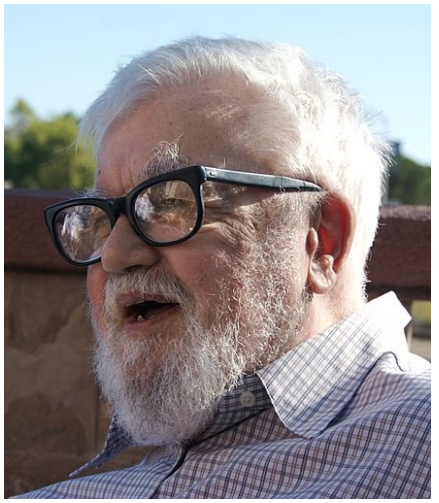
Faire Computer? Persönliche und gesellschaftliche Folgen beim Einsatz von KI.

Prof. Dr. Kai Eckert, Hochschule der Medien
Open Up!
23.06.2020

Was ist Künstliche Intelligenz?



John McCarthy

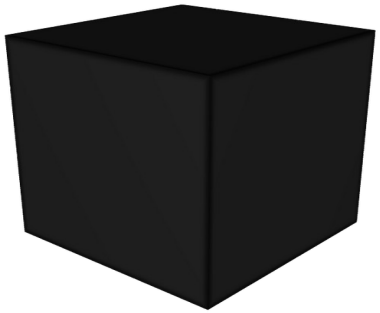
"As soon as it works,
no one calls it AI any more."

Künstliche Intelligenz

- Programme, die **intelligentes Verhalten** nachahmen.
- Verhalten wird bestimmt von **externen Wahrnehmungen**.
- Verhalten basiert auf **vorhandenen Daten** (gelerntes Verhalten) oder einer **formalen Repräsentation** von Hintergundwissen.
- Wir konzentrieren uns auf die aktuell überwiegend entwickelten **maschinellen Lernverfahren.**

Maschinelles Lernen

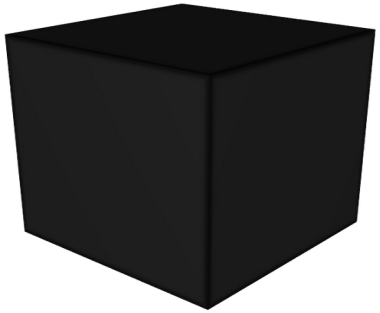
Lernen
aus Daten



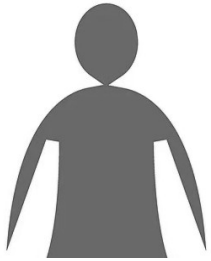
Gelerntes
Modell

Model

Neuer
Fall



Ergebnis



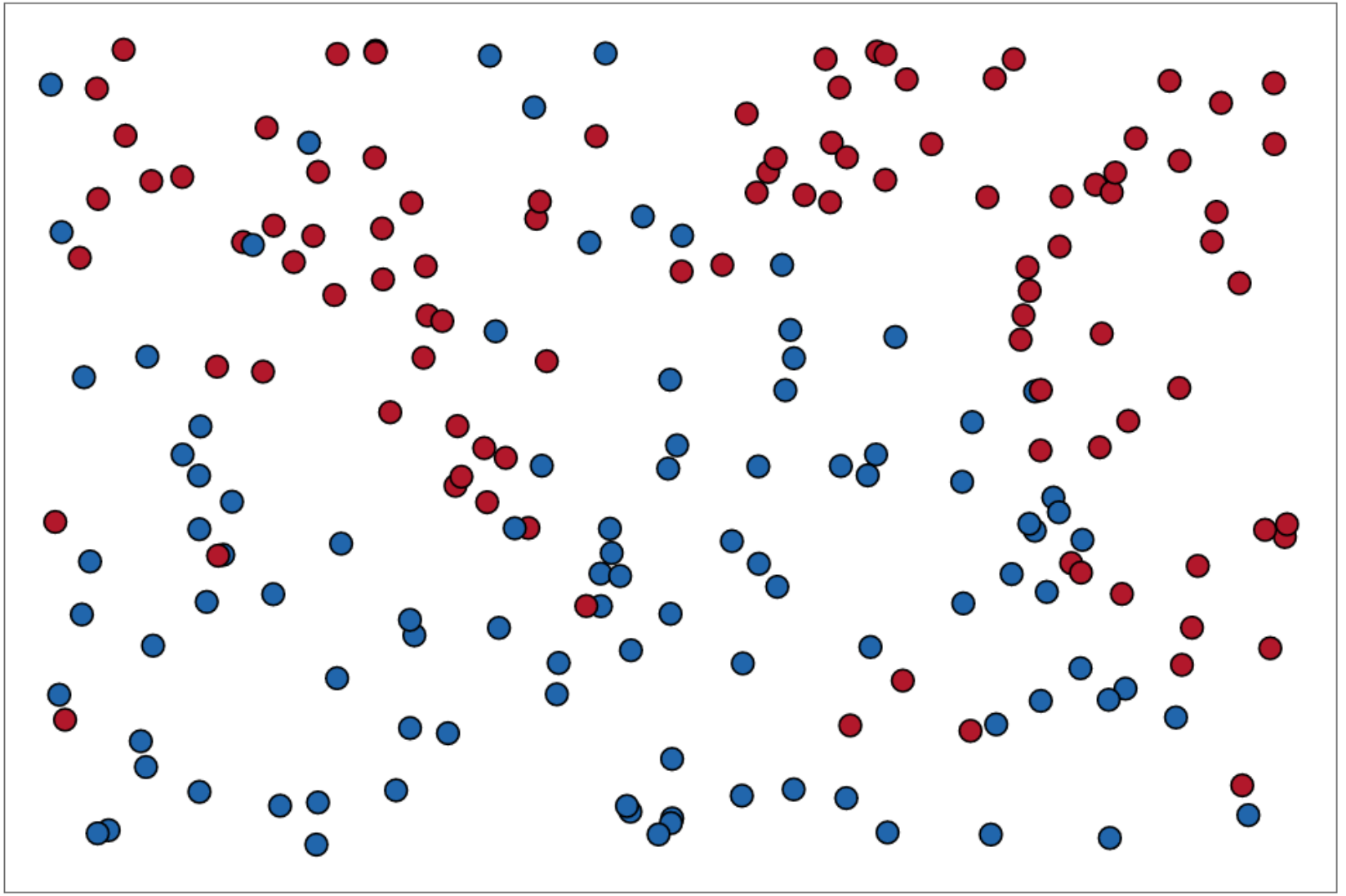
Intuitives Beispiel: Fallbasiertes Schließen

- Das Modell besteht aus den Daten selbst, für die die gewünschten Ergebnisse oder Erkenntnisse vorliegen, z.B. aus Daten zu Personen plus der Information, ob sie straffällig geworden sind.
- Für neue Fälle sucht man den ähnlichsten bekannten Fall aus der Datenbasis heraus und gibt das Ergebnis dieses Falls zurück.

Wo ist das Problem?

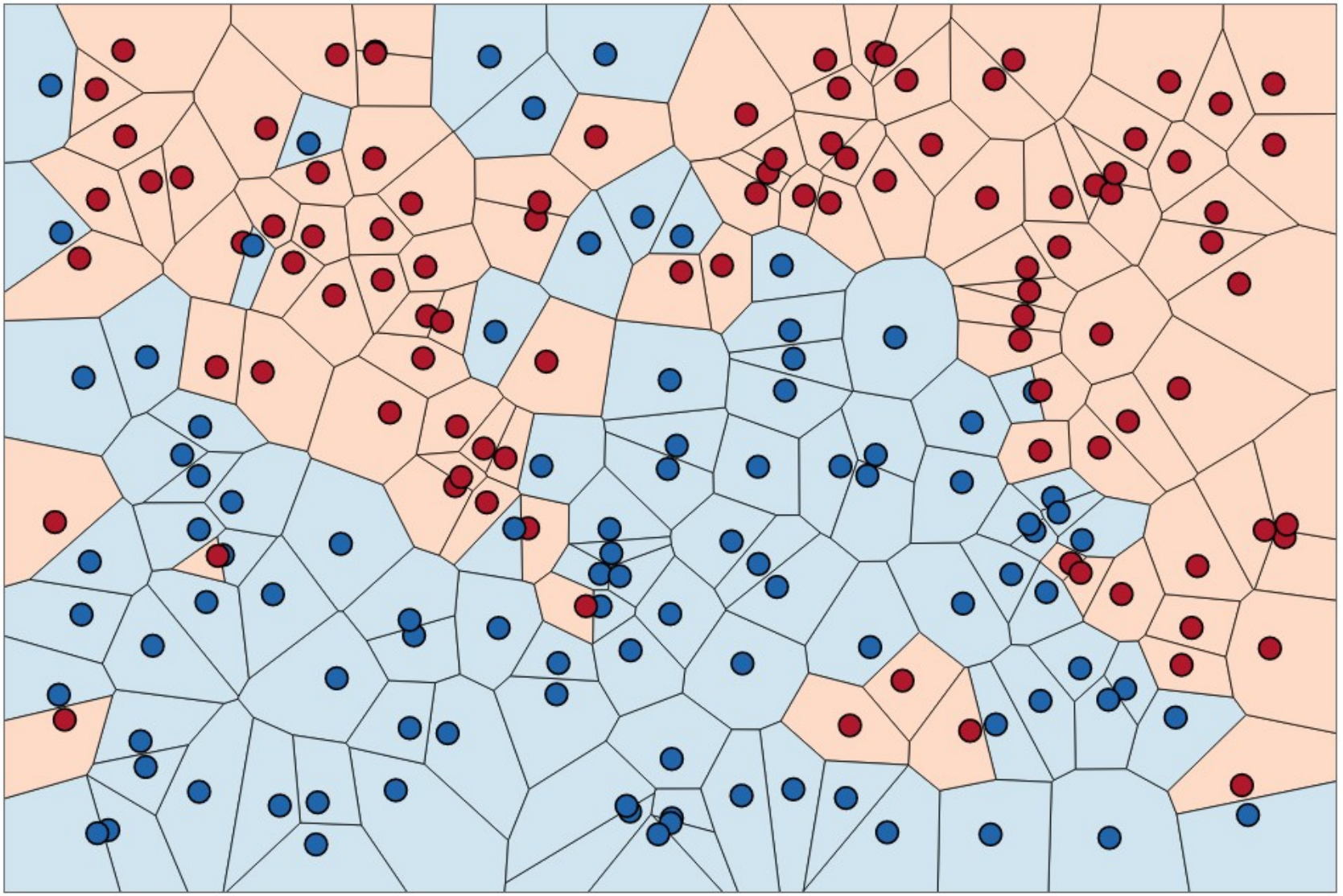
Fußballerin?

Wähler nach Religiösität und Wohlstand



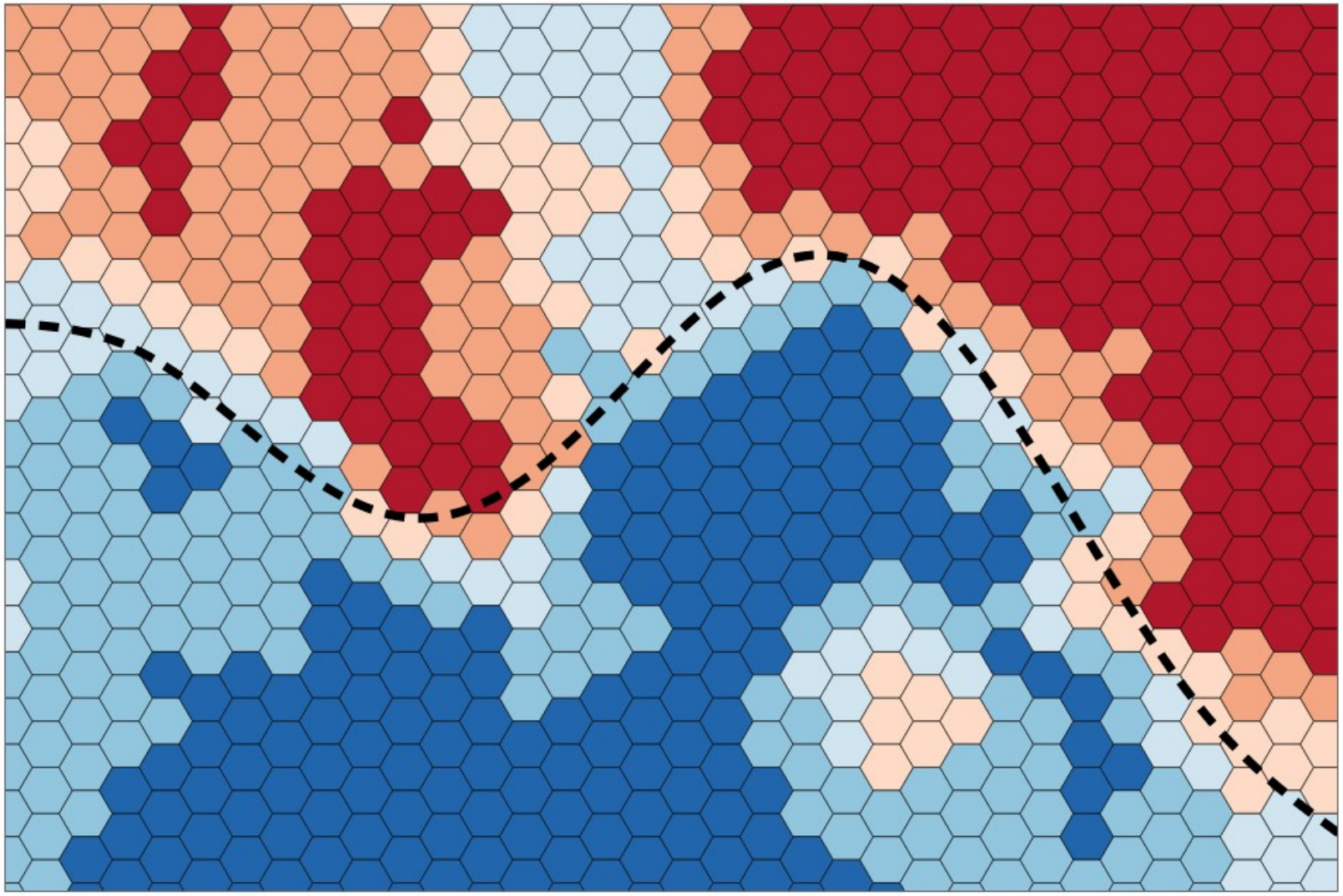
Quelle: http://scott.fortmann-roe.com/docs/BiasVariance.html

Klassifikation nach nächstem Nachbarn



Quelle: http://scott.fortmann-roe.com/docs/BiasVariance.html

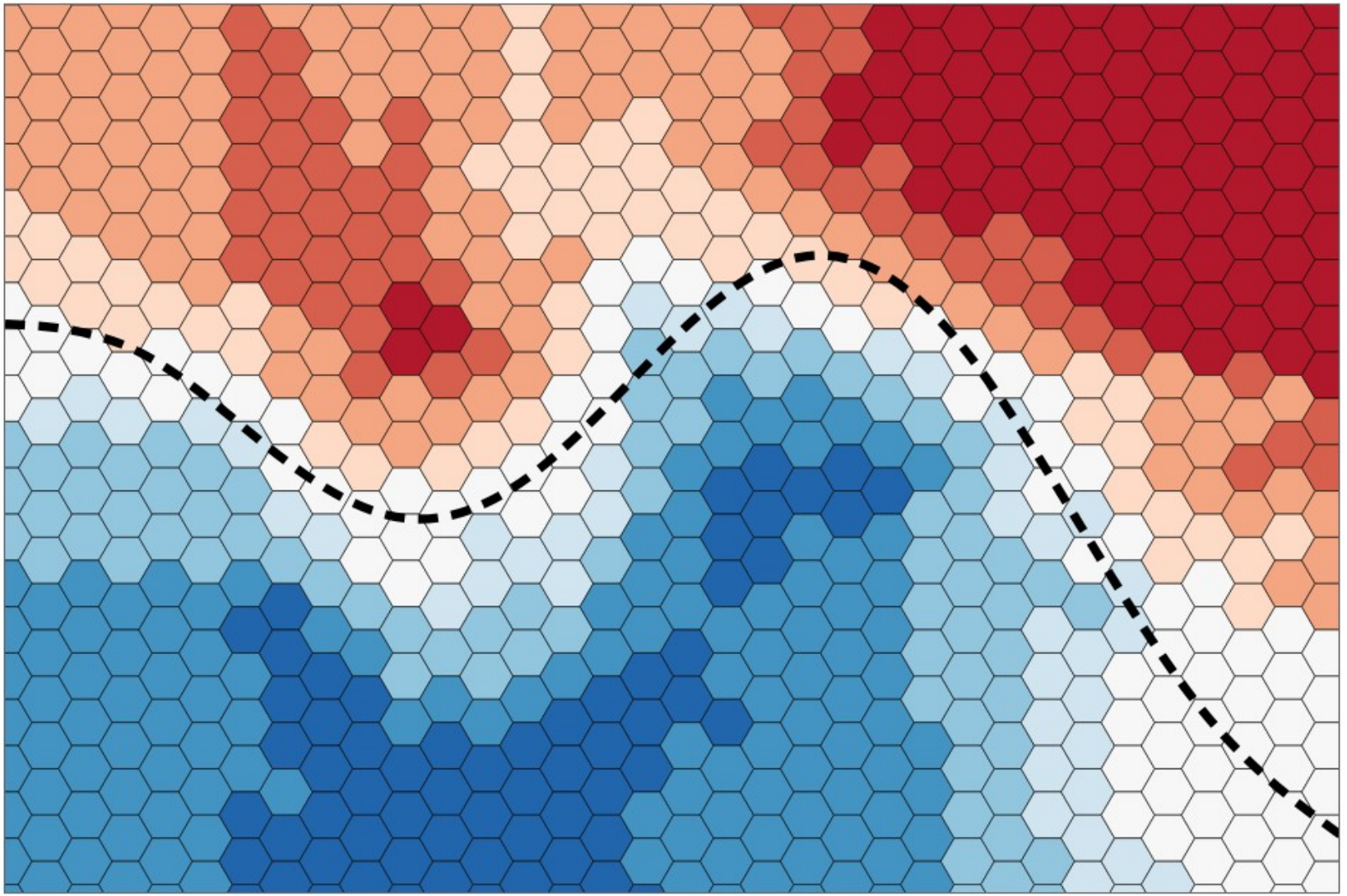
K-Nearest Neighbors: 5



k-Nearest Neighbors: 5

Quelle: http://scott.fortmann-roe.com/docs/BiasVariance.html

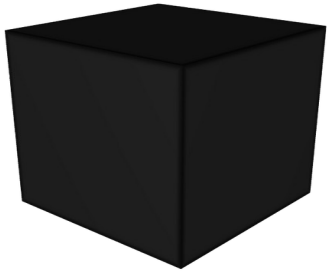
K-Nearest Neighbors: 5



k-Nearest Neighbors: 20

Quelle: http://scott.fortmann-roe.com/docs/BiasVariance.html

Komponenten eines Machine Learning Systems
Wo kann etwas schief laufen, wo kann man eingreifen?



Programmierung des Verfahrens

Datenauswahl für das Training

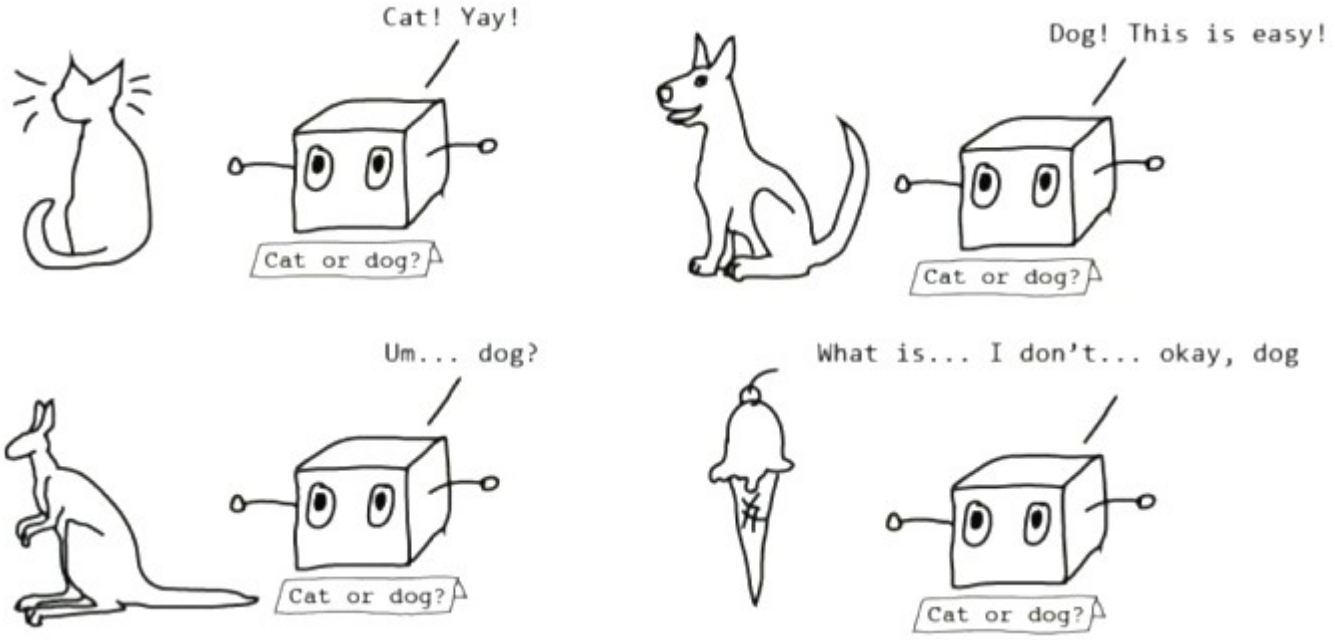
Implementierung des Systems, Anwendung

Nicht immer wird sorgfältig gearbeitet

Quelle: https://xkcd.com/1838/

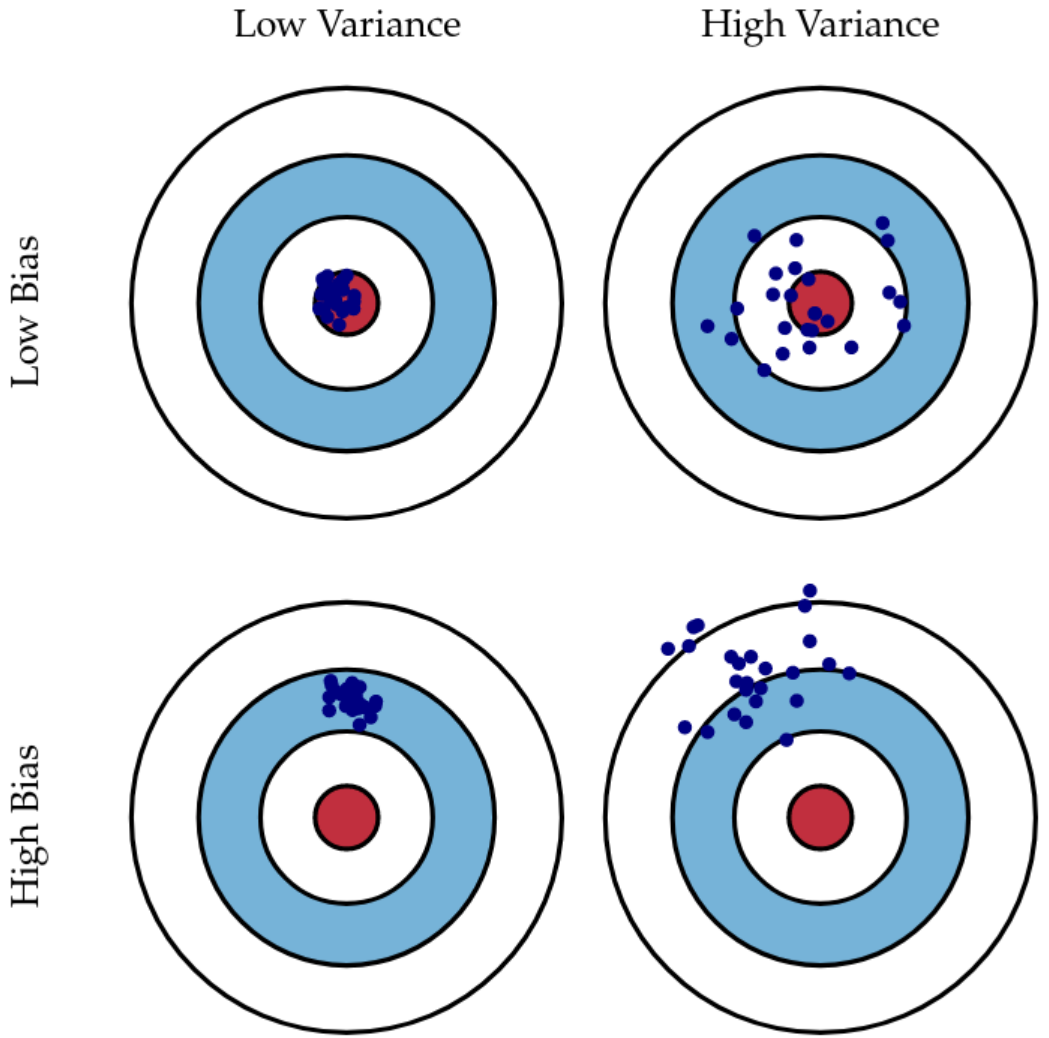
Bias in den Daten

Sind die Daten fair?
→ Repräsentieren die Daten die Realität?



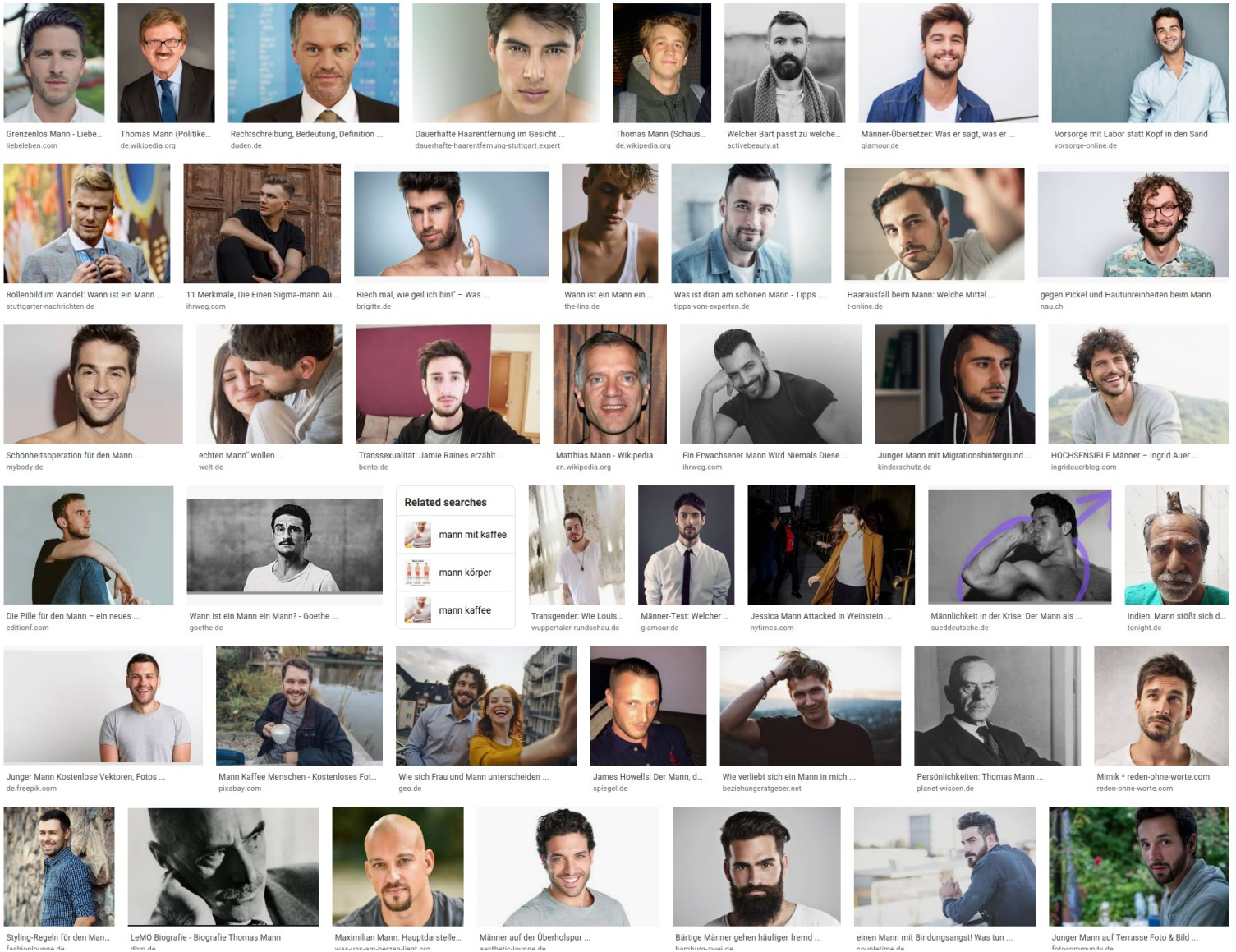
Quelle: Janelle Shane, aiweirdness.com

Bias und Varianz

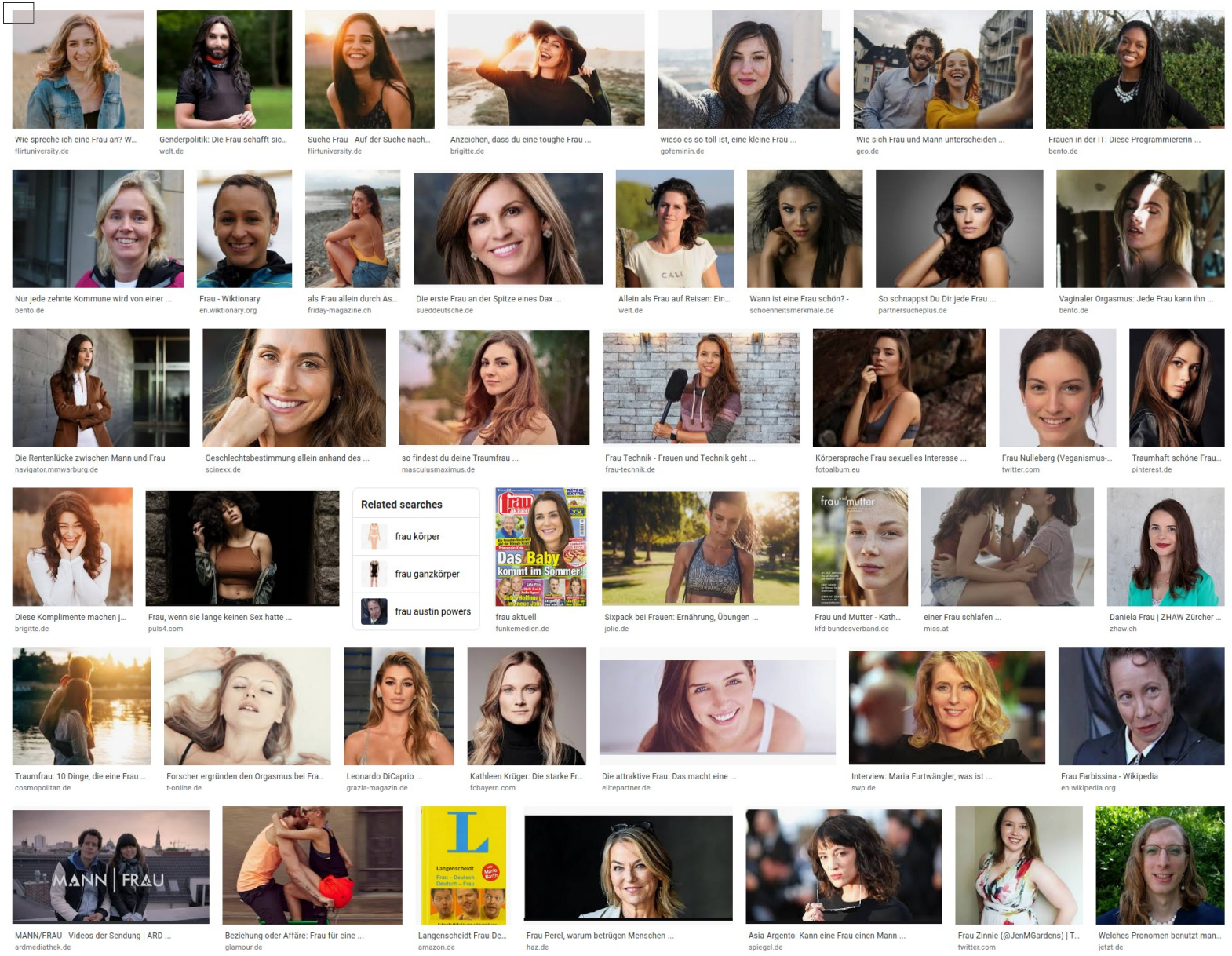


Quelle: http://scott.fortmann-roe.com/docs/BiasVariance.html

Google Bildersuche: Mann

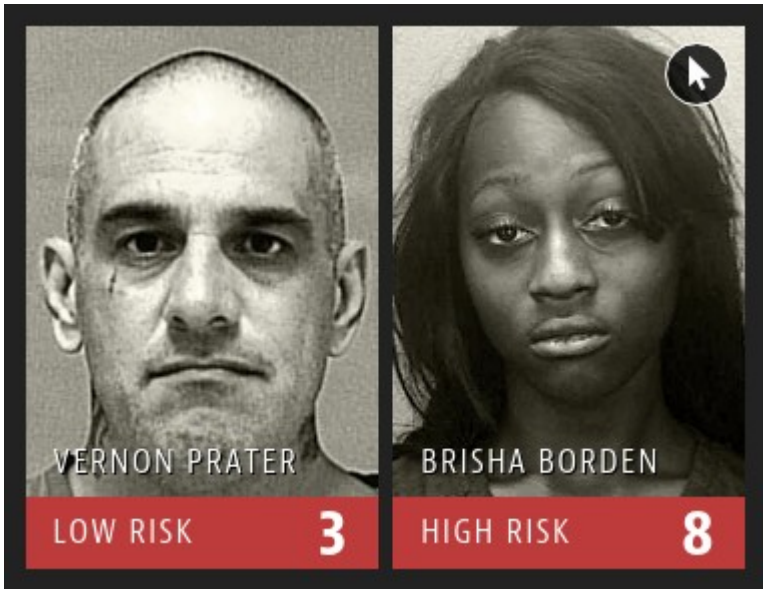


Google Bildersuche: Frau



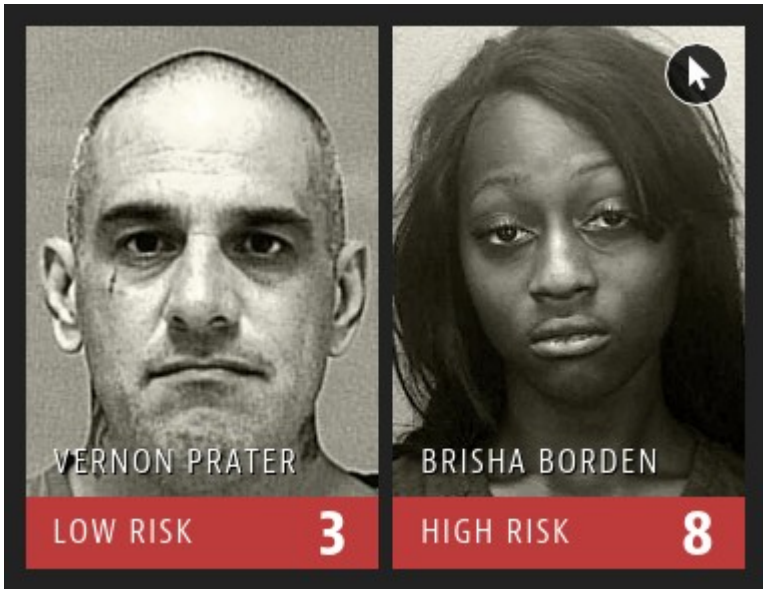
Bias in den Daten

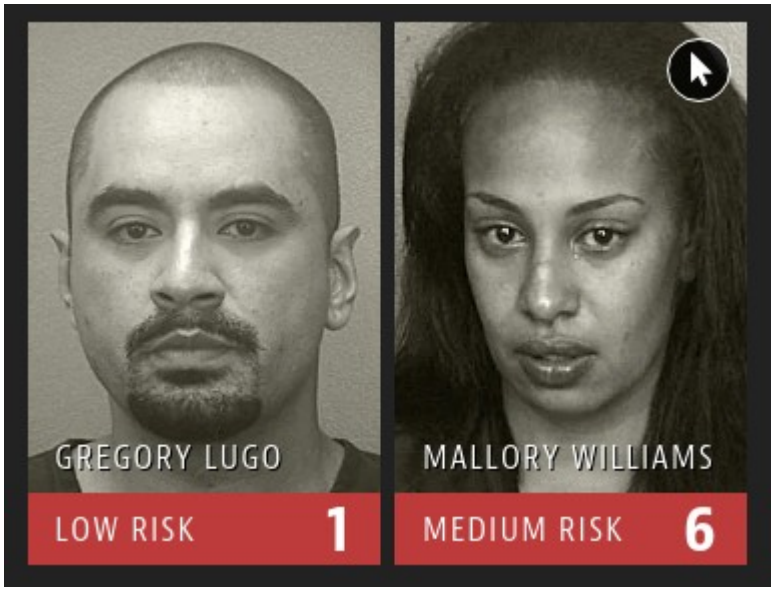
Ist die Gesellschaft fair?
→ Gibt es Vorurteile und systematische Ungleichheit?

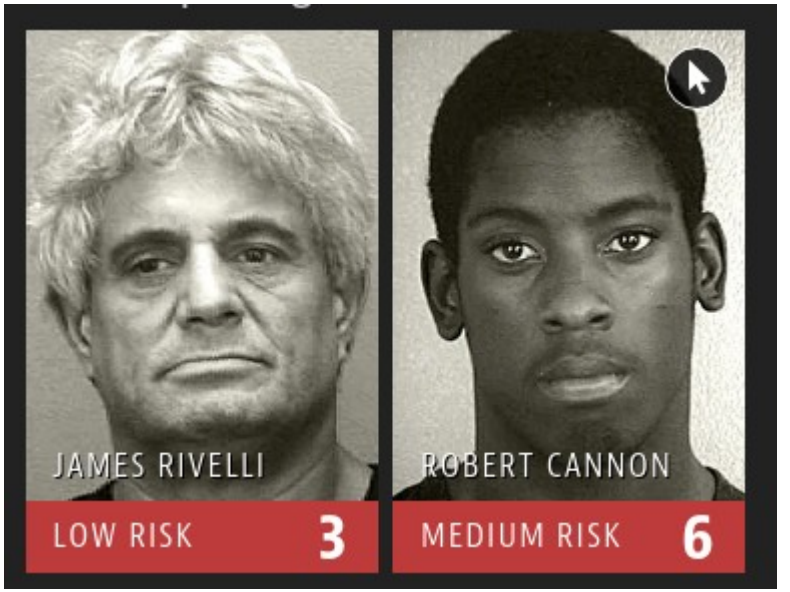


COMPAS
Correctional Offender Management Profiling for Alternative Sanctions

- Kommerzielle Software, die in den USA verbreitet zur Risikobewertung für mögliche Bewährungsstrafen eingesetzt wird.
- Bewertung auf einer Skala von 1 (Geringes Risiko für weitere Straftaten) bis 10 (hohes Risiko).
- Trainingsdaten basieren auf Fragebögen und Daten aus den Polizeiakten.
- Nach der Hautfarbe wird nicht gefragt, aber sie ist über Proxies implizit enthalten (z.B. Sozioökonomischer Status, Familienverhältnisse, Kriminalität am Wohnort).







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

Verteilung der Risikoklassen

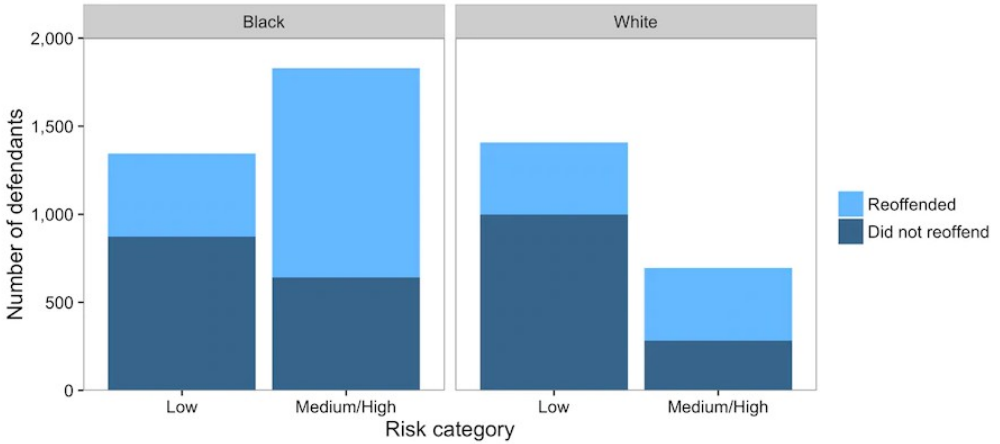
Schwarze Angeklagte

Weiße Angeklagte

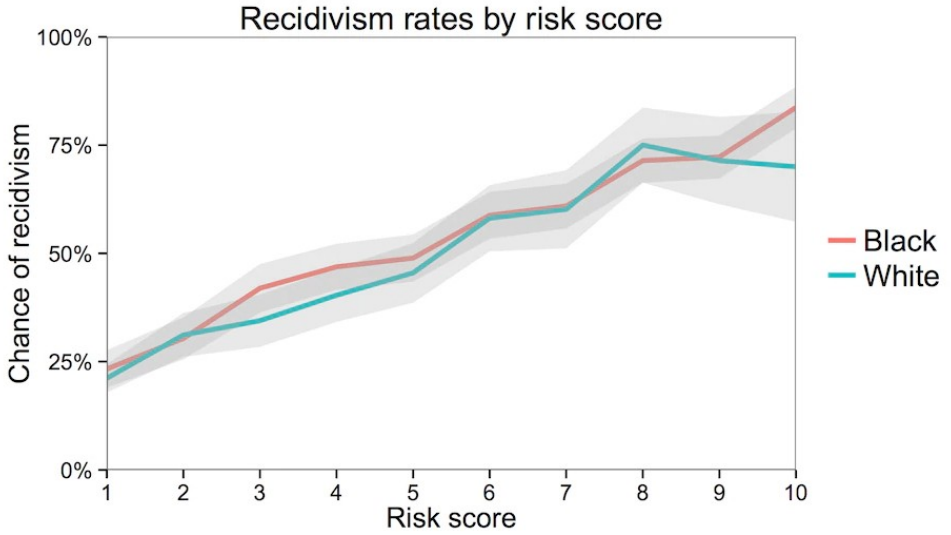
Quelle: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Was ist fair?

- In jeder Kategorie ist der **Anteil der Wiederholungs-täter** für beide Hautfarben **etwa gleich** (fair).
- Die **Wiederholungsrate** ist für schwarze Angeklagte höher als für weiße Angeklagte.
- Schwarze werden eher als **Hochrisiko** eingestuft als Weiße.
- Schwarze, die keine Wiederholungstäter sind, werden eher als Hochrisiko eingestuft als Weiße (unfair).



Distribution of defendants across risk categories by race. Black defendants reoffended at a higher rate than whites, and accordingly, a higher proportion of black defendants are deemed medium or high risk. As a result, blacks who do not reoffend are also more likely to be classified higher risk than whites who do not reoffend.



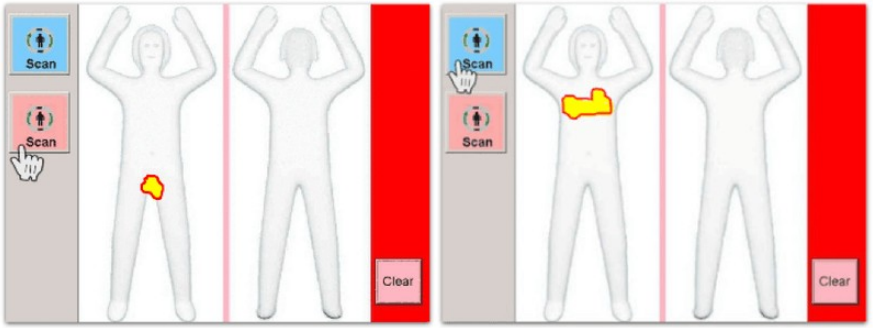
Recidivism rate by risk score and race. White and black defendants with the same risk score are roughly equally likely to reoffend. The gray bands show 95 percent confidence intervals.

Quelle: https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

Lawrence Lessig

“Code is Law.”

Body-Scanner am Flughafen



- Vorauswahl des Geschlechts erforderlich.
- Entsprechend lösen die Systeme regelmäßig Fehlalarme aus, z.B. bei transsexuellen Personen.
- Nicht leicht zu verbessern, denn per Definition sucht die Software in den Scannern nach Abweichungen von einer durch Trainingsdaten vorgegebenen „Norm“.
- Hier kommen allerdings durch die Auswahlknöpfe bewusste Design-Entscheidungen hinzu.

Selbstfahrende Autos

Das Trolley-Problem

Utilitarismus vs Humanismus

Eine Straßenbahn rollt auf eine **Gruppe Menschen** zu, die nicht ausweichen können.

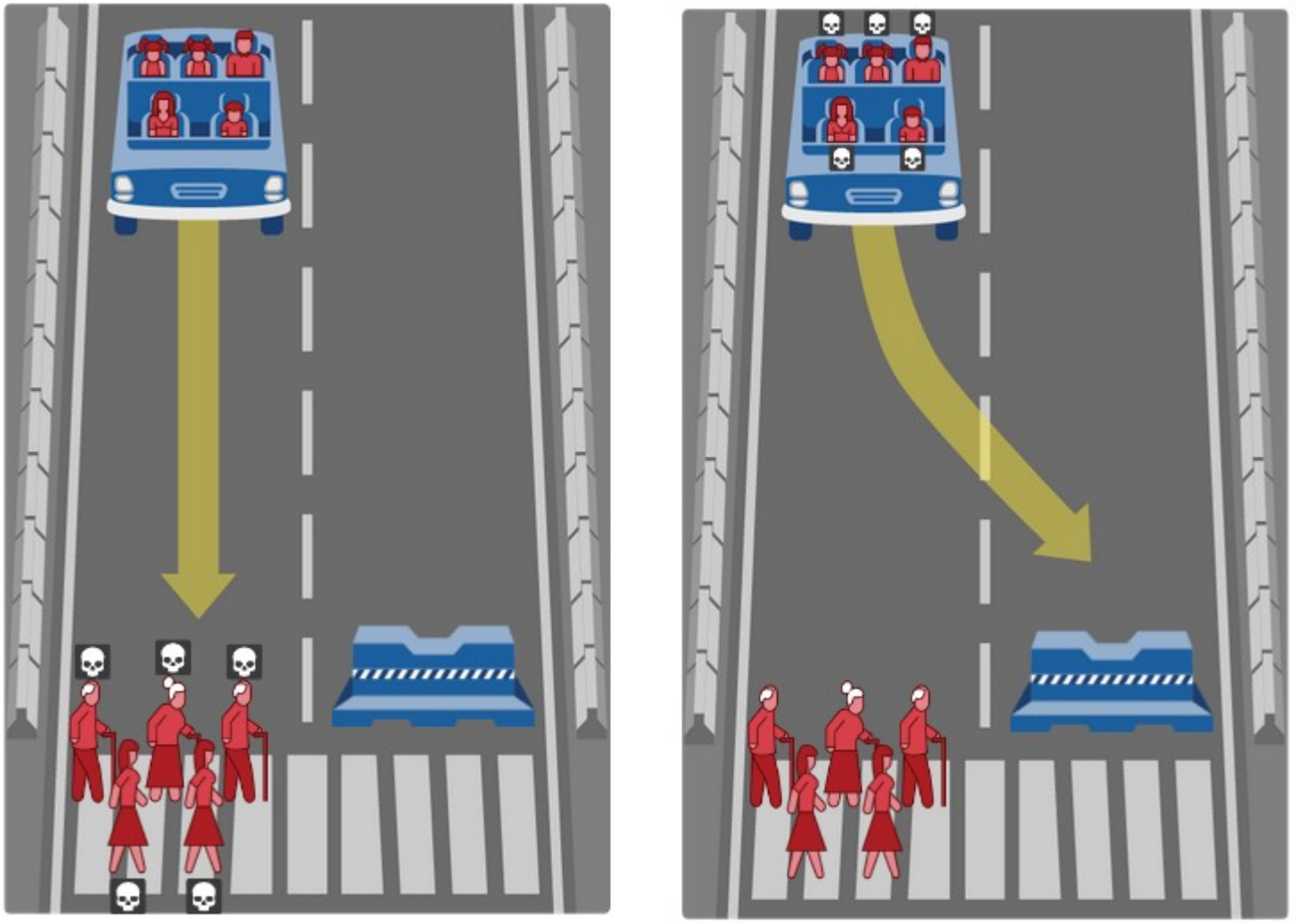
Wenn Sie den **Hebel umlegen**, können Sie die Bahn umleiten und nur **ein Mensch** stirbt.

Was würden Sie tun?

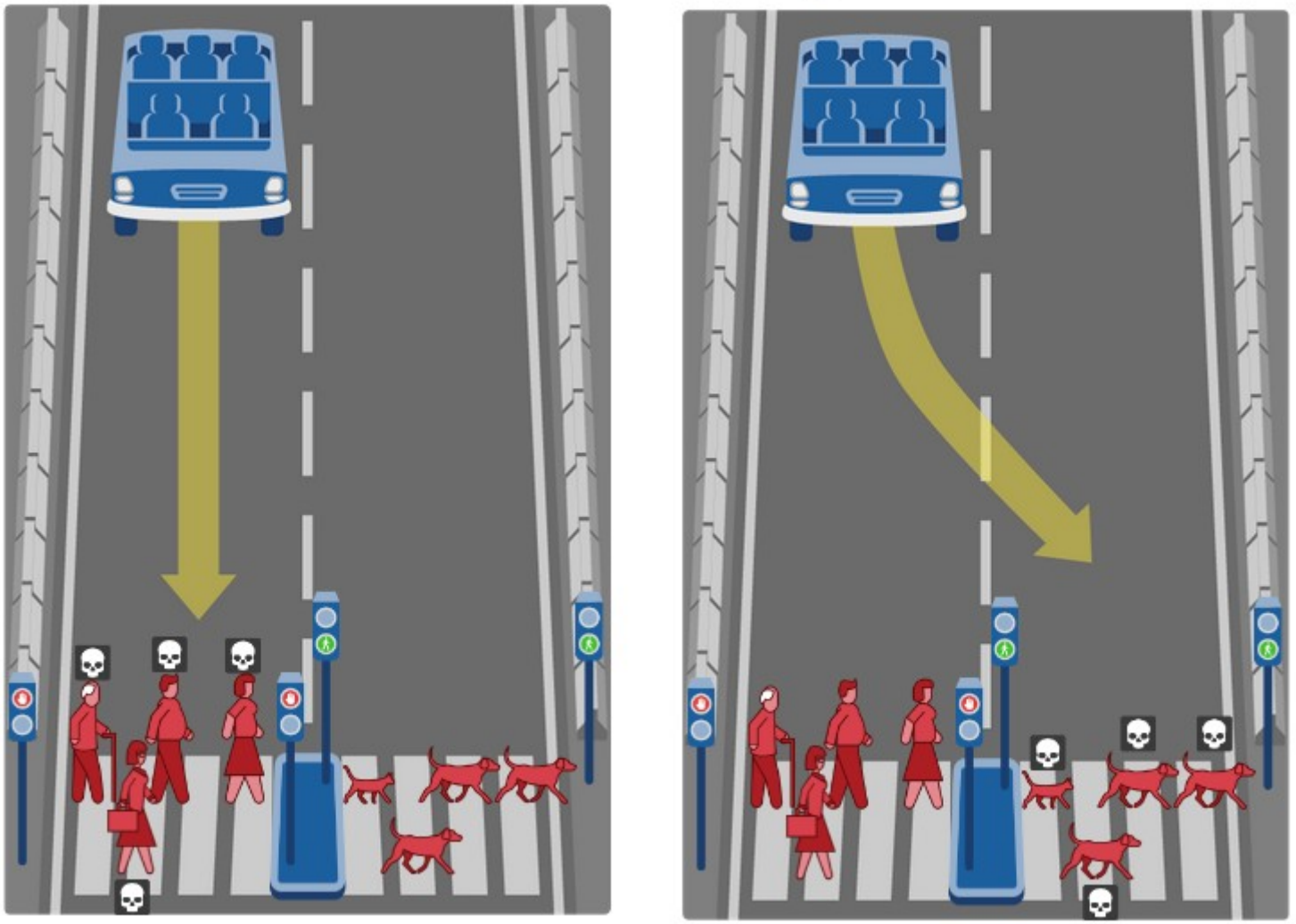
Moral Machine (MIT)

Idee: Erfassen, wie Menschen zu moralischen Entscheidungen stehen, die von intelligenten Maschinen, wie z.B. selbstfahrenden Autos, getroffen werden.

Awad, E., Dsouza, S., Kim, R. et al.
The Moral Machine experiment.
Nature 563, 59–64 (2018).
https://doi.org/10.1038/s41586-018-0637-6



Quelle: https://moralmachine.mit.edu/



Quelle: https://moralmachine.mit.edu/

Kriterien

Mehr Leben retten

Präferenz einer Spezies

Mitfahrer beschützen

Alterspräferenz

Einhalten des Gesetzes

Präferenz von Sportlichkeit

Eingreifen verhindern

Präferenz von sozialem Wert

Geschlechterpräferenz

Quelle: https://moralmachine.mit.edu/

Spock

“The needs of the many outweigh the needs of the few, or the one.”

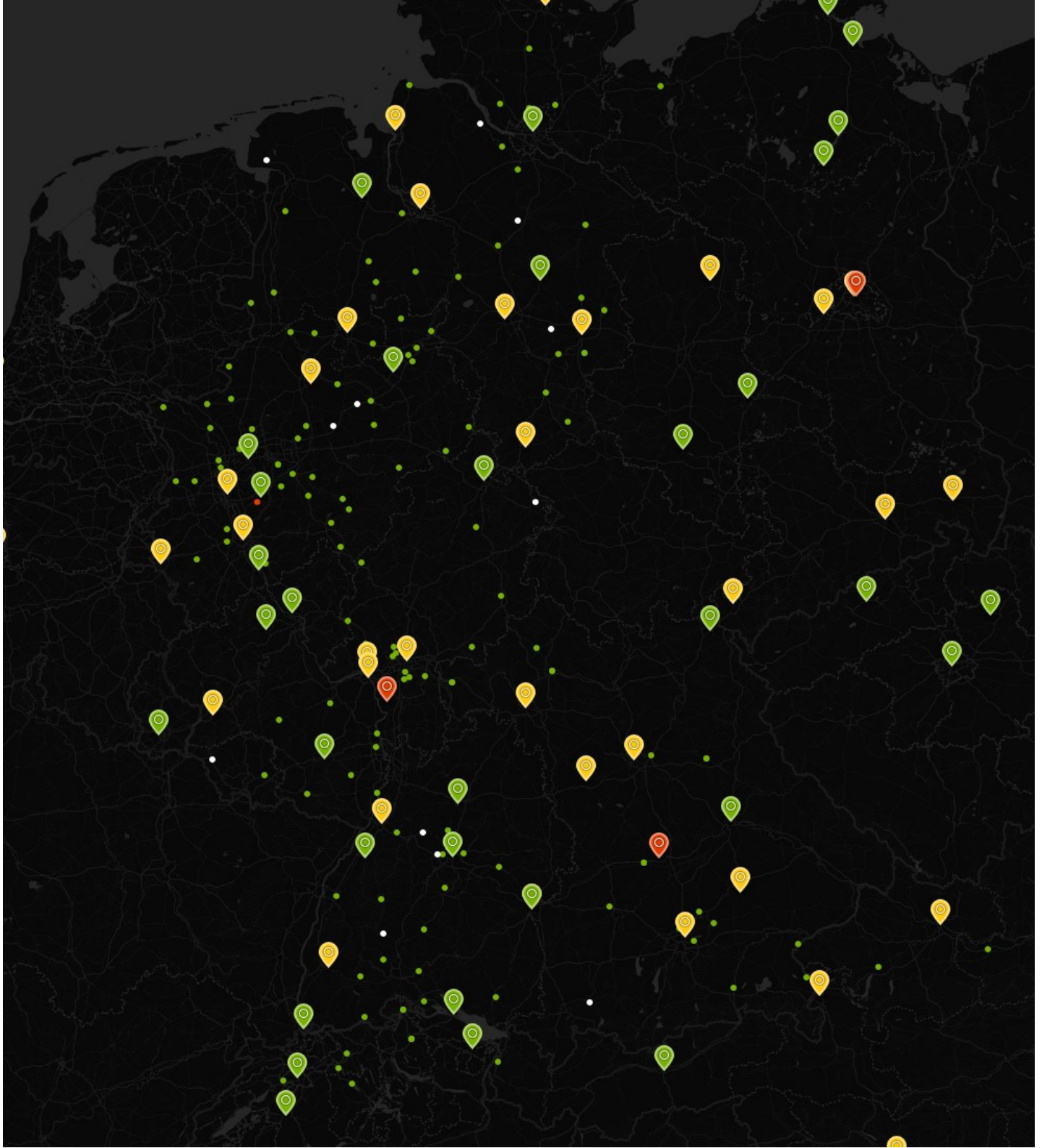
FERDINAND VON SCHIRACH

TERROR

70.000 im Stadion.
160 Passagiere im Flugzeug.

Ein Kampfpilot schießt das Flugzeug ab.
Er wird angeklagt.

Schuldig oder nicht schuldig?



Quelle: https://terror.theater/

Der Film: Terror – Ihr Urteil

Freispruch: 86,9 Prozent

Schuldig: 13,1 Prozent

Wenn man Menschen befragt, tendieren sie dazu, Utilitarier zu sein.

Was spricht dagegen?

Grundgesetz für die Bundesrepublik Deutschland

Art. 1, Abs. 1: Die **Würde des Menschen** ist unantastbar.

Art. 2 Abs. 2: Jeder hat das **Recht auf Leben** und körperliche Unversehrtheit.

Die **Ermächtigung** der Streitkräfte, gemäß § 14 Abs. 3 des Luftsicherheitsgesetzes durch unmittelbare Einwirkung **mit Waffengewalt ein Luftfahrzeug abzuschießen**, das gegen das Leben von Menschen eingesetzt werden soll, **ist mit dem Recht auf Leben** nach Art. 2 Abs. 2 Satz 1 GG in Verbindung mit der Menschenwürdegarantie des Art. 1 Abs. 1 GG **nicht vereinbar**, soweit davon tatunbeteiligte Menschen an Bord des Luftfahrzeugs betroffen werden.

Bundesverfassungsgericht, Urteil von 2006

Roboter können keine Moral

Voll automatisierte Robo-Cars werden als die Zukunft angesehen. Folglich sollen Ethiker sich Gedanken darüber machen, wie voll automatisiert fahrende Autos in einer Dilemma-Situation entscheiden sollen, wenn sich der Tod von Menschen nicht verhindern lässt. Wer soll überleben und wer überfahren werden? Kinder oder Rentner? Hunde, Kranke, Straffällige?

Dass das Grundgesetz mit seiner Definition der Menschenwürde solche Überlegungen schon im Ansatz nicht zulässt, hat die deutsche Automobilindustrie bislang zu keinem klaren Bekenntnis gegen ein solches Aufrechnen von Lebenswert genötigt.

Richard David Precht, ZEIT, 18. Juni 2020

Roboter können keine Moral

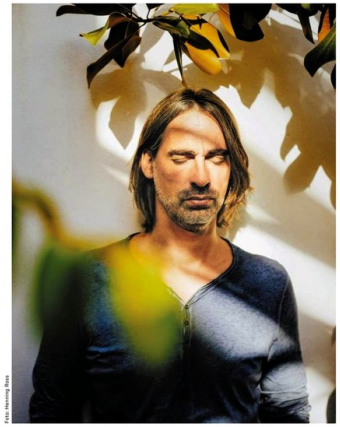
Warum das Gerede von superintelligenten, allmächtigen Maschinen nur ein großes Ablenkungsmanöver ist VON RICHARD DAVID PRECHT

Für viele ist es nur eine Frage der Zeit: Ob in zehn, zwanzig oder fünfzig Jahren – irgendwann in diesem Jahrhundert wird die Superintelligenz geboren, eine »starke« künstliche Intelligenz (KI), die die kognitive Leistungsfähigkeit des Menschen in allen Bereichen weit übersteigt. Das Schicksal des Menschen liege dann in ihrer Hand. Die spannende Frage, so heißt es dann, sei nur, ob diese Superintelligenz gut oder böse sei – und ob es den Menschen rechtzeitig gelingt, dies sicher festzulegen.

Die KI-Branche befeuert solche Fantasien: Sie verkauft Anfangserfolge als Durchbrüche und schwelgt in dunklen Allmachtsfantasien. Doch die Wahrscheinlichkeit, dass in absehbarer Zeit eine Superintelligenz dem menschlichen Gehirn in allen Bereichen überlegen sein wird, ist verschwindend gering. Der Begriff »künstliche Intelligenz« wirkt auf viele IT-Experten wie ein Marketingtrick. Dass Maschinen überschaubare Aufgaben blitzschnell auszuführen lernen, ist das eine – eine allumfassende Superintelligenz etwas anderes. Wer glaubt, dass tote Rechenmaschinen Geist, Bewusstsein oder Gefühle entwickeln, hat nicht mal entfernt verstanden, was Geist, Bewusstsein und Gefühle sind.

Nein, künstliche Intelligenz wird Homo sapiens nicht in den Schatten stellen. Und sie wird auch keinen eigenen bösen Willen entwickeln, um die Menschheit zu knechten oder auszulöschen. In der Evolution führte der lange Weg vom Überlebenswillen vieler zur Intelligenz sehr weniger Lebewesen. Die umgekehrte Richtung, dass aus Intelligenz ein Wille erwächst, ist unmöglich. Auch menschliche Intelligenz besitzt keinen Willen – sondern unser Wille geht unseren höheren kognitiven Leistungen immer voraus. Und außerdem: Wer sagt denn, dass eine Superintelligenz ausgerechnet über Machtgelüste verfügte? Menschliches Machtstreben ist unseren animalischen Leibern erwachsen und ist beileibe nicht unser wichtigster Antrieb. Ein kurzer Blick an einen Strand genügt. Menschen haben so viele andere Gelüste als Macht – Nichtstun, Sex, Alberntein und Alkohol, kurzum: den Willen zur Ohnmacht!

Eine Superintelligenz, setzte sie eigenständig ihre Ziele, hätte wohl niemals biologisch-menschliche Bedürfnisse: keine Ressourcensicherung, keinen Expansionswillen, keinen Machttrieb. Wenn überhaupt, gliche sie eher Sokrates als Jeff Bezos. Die Ängste, die Hightech-Gurus vor bösen Robotern schüren, sind bisweilen Selbstbespiegelung oder Paranoia – meist sind sie aber ein gigantisches Ablenkungsmanöver. Denn KI, falsch eingesetzt, kann Menschen, Kulturen und Demokratien durchaus gefährlich werden – allerdings nicht durch ihren bösen Willen.



Peter Hartwig Rees

»Künstliche Intelligenz wird Homo sapiens nicht in den Schatten stellen«

Der Philosoph und Bestsellerautor Richard David Precht

Das Gefahrenpotenzial künstlicher Intelligenz entlädt sich nicht auf einmal. Es hat nichts Bombastisches, sondern steckt in unfreiwilligen Diskriminierungen, wie unlängst bei einer Gesichtserkennungssoftware zu sehen, die Schwarze benachteiligt; es steckt in programmierten Unzulänglichkeiten und unbeabsichtigten Folgen, eigentlich überall da, wo Maschinen über Menschen richten. Am gefährlichsten wird es dort, wo Computer moralische Entscheidungen treffen sollen. Etwa bei voll automatisiert fahrenden Autos, bei Waffen, bei Bewertungen über Straffälligе oder bei der Besetzung von Arbeitsstellen.

Die Idee einer »ethischen Programmierung« ist eine Illusion. Denn bereits ein kleiner Blick auf die gravierenden Unterschiede zwischen menschlichem und maschinellem Problemlösen zeigt, dass sie nicht realistisch ist. Das, was sich Menschen von »ethischer Programmierung« versprechen, ist höchst widersprüchlich. Denn der Computer oder Roboter soll einerseits ethisch »gut« sein und andererseits in konkreten Dilemma-Situationen präzise funktionieren. Er soll moralisch und zielgerichtet zugleich sein.

Doch je mehr man von menschlicher Ethik und Moral versteht, umso deutlicher wird, dass dieser Anspruch unerfüllbar ist. Lernen kann man das bereits bei David Hume. Der schottische Philosoph analysierte im 18. Jahrhundert, dass unser Gefühl unserem Verstand vorausgeht. Er wird heute von der Psychologie umfassend bestätigt. Weil wir bestimmte gefühlte Überzeugungen, Haltungen und Weltanschauungen haben, suchen wir uns die dazu passenden vernünftigen Argumente. Abtreibungsgegner sind zumeist nicht deshalb gegen Abtreibung, weil vernünftige Argumente sie überzeugen. Sondern sie suchen sich mehr oder weniger vernünftige Argumente, weil sie gegen Abtreibung sind.

Tatsächlich richten wir uns in der Moral zumeist nach unserer Intuition und entscheiden nach Gefühl. Die menschliche Moral ist ein Ensemble von irgendwie nützlichen, instinktiven Handlungen und Haltungen. Vernunft allein gebietet dagegen keine Moral. Denn ohne soziale Gefühle wie Liebe, Zuneigung, Respekt, Mitleid, Furcht, Unbehagen, Ablehnung, Ekel, Scham weiß unsere Vernunft nicht, was Gut und Böse ist. Sollte man also daran denken, moralisch wichtige Intuitionen bei der »ethischen Programmierung« zu berücksichtigen? Ein schräger Einfall! Denn damit würde man etwas normieren, das nicht normierbar ist.

Über Jahrtausende hinweg hat sich der Mensch als das »Andere der Natur« definiert. Angesichts von Maschinen, die bestimmte kognitive Leistungen von Menschen übertreffen, müssen wir nun lernen, den Menschen als das »Andere der künstlichen Intelligenz« zu sehen. Nirgendwo zeigt sich das so deutlich wie in der Moral. Moral ohne Subjektivität ist keine Moral und Subjektivität ohne Moral keine Subjektivität. Moralische Urteile bestehen nicht nur aus Ergebnissen oder gar »Lösungen«, sondern der Weg, der Akt der Entscheidung, ist selbst von größter Bedeutung. Wenn die antiken Griechen ihre Tugendethik umzusetzen versuchten, wollten sie sich dabei im moralischen Handeln selbst kultivieren. Die Anstrengung gehört ebenso zur Moral wie die Handlung. In Immanuel Kants Pflichtenethik ist das nicht anders: Wir sollen unserem guten Willen folgen und lernen, das, was wir sollen, auch tatsächlich zu wollen. Die Anstrengung und der Akt des moralischen Handelns gehören also zur Moral entscheidend mit dazu.

Die präziseste Analyse dieses Aktes lieferte zu Beginn des 20. Jahrhunderts der amerikanische Philosoph, Psychologe und Soziologe George Herbert Mead. Menschen sortieren ihre Umwelt, indem sie sie verdinglichen. Personen, Dinge, Einsichten, ja sogar ganze Weltanschauungen werden so zu fest umrissenen und bewerteten Objekten. Wie das Design dieser Objekte aussieht, bestimmen die subjektiven Umstände, durch die sie in mein Leben treten. Nur so komme ich zu festen Überzeugungen über Hunde, Donald Trump, Schalke 04 oder den Kommunismus. Der Wert eines Objekts ist niemals davon zu lösen, wie er zustande kommt. Entscheidend ist am Ende nicht die Sache, sondern der Akt, in dem das Objekt geformt und damit bewertet wird.

Das hat gewaltige Folgen für die Moral. Auch hier geht es vor allem um das Situative und um den Kontext. Wirkliche moralische Entscheidungen werden eigentlich selten getroffen. Denn meistens wissen wir in sozialen Situationen ziemlich genau, was wir tun werden. Wir überlegen nur dann, wenn wir mit etwas überfordert sind und ein Konflikt entsteht: Wie bewerte ich das, was ich will, im Hinblick auf das, von dem andere wollen, dass ich es tue? In einer solchen Dilemma-Situation hilft mir meine Ausstattung mit Fürsorgeverhalten nicht weiter. Und es geht auch nicht darum, die eigene Lust oder den eigenen Schaden mit der Lust oder dem Schaden anderer abzugleichen. Viel wichtiger in Konfliktsituationen ist, die eigene Identität zu wahren. Wenn ich nicht weiß, was ich tun soll, treibt mich dies in eine Persönlichkeitskrise; ein unschöner Zustand, der mich dazu anhält, die Konsistenz meines Selbst wiederherzustellen. Menschen handeln also nicht wie Maschinen. Statt neutral auf die Umwelt reagieren sie immer auch auf ihre eigenen sozialen Erfahrungen. Sie beschäftigen sich mit ihrem Selbstkonzept. Was den Menschen zum Menschen macht, liegt nicht in apriorischen Festlegungen oder Programmierungen. Es ist die besondere Weise, wie wir mit unserer Umwelt kommunizieren. Wir reagieren nicht einfach auf sie, sondern wir konstruieren sie uns als unsere Welt.

Die menschliche Moral ist irrational, von nicht generalisierbaren sozialen Intuitionen durchzogen, hochgradig situativ, abhängig vom Kontext und aufs Engste verbunden mit unserem Selbstwertgefühl und unserem Selbstkonzept. Wiewohl sie sich an Leitplanken orientieren mag, die Philosophen und Gesetzgeber angebracht haben, bleibt sie doch eine emotional komplexe Angelegenheit. Der Gebrauch an Moral lässt sich nicht normieren wie die Größe von Dübeln. Eine Moral ohne subjektive Haltungen, reduziert auf allgemeine Bewertungen des Lebens, bleibt sinnleer. Philosophie der Moral kann deshalb kein Werkzeugkasten sein, um moralische Probleme ein für alle Mal zu lösen. Vielmehr ist sie ein Reflexionsmedium, das moralische Themen in einen Horizont stellt. Man erfindet keine zu Menschen oder Robotern passende Moral, so wie man passende Technik erfindet.

Vor diesem Problem stehen heute die Philosophen, Theologen und Gesellschaftswissenschaftler, die in den vielen Kommissionen und Ethikräten sitzen, um über die Moral des Zweiten Maschinenzeitalters oder die »ethische« Programmierung von KI nachzudenken. Bedroht vom Ziel, klare Analysen und Empfehlungen zu liefern, pflegen sie einen notgedrungen unphilosophischen Umgang mit philosophischen Fragen. Und statt eines ethischen Blicks auf Technik mühen sie sich um einen technischen Blick auf Ethik. Ärgerlich, ärgerlich, dass man Ethik nicht begradigen kann wie einen Wildwasserfluss, den man zum Kanal macht. Moralische Intuition fließt nicht in geregelten Bahnen. Und was kein Programm ist, kann auch nicht programmiert werden.

Bei den in Ethik-Kommissionen zur KI verhandelten Fragen handelt es sich meist nicht um »Probleme«, die sich mithilfe der Wissenschaft, einer Technikfolgenabschätzung oder neuerdings »Sozioinformatik« lösen lassen. In der Welt der Moral gibt es keine objektivierenden Verfahren wie in den Naturwissenschaften. Man kann das für belanglos halten; es ist aber gefährlich: Denn je technisierter und instrumenteller man menschliche Moral zerlegt und zergliedert, umso zahnloser wird sie zugleich. Von der ehemaligen Kraft einer philosophisch begründeten Haltung bleibt nur noch ein Gefilz von Einzelproblemen übrig, das nirgendwo einen passenden Ort findet, um sichtbar ein Stoppschild aufzustellen.

Tatsächlich verlangt die Maschinenethik, Maschinen nicht »ethisch« zu programmieren, also Entscheidungen treffen zu lassen, die über Menschen richten. Ethisch mit Maschinen umzugehen ist genau das Gegenteil davon, sie »ethisch« zu programmieren. Ein Beispiel dafür ist das sogenannte »autonome« Fahren. Voll automatisierte Robo-Cars werden als die Zukunft angesehen. Folglich sollen Ethiker sich Gedanken darüber machen, wie voll automatisiert fahrende Autos in einer Dilemma-Situation entscheiden sollen, wenn sich der Tod von Menschen nicht verhindern lässt. Wer soll überleben und wer überfahren werden? Kinder oder Rentner? Hunde, Kranke, Straffällige?

Dass das Grundgesetz mit seiner Definition der Menschenwürde solche Überlegungen schon im Ansatz nicht zulässt, hat die deutsche Automobilindustrie bislang zu keinem klaren Bekenntnis gegen ein solches Aufrechnen von Lebenswert genötigt. So steht zu befürchten, dass auch bei uns daran gearbeitet wird, Autos »ethisch« zu programmieren. Da diese Ethik, wie gezeigt, keine ist und zudem dem Grundgesetz widerspricht, müssten sämtliche Ethiker in Deutschland hier laut aufschreien. Stattdessen sitzen in Kommissionen (neben Ausnahmen) hinreichend Fachleute, die leiser kaum sein könnten. Und das, obwohl das zukünftige »autonome« Fahren in unseren Großstädten nicht nur den Lebenswert von Menschen verrechnen soll, sondern auch jeden Verkehrsteilnehmer uneingeschränkt überwachen wird. Robo-Cars in den Innenstädten sind vielleicht die Lösung für unsere Verkehrsprobleme. Viel wahrscheinlicher sind sie nicht einmal das. In jedem Fall aber gehen sie einher mit einem Verstoß gegen das Grundgesetz und einer drastischen Einschränkung unserer Freiheit. Wer das will, will nicht nur einen anderen Verkehr – er will einen anderen Staat.

www.zeit.de/audio

Trotzdem Lösungsansätze?

Einbeziehen von Fairness-Kriterien

- Beispiel Bewerbung
- „Neutrales“ Ranking nach objektiven Kriterien.
- Fair Ranking: Sicherstellen, dass das Ranking unterrepräsentierte Gruppen angemessen berücksichtigt.
- Der „Verlust“ gegenüber einem Ranking ohne Fairnesskriterien ist sehr gering.
- Dagegen: deutlicher Gewinn an Diversität.

FA*IR: A Fair Top-k Ranking Algorithm

Meike Zehlike
TU Berlin
Berlin, Germany
meike.zehlike@tu-berlin.de

Francesco Bonchi
ISI Foundation
Turin, Italy
francesco.bonchi@isi.it

Carlos Castillo
Universitat Pompeu Fabra
Barcelona, Catalunya, Spain
chato@acm.org

Sara Hajian
NTENT
Barcelona, Catalunya, Spain
shajian@ntent.com

Mohamed Megahed
TU Berlin
Berlin, Germany
mohamed.megahed@campus.tu-berlin.de

Ricardo Baeza-Yates
Universitat Pompeu Fabra
Barcelona, Catalunya, Spain
rbaeza@acm.org

ABSTRACT

In this work, we define and solve the Fair Top-k Ranking problem, in which we want to determine a subset of k candidates from a large pool of $n \gg k$ candidates, maximizing utility (i.e., select the "best" candidates) subject to group fairness criteria.

Our ranked group fairness definition extends group fairness using the standard notion of protected groups and is based on ensuring that the proportion of protected candidates in every prefix of the top-k ranking remains statistically above or indistinguishable from a given minimum. Utility is operationalized in two ways: (i) every candidate included in the top-k should be more qualified than every candidate not included; and (ii) for every pair of candidates in the top-k, the more qualified candidate should be ranked above.

An efficient algorithm is presented for producing the Fair Top-k Ranking, and tested experimentally on existing datasets as well as new datasets released with this paper, showing that our approach yields small distortions with respect to rankings that maximize utility without considering fairness criteria. To the best of our knowledge, this is the first algorithm grounded in statistical tests that can mitigate biases in the representation of an under-represented group along a ranked list.

KEYWORDS

Algorithmic fairness, Bias in Computer Systems, Ranking, Top-k selection.

1 INTRODUCTION

People search engines are increasingly common for job recruiting and even for finding companionship or friendship. A top-k ranking algorithm is typically used to find the most suitable way of ordering items (persons, in this case), considering that if the number of people matching a query is large, most users will not scan the entire list. Conventionally, these lists are ranked in descending order of some measure of the relative quality of items.

The main concern motivating this paper is that a biased machine learning model that produces ranked lists can further systematically reduce the visibility of an already disadvantaged group [10, 31] (corresponding to a legally protected category such as people with disabilities, racial or ethnic minorities, or an under-represented gender in a specific industry).

According to [14] a computer system is *biased* "if it systematically and unfairly discriminate[s] against certain individuals or groups of individuals in favor of others. A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or a group of individuals on grounds that are unreasonable or inappropriate." Yet "unfair discrimination alone does not give rise to bias unless it occurs systematically" and "systematic discrimination does not establish bias unless it is joined with an unfair outcome." On a ranking, the desired good for an individual is to appear in the result and to be ranked amongst the top-k positions. The outcome is unfair if members of a protected group are systematically ranked lower than those of a privileged group. The ranking algorithm discriminates unfairly if this ranking decision is based fully or partially on the protected feature. This discrimination is systematic when it is embodied in the algorithm's ranking model. As shown in earlier research, a machine learning model trained on datasets incorporating *preexisting bias* will embody this bias and therefore produce biased results, potentially increasing any disadvantage further, reinforcing existing bias [28].

Based on this observation, in this paper we study the problem of producing a fair ranking given one legally-protected attribute,[1] i.e., a ranking in which the representation of the minority group does not fall below a minimum proportion p *at any point in the ranking*, while the utility of the ranking is maintained as high as possible.

We propose a post-processing method to remove the systematic bias by means of a *ranked group fairness criterion*, that we introduce in this paper. We assume a ranking algorithm has given an undesirable outcome to a group of individuals, but the algorithm itself cannot determine if the grounds were appropriate or not. Hence we expect the user of our method to know that the outcome is based on unreasonable or inappropriate grounds and provide p as input which can originate in a legal mandate or in voluntary commitments. For instance, the US Equal Employment Opportunity Commission sets a goal of 12% of workers with disabilities in

CIKM'17, November 6–10, 2017, Singapore.

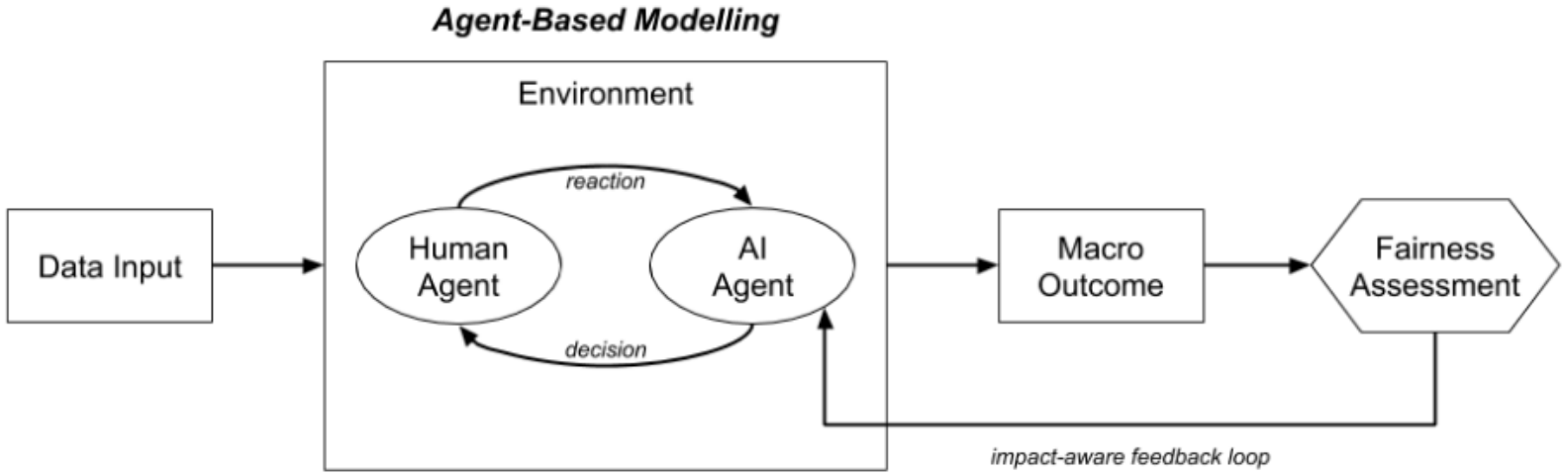
[1] We make the simplifying assumption that there is a dominant legally-protected attribute of interest in each case. The extension to deal with multiple protected attributes is left for future work.

Consequences of Artificial Intelligence on Urban Societies (CAIUS)

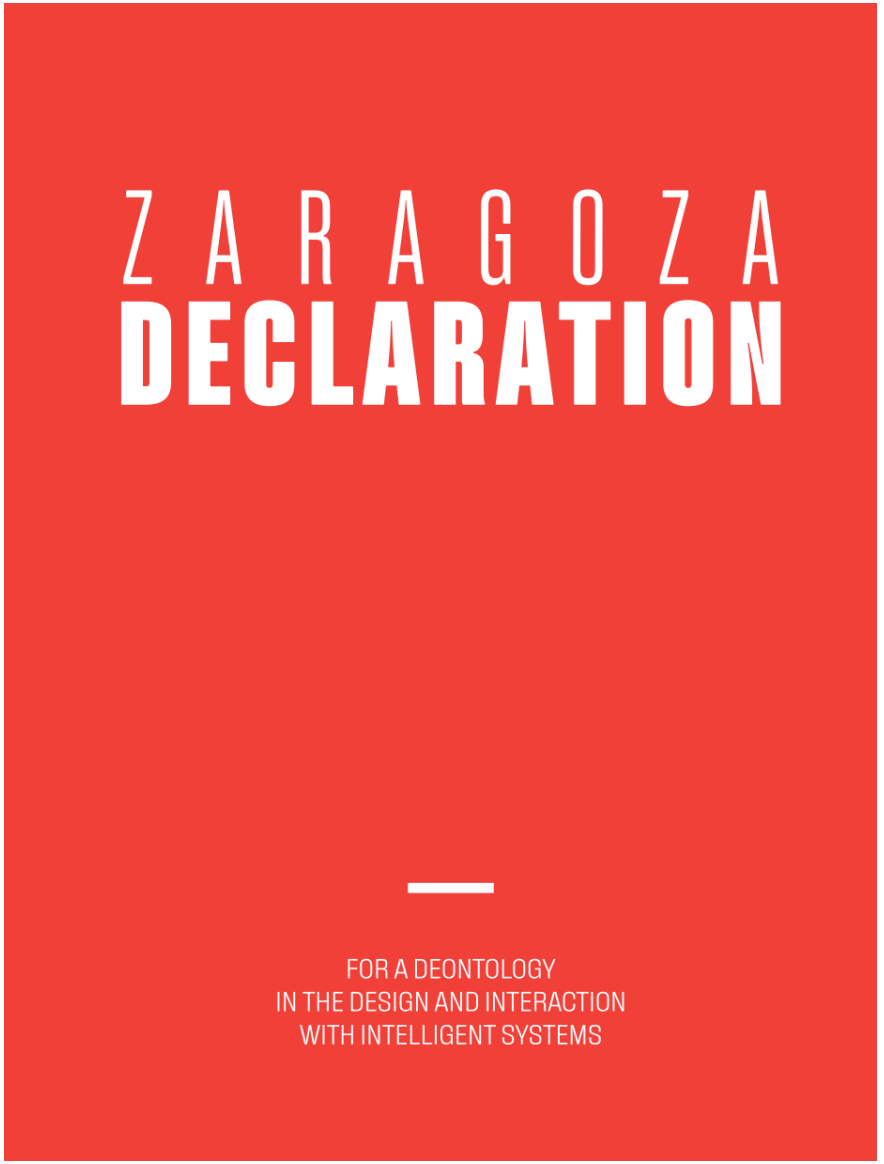
- Hintergrund: Automatische Entscheidungsfindung in Smart Cities
- Frage: Führen „smarte“ Lösungen zu einer Verringerung der Solidarität und Gleichheit in der Gesellschaft.
- Motivierendes Beispiel: Parkplatz-Leitsystem über das Smartphone.
- KI kann beeinflussen: Preis, Reservierungen
- Problem: Eigentlich faire Verteilung (first come first serve) wird tendenziell weniger solidarisch (Menschen ohne Smart Phone, Menschen mit weniger Einkommen).
- Was ist das Ziel? Weniger Verkehr? Weniger Autos? Faire Verteilung? Was ist fair?

Ansatz

- Fairness als Zielfunktion nutzen (einfacher gesagt als getan, was ist fair?)
- Simulation des Einsatzes der KI auf Echtdaten nutzen, um die Fairness von Entscheidungen zu prüfen. Feedback aus Simulation auf das gelernte Modell der KI.
- Vergleichbar mit Fa*ir, klassische innere Optimierung, Fairness als übergeordnetes Ziel.



Zaragoza Declaration



1

All A.I. technologies
must be socially and
environmentally responsible

2

Traceability and testability
are essential elements
of any AI technologies

3

Fundamental rights cannot be limited by A.I. technologies that cannot be explained, or via methods that are not reproducible.

HOCHSCHULE DER MEDIEN

4

Teams that develop A.I. technologies should be transdisciplinary, integrating technoscientific and humanistic knowledge.

HOCHSCHULE
DER MEDIEN

5

The development of any A.I.
technologies must adhere to a
deontological code of social responsibility
for professionals and companies
that respects the previous points

Über- oder Unterschätzung?

Joseph Weizenbaum

Der meiste Schaden, den der Computer potentiell zur Folge haben könnte, hängt **weniger davon ab, was der Computer tatsächlich machen kann** oder nicht kann, als **vielmehr von den Eigenschaften, die das Publikum dem Computer zuschreibt.**

ZEIT Nr. 03/1972

Literatur

- Machine Bias: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
  Gegendarstellung: http://www.crj.org/assets/2017/07/9_Machine_bias_rejoinder.pdf
- Code is Law, Codev2: http://codev2.cc/
- Design Justice, A.I., and Escape from the Matrix of Domination:
  https://jods.mitpress.mit.edu/pub/costanza-chock/release/4
- The Moral Machine experiment:
  https://ore.exeter.ac.uk/repository/bitstream/handle/10871/39187/Moral%20Machine.pdf
  https://moralmachine.mit.edu/
- Roboter können keine Moral: https://www.zeit.de/2020/26/kuenstliche-intelligenz-roboter-moral-gefahr-ethik
- FA*IR: A Fair Top-k Ranking Algorithm: https://arxiv.org/abs/1706.06368
- Zaragoza Declaration: https://www.fundacionzcc.org/estaticos/upload/0/001/1914.pdf