



CATÓLICA
FACULTY
OF LAW

ESCOLA DO PORTO



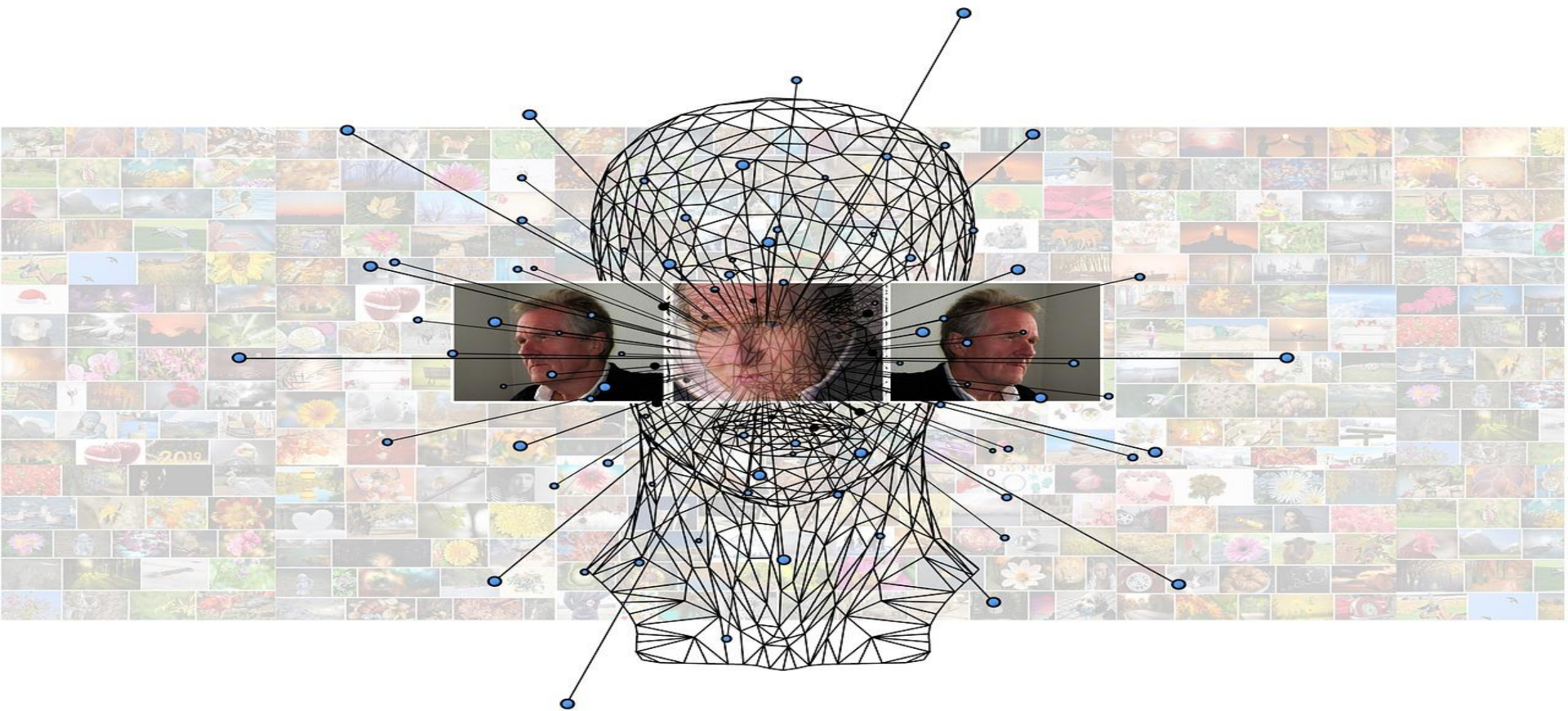
Funded by
the European Union

Das Ende der Täuschung? - Der algorithmischen Diskriminierung im digitalen Zeitalter entgegenwirken

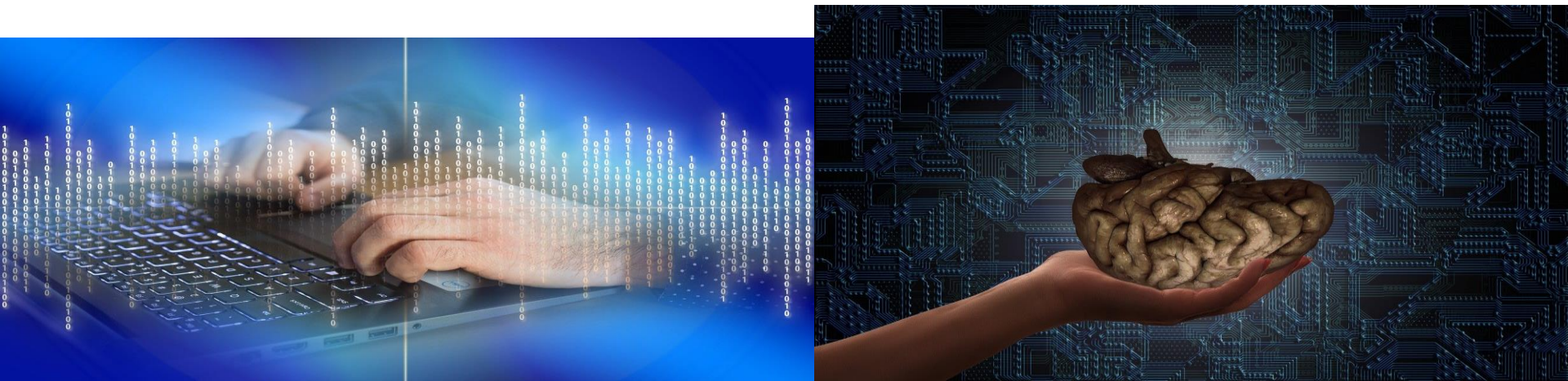
Catarina Santos Botelho
cbotelho@ucp.pt

Was sind Algorithmen?

- Algorithmen sind kodierte Verfahren, bei denen Eingabedaten in einer logischen und mathematischen Abfolge in ein Ergebnis umgewandelt werden.



- In diesem Stadium sind die Algorithmen komplex und entziehen sich dem menschlichen Verständnis.
- Algorithmen können in wissensbasierte und maschinelle Lernalgorithmen (ML) unterteilt werden.
- **Wissensbasierte** Systeme sind die traditionellen Flussdiagramm-Algorithmen, die eine manuelle Kodierung erfordern.
- **ML** sind Algorithmen, die große Datenmengen mit Ausgangsvariablen füttern, damit der Algorithmus "lernt, ohne explizit programmiert zu werden".



Warum ist dieses Thema wichtig?

Automatisierte Entscheidungsprozesse sind eine zunehmende Realität



- Bestimmungen für das Gesundheitswesen,
- Zuteilung von Impfstoffen,
- Haftstrafen,
- polizeiliche Kontrolle,
- Auswahl des IRS-Audits,
- Auszählung der Stimmen,
- Kreditentscheidungen,
- Bewerbungen,
- Arbeitsverteilung,
- Anträge auf Sozialleistungen,
- die Erteilung von Einwanderungsvisa,
- Auswahlverfahren für die Zulassung zum Kindergarten oder zur Universität,
- Preise,
- Überwachung der Verbraucher,
- zielgerichtetes Mikro-Marketing.



Was sind die großen Fragen?

- Wie geht es weiter mit der Demokratie und dem Schutz der Menschenrechte in einer neuen "sozio-technischen Internet-Architektur"?
- Können Algorithmen die demokratische Gegenwehr verstärken
- Ist die Mitwirkung des Menschen an Entscheidungsprozessen für den Schutz der Menschenrechte noch wichtig oder eine unnötige Vorsichtsmaßnahme?



Unmittelbare Diskriminierung / mittelbare Diskriminierung

- **Unmittelbar:** wenn eine Person aufgrund eines geschützten Merkmals in einem geschützten Bereich eine weniger günstige Behandlung erfährt als eine andere.
- **Mittelbar:** wenn eine scheinbar neutrale Bestimmung, ein Kriterium oder eine Praxis eine Person aus einer geschützten Gruppe ohne akzeptable Rechtfertigung benachteiligt.



Ungleiche Behandlung / Unterschiedliche Auswirkungen

- Titel VII des Bürgerrechtsgesetzes von 1964 (USA)
- *Ungleiche Behandlung*: wenn der Arbeitgeber den Arbeitnehmer aufgrund eines geschützten Merkmals anders behandelt.
- *Unterschiedliche Auswirkungen*: eine Praxis, die auf den ersten Blick neutral und nicht diskriminierend erscheint, in Wirklichkeit aber eine geschützte Gruppe unverhältnismäßig stark benachteiligt.



Kriterien für geschützte Gruppen

- Sind geschützte Gruppen festgelegt oder können andere „sozial hervorstechende Identitätsachsen“ hinzugefügt werden?
- Kann der sozioökonomische Status oder das Bildungsniveau bei Antidiskriminierungsvorschlägen berücksichtigt werden? Und was ist mit Fettleibigkeit, chronischen Krankheiten oder dem Browsertyp?
- Sollte die Gestaltung des Antidiskriminierungsrechts, insbesondere die Liste der Gründe und/oder Ausnahmen, offen, geschlossen oder gemischt sein?



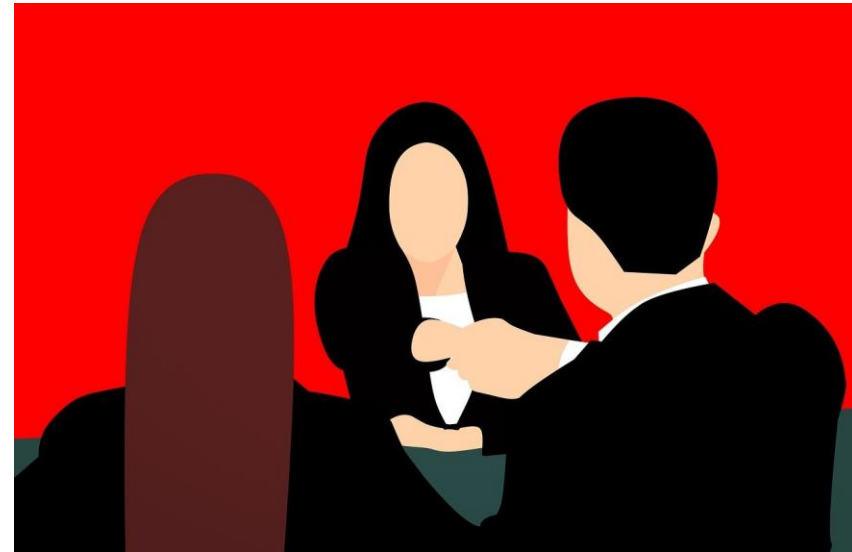
Vorurteile gegenüber ML-Algorithmen:

- *die daraus resultierende Voreingenommenheit*, wenn die Ersteller des Algorithmus absichtlich oder unabsichtlich voreingenommen sind und sich diese Voreingenommenheit im Datenverarbeitungsmechanismus fortsetzt;
- *verzerrte Stichprobendaten*, bei denen die bereitgestellten Stichproben- oder Trainingsdaten verzerrt sind;
- *Ergebnisverzerrung oder Teufelskreisverzerrung*, die selbst dann auftritt, wenn der Algorithmus mit repräsentativen Daten trainiert wurde, und zwar aufgrund der in den Daten enthaltenen sozialen Ungleichheiten.

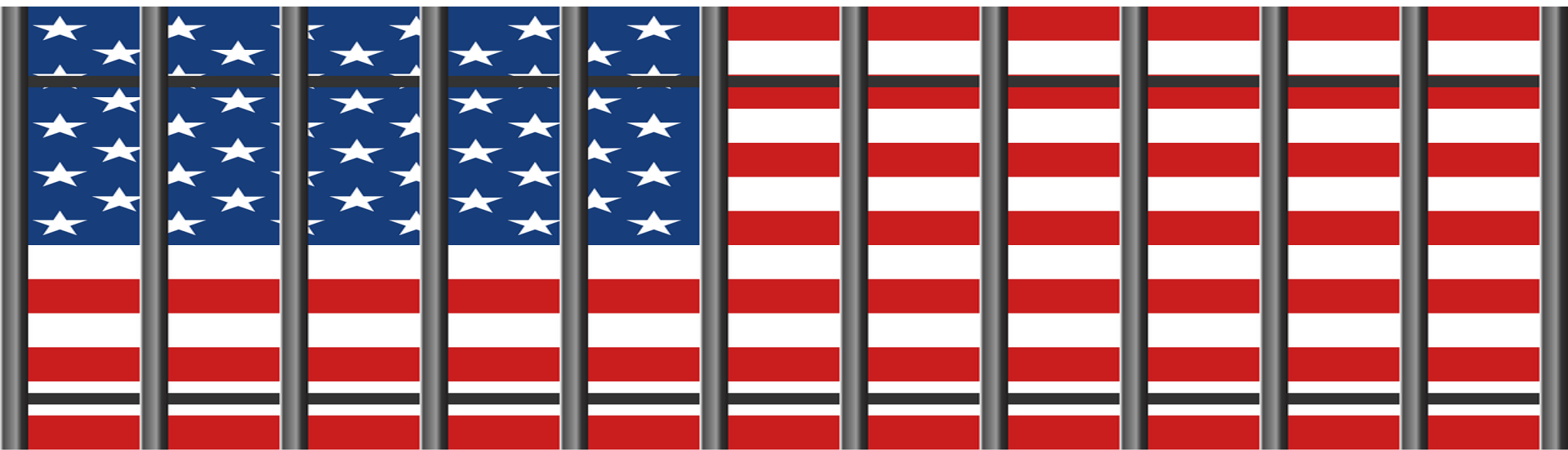


Beispiele für algorithmische unterschiedliche Auswirkungen und statistische Diskriminierung

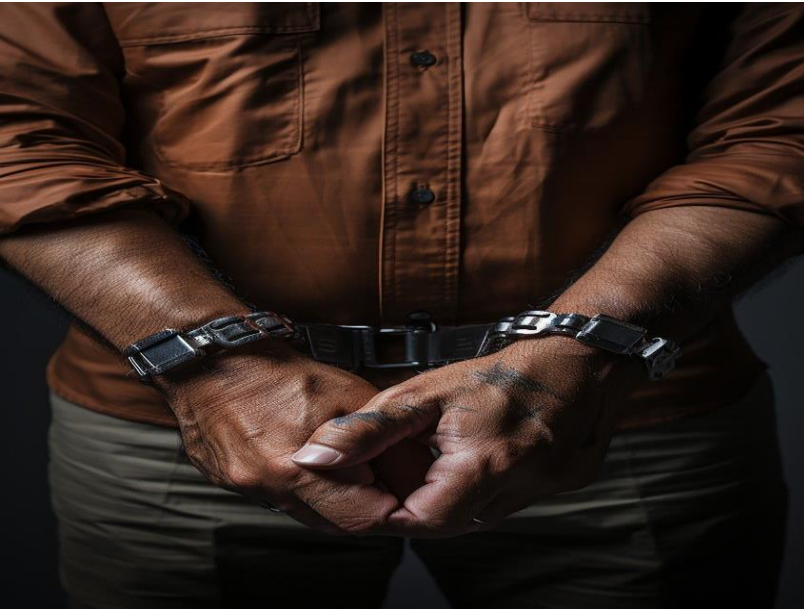
- **Amazon Fall:**
- Amazon hat ein ML-System entwickelt, um die zukünftigen Softwareentwickler-Mitarbeiter zu identifizieren. Der Algorithmus stufte die Bewerber nach ihrer erwarteten Produktivität ein. Da der Algorithmus anhand vorhandener Einstellungsdaten trainiert wurde, die die bestehende männliche Dominanz in der Technologiebranche widerspiegeln, wurden Bewerber mit weiblichen Indikatoren in ihren Lebensläufen (z. B. Frauentennisverein) ungeachtet ihrer Qualifikation für die Stelle diskriminiert.



- ***Fall COMPAS (Correctional Offender Management Profiling for Alternative Sanctions):***
- Auf der Grundlage von Daten aus einem US-Bezirk entwickelte ein Unternehmen ein als COMPAS bekanntes Risikobewertungsinstrument zur Vorhersage der Rückfälligkeit von Häftlingen und Angeklagten.
- Bei der Frage „Wird diese Person rückfällig, wenn sie auf Bewährung entlassen wird?“ verglich der Algorithmus die Merkmale der Person mit den Merkmalen derjenigen, die während der Bewährung rückfällig wurden.



- Das Problem, über das ProPublica im Jahr 2016 berichtete, bestand darin, dass COMPAS seine Vorhersage der erneuten Verhaftung als Ersatz für die erneute Straffälligkeit verwendete.
- COMPAS wurde vorgeworfen, rassistisch verzerrte Ergebnisse zu liefern, da die falsch-positiven Ergebnisse (Vorhersage eines hohen Rückfallrisikos, obwohl der Täter tatsächlich nicht rückfällig wird) bei afroamerikanischen Angeklagten doppelt so hoch waren wie die falsch-positiven Ergebnisse bei Angeklagten kaukasischer Herkunft.



Die Illusion der algorithmischen Objektivität

- "Algorithmus-Aversion" / "KI-als-Monster"

gegen

- Herkulische Superkräfte / Meister der Objektivität



Endogene algorithmische Diskriminierung

- Die Diskriminierung liegt in der algorithmischen Gestaltung selbst.
- Mehr Transparenz: Verhinderung der Verlagerung von diskriminierenden menschlichen Entscheidungen auf Algorithmen.
- Ist Transparenz die neue Zustimmungsfalle?



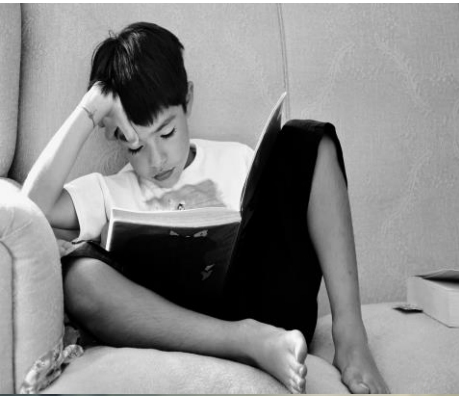
Exogene algorithmische Diskriminierung

- Das Problem ist nicht die "Black Box" selbst, sondern die *Welt*, in der die Algorithmen arbeiten.
- Diskriminierung ist das Ergebnis impliziter Voreingenommenheit, statistischer Verzerrungen oder historischer Falschdarstellungen von Gruppen, die ihren Weg durch das vermeintlich aseptische ML gefunden haben.
- Es geht darum, "über die bloße *Nicht-Diskriminierung* hinaus zu aktiv *antidiskriminierenden* Algorithmen zu gelangen" (Bornstein).



Substanzielle Gleichstellung

- **Algorithmische positive Diskriminierung** kann durch verschiedene Maßnahmen erreicht werden, z. B. durch die Einbeziehung von Gleichheitsmetriken in die Modellauswahlprozesse oder durch Gruppenfairness unter Verwendung von Ansätzen der "demografischen Parität" oder "statistischen Parität".



Algorithmen zur Regulierung und Entschärfung

- Da Algorithmen für den Menschen unverständlich sind, ist eine zusätzliche Anstrengung erforderlich, um diskriminierende Ergebnisse zu entdecken.
- Unter diesen Umständen sollten politische Entscheidungsträger, Gerichte, Agenturen, Behörden und Arbeitgeber keine romantisierende Haltung gegenüber Algorithmen einnehmen und diesen nicht unkritisch und mit übermäßig viel Vertrauen gegenüberstehen.
- Eine aktive Wachsamkeit ist daher sehr zu empfehlen.



1. Algorithmische Kontrolle



- Die europäische KI-Verordnung (KI-VO) ist bahnbrechend und zielt auf eine menschenzentrierte und ethische KI ab, bei der die Ergebnisse von KI-Systemen nachvollziehbar, erklärbar und transparent sind.
- KI-Systeme werden nach dem Grad ihres potenziellen Risikos eingestuft (Artikel 5):
 - Von vornherein untersagt
 - Hohes Risiko
 - Geringes Risiko

- Prüfung der Verhältnismäßigkeit.
- Um die potenziellen Auswirkungen von KI-Systemen mit hohem Risiko zu minimieren, müssen die Betreiber vor dem Einsatz eines KI-Systems mit hohem Risiko **eine Folgenabschätzung für die Grundrechte** durchführen (Artikel 27).



Artikel 6 (KI-VO)

(3) Abweichend von Absatz 2 gilt ein in Anhang III genanntes KI-System nicht als hochriskant, wenn es kein erhebliches Risiko der Beeinträchtigung in Bezug auf die Gesundheit, Sicherheit oder Grundrechte natürlicher Personen birgt, indem es unter anderem nicht das Ergebnis der Entscheidungsfindung wesentlich beeinflusst.

Unterabsatz 1 gilt, wenn eine der folgenden Bedingungen erfüllt ist:

- a) das KI-System ist dazu bestimmt, eine eng gefasste Verfahrensaufgabe durchzuführen;*
- b) das KI-System ist dazu bestimmt, das Ergebnis einer zuvor abgeschlossenen menschlichen Tätigkeit zu verbessern;*
- c) das KI-System ist dazu bestimmt, Entscheidungsmuster oder Abweichungen von früheren Entscheidungsmustern zu erkennen, und ist nicht dazu gedacht, die zuvor abgeschlossene menschliche Bewertung ohne eine angemessene menschliche Überprüfung zu ersetzen oder zu beeinflussen; oder*
- d) das KI-System ist dazu bestimmt, eine vorbereitende Aufgabe für eine Bewertung durchzuführen, die für die Zwecke der in Anhang III aufgeführten Anwendungsfälle relevant ist.*

2. Blockieren von Informationen: "Fairness durch Blindheit?"

- Verhindert die Sperrung persönlicher Informationen über geschützte Kategorien in Schulungsdaten diskriminierende Ergebnisse?
- Um Ethnie-, Behinderten-, Geschlechts-, Einkommens- oder Postleitzahlen-blinde Ergebnisse zu erzielen, reicht es nicht aus (und ist auch nicht ratsam), sensible Merkmale einfach aus der Betrachtung herauszunehmen.



- ML-Systeme zeichnen sich durch die "Erkennung von Mustern und die Identifizierung und Nutzung von nicht offensichtlichen Proxies für ausgelassene sensible Variablen" aus.



- Wie können wir "Schichten über Schichten von Spiegeln und Stellvertretern" offenlegen, und woher wissen wir, welche *Stellvertreter* (funktionale Substitute) wertvolle Informationen enthalten?
- Erschwerend kommt hinzu, dass die Gleichstellung in Bezug auf Ethnie und Geschlecht möglicherweise immer noch übergreifende Unterschiede innerhalb dieser Gruppen in Bezug auf andere Dimensionen wie Behinderung, Alter oder Nationalität vernachlässigt.



3. Regulierung und Formung von Daten durch algorithmische Gestaltung

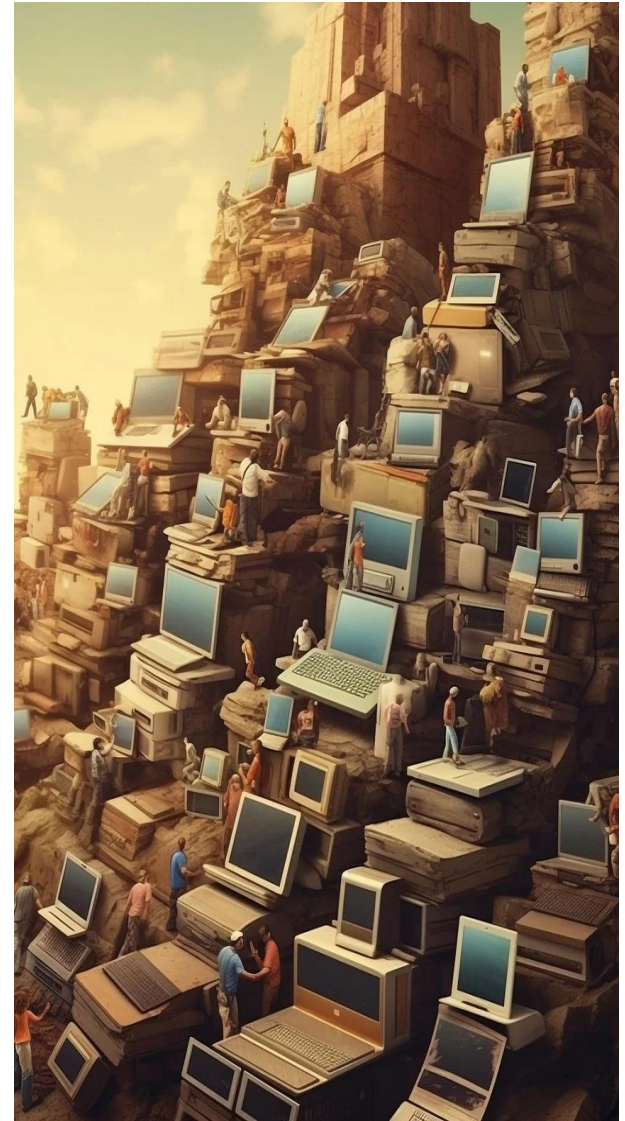
- