Künstliche Intelligenz (KI)

Lutz Prechelt

- Definitionen für KI
- Funktionsweise symbolische KI
  - Erfolgsbeispiel
- Funktionsweise subsymbolische KI (maschinelles Lernen, Neuronale Netze)
  - Erfolgsbeispiel
- Quellen von KI-Technikfolgen:
  - Robustheitsprobleme
  - Fairnessprobleme
    - Objektivitäts-Pseudodiskussion
  - andere Verzerrungen

These "Definition"

Wir haben keine zufriedenstellende Definition von

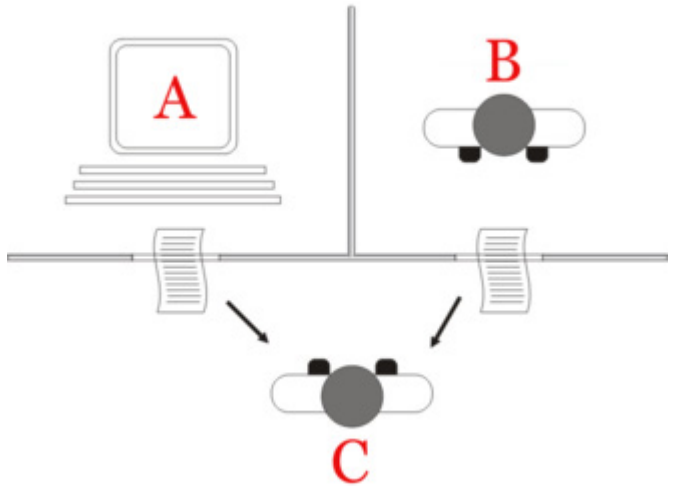
"Künstlicher Intelligenz",

die operationalisierbar ist.

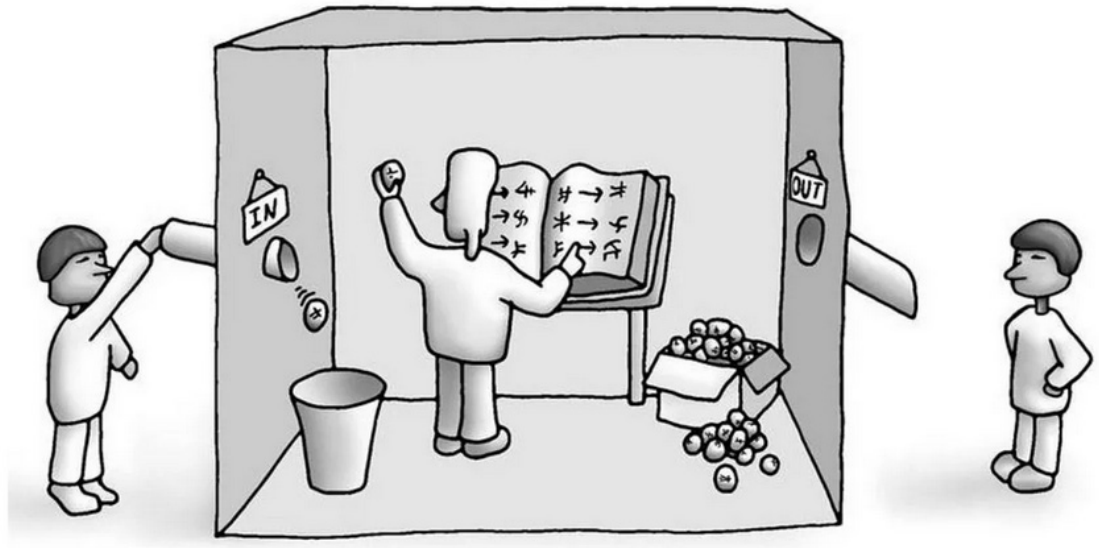
Operationalisierung: Festlegung, wie (unabhängig vom Beobachter) gemessen wird.

Definition "Künstliche Intelligenz" (KI) (artificial intelligence, AI)

- [Alan Turing, 1950]: "The Imitation Game"
  - Heute meist genannt "Turing-Test"
  - Vorschlag eines Verfahrens als Ersatz für "Can the machine think?"



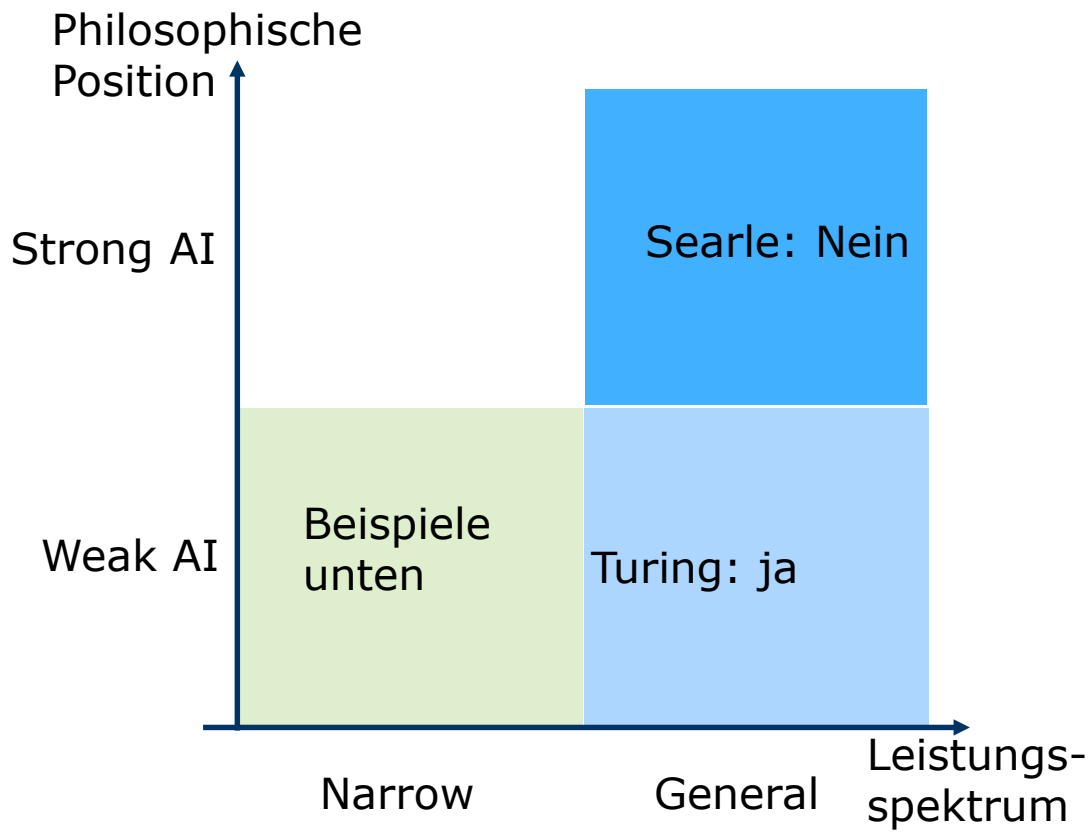
- Einwand: (z.B.) [John Searle, 1980]
  - Das beweist kein Verstehen ("intentionality")



Offene Frage: Braucht KI ein Bewusstsein?

- **"Schwache" KI (weak AI)**:
  Computer ist Problemlösungswerkzeug
- **"Starke" KI (strong AI)**:
  Computer kann ein Bewusstsein haben wie ein Mensch

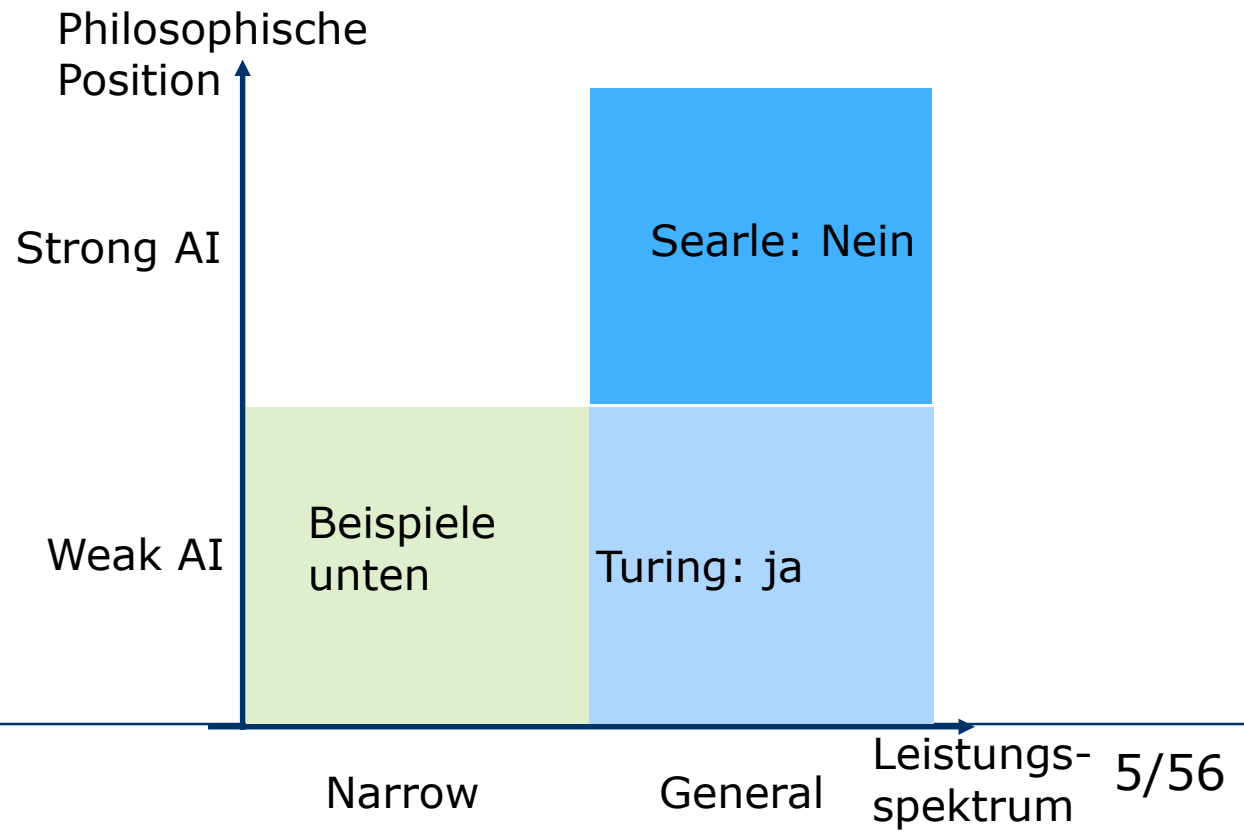
- **Narrow AI**:
  Löst spezifisches, eng umgrenztes Problem
- **Artificial General Intelligence, AGI**:
  Zeigt intelligentes Verhalten auf allen Sektoren, wie ein Mensch (oder darüber hinaus: Artificial Superintelligence)



Definition "Künstliche Intelligenz" (KI)
2. Versuch: Nichtprozedurale Definition

- [Patrick Henry Winston, 1981]
  - *"Artificial Intelligence is the study of ideas that enable computers to be intelligent."*
    - Tätigkeitsbereich, nicht Produkt!
    - Er gesteht ein, dass schwierig festzulegen ist, was "intelligent" bedeutet.
  - Ziel 1: *"make computers more useful"*
  - Ziel 2: *"understand the principles that make intelligence possible"*

- Einwand: "more useful" verglichen womit?
  - Konsequenz: Sobald wir etwas gut hinbekommen haben, ist es keine KI mehr.

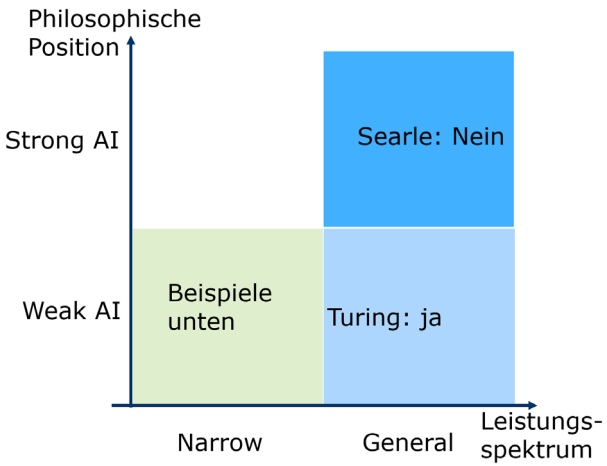


Definition "Künstliche Intelligenz" (KI): Noch 4 Arten von Künstlicher Intelligenz [RusNor16]

Freie Universität Berlin

Thinking Humanly	**Thinking Rationally**
"The exciting new effort to make computers think . . . *machines with minds*, in the full and literal sense." (Haugeland, 1985)	"The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985)
Strong AI	"The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
Acting Humanly	**Acting Rationally**
"The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990)	"Computational Intelligence is the study of the design of intelligent agents." (Poole *et al.*, 1998)
"The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	"AI . . . is concerned with intelligent behavior in artifacts." (Nilsson, 1998)

Whitebox AI
(Prozess verstehen)



Blackbox AI
(nur Problem lösen)

Menschenähnlich

Verfolgung präziser Ziele

Problem des Whitebox-KI-Ansatzes:

Die Hirnfunktion ist sehr unvollständig verstanden:

- Thinking Humanly:
  Die Gehirn-Hardware funktioniert ganz anders als die von Digitalrechnern
  - z.B. meist zeitlich nicht getaktet
- Thinking Rationally:
  Ob komplexe Kognition über Alltagsdinge überhaupt scharf rational funktionieren *kann*, ist unklar.
  - Falls nicht, was ist mit "thinking" gemeint?

Problem des rationalen KI-Ansatzes:

- Rationalität verlangt eine präzise bekannte Zielfunktion
- Die hat man in der Praxis nur in einfachen Fällen
  - z.B. für "Schachspiel gewinnen"
  - aber z.B. schon nicht mehr im Alltag für "Schach spielen"
  - geschweige denn für z.B. "eine Straße überqueren"

Definition "Künstliche Intelligenz" (KI), 3. Versuch

- "AI is a science of computer technologies developed to achieve and explain *intelligent* behavior in machines.
- By *intelligent* I refer here to a collection of abilities that enable an entity to
  - (i) solve or learn how to solve certain (difficult) domain specific problems,
  - (ii) master the known and the unknown, i.e., to act successfully in known, unknown, and dynamic environments (which requires perception, planning, agency, etc.),
  - (iii) think rationally, avoid contradictions, and explore abstract theories,
  - (iv) self-reflect, recognize self-contradictions, and reconcile one's reasoning with higher-level goals and norms, and
  - (v) interact socially with other entities and reconcile one's goals and norms with those of a community (for a greater good)."

[Benzmüller22]

Einiges davon deckt das Imitation Game mit ab, aber die Problemklasse ist weiter. "perception" und "community" sprengen seinen Rahmen z.B. gänzlich.

Viele obige Begriffe sind interpretationsbedürftig.

These "Definition"

Wir haben keine zufriedenstellende Definition von

"Künstlicher Intelligenz",

die operationalisierbar ist.

(Es ist offen, wie weit das Imitation Game reicht.)

2 Sorten von Verfahren: Symbolische vs. subsymbolische KI

Freie Universität Berlin

Symbolische KI

- operiert mit Zwischenergebnissen mit wohlverstandener Bedeutung
  - Variablen, für die wir sinnvolle Namen haben
- eignet sich für "acting rationally"

- Beispiele:
  - Brettspiele wie Dame, Reversi, Schach
  - Such- und Optimierungsverfahren
  - Expertensysteme
  - Theorembeweiser

Subsymbolische KI (≈Machine Learning)

- operiert mit Zwischenergebnissen, deren Rechenregeln aus Daten gelernt wurden
  - und meist kaum zu verstehen sind
- eignet sich für viele Fälle von "acting humanly"
  - **vor allem, wenn die Domäne nicht scharf umgrenzt ist (physische Welt)**

- Beispiele:
  - Computer Vision
  - Spracherkennung
  - Maschinelles Übersetzen
  - allgemeine Antwortsysteme
  - Robotik

Symbolische KI: Beispiel: Reversi/Othello

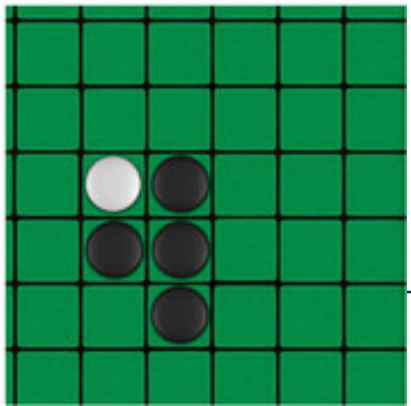
- 2 Spieler_innen (schwarz/weiß), setzen abwechselnd je 1 Stein (vorn/hinten = schwarz/weiß)
  - 8 x 8 Spielbrett (64 Felder)
  - Ziel: Mehr Steine als Gegner_in am Ende
  - Anfangs liegen 2x2 in der Mitte
    - nächster Stein muss immer angrenzen
  - Regel: Wer gegnerische(n) Stein(e) in einer Linie einschließt, darf sie umdrehen
    - Spiel dauert 60 Züge
- Triviales KI-Programm dafür:
  - Alle künftigen Spielstellungen untersuchen
    - nach 60 Zügen ist ja klar, wer gewinnt
  - Günstigsten Zug finden, auf den Gegner_in keine gute Antwort haben kann

- Dauert zu lange:
  - Bei je 10 Möglichkeiten gibt es 10^60 Stellungen
  - Optimistische Annahmen: 10^9 Stllg/sec, 10^5 sec/t, 10^3 t/Jahr ➔ 10^(60-9-5-3) = 10^43 J
  - Das Universum ist 10^10 J alt ➔ 10^33 Universen
    - mit 1000x schnellerer Hardware 10^30 Universen
- Das KI-Problem:
  - Wie verkleinert man diesen Suchraum* genügend weit?
    - Die meiste symbolische KI besteht aus Suchraum-verkleinerungen

1

2

3



*Der Suchraum sind die 10^60 möglichen Stellungen.
Verkleinern heißt, ganze Teile davon zugleich als ungünstig zu erkennen.

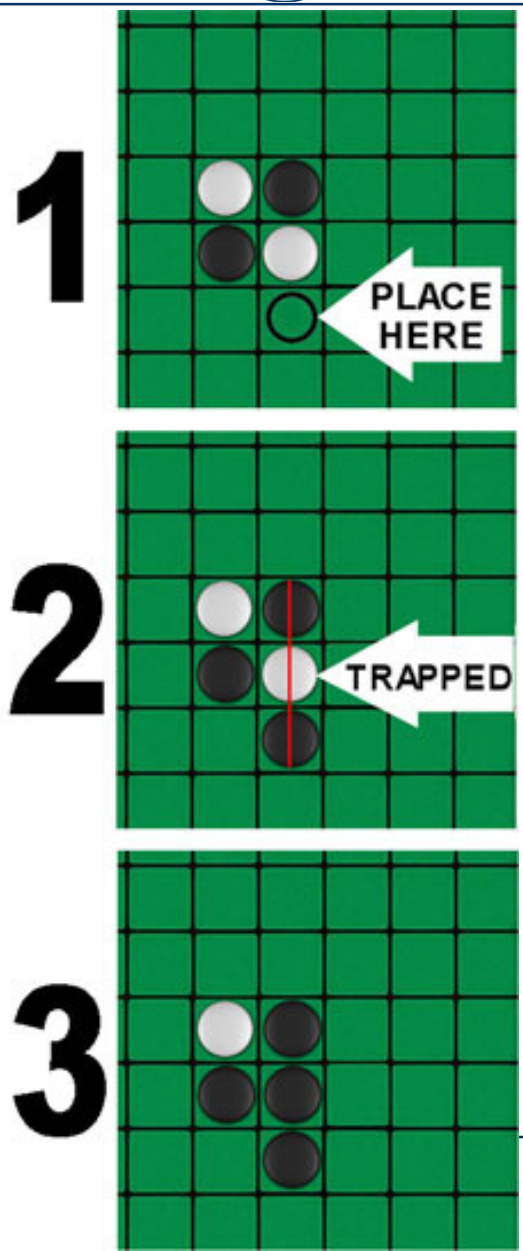
Symbolische KI: Beispiel: Reversi/Othello
Algorithmenideen

- Das KI-Problem:
  - Wie verkleinert man diesen Suchraum genügend weit?

Ideen:

1. Symmetrie ausnutzen
2. Nur k Züge weit die Stellungen entwickeln, dann bewerten
   - z.B. Zahl eigener Steine, Steine am Rand, Steine in Ecken
3. Sobald ein Zug Z2 als garantiert schlechter erkannt ist als Z1:
   - alle anderen Folgestellungen von Z2 nicht mehr betrachten
4. Vielversprechendste Züge zuerst erkunden

- Ideen 2-4 sind für viele solche Brettspiele anwendbar
- Algorithmus: Alpha-Beta-Suche



Symbolische KI: Beispiel: Reversi/Othello
Alpha-Beta-Suche


```
function alphabeta(node, depth, α, β, maximizingPlayer) is
    if depth = 0 or node is a terminal node then
        return the heuristic value of node
    if maximizingPlayer then
        value := −∞
        for each child of node do
            value := max(value, alphabeta(child, depth − 1, α, β, FALSE))
            if value > β then
                break (* β cutoff *)
            α := max(α, value)
        return value
    else
        value := +∞
        for each child of node do
            value := min(value, alphabeta(child, depth − 1, α, β, TRUE))
            if value < α then
                break (* α cutoff *)
            β := min(β, value)
        return value
```

Rekursiver Alg.

Idee 2 → Unterprogramm

Idee 4 → Unterprogramm

Idee 3

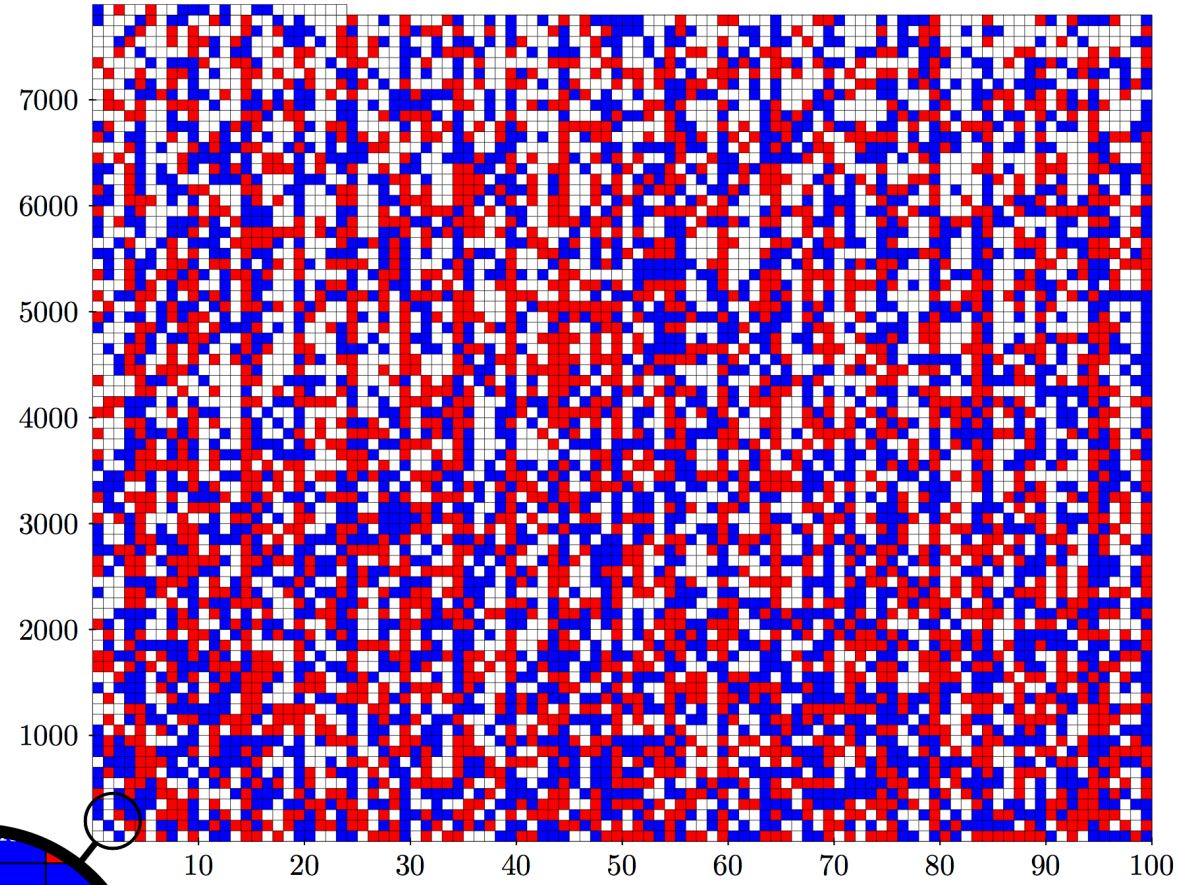
Idee 4

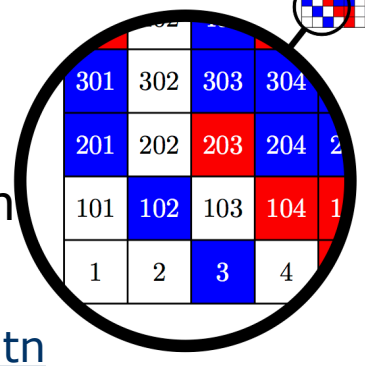
Idee 3

https://en.wikipedia.org/wiki/Alpha%E2%80%93beta_pruning

Symbolische KI: Erfolgsbeispiel

- Pythagoräisches Tripel:
  - Ganze Zahlen a, b, c, so dass $a^2+b^2=c^2$
- Frage: Kann man die nat. Zahlen so in blau oder rot färben, dass kein pythag. Tripel nur eine Farbe hat?
- Antwort: Nein
  - Für 1..7824 geht es; für 1..7825 nicht
- Zu betrachten sind 2^{7825} Färbungen
  - $\approx 10^{2355}$, unfassbar viele
- Der Beweis ist 200 Terabytes lang
  - Basisverfahren: SAT-Solving
    - Prüft, ob eine riesige aussagenlogische Formel erfüllbar ist (satisfiable)
  - 2 Tage Rechenzeit auf 800 CPU-Kernen





Einige der vielen möglichen Färbungen für 1..7824
(weiß = beliebig; jede weiße Zelle verdoppelt die Zahl der Lösungen)

Quellen s. https://www.cs.utexas.edu/~marijn/ptn
Auf die Antwort war seit 1980 ein Preis ausgesetzt

- 3 Arten:
  - **überwacht** (supervised): Daten enthalten gewünschte Ergebnisse
    - Lernt viel, aber Beschaffung der "Trainingsdaten" ist aufwändig
      - Bei Klassifikation heißen die Ergebnisse "Labels"
  - **unüberwacht** (unsupervised): Daten enthalten keine Ergebnisse, Algorithmus bestimmt Ähnlichkeitsgruppen
    - Trainingsdaten sind kein Problem, bestimmt aber nur "irgendwelche" Gruppen
  - **verstärkend** (reinforcement learning): Algorithmus erhält zu Ausgaben eine globale Bewertung (Gütemaß)
    - Algorithmen sind sehr viel schwieriger hinzubekommen
    - Oft hilfreich, wenn Problemlösung mehrere Schritte hat und Bewertung am Ende klar ist
      - z.B. bei Reversi
- Diverse Variationen, z.B. teilüberwacht (semi-supervised) mit nur wenigen Labels

Beispiel für maschinelles Lernen (sehr vereinfacht): Fishers "iris"-Datensatz

Freie Universität Berlin

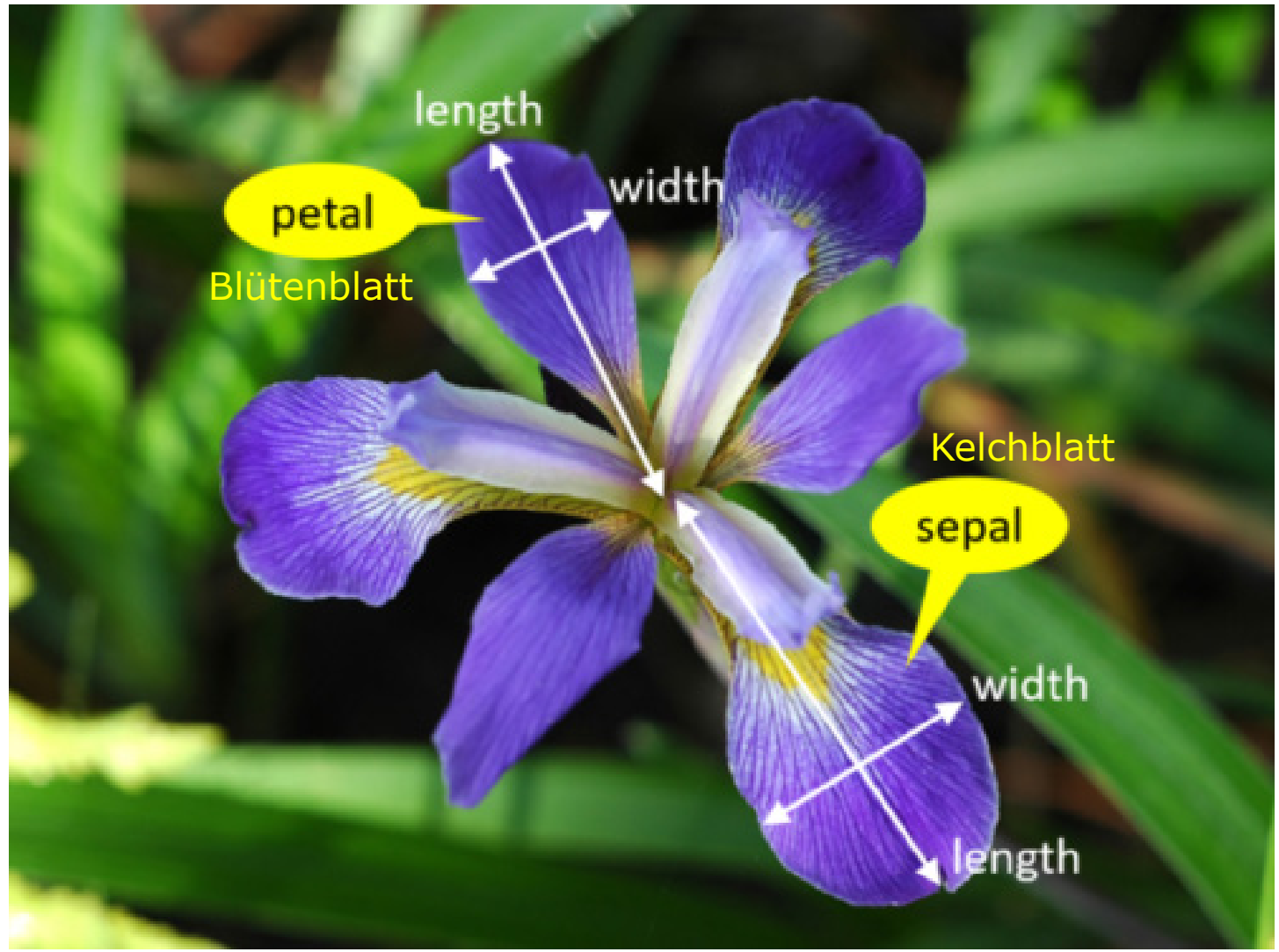
- Wir wollen aus Daten lernen, drei Spezies von Lilien (engl: iris) zu unterscheiden:
  - Iris Versicolor, Iris Virginica, Iris Setosa

- Dazu benötigen wir Merkmale, die sich leicht und verlässlich ermitteln lassen.
  - Geeignet sind z.B. die Maße der Blütenblätter und der Kelchblätter

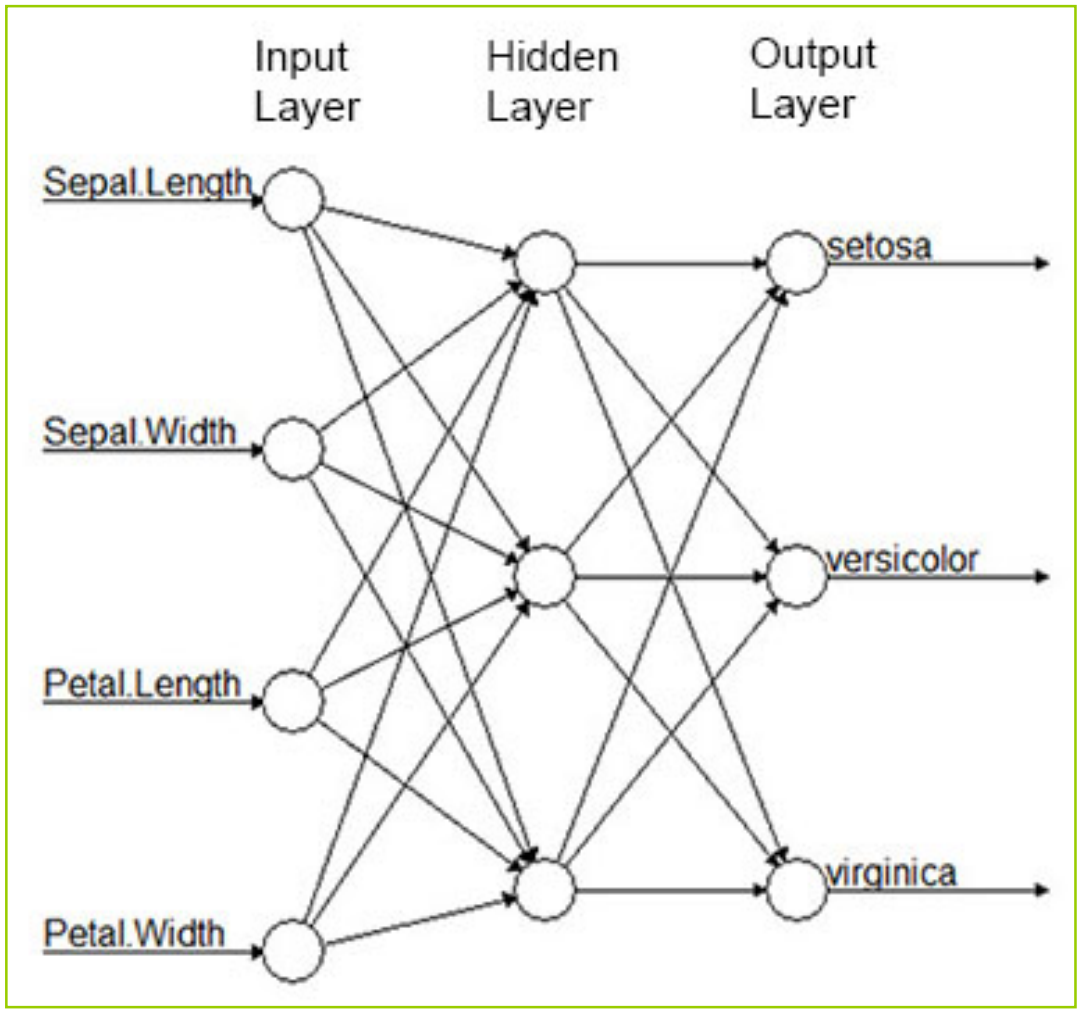
Achtung: Wir benutzen das Problem nur der Anschaulichkeit halber.
Lösen kann man es gut ohne ML.

Beispiel für maschinelles Lernen (sehr vereinfacht): Fishers "iris"-Datensatz: Merkmale

- Wir vermessen an vielen Blüten diese vier Merkmale
  - das wird die Eingabe
- und bestimmen dazu die Spezies
  - Die Bestimmung braucht eine Expert_in!
    - Dadurch sind Trainingsdaten tendenziell knapp
  - Das wird die zu lernende Ausgabe



Beispiel für maschinelles Lernen (sehr vereinfacht): Fishers "iris"-Datensatz: Neuronales Netz (NN)



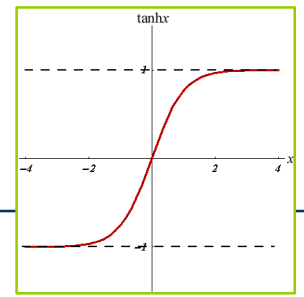
Hidden Layer könnte auch 2 oder 20 Knoten haben.
Deep Learning: Viele Hidden Layers (z.B. 10).

-  Verbindung (connection):

  Multipliziere mit einem "Gewicht" (weight)
  - Parameter des statistischen Modells, der beim Lernen angepasst wird

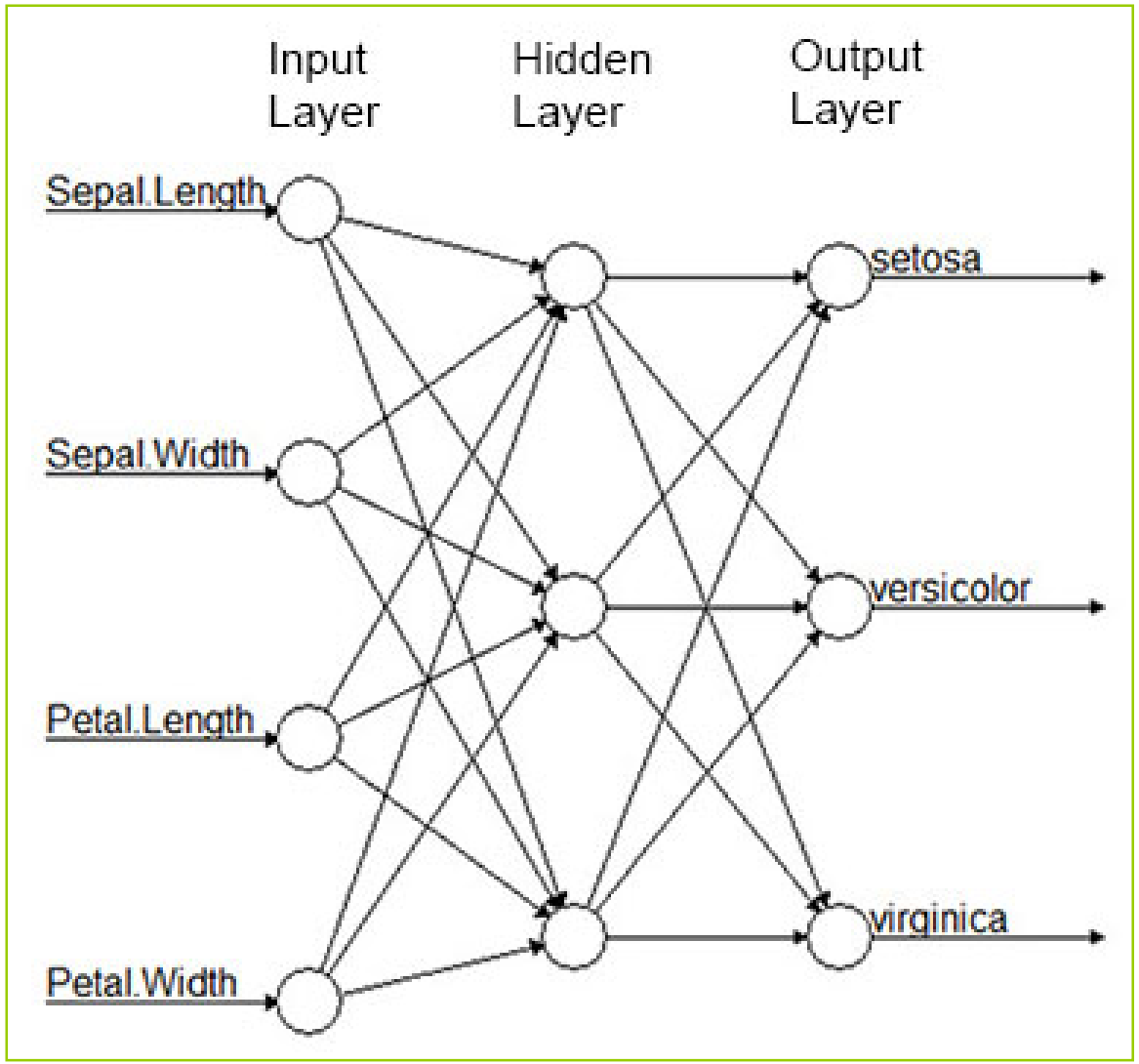
-  Knoten (node):

  Addiere Eingaben, wende "Aktivierungsfunktion" an
  - eine Quetschfunktion (squashing function), z.B. tanh, die die Daten in den Bereich z.B. -1…1 zwingt und differenzierbar ist
    - Nichtlinear!



Nicht im Bild: Eine zusätzliche feste Eingabe von 1 ("bias") mit Verbindung zu in jedem Knoten, die den Lernvorgang stark erleichtert.

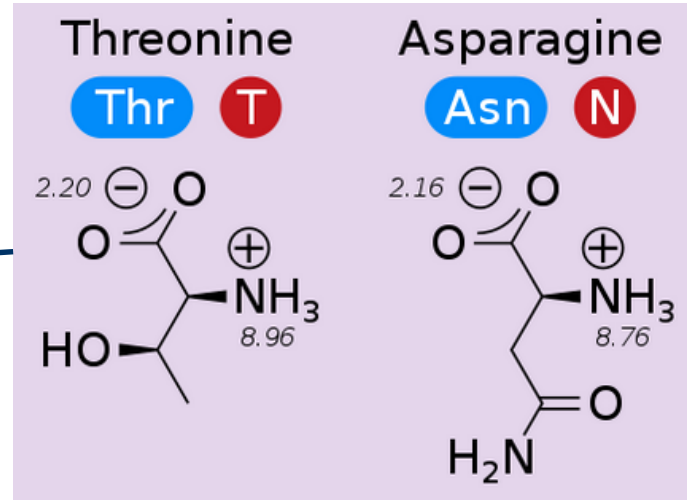
Beispiel für maschinelles Lernen (sehr vereinfacht): Fishers "iris"-Datensatz: Kodierung, Lernverfahren



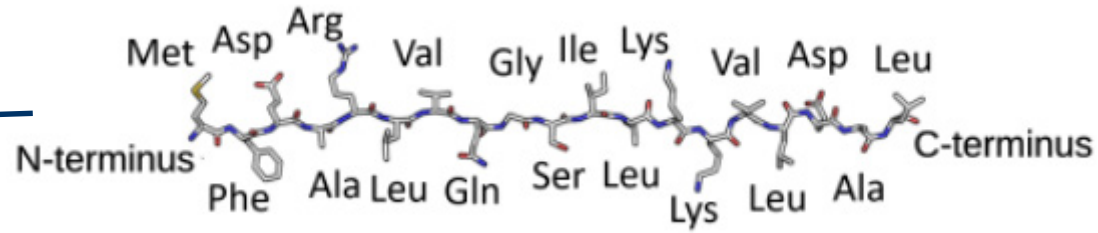
- Soll-Ausgabe ist z.B. (0, 0, 1) für eine Eingabe, die eine virginica beschreibt.
- Algorithmus (Grundidee):
  - Anfangsgewichte sind zufällig.
  - Iteriere viele Male:
    - Berechne den mittleren quadratischen Fehler E über alle Eingabebeispiele
      - z.B.: Ausgabe = (0.3, 0.1, 0.8) → E = (0.09, 0.01, 0.04)
      - E wollen wir minimieren, dann beschreiben die Gewichte eine Näherungslösung für das Lernproblem
    - Berechne den Gradienten von E für jedes Gewicht (Kettenregel der Differentiation)
    - Ändere jedes Gewicht ein bisschen, proportional zu seinem Gradienten
- (Die Details spielen für uns keine Rolle)

Subsymbolische KI: Erfolgsbeispiel AlphaFold

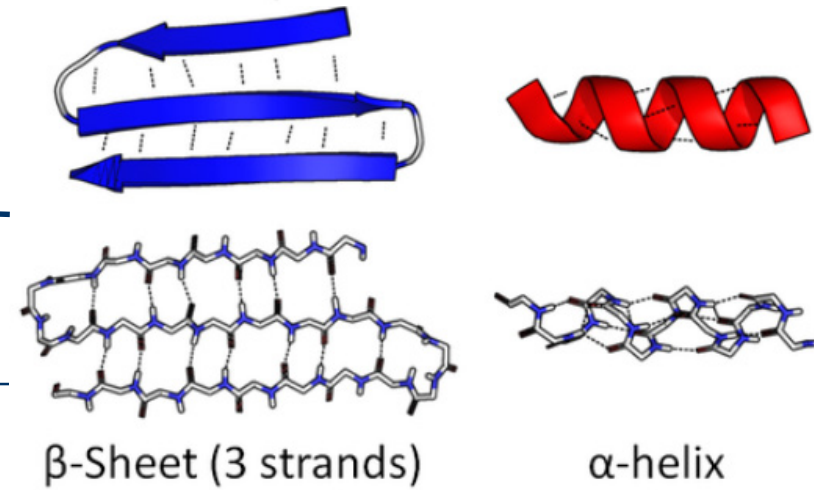
- Die meisten biologischen Funktionen werden durch Proteine (Eiweiße) vermittelt.
  - Proteine bestehen aus Aminosäuren
    - und diese fast nur aus C, N, O, H
  - Es gibt nur ~20 verschiedene Aminosäuren



- Proteine sind Ketten von Aminosäuren
  - "Primärstruktur"



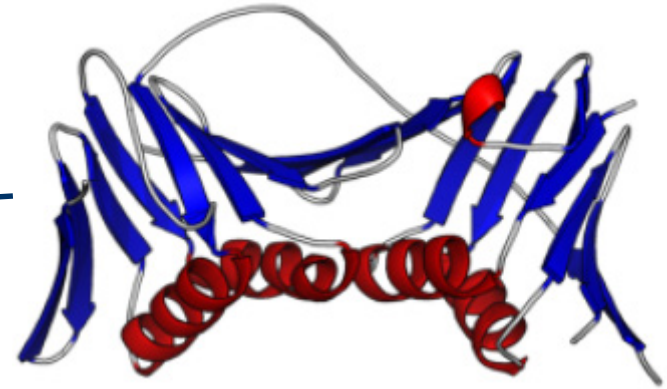
- Diese falten sich streckenweise zu 2 räumlichen Strukturen
  - "Sekundärstruktur":
    - alpha-Helix
    - beta-Sheet



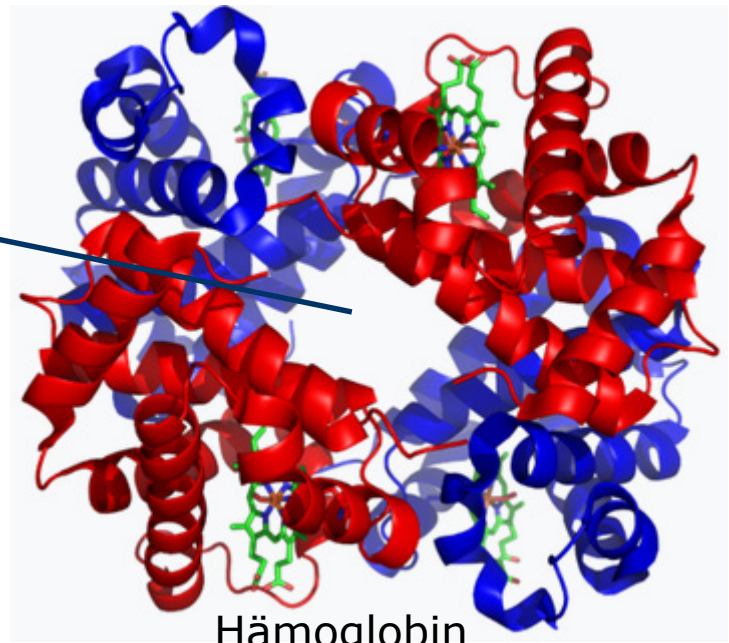
Lehrvideo: https://www.youtube.com/watch?v=wvTv8TqWC48

Subsymbolische KI: Erfolgsbeispiel AlphaFold (2)

- Die Abschnitte der Sekundärstruktur falten sich in eine komplizierte räumliche Form
  - "Tertiärstruktur"



- Diese ist entscheidend für die biologische Funktion
  - z.B. Hämoglobin bildet eine "Tasche" für ein Sauerstoffmolekül

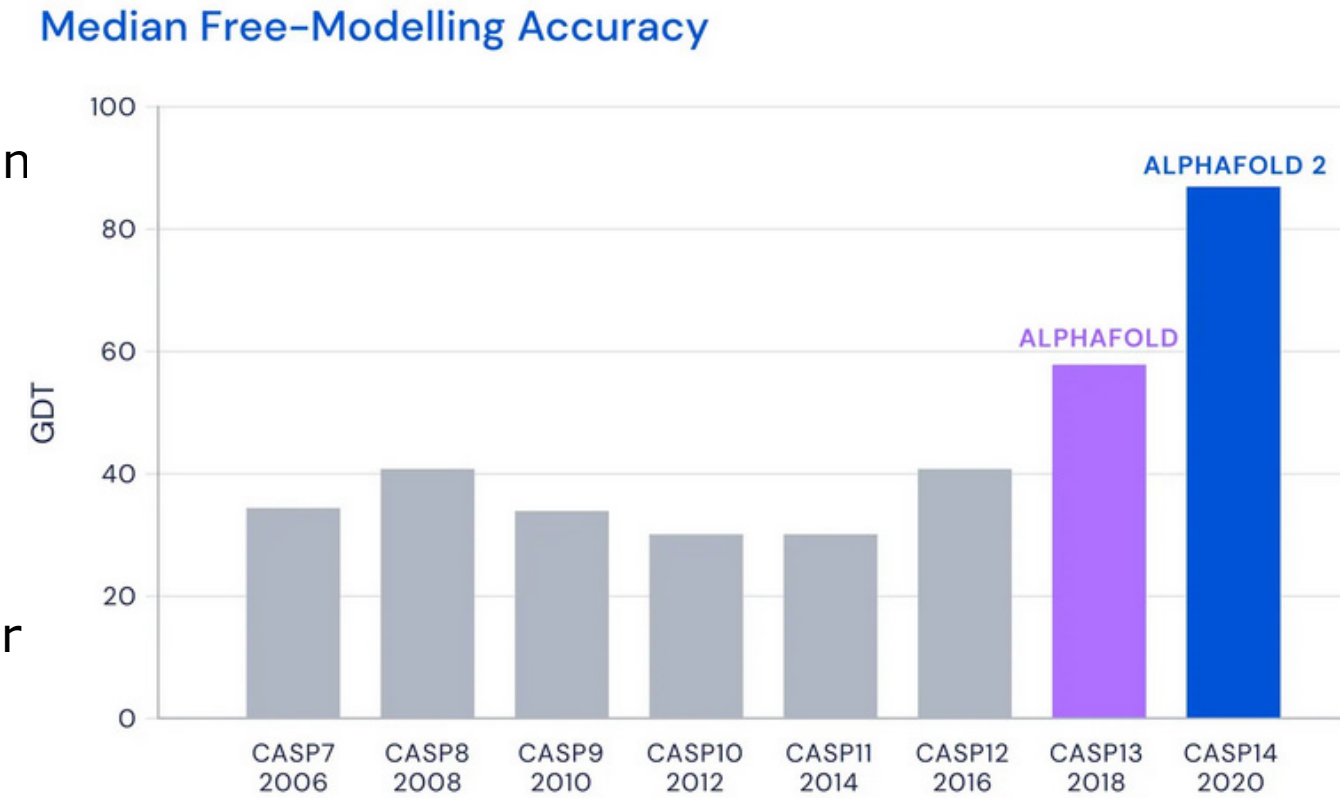


Hämoglobin

- Theoretisch lässt sich die Struktur berechnen (elektrische Kräfte etc.)
  - Das ist aber seit Jahrzehnten kaum gelungen
  - Messung im Labor ist extrem aufwendig

Subsymbolische KI: Erfolgsbeispiel AlphaFold (3)

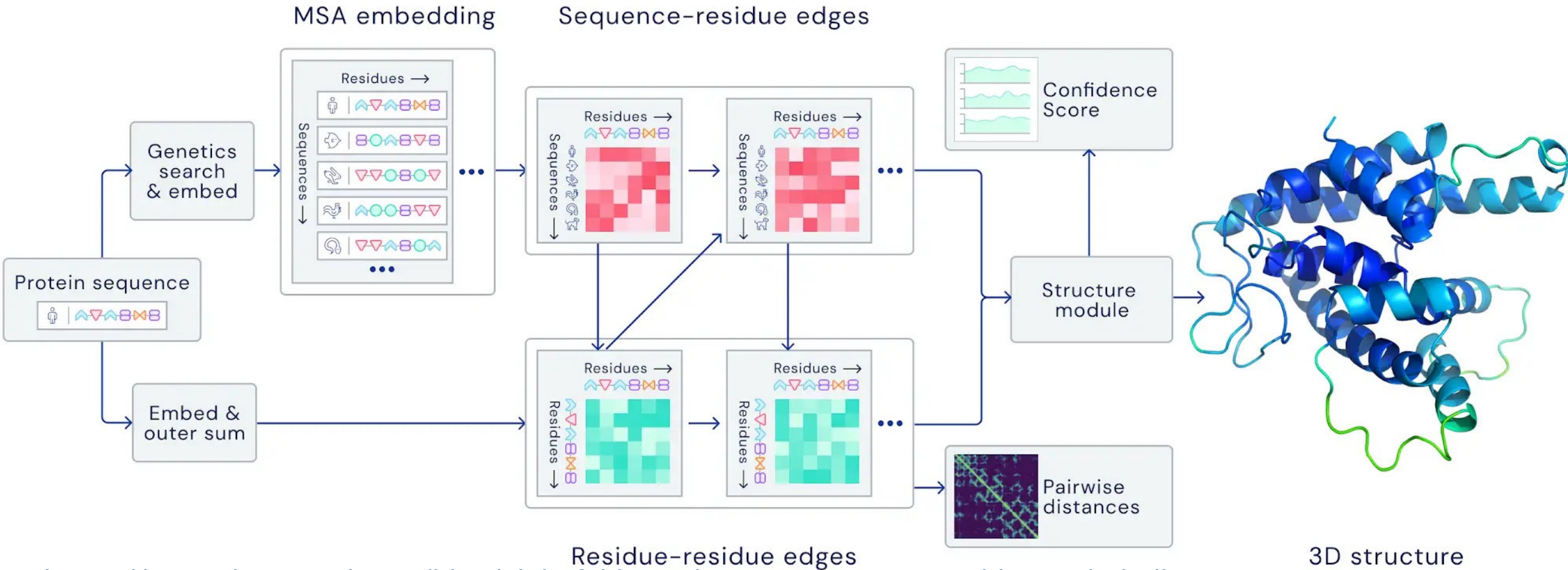
- AlphaFold2:
  - NN, trainiert mit den ~170 000 Strukturen der rcsb.org Protein-Datenbank
  - plus vielen Sequenzen unbekannter Struktur (teilüberwachtes Lernen)
- CASP-Wettbewerb:
  - Sage Tertiärstruktur neuer, noch unbekannter Proteine voraus
  - Lange schafften die besten Teams das nur für ~40% der Abschnitte gut
  - AlphaFold2 erreichte die beste erwartbare Leistung von ~90%
    - (und genauer sind die Labordaten gar nicht)
  - Bio-Forscher_innen sind begeistert



Können Menschen das? Nein.
Ist es KI? Ja:
"… (i) solve or learn how to solve certain (difficult) domain specific problems"

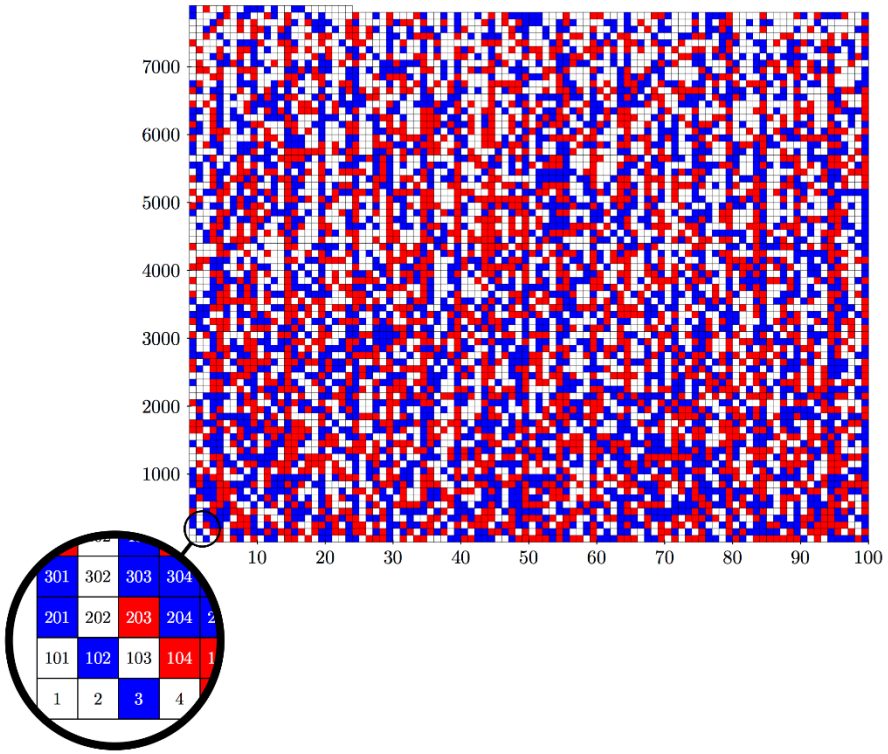
Subsymbolische KI: Erfolgsbeispiel AlphaFold (4)

