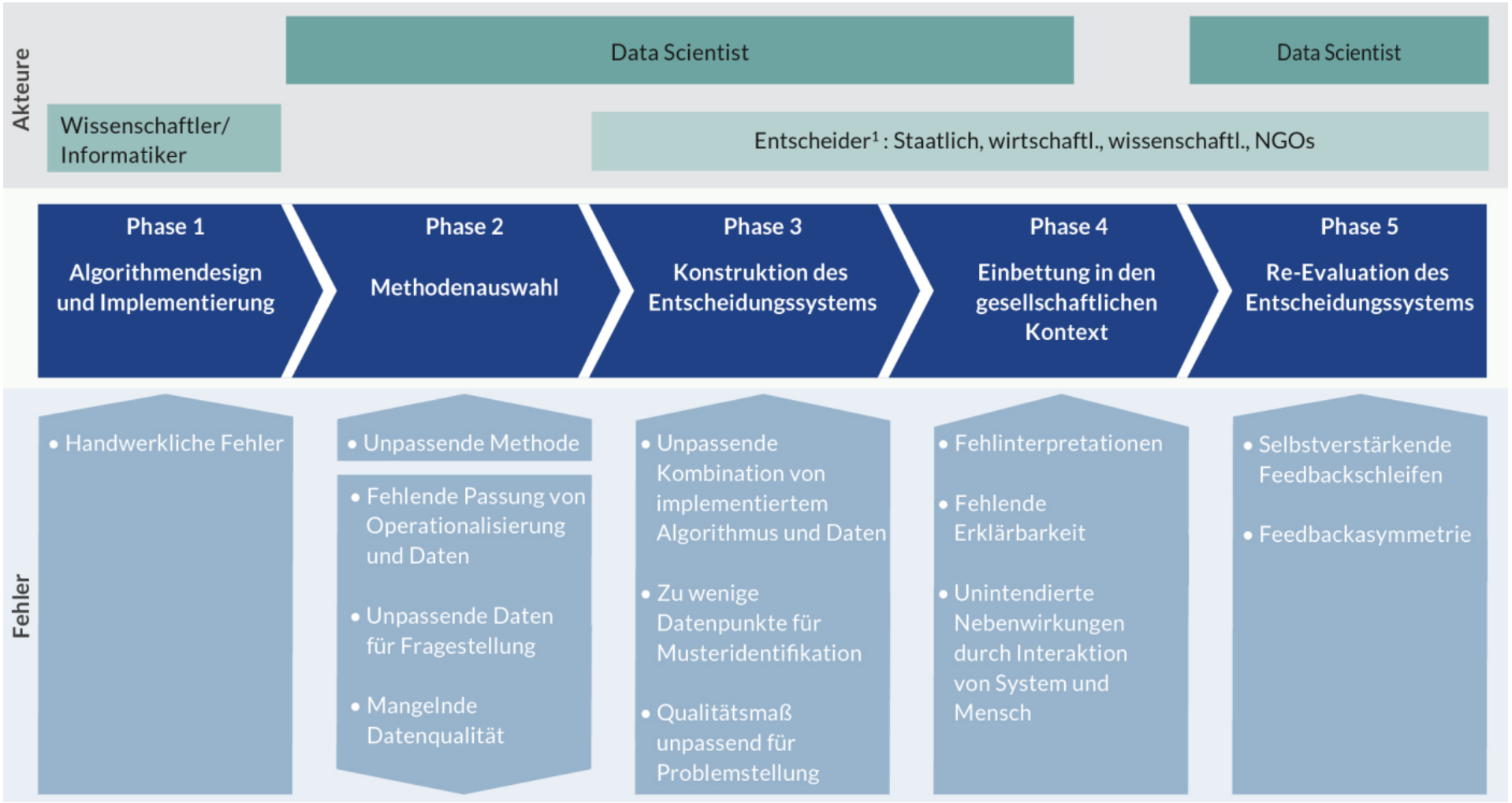
„Ich würde ja gern …. Aber der Computer sagt NEIN“

Absichtliches Verstecken hinter dem Algorithmus ….
durch Rausreden auf „wertfreien“, „objektiven“ Algorithmus. Siehe Facebook, Google, etc.: Wir sind ja nur eine Plattform und wollen keine Verantwortung für Inhalte.

Unabsichtliches Verstecken hinter dem Algorithmus ….
Targeted Advertising (gezielte Anzeigen) die manchmal eine Re-Traumatisierung auslösen (z.B. Diätmittel für Anorektiker, Glückspiel für Süchtige, etc.)

Den Begriff „Mathwashing“ hat Fred Benenson geprägt.
https://technical.ly/brooklyn/2016/06/08/fred-benenson-mathwashing-facebook-data-worship/
Quelle viele Materialien: Bertelsmann Stiftung https://algorithmenethik.de/mathwashing/

Mögliche Fehlerquellen bei algorithmischen Systemen



Prof. Dr. Katharina A. Zweig, TU Kaiserslautern: Wo Maschinen irren können
Fehlerquellen und Verantwortlichkeiten in Prozessen algorithmischer Entscheidungsfindung

Mögliche Fehlerquellen bei Algorithmen

Phase 1: Algorithmendesign und Implementierung (Wissenschaftler oder Informatiker)

Als Angestellte, als Auftragnehmer oder sogar kostenlos bei Open Source Projekten.

Bei der Entwicklung und Programmierung können / werden Fehler passieren.

https://algorithmenethik.de/wp-content/uploads/sites/10/2018/02/ZweigADMFehlerquellen_final_ohneCover.pdf

Mögliche Fehlerquellen bei Algorithmen

Phase 2: Methodenauswahl - Datenauswahl
(Data Scientist)

Entscheidung über den Algorithmus und Datensammlung und Auswahl von (Trainings-)Daten. Bei den Daten gilt „garbage in – garbage out“ – „Vorurteil rein – Vorurteil raus“.

Teil der Datensammlung ist die sog. **Operationalisierung** (abstrakte Konzepte wie „Kreditwürdigkeit“, „Anfälligkeit für terroristische Ideen“, „Rückfallswahrscheinlichkeit“ wird auf einige messbare Variable reduziert). Dabei müssen weitgehende Annahmen getroffen werden die naturbedingt mit großen Fehlern behaftet sind.

https://algorithmenethik.de/wp-content/uploads/sites/10/2018/02/ZweigADMFehlerquellen_final_ohneCover.pdf

Mögliche Fehlerquellen bei Algorithmen

Immer noch Phase 2: Datenauswahl
(Data Scientist)

Diskriminierung liegt bereits in der Auswahl der Datenquellen vor

2 typische Probleme bei der Datenauswahl:

- Wichtige Variable fehlen in den Datensätzen die besser geeignet werden liegen nicht vor
- Die verfügbaren Datensätze kommen aus bestimmten Gruppen oder fehlen für andere Gruppen –
  false positive + false negative Probleme: Gruppen sind nicht repräsentiert oder sie sind überrepräsentiert

Automating Inequality - How High-Tech Tools Profile, Police, and Punish the Poor
https://cyber.harvard.edu/events/automating-inequality

Mögliche Fehlerquellen bei Algorithmen

Phase 3: Konstruktion des Entscheidungssystems (Data Scientist)

Im Entscheidungssystem wird eine Methode des maschinellen Lernens mit den ausgewählten Trainingsdaten zusammengeführt.

Dabei werden die im Rahmen der Operationalisierung ausgewählten Datenelemente berücksichtigt und Qualitätskriterien gesetzt (z.B. false negative / false positive Rate)

- es sind primär politische Entscheidungen welche Typen von Fehlern toleriert werden sollen (im Zweifelsfall keine Förderungen/Schulungen anbieten oder im Zweifelsfall schulen/fördern)

https://algorithmenethik.de/wp-content/uploads/sites/10/2018/02/ZweigADMFehlerquellen_final_ohneCover.pdf

Mögliche Fehlerquellen bei Algorithmen

Immer noch Phase 4: Einbettung in den gesellschaftlichen Prozess (Data Scientist + Auftraggeber)

Speziell bei Systemen im staatlichen Bereich:

Die Politik gibt vor ob z.B. bei der „Verwaltung“ von knappen Ressourcen (finanzielle Förderungen, Schulungen, Coaching, Krediten, …) die Fairness oder die Reduzierung der Ausgaben im Vordergrund steht.

Über die Mittel für Schulungen entscheidet die Politik, die Parameter des Algorithmus bilden diese Politik nur einfach ab.

