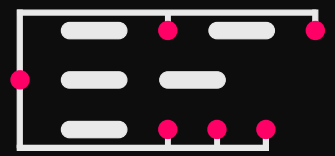
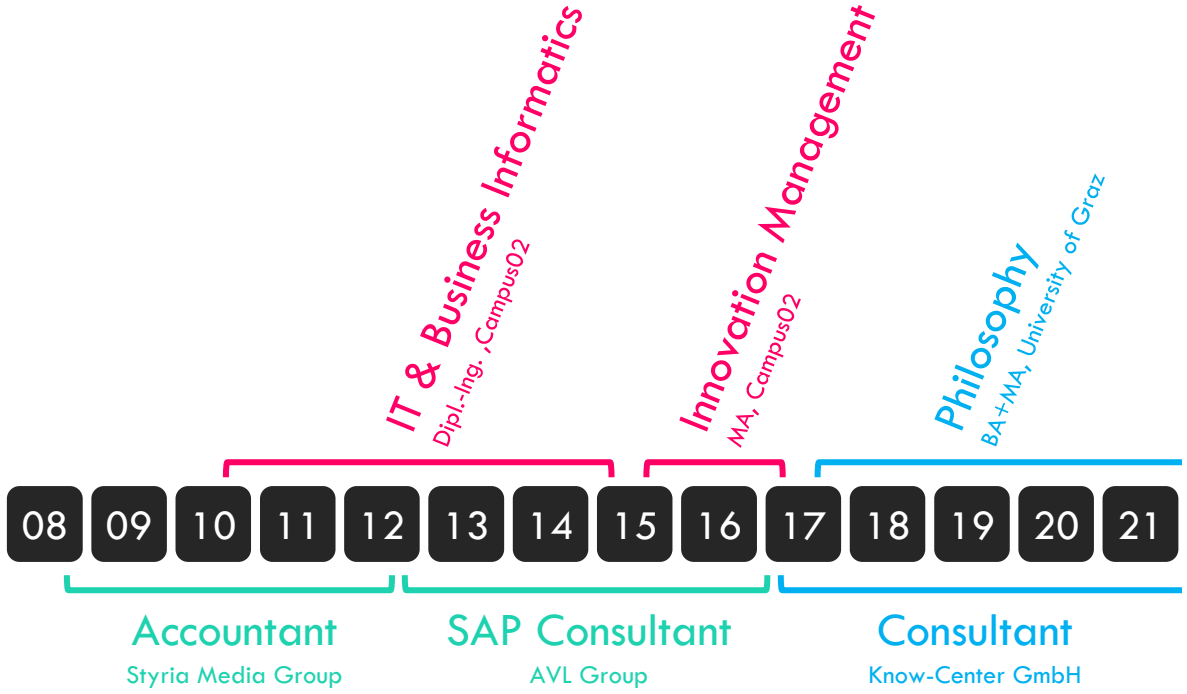
Nachhaltige KI

eine technische Perspektive



14.10.2022, Christof Wolf-Brenner

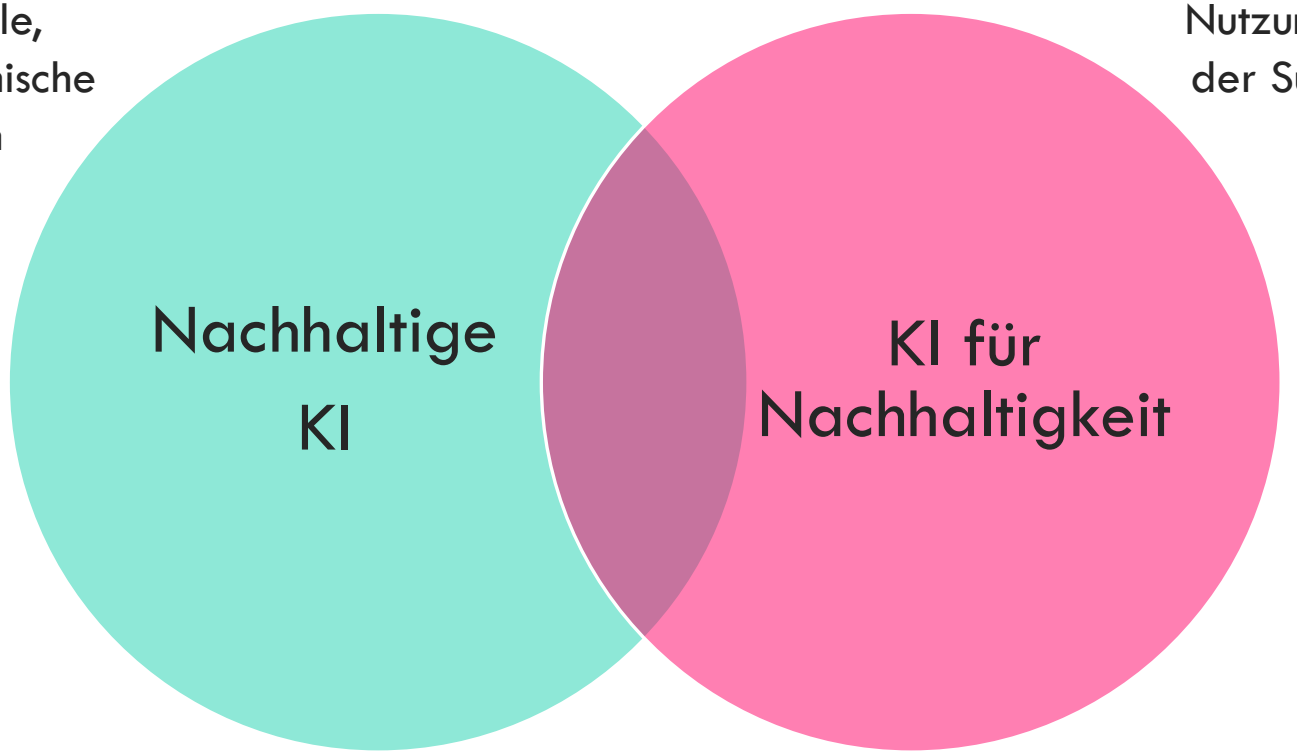


Christof Wolf-Brenner

brennechr@gmail.com

NACHHALTIGE KI ODER KI FÜR NACHHALTIGKEIT?

Verantwortung für soziale, ökologische und ökonomische Auswirkungen durch den Einsatz von KI



Nutzung von KI zur Erreichung der Sustainable Development Goals (AI4SDG)

Quelle: Rohde et al (2021):Nachhaltigkeitskriterien für künstliche Intelligenz. Entwicklung eines Kriterien- und Indikatorensets für die Nachhaltigkeitsbewertung von KI-Systemen entlang des Lebenszyklus, Schriftenreihe des IÖW 220/2021, Berlin

KÜNSTLICHE INTELLIGENZ: HEUTE UND MORGEN

Starke KI, Artificial General Intelligence

- Maschinen sind intellektuell on par mit Menschen

20xx

2020

Schwache KI, angewandte KI

- Logik- und musterbasiert/datengetrieben AI
- Ahmt intelligentes Verhalten nach
- Assistiert Menschen bei relativ einfachen Aufgaben

20xx-21xx

Artificial Superintelligence

- Maschinen übertreffen Menschen

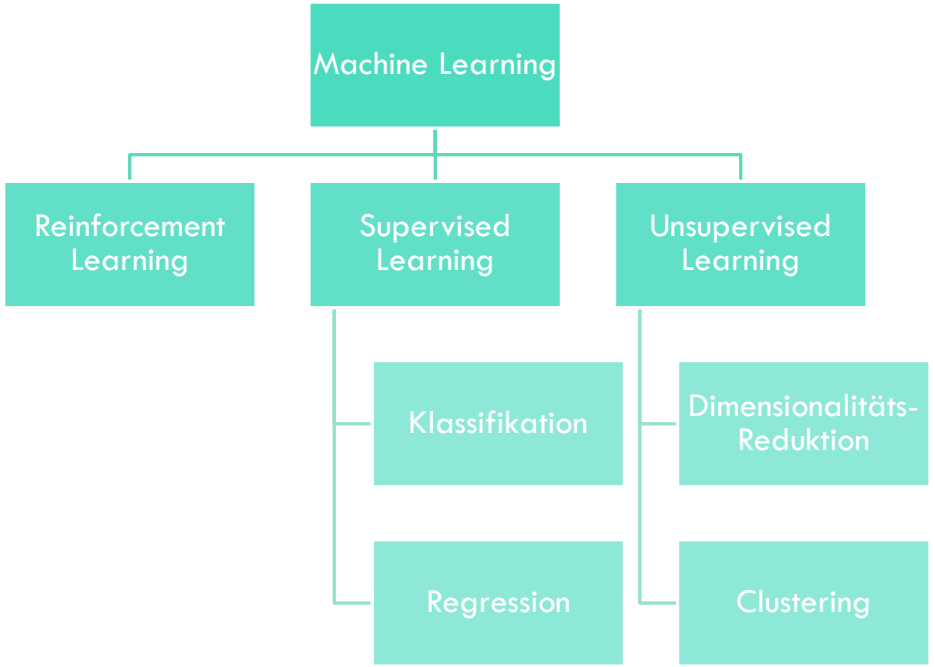
KI GENERATIONEN

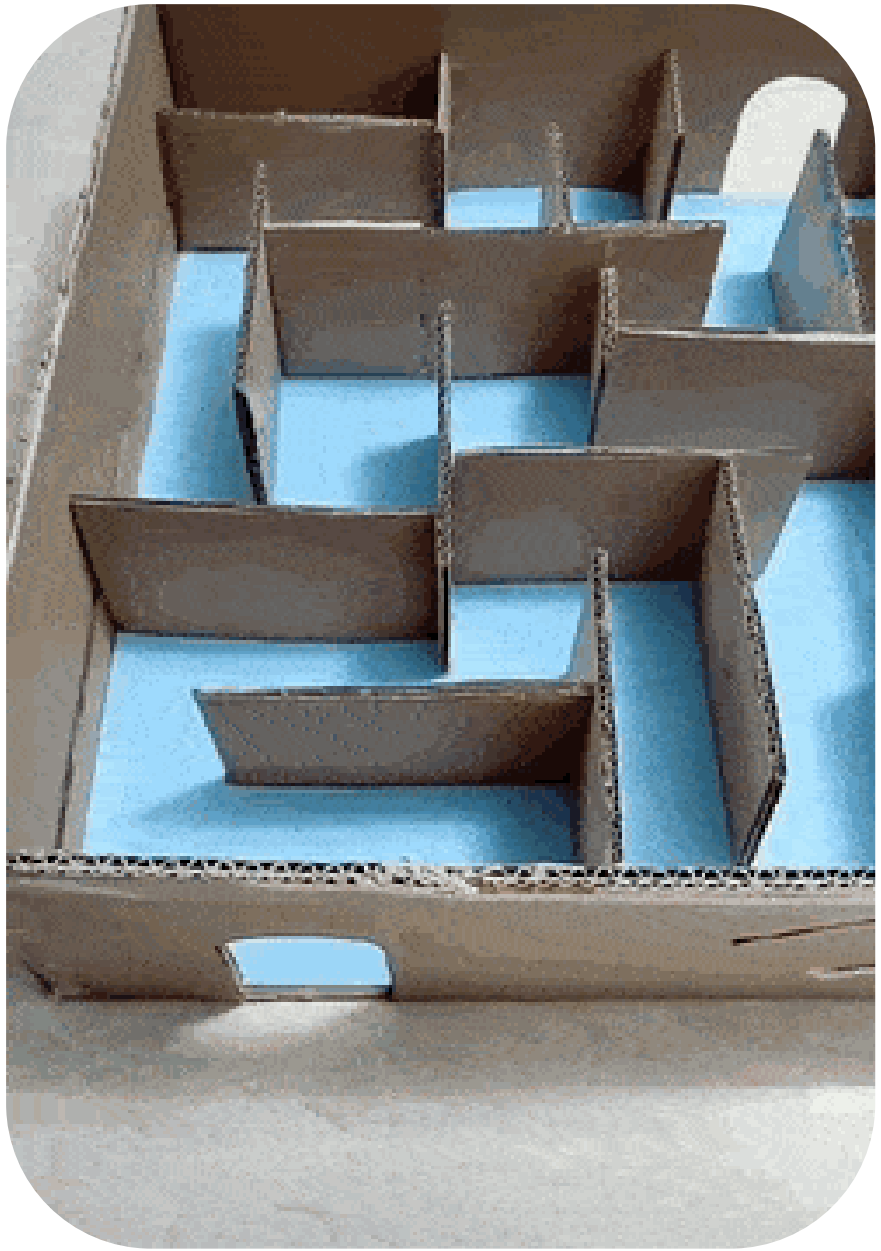
GENERATION 1 (1980ER)

EXPERTENSYSTEME

GENERATION 2 (AB ETWA 2006)

MASCHINELLE LERNSYSTEME





TECHNOLOGISCHER LAMPENGEIST

I. WORTWÖRTLICHKEIT

II. HYPEREFFIZIENT

DREI AKTUELLE PROBLEME

Biases

#1

Black Boxes

#2

Privacy

#3

/ˈbʌɪəs/

(Vorurteil, Voreingenommenheit; frz.: schief, schräg)

Implizite oder explizite Bevorzugung ohne sachliche Grundlage oder nicht auf der Grundlage eines fairen Urteils, bspw. von

- einer Gruppe von Menschen oder
- einer Seite in einem Streit oder
- einer Sache gegenüber einer anderen

→ Voreingenommenheit ist nicht automatisch falsch oder schlecht.

→ Sich nicht bewusst zu sein, dass man voreingenommen ist, schon.

BIAS IN DATEN UND ALGORITHMEN

TRAINING DATA

Human Bias in Data

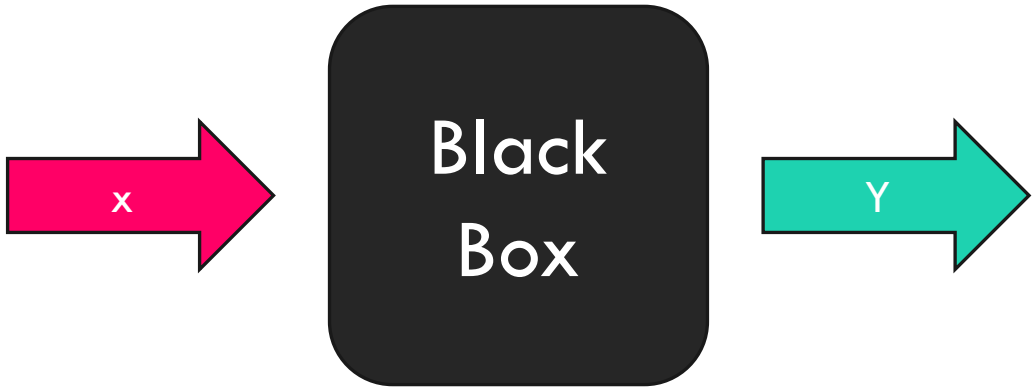
- Reporting Bias
- Overgeneralization
- Halo Effect
- Prejudice
- …

Human Biases in Data Collection

- In-group bias
- Confirmation Bias
- Anecdotal Fallacy
- Experimenter‘s Bias
- …

BLACK BOX ALGORITHMEN

Manchmal lässt sich nicht feststellen, wie Algorithmen bei einem bestimmten Input zu einem bestimmten Output kommen.

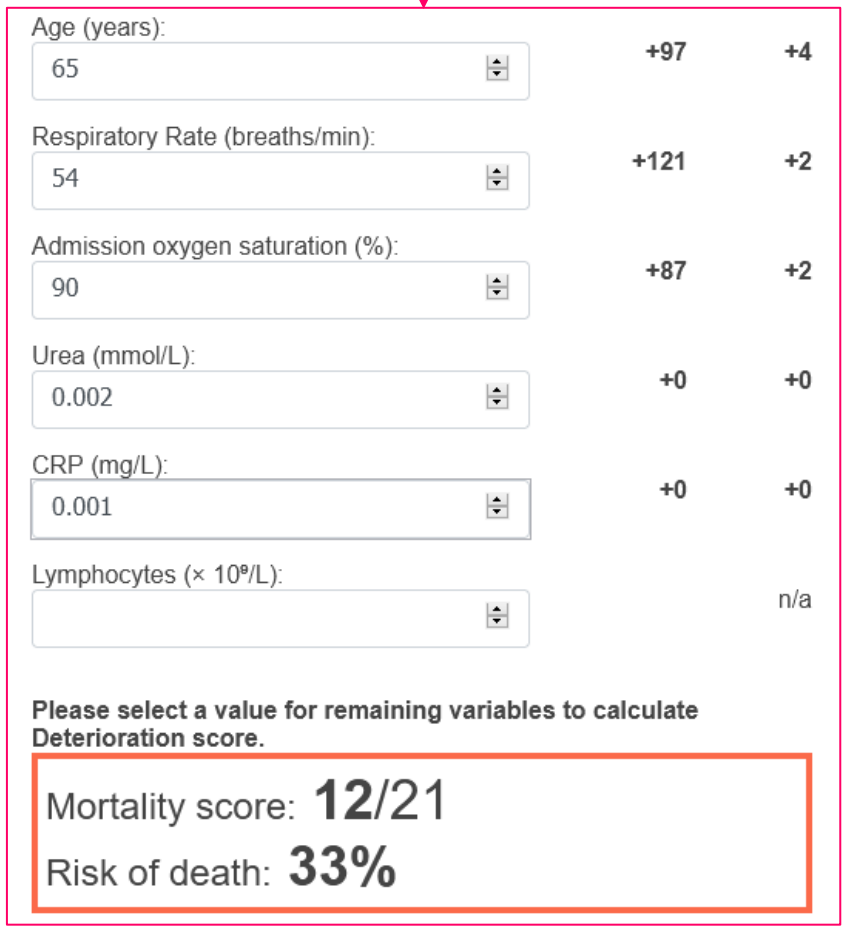


Forschungsaufgabe: Finde eine Erklärung für die von einer Blackbox erzeugten Ergebnisse.

Naiver Lösungsansatz: Man nehme eine White Box, um die Funktion der Black Box nachzuahmen.

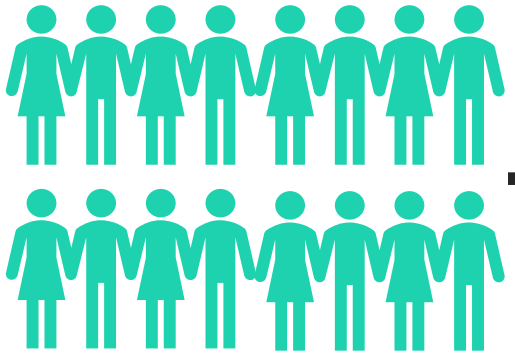
Quellen:
- https://www.researchgate.net/publication/322976218_A_Survey_of_Methods_for_Explaining_Black_Box_Models

WHITE BOX ALGORITHMEN



Quelle: https://isaric4c.net/risk/

AMS ALGORITHMUS



Arbeitssuchende

Chance auf Reintegration in den Arbeitsmarkt innerhalb von 6 Monaten	Konsequenzen
> 66 % Chance	Keine oder wenig Unterstützung
< 25 % Chance	Keine oder wenig Unterstützung
Übrige	Volle Unterstützung

Überstimmung/Bestätigung durch Mensch(en)

PRIVATSPHÄRE/DATENSCHUTZ

Oft undefiniert, aber meist verstanden als oder in Zusammenhang mit Datenschutz, Datensicherheit, manchmal auch Freiheit und Vertrauen.

Wie kann man Datenschutz erreichen?

- Technische Lösungen: z. B. Beschreibung von Gruppen anstelle von Einzelpersonen, z. B. durch Muster, Datenschutz durch Technik, Datenminimierung, Zugriffskontrolle usw.
- Regulierungsansätze: z. B. Zertifikate, Anpassung von Gesetzen und Vorschriften an die Besonderheiten des KI-Datenschutzes usw.

Quelle: Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

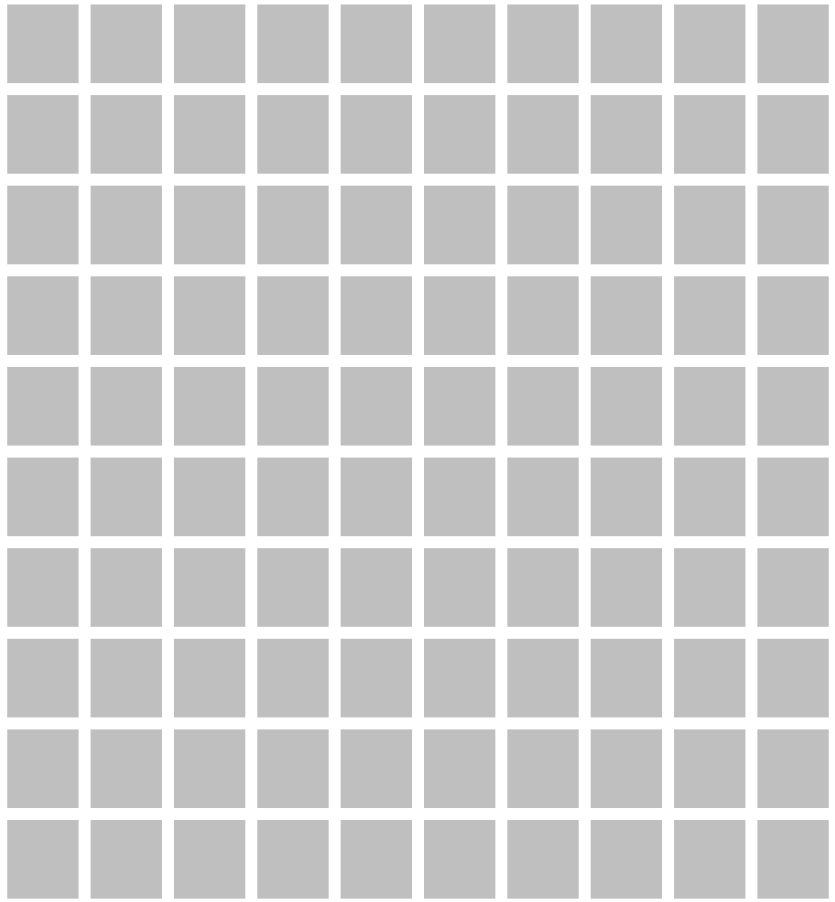
WIE KANN MAN DEN PROBLEMEN ENTGEGENWIRKEN?

Aktuelle Problemstellungen bergen ein immenses Schadenspotenzial.

1. Wie viel Zeit und Ressourcen verschlingt ein Rechtsstreit?
2. Wie viel Schadenersatz muss gezahlt werden?
3. Wie viel kostet es, das verlorene Vertrauen der Kunden wiederzugewinnen?

→ Das sind Risiken, die es zu managen gilt!

INWIEWEIT SETZT IHR UNTERNEHMEN KI EIN BZW. PLANT UND DISKUTIERT DEN EINSATZ (09/2022)?



KI kein Thema

Geplant oder diskutiert

KI im Einsatz

Quelle: https://www.bitkom.org/Presse/Presseinformation/Kuenstliche-Intelligenz-2022

WIE ETHISCHE RISIKEN AKTUELL BEHANDELT WERDEN

- Daumendrücken und/oder reine Befolgung der Gesetze (insbesondere in Bezug auf den Datenschutz)
- Primärer Fokus auf die Reduzierung von Biases
  - Es gibt technische Lösungen. Sie prüfen, ob geschützte Personen speziell benachteiligt werden, und ändern dann das Ergebnis für jede Gruppe, bis eine Fairness-Metrik "passt".
  - Es werden nicht alle Formen der Voreingenommenheit berücksichtigt, und es ist nicht klar, welche Metrik zur Messung der Fairness verwendet werden sollte.
  - Datenwissenschaftler verfügen in der Regel nicht über die nötigen Fähigkeiten, um diese Fragen zu beantworten.
- Sekundärer Fokus auf Erklärbarkeit
  - Erklärbarkeit ist ein sehr flexibles Konzept
  - Wird derzeit intensiv erforscht
  - Klassischer Ansatz: Wir bauen ein zweites Modell, um das erste zu erklären

WIE ETHISCHE RISIKEN AKTUELL BEHANDELT WERDEN

a. INHALT: Mit welchen Risiken im Kontext von Ethik & KI wird überhaupt gerechnet?

b. PROZESS: Wie werden Risiken im Kontext von Ethik & KI im Unternehmen systematisch identifiziert und behandelt?

→ Fast niemand kümmert sich um b!

ETHISCHE PRINZIPIEN: METASTUDIE

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity

Ethical principle	Number of documents	Included codes
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

Quelle: Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

WEITERE RISIKEN

Beim autonomen Fahren z. B. sind nicht Voreingenommenheit, Erklärbarkeit oder Datenschutz die primären ethischen Risikobereiche, sondern die Verletzung oder Tötung von Menschen!

→ Viele Risiken sind anwendungsspezifisch und bedürfen einer besonderen Behandlung

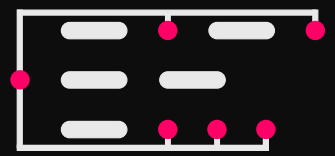
→ Viele Risiken basieren auf sogenannten „ill-structured problems“

WIE STARTEN?

1. Aufklären. Durchführung von Seminaren, um das Bewusstsein und das Verständnis für die Problematik zu schärfen, was den Weg für die Übernahme des Themas durch die Organisation und für einen strategischen Ansatz ebnet.
2. Verfassung einer (KI-)Ethikerklärung unter Einbezug leitender Angestellter. Eine gute Erklärung kann in relativ kurzer Zeit sozialisiert und operationalisiert werden.
3. Durchführung einer Risikoanalyse. Identifikation von Ressourcen für ein KI-Ethikprogramms, inkl. einem Kernteam mit Management-Backing.

Nachhaltige KI

eine technische Perspektive



14.10.2022, Christof Wolf-Brenner