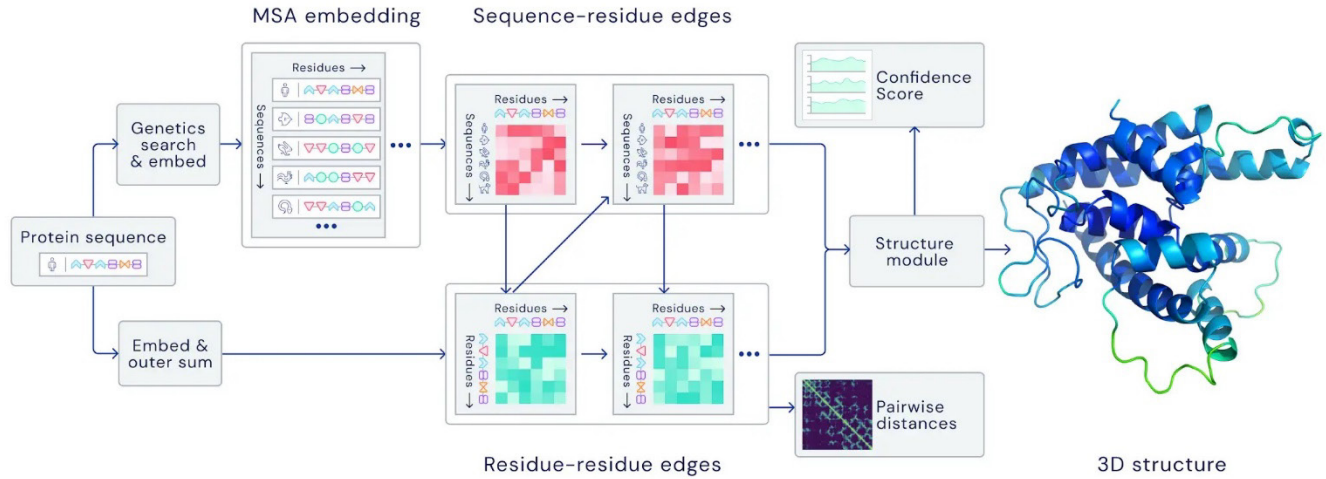
- Extrem komplexe NN-Architektur: Ähnlichkeitssuche mit bekannten Proteinen, Sequenz als räumlicher Graph möglicher Anziehungen, iterative Funktionsweise.



https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

Welche Leistung ist beeindruckender? Warum?

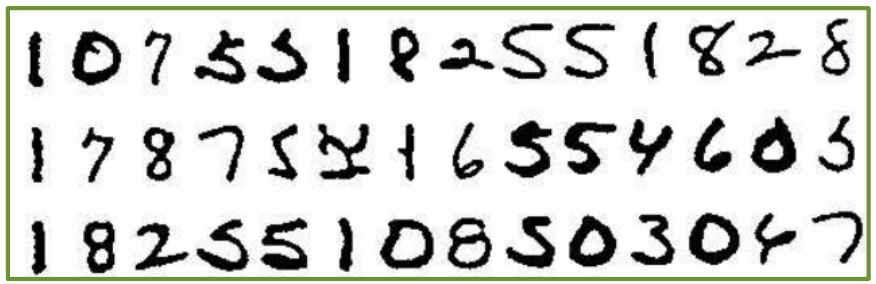


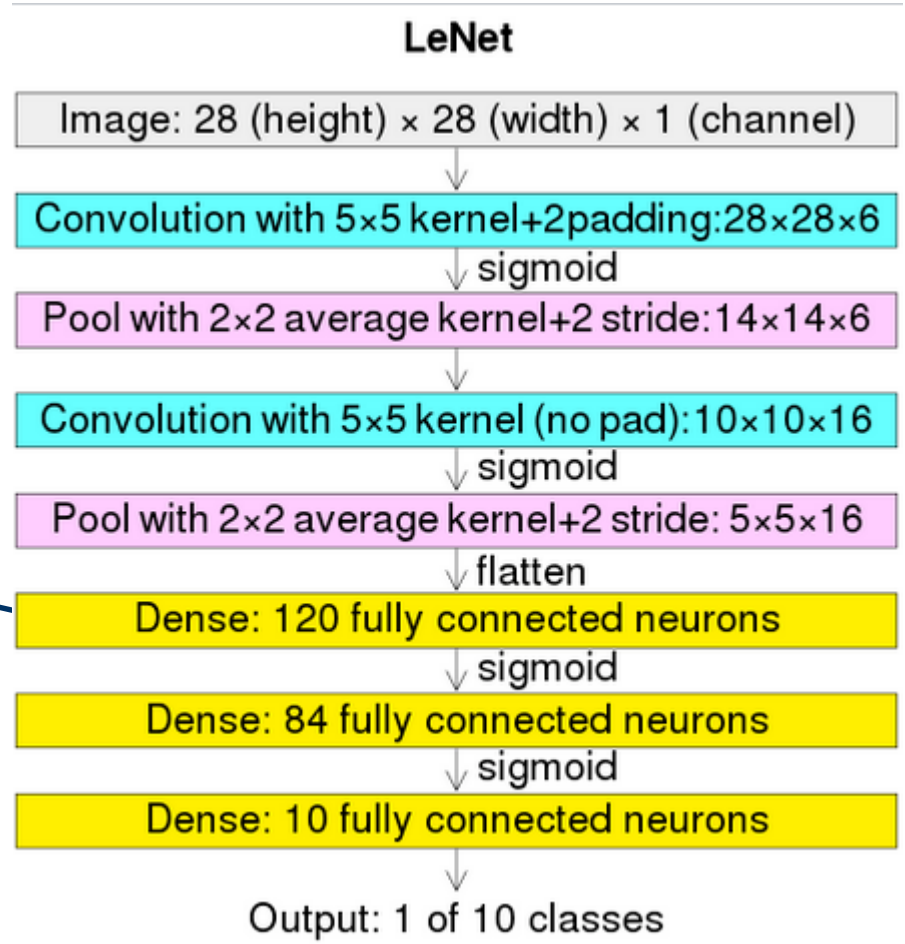


- Unfassbar großen Suchraum komplett durchsuchen
- Komplexes Molekülverhalten im Rechner nachbilden
  - **Diese.** Weil man nicht versteht, wie das eigentlich geht

Neuronale Netze: Warum funktioniert das so gut?

- Beispiel "LeNet", 1989: Erkennt hand-geschriebene Ziffern von echten Briefen
  - Netzwerkstruktur wie rechts: 7 hidden layers.
  - Frage: Was bedeutet die Ausgabe von Neuron 71 im Layer 5?
  - Antwort: Wir müssen es nicht wissen, das entwickelt sich beim Lernen von allein
    - Jedenfalls wenn wir genug Daten haben
    - **Wenn es gut klappt, spiegeln diese internen Features alle relevanten Eigenschaften der Daten wider**





Thesen (2)

- These "Robustheit":
  - KI-Entscheidungen sind robuster gegen Störungen als menschliche.
- These "Fairness":
  - Damit KI-Entscheidungen fair sind, muss und kann man für geeignete Trainingsdaten sorgen.
- These "Objektivität":
  - KI-Entscheidungen sind objektiv und deshalb ethisch günstiger als menschliche.
- These "Verzerrungen":
  - Verzerrungen bei KI-Entscheidungen kann man vermeiden, aber dazu sind organisatorische Maßnahmen nötig.

Wichtiges Phänomen der KI (vor allem ML-basierter): Emergenz

Herkömmliche Software (im Prinzip):

1. Spezifikation aufstellen
2. Korrekt implementieren

➔ Funktionale Eigenschaften bekannt

ML-basiertes System:

1. Daten der gewünschten Art sammeln
2. System damit trainieren

➔ Emergente Eigenschaften sind nur f. d. Trainingsbeispiele bekannt

Abweichungen in der Praxis:

- Statt genauer Spezifikation nur ungefähre Wünsche
- Implementierung enthält Defekte

➔ Funktionale Eig. nur ungefähr bekannt

- Aber immerhin meistens!
- Unerwünschte Technikfolgen halbwegs absehbar

Folgen in der Praxis:

1. Zerbrechlichkeit
2. Verletzungen der Fairness
3. diverse andere Verzerrungen

➔ Risiken unerwünschter Technikfolgen

(Folgen 2&3 gibt es auch bei such-basierter KI, sowie auch aus anderen Quellen)

Näheres über funktionale vs. nichtfunktionale (emergente) Eigenschaften siehe Modul "Softwaretechnik"

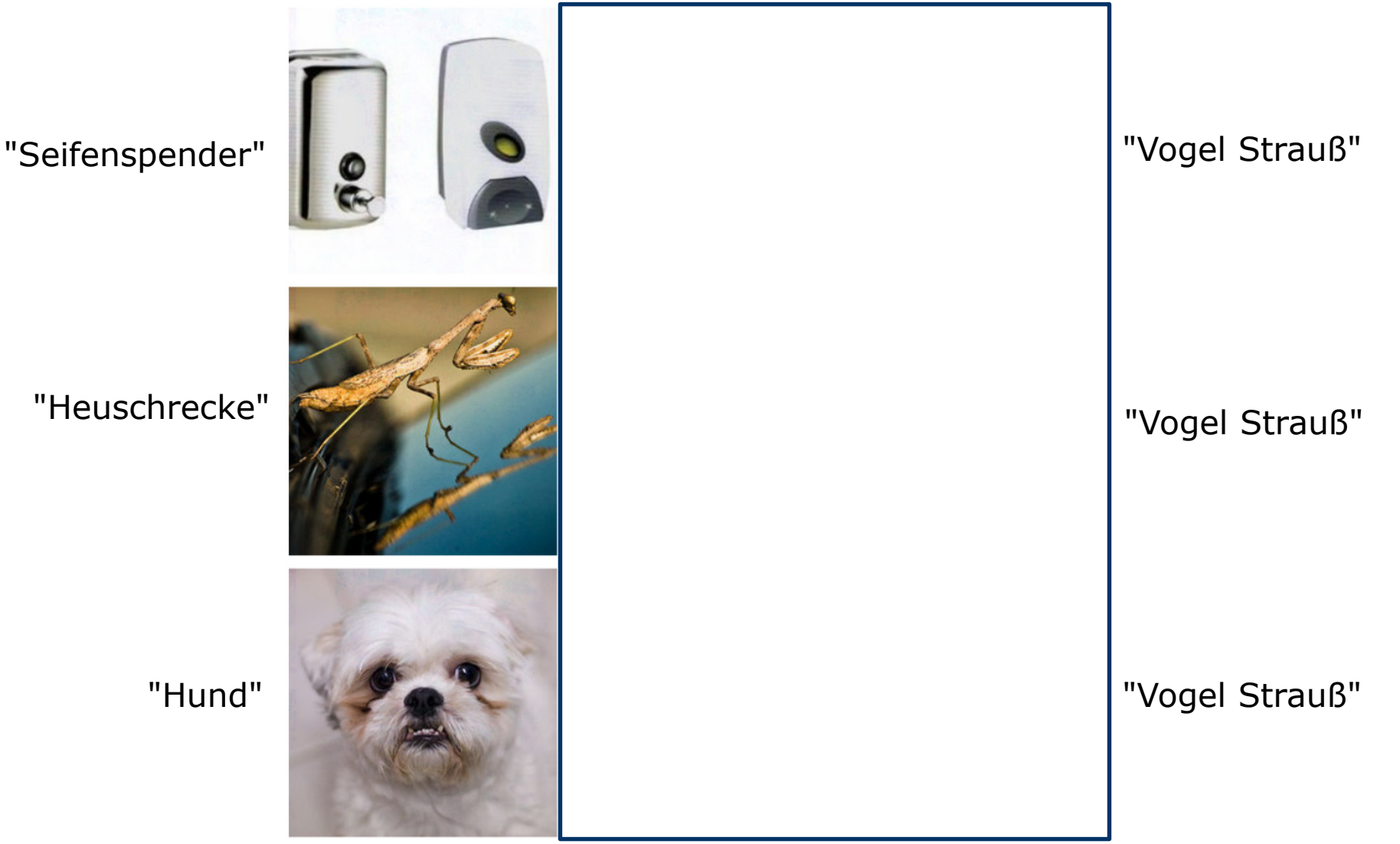
Problem 1: Zerbrechlichkeit (Nicht-Robustheit, brittleness): Feindliche Eingaben ("adversarial examples") bauen

- Beim Training eines NN optimiert man die Gewichte, um einen niedrigen Klassifikationsfehler zu erreichen
  - z.B. Objekte auf Bildern erkennen

Dann, mit dem fertigen Klassifikator:

- Rauschen zu einem Bild zufügen und so optimieren, dass eine *bestimmte* Fehlklassifikation passiert.
  - Das modifizierte Bild sieht für Menschen immer noch fast genau gleich aus.
    - Beispiele folgen
  - Die Klassifikation ist also "zerbrechlich" (nicht robust) angesichts kleiner Bildänderungen.
- Das geht grundsätzlich mit jedem Klassifikator-NN
- Mit modifizierten Trainingsalgorithmen kann man den Effekt abschwächen
  - beseitigen aber vermutlich nicht

Problem 1: Zerbrechlichkeit (brittleness): Feindliche Eingaben bauen [AlexNet mit 1000 Klassen]



Ist das als Technikfolge ein Problem?

Es braucht dafür ja ganz präzise dieses eine Bild.

Man kann nicht einen echten Seifenspender so manipulieren, dass er stets als Vogel Strauß erkannt wird.

Oder doch??

Szegedy et al.: Intriguing properties of neural networks
https://arxiv.org/abs/1312.6199

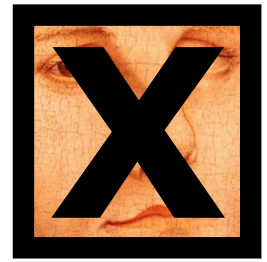
Problem 1: Zerbrechlichkeit (brittleness): Geht das auch im wirklichen Leben? [EykEvtFer18]

100% 73.33% 66.67% 100% 80% Täuschungsquote

wird erkannt als

Leider wohl ja.

Thesen (2)



- These "Robustheit":
  - KI-Entscheidungen sind robuster gegen Störungen als menschliche.
    - **Vielleicht manchmal, aber generell sicher nicht.**
- These "Fairness":
  - Damit KI-Entscheidungen fair sind, muss und kann man für geeignete Trainingsdaten sorgen.
- These "Objektivität":
  - KI-Entscheidungen sind objektiv und deshalb ethisch günstiger als menschliche.
- These "Verzerrungen":
  - Verzerrungen bei KI-Entscheidungen kann man vermeiden, aber dazu sind organisatorische Maßnahmen nötig.

Problem 2: Verletzung der Fairness
COMPAS [CPkrit][CP]

Freie Universität Berlin

- Klassifikationsverfahren machen unvermeidlich manchmal Fehler
- Lehrreiches Beispiel: COMPAS
  - Kommerzielle Software
  - "Correctional Offender Management Profiling for Alternative Sanctions"
    - Zweck: Maßnahmen für Straftäter auswählen
      - Eingesetzt in mehreren US-Bundesstaaten
    - Verwendet 137 Merkmale und bewertet damit ca. 20 "criminogenic needs", z.B. "criminal personality", "social isolation", "substance abuse", "residence/stability".
  - Berechnet auch 2-Jahres-Rückfall-Risiko
    - $CP_{10} \in 1..10$, $CP_3 \in \{niedrig,mittel,hoch\}$, $CP_2 \in \{nichthoch,hoch\}$
    - für Gewaltverbrechen $CP_{k,gew}$
    - für Straftaten insgesamt $CP_{k,allg}$

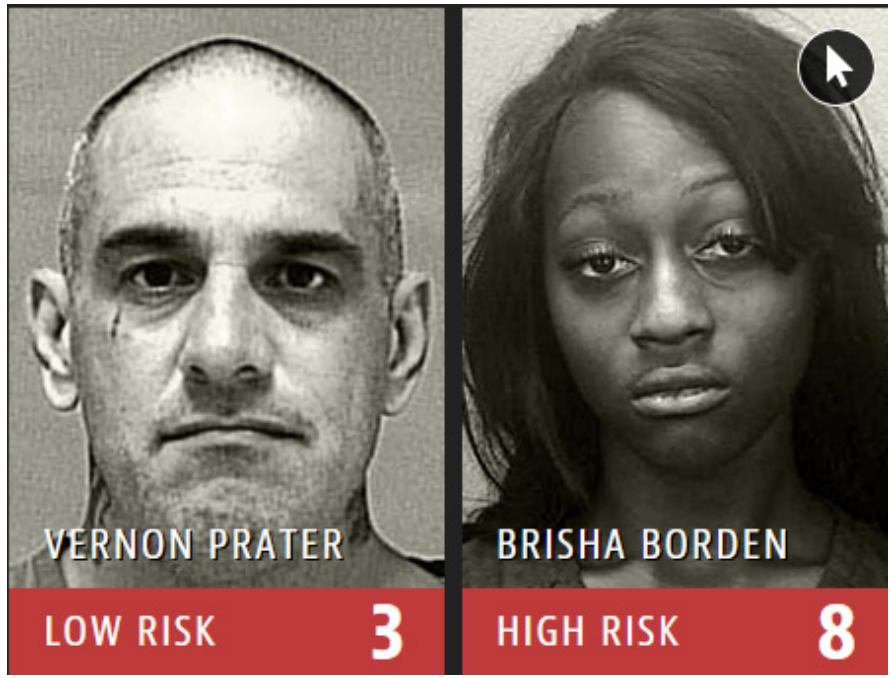
- Brisha Borden fuhr aus Spaß spontan mit einem fremden Kinderroller los
  - ließ ihn fallen, als die Mutter ihr hinterherrief
  - Vorstrafen: mehrere Vergehen als Jugendliche
  - Rückfallrisiko für Straftat lt. COMPAS:

    $CP_{10,allg} = 8$ $CP_{3,allg} =$ **hoch**
- Vernon Prater stahl in einem Baumarkt Werkzeug im Wert von 86 Dollar
  - Vorstrafen: 5 J. Freiheitsstrafe für 2x bewaffneten Raubüberfall und 1x versuchten bewaffneten Raubüberfall
  - Rückfallrisiko für Straftat lt. COMPAS:

    $CP_{10,allg} = 3$ $CP_{3,allg} =$ **niedrig**

Problem 2: Verletzung der Fairness
COMPAS (2)

- Tatsächlich war das Ergebnis umgekehrt:
  - Borden wurde nicht rückfällig
  - Prater verurteilt zu 8 J. Freiheitsstrafe für Einbruchdiebstahl (Elektronik-Lagerhaus)
- Einzelfall? Leider nein, sondern eine Tendenz (laut [CPkrit]):
  - $CP_{2,allg}$ = hoch ➔ 61% wirklich rückfällig
  - $CP_{2,gew}$ = hoch ➔ 20% wirklich rückfällig
    - COMPAS ist also pessimistisch
  - Schwarze wurden dabei öfter zu negativ eingeschätzt:
    - Fälschlich "hohes" Risiko ist bei ihnen doppelt so häufig wie bei Weißen
      - Basis: 7000 Fälle aus Florida

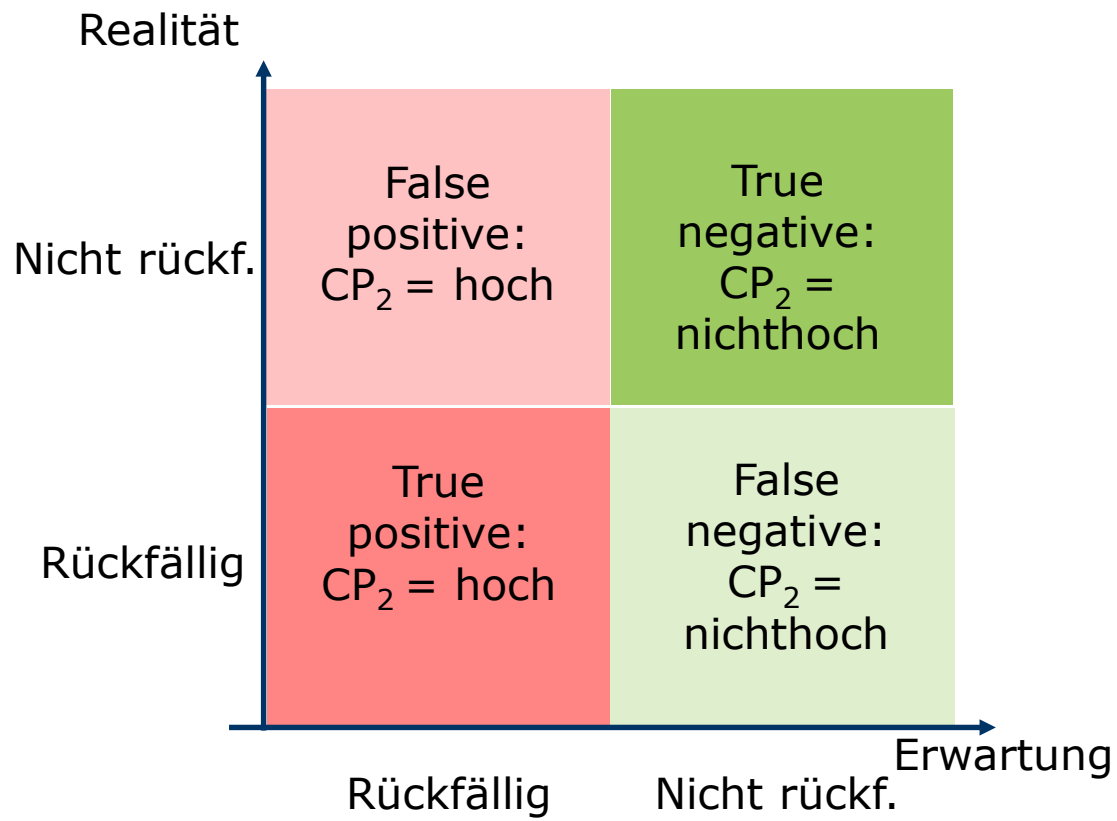


- Rechnet man das Vorstrafenregister heraus, nimmt der Effekt ab, geht aber nicht weg:
  - Fälschlich "hohes" Risiko bei Schwarzen immer noch 77% höher als bei Weißen

Exkurs: Terminologie der binären Klassifikation

- Ziel: Klasse X erkennen
  - COMPAS: X = "wird rückfällig"
    - diese Klassifikation wird "positiv" genannt.

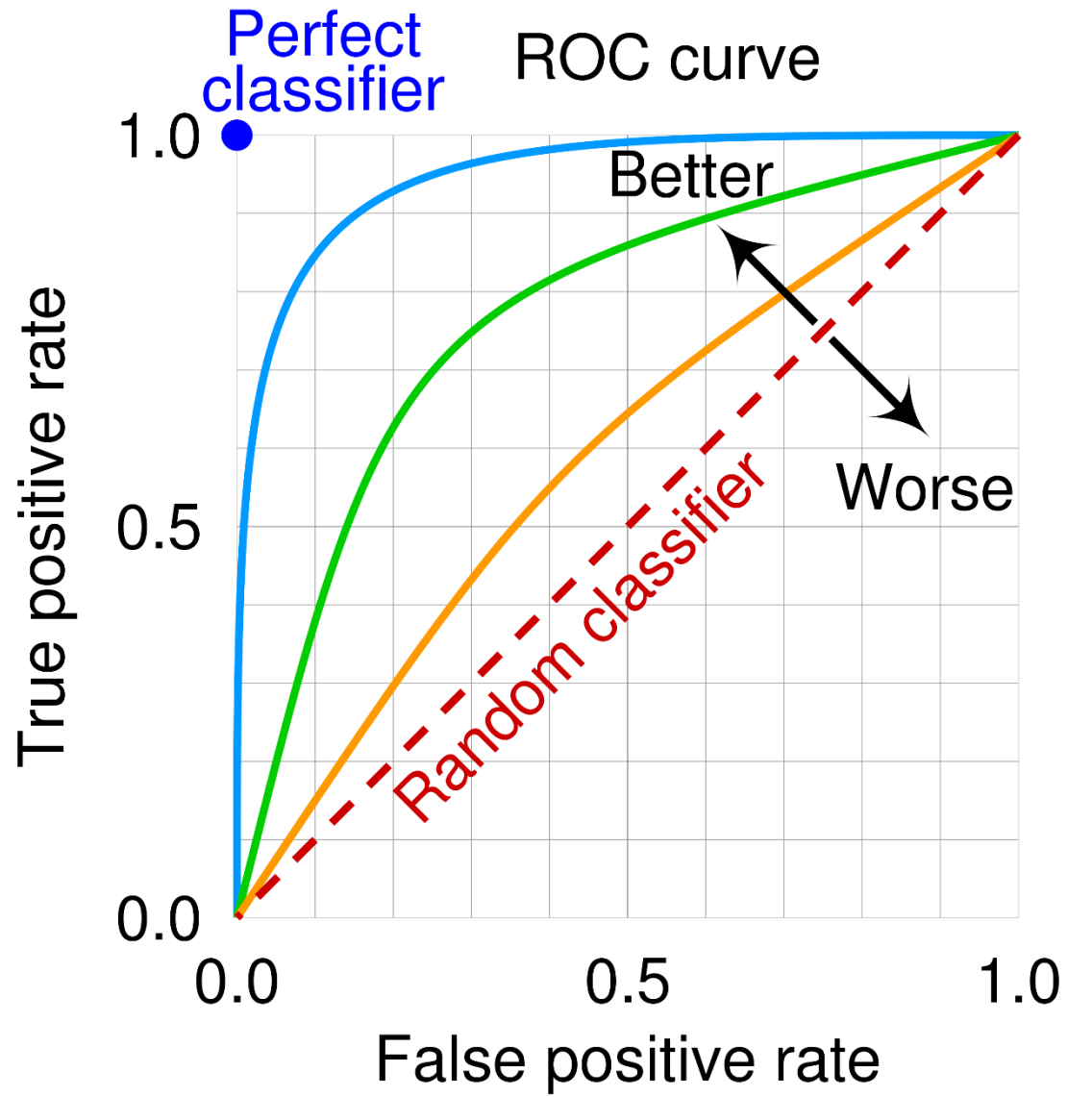
- Binäre Klassifikationen fallen in 4 Gruppen:
  - positiv oder negativ vorausgesagt
    kombiniert mit
    korrekt (true) oder falsch (false) vorausgesagt.
- Deren Häufigkeiten heißen True Positive Rate (TPR), etc.
  - TPR + FNR = TNR + FPR = 100%
    - Diese vier Werte beschreiben summarisch die Qualität des Klassifikators



https://de.wikipedia.org/wiki/Beurteilung_eines_bin%C3%A4ren_Klassifikators

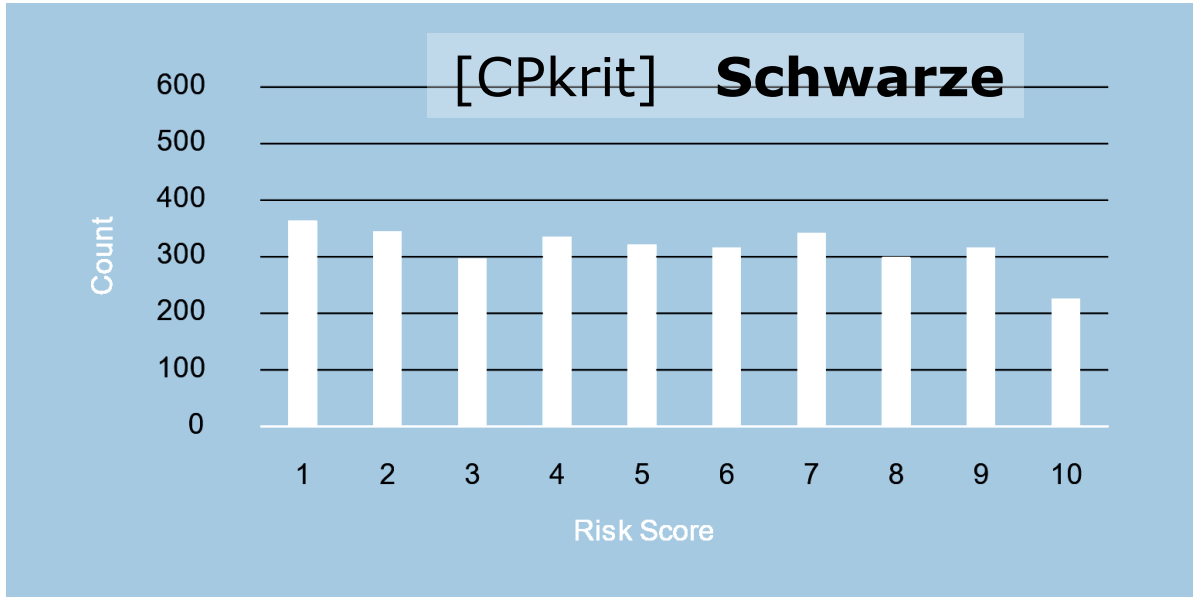
Problem 2: Verletzung der Fairness
COMPAS (3)