Automating Inequality - How High-Tech Tools Profile, Police, and Punish the Poor
https://cyber.harvard.edu/events/automating-inequality

Mögliche Fehlerquellen bei Algorithmen

Phase 5: Re-Evaluation (optional)
(Data Scientist oder Entscheider, Auftraggeber)

Hierbei können die erzielten Ergebnisse bewertet werden und dann Änderungen an Trainingsdaten, der Operationalisierung, der Methode und ihren Parametern oder des Entscheidungssystem getroffen werden.

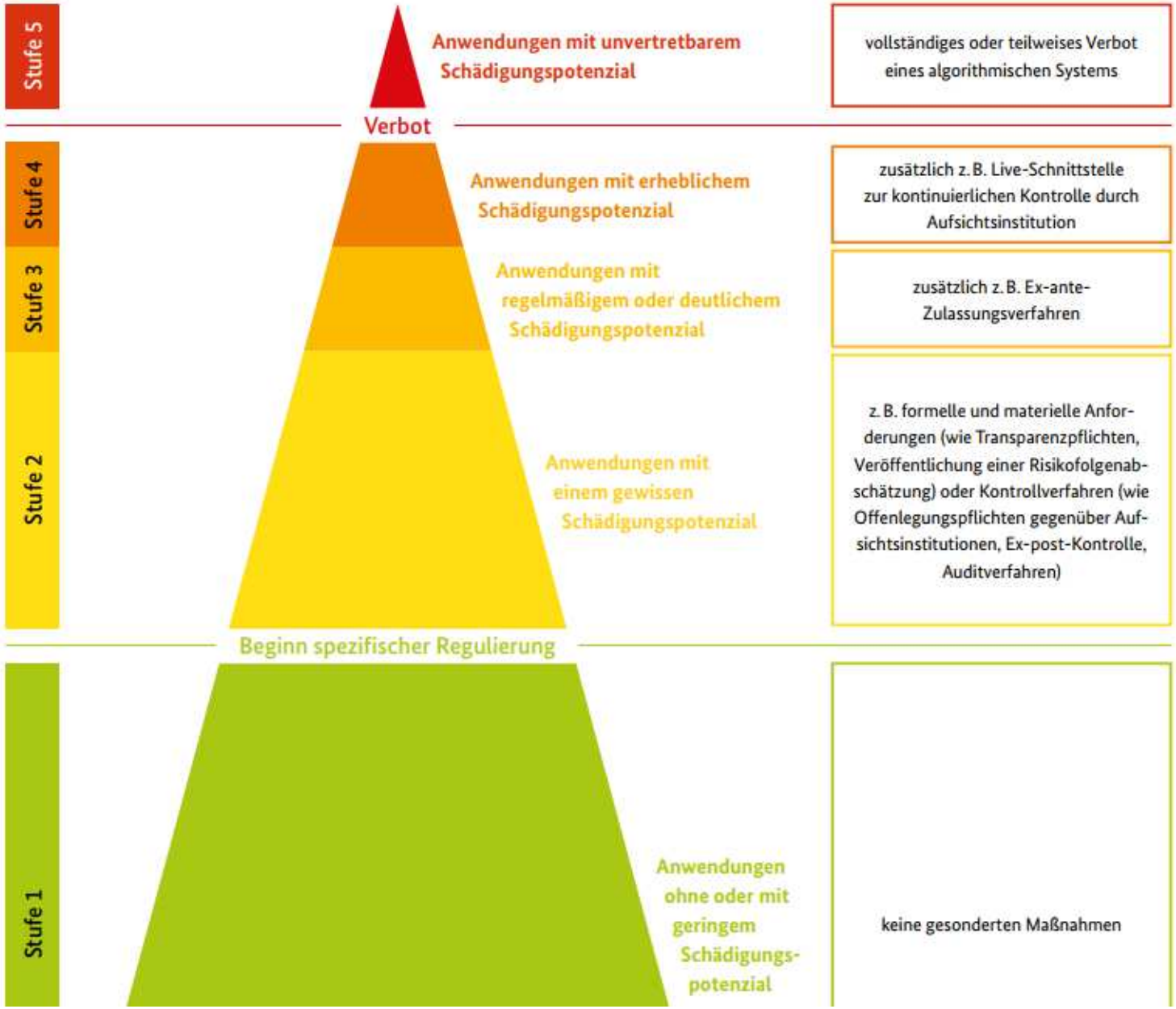
https://algorithmenethik.de/wp-content/uploads/sites/10/2018/02/ZweigADMFehlerquellen_final_ohneCover.pdf

Gutachten der deutschen Datenethik-Kommission (DEK)

- „Risikoadaptiertes Regulierungssystem für den Einsatz algorithmischer Systeme“
- Einteilung aller Algorithmen in Kritikalitäts-Stufen 1 bis 5
- Stufe 1 = harmlos
- Stufe 5 "Anwendungen mit unvertretbarem Schädigungspotenzial“ → gehören verboten (Totalüberwachung, die Integrität der Persönlichkeit verletzende Profilbildung oder gezieltes ausnutzen von Schwachstellen und Angriffspunkten)
- Ziel ist eine europäische Verordnung für algorithmische Systeme (EUVAS) vor

https://datenethikkommission.de/wp-content/uploads/191023_DEK_Gutachten_Kurzfassung_dt_bf.pdf

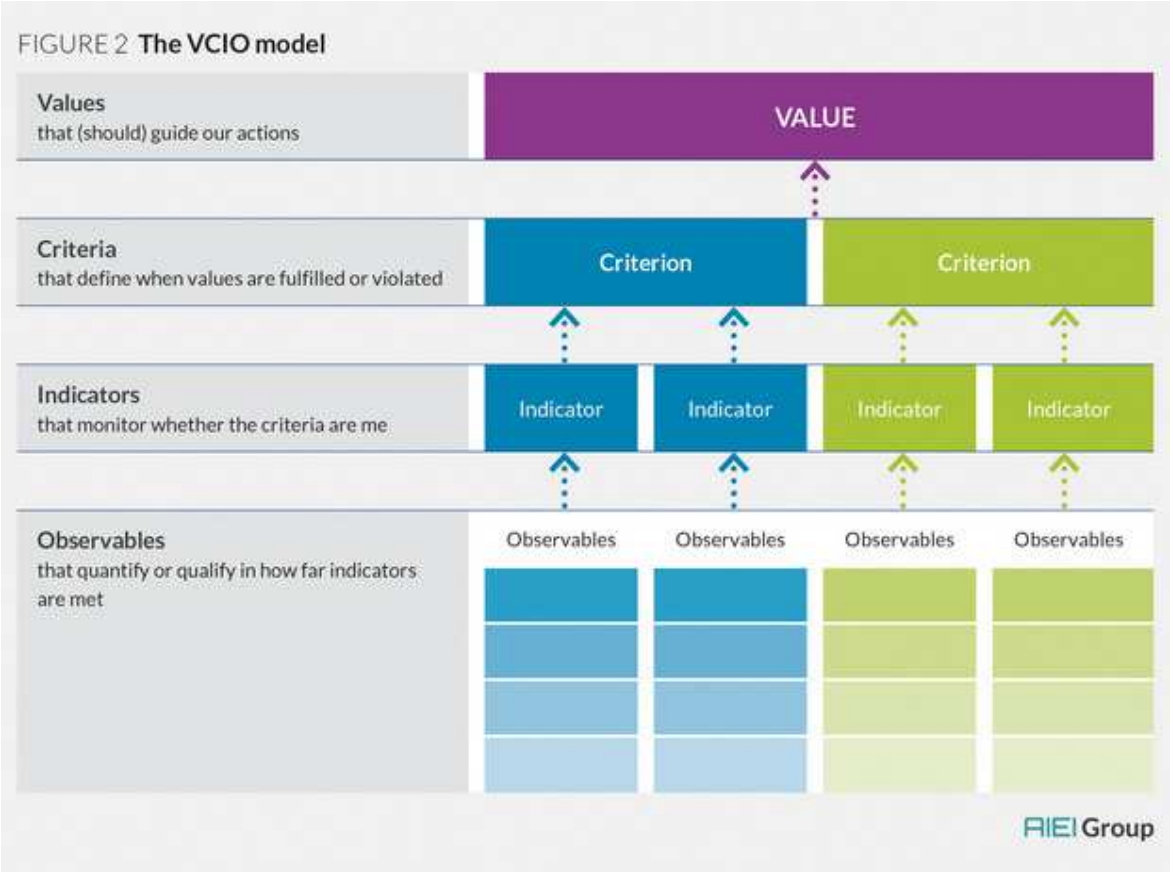
Kritikalitäts-Pyramide nach DEK



From principles to practice: Wie wir KI-Ethik messbar machen können

Das Ziel:

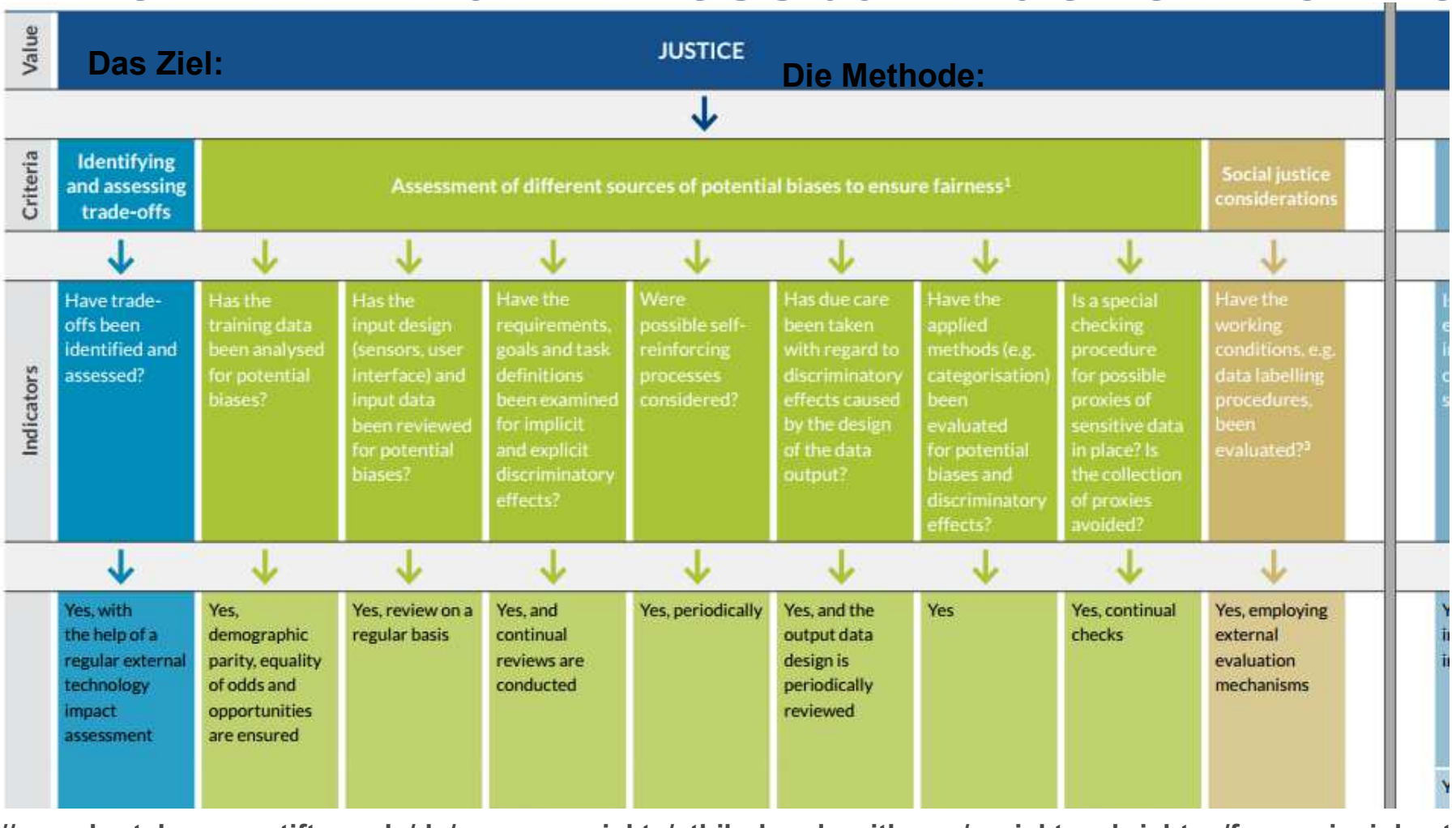
Die Methode:



https://www.bertelsmann-stiftung.de/de/unsere-projekte/ethik-der-algorithmen/projektnachrichten/from-principles-to-practice-wie-wir-ki-ethik-messbar-machen-koennen

https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf

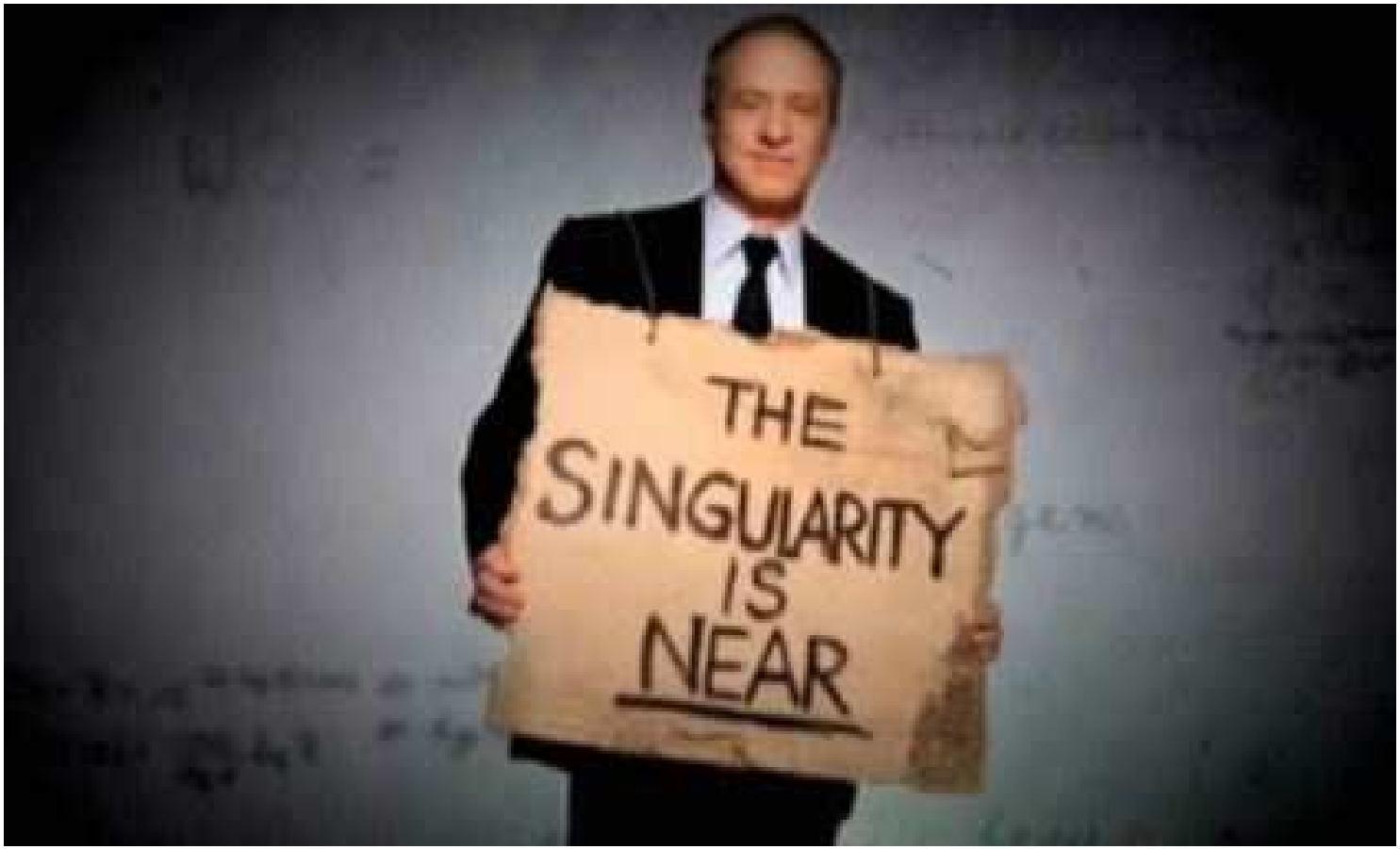
From principles to practice: Wie wir KI-Ethik messbar machen können



https://www.bertelsmann-stiftung.de/de/unsere-projekte/ethik-der-algorithmen/projektnachrichten/from-principles-to-practice-wie-wir-ki-ethik-messbar-machen-koennen

https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf

Danke



Ray Kurzweil (Google)

Mehr dazu:
http://sicherheitskultur.at/Manipulation.htm
http://philipps-welt.info/robots.htm#asimovlaws