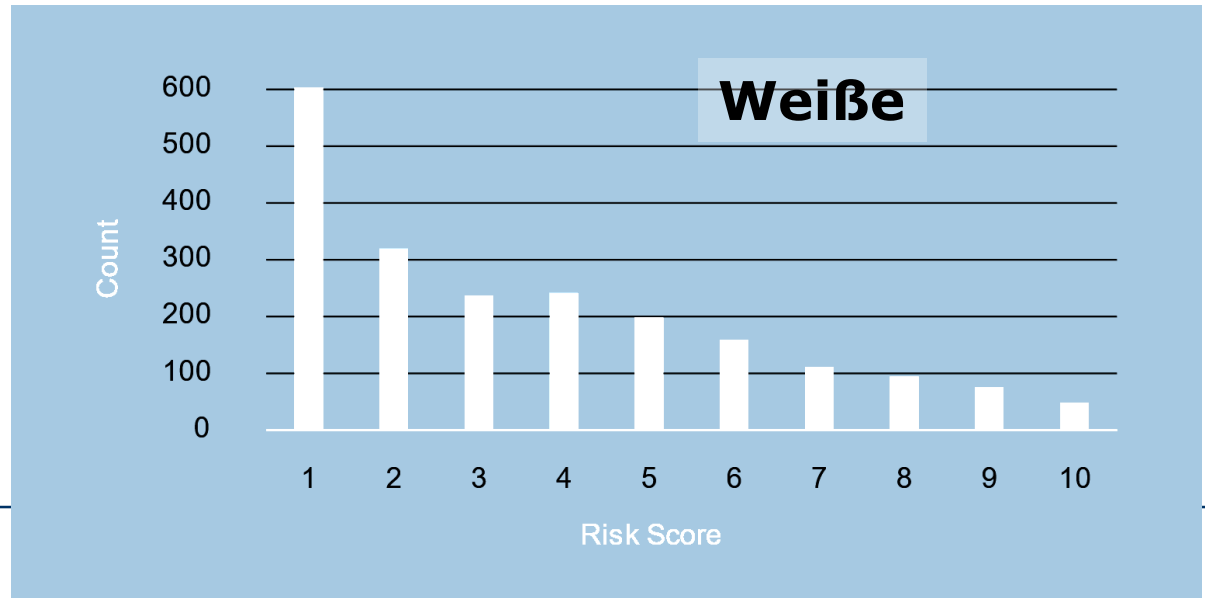
- Was steckt hinter der Benachteiligung von Schwarzen?
  - Das COMPAS-Modell wurde von einem Statistik-Professor entwickelt
    - (es ist wohl kein Neuronales Netz)
    - Das Handbuch [CP] begründet viele Entscheidungen mit dem Stand der kriminologischen Forschung
  - Hautfarbe kommt unter den 137 Merkmalen nicht vor
    - Geschlecht auch nicht
  - Die Vorhersage-Qualität wurde separat für verschiedene Untergruppen untersucht: Men, Women, White, Black, Hispanic, All
    - lag für $CP_{2,allg}$ laut [CP] immer so um AUC ~0.7
      - AUC: *Area Under the Curve* (für ROC-Kurve)



Problem 2: Verletzung der Fairness
COMPAS (4)

- Kann das beides zugleich sein?:
  1. AUC ähnlich für Weiße und Schwarze
  2. Schwarze doppelt so oft falsch positiv "high risk" wie Weiße
- Ja, kann es:
  - Dem Pessimismus bei den Schwarzen kann ein Optimismus bei den Weißen gegenüberstehen
  - Beides könnte die AUC etwa gleich stark beschränkt haben.
    - Aber bei "high" war es relevanter
- Und so scheint es bei COMPAS zu sein*:
  - Häufigkeit der Risiken von 1 (links) bis 10 (rechts)
- **COMPAS ist nicht fair**





*bei Weißen sind die Risikoschätzungen tatsächlich zu niedrig

Exkurs: Asymmetrische Fehlerkosten

- Für den Zweck von COMPAS sind falsch-positive Ergebnisse viel schlimmer als falsch-negative
  - denn es sollte gelten: Im Zweifel für den Angeklagten
    - Das folgt aus den Menschenrechten
- Das ist bei vielen Klassifikationsaufgaben ähnlich
  - evtl. andersherum
- Der AUC-Wert ignoriert diese Asymmetrie
  - Also "Augen auf beim Maßekauf!"

[MehMorSax21] sammelt 10 Definitionen:

1. Equalized Odds
2. Equal Opportunity
3. Demographic Parity
4. Fairness through Awareness
5. Fairness through Unawareness
6. Treatment Equality
7. Test Fairness
8. Counterfactual Fairness
9. Fairness in Relational Domains
10. Conditional Statistical Parity

- Nummer 5 bedeutet:
  "Die Hautfarbe darf im Modell nicht explizit benutzt werden"
  - Das wird von COMPAS erfüllt!
  - COMPAS ist 5-fair

Problem 2: Verletzung der Fairness
Warum diskriminiert COMPAS Schwarze?

- Wie kommt es zu der Verzerrung, wenn gar kein Hautfarbe-Merkmal vorhanden ist?

- Es gibt viele Merkmale, die stark mit Schwarzsein (in den USA) korrelieren:

  Themenkreise der 138 Merkmale sind
  - Kriminalität in der Familie (8 Merkmale)
  - Kriminalität im Freundeskreis (5)
  - Stabilität des Lebensumfelds (11)
  - Kriminalität im Lebensumfeld (6)
  - Schullaufbahn (9)
  - etc.

- Grund für die Wirksamkeit dieser Merkmale ist vermutlich die starke Segregation der Hautfarben zwischen Wohnvierteln in den USA

[MehMorSax21] sammelt 10 Definitionen:

1. Equalized Odds
2. Equal Opportunity
3. Demographic Parity
4. Fairness through Awareness
5. Fairness through Unawareness

Angenommen, wir hätten

- einen perfekten COMPAS-Klassifikator
  - also FPR = FNR = 0
- Und die Gruppen A, B würden rückfällig zu 80%, 90% also
  - TPR(A) = 80% TNR(A) = 20%
  - TPR(B) = 90% TNR(B) = 10%

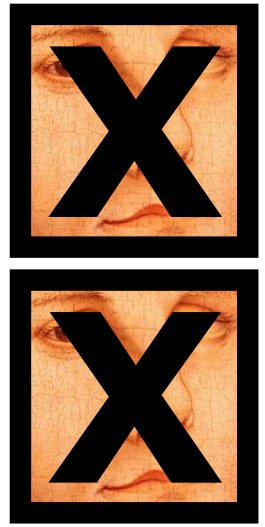
- 2. Equal Opportunity:
  "Schwarze und Weiße, die nicht rückfällig werden, haben gleiche Chancen auf einen günstigen Ausgang"
  - hier: beide 100%
  - Der perfekte Klassifikator ist 2-fair

- 3. Demographic Parity:
  "Schwarze und Weiße haben gleiche Chancen auf einen günstigen Ausgang"
  - hier jedoch: einmal 10%, einmal 20%
  - Der perfekte Klassifikator ist nicht 3-fair, er verletzt die demografische Parität

➔ nicht jeder Fairnessbegriff ist in einem gegebenen Fall auch sinnvoll!

1. Es gibt zahlreiche präzise Definitionen von Fairness
2. Manche davon widersprechen sich in einem gegebenen Fall
3. Selbst ein perfekter Klassifikator hält sie nicht alle ein
4. Selbst ein hochgradig nicht fairer Klassifikator hält u.U. manche davon ein
5. Manche sind in einem gegebenen Fall gar nicht sinnvoll,
   in anderen Fällen wären sie es aber sehr wohl

Thesen (2)



- These "Robustheit":
  - KI-Entscheidungen sind robuster gegen Störungen als menschliche.
- These "Fairness":
  - Damit KI-Entscheidungen fair sind, muss und kann man für geeignete Trainingsdaten sorgen.
    - **Fairness ist nicht allgemein lösbar, weil man über die Definition streiten kann**
- These "Objektivität":
  - KI-Entscheidungen sind objektiv und deshalb ethisch günstiger als menschliche.
- These "Verzerrungen":
  - Verzerrungen bei KI-Entscheidungen kann man vermeiden, aber dazu sind organisatorische Maßnahmen nötig.

Exkurs: Objektivität

- Wikipedia: "**Objektivität** […] bezeichnet die Unabhängigkeit der Beurteilung oder Beschreibung einer Sache, eines Ereignisses oder eines Sachverhalts vom Beobachter"

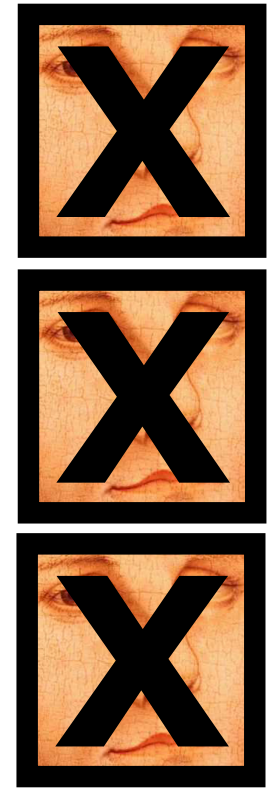
- Jede Beurteilung, die vollautomatisch von einem Computer vorgenommen wird, ist also objektiv.

- **Die übliche Annahme "objektiv = gut" ist Unfug**

- Gedankenspiel: Wir bauen COMPAS2
  - Wir verwenden nur eines der 138 Merkmale von COMPAS zur Klassifikation:
    - Nr 67: *"In your neighborhood, have some of your friends or family been crime victims?"*
  - COMPAS2 funktioniert so:
    - Ja → High risk
    - Nein → Low risk
  - COMPAS2 ist objektiv, aber die Einschätzungen werden sehr oft falsch sein
- Was wir uns tatsächlich wünschen:
  - Die Klassifikation soll "valide" (gültig) sein
  - Was das jeweils bedeutet, ist aber meist nicht objektiv zu klären
  - Auch bei COMPAS nicht. Warum?

https://de.wikipedia.org/wiki/Objektivität

Thesen (2)



- These "Robustheit":
  - KI-Entscheidungen sind robuster gegen Störungen als menschliche.
- These "Fairness":
  - Damit KI-Entscheidungen fair sind, muss und kann man für geeignete Trainingsdaten sorgen.
- These "Objektivität":
  - KI-Entscheidungen sind objektiv und deshalb ethisch günstiger als menschliche.
    - **Objektiv hat mit "ethisch günstig" wenig zu tun!**
- These "Verzerrungen":
  - Verzerrungen bei KI-Entscheidungen kann man vermeiden, aber dazu sind organisatorische Maßnahmen nötig.

Problem 3: Verzerrungen (bias)

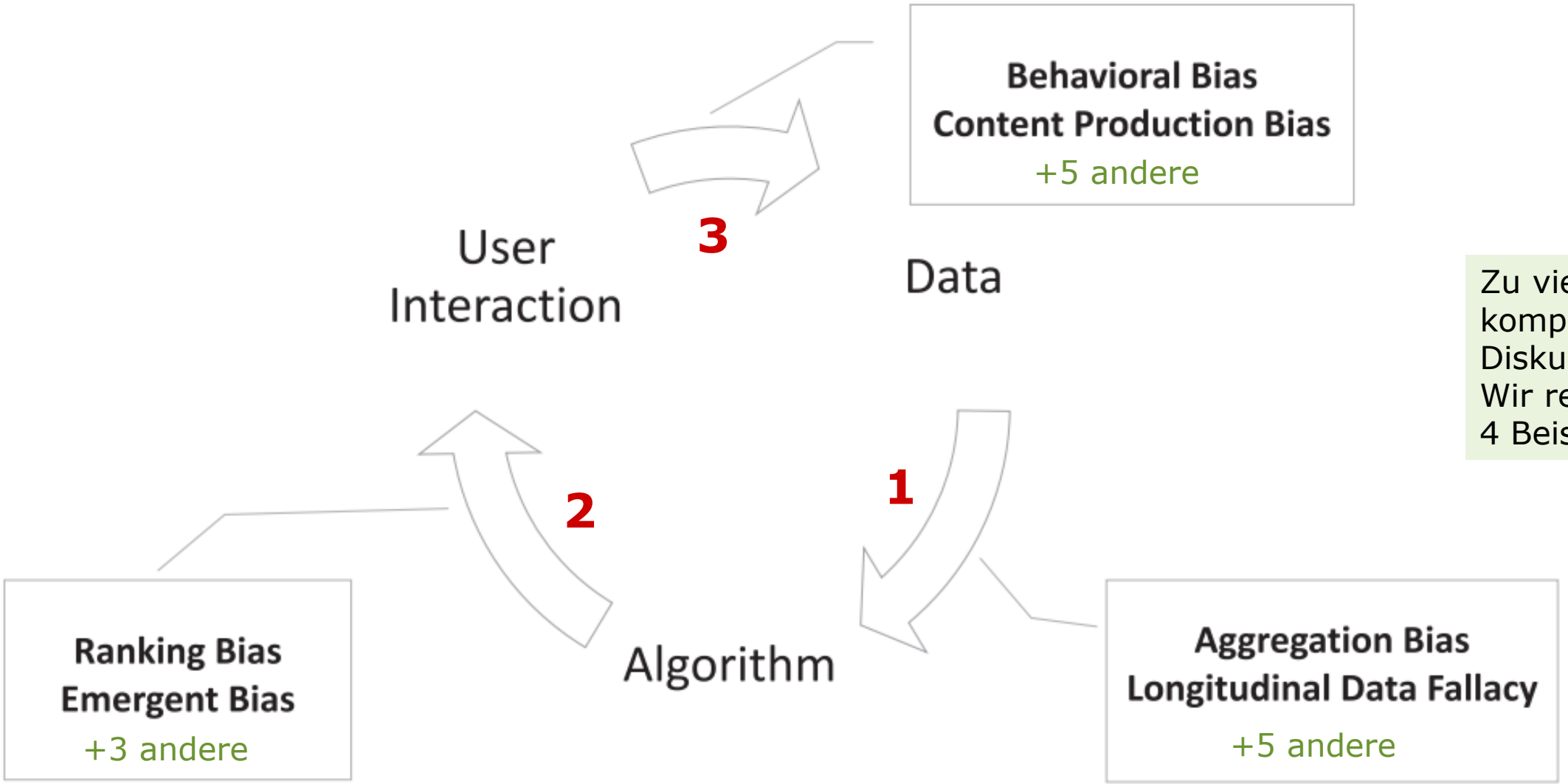
- Verletzungen der Fairness sind eine Form von "bias"
  - es gibt aber noch viele andere
- "bias" heißt auf Deutsch "Neigung"
  - gemeint ist eine *unerwünschte* Neigung
    - dafür ist "Verzerrung" klarer
- Neigung/Verzerrung impliziert die Existenz einer idealen Lösung
  - wir haben oben für Fairness gesehen, dass es die nicht unbedingt gibt
  - Also bitte Vorsicht mit dem Begriff!
- Es folgen zwei Listen von Arten von Verzerrungen aus der Literatur

1. Vor-existierende Verzerrung:
   - Das System spiegelt Verzerrungen wider, die schon vor ihm existiert haben (in Institutionen, Praktiken, Haltungen)

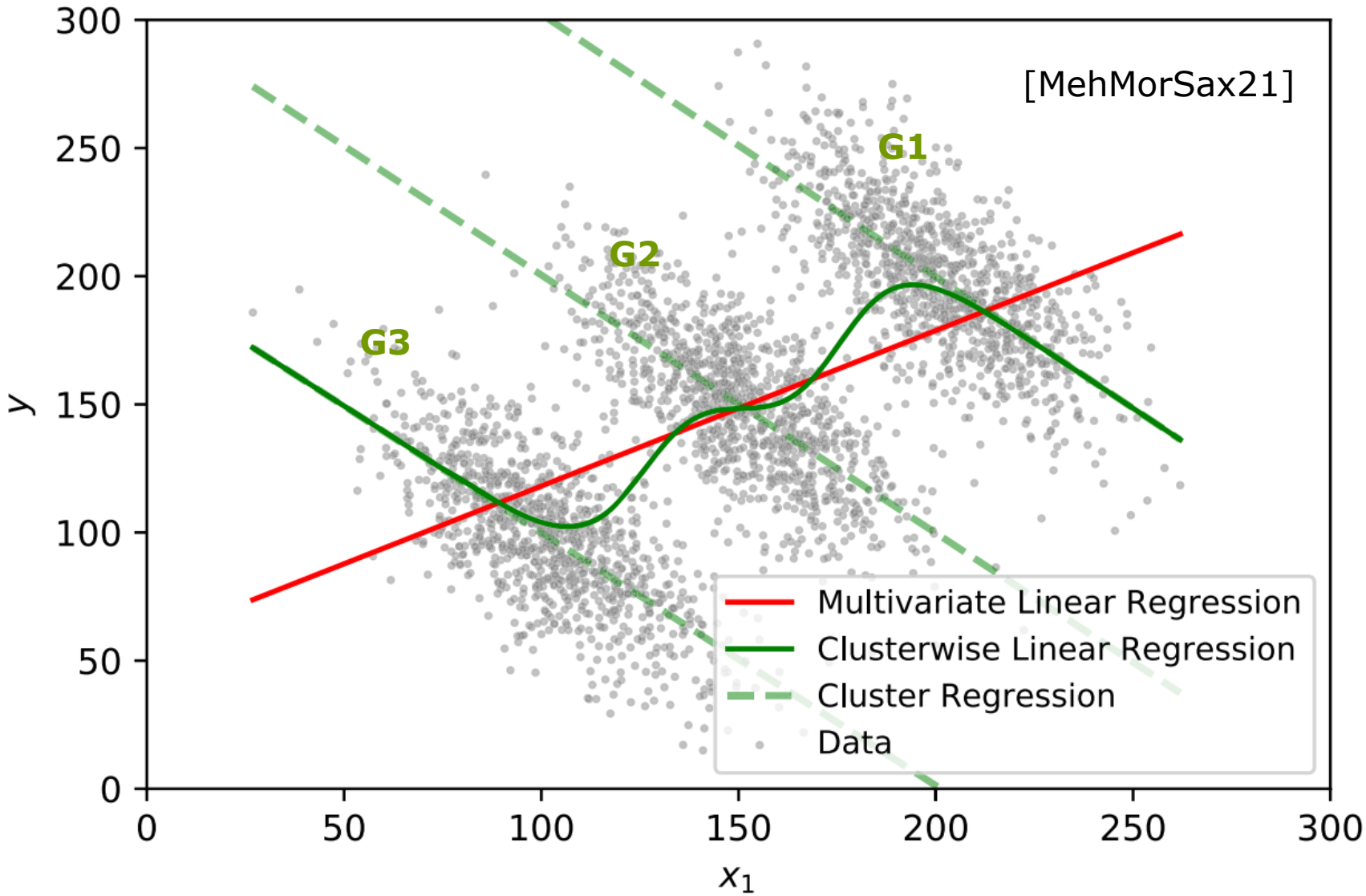
2. Technisch bedingte Verzerrung
   - Fehlende Randomisierung
     - z.B. gleichrangige Suchtreffer alphabetisch anstatt randomisiert auflisten
   - Falsche vereinfachende Annahmen
     - z.B. unpassende, feste Anwortkategorien

3. Emergente Verzerrung
   - entsteht aus dem Nutzungskontext
   - z.B. System ist nicht passend für die Benutzer_innengruppe gestaltet
     - z.B. keine Vorkehrungen für Blinde

Problem 3: Verzerrungen (bias)
Kategorien für ML-Systeme nach [MehMorSax21]



Problem 3: Verzerrungen (bias)
Beispiel für Aggregation Bias: Simpsons Paradoxon



Der globale Trend in Daten kann ein ganz anderer sein als der (immer gleiche) Trend in jeder Untergruppe.

Wer die Untergruppen ignoriert, bekommt Unsinn heraus.

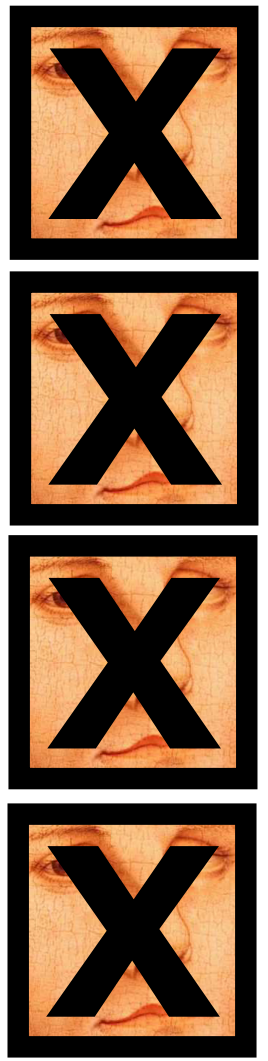
1. Data-to-Algorithm:
   - Die Art, wie Daten beim Lernen benutzt werden, kann Verzerrungen verursachen
   - siehe Simpsons Paradoxon eben

2. Algorithm-to-User:
   - Die Art, wie die KI mit Benutzer_innen interagiert, kann Verzerrungen verursachen
   - Ranking Bias und Emergent Bias sind uns von [FriNis96] schon bekannt

3. User-to-Data:
   - Die Art, wie neue Daten aus Interaktionen gewonnen werden, kann Verzerrungen verursachen
   - Beispiel: Reaktion auf ein Emoji kann stark davon abhängen, wie es dargestellt wurde:
     - [MilTheCha16]
     - Unicode-Zeichen U+1F601 ("grinning face with smiling eyes")
     - Googles Font:
       - Wahrnehmung: "superglücklich"
     - Apples Font:
       - Wahrnehmung: "kampfbereit"

Thesen (2)

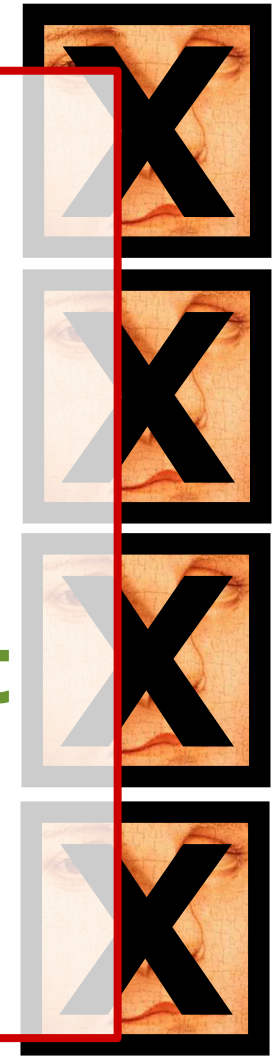
- These "Robustheit":
  - KI-Entscheidungen sind robuster gegen Störungen als menschliche.
- These "Fairness":
  - Damit KI-Entscheidungen fair sind, muss und kann man für geeignete Trainingsdaten sorgen.
- These "Objektivität":
  - KI-Entscheidungen sind objektiv und deshalb ethisch günstiger als menschliche.
- These "Verzerrungen":
  - Verzerrungen bei KI-Entscheidungen kann man vermeiden, aber dazu sind organisatorische Maßnahmen nötig.
    - **Es gibt zu viele Quellen von Verzerrungen, um sie alle vermeiden zu können**



- These "Robustheit":
  - KI-Entscheidungen sind robuster gegen Störungen als menschliche.
- These "Fairness":
  - Damit KI-Entscheidungen fair sind, muss und kann man für geeignete Trainingsdaten sorgen.
- These "Objektivität":
  - KI-Entscheidungen sind objektiv und deshalb ethisch günstiger als menschliche.
- These "Verzerrungen":
  - Verzerrungen bei KI-Entscheidungen kann man vermeiden, aber dazu sind organisatorische Maßnahmen nötig.

KI hat viel (und vielfältiges) Potential für problematische Technikfolgen

Viele davon hängen gar nicht direkt an der KI, sondern werden uns nur durch die KI stärker bewusst.



Literatur

- [BarHarNar22] Solon Barocas, Moritz Hardt, Arvind Narayanan:
"Fairness and Machine Learning: Limitations and Opportunities", https://fairmlbook.org
- Christoph Benzmüller:
"Symbolic AI and Gödel's Ontological Argument", 2022
https://doi.org/10.1111/zygo.12830
- [CP] Northpointe Inc:
"Practitioner's Guide to COMPAS Core",
https://www.documentcloud.org/documents/2840784-Practitioner-s-Guide-to-COMPAS-Core.html
- [CPkrit] Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner:
"Machine Bias", 2016,
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- [EykEvtFer18] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes et al.:
"Robust physical-world attacks on deep learning visual classification",
Proc. Conf. on Computer Vision and Pattern Recognition. 2018,
https://tinyurl.com/rpwadlvc18
- [FriNis96]: Batya Friedman, Helen Nissenbaum:
"Bias in Computer Systems",
ACM Trans. on Information Systems 14(3), 1996,
https://doi.org/10.1145/230538.230561

Literatur (2)

- [MehMorSax21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan:
"A Survey on Bias and Fairness in Machine Learning",
ACM Computing Surveys 54(6), 2021,
https://doi.org/10.1145/3457607
- [MilTheCha16] Hanna Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, Brent Hecht:
"'Blissfully happy' or 'ready to fight': Varying interpretations of emoji"
Proc. AAAI Conf. on Web and Social Media 10(1), 2016,
https://ojs.aaai.org/index.php/ICWSM/article/download/14757/14606
- [RusNor16] Stuart Russell, Peter Norvig:
Artificial Intelligence: a Modern Approach",
Pearson Education 2016
- John Searle: "Minds, Brains, and Programs". Behavioral and Brain Sciences 3(3), 1980,
https://doi.org/10.1017/S0140525X00005756
- Alan Turing: "Computing Machinery and Intelligence", Mind LIX(236), 1950,
https://www.abelard.org/turpap/turpap.php

Quellen der Abbildungen

- The Imitation Game: https://en.wikipedia.org/wiki/Turing_test
- Chinese Room: https://medium.com/acing-ai/what-is-the-chinese-room-argument-in-artificial-intelligence-d914abd02601
- Reversi: https://playpager.com/wp-content/uploads/2018/11/howto1.jpg
- Lilien-Arten: https://onesixx.com/data-iris/
- Lilienblüte, annotiert: https://www.integratedots.com/determine-number-of-iris-species-with-k-means/
- Iris-Klassifikator-NN: https://medcraveonline.com/BBIJ/neural-networks-for-classification-and-regression.html
- tanh-Plot: http://www.efunda.com/math/hyperbolic/images/tanh_plot.gif
- Aminosäuren: https://en.wikipedia.org/wiki/Amino_acid
- Aminosäure-Kette: https://en.wikipedia.org/wiki/Protein_primary_structure
- alpha-Helix, beta-Sheet: https://en.wikipedia.org/wiki/Protein_secondary_structure
- Tertiärstruktur: https://en.wikipedia.org/wiki/Protein_tertiary_structure
- Hämoglobin: https://en.wikipedia.org/wiki/Hemoglobin

Quellen der Abbildungen (2)

- LeNet: https://en.wikipedia.org/wiki/LeNet
- LeNet-Ziffern: http://yann.lecun.com/exdb/publis/pdf/lecun-89e.pdf
- Speed Limit 45: http://www.streetsignusa.com/shop/item.aspx?itemid=474
- ROC-Kurve: https://en.wikipedia.org/wiki/Receiver_operating_characteristic
- COMPAS-Porträts, COMPAS-Risikohäufigkeiten: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Emojis: [MilTheCha16]

Danke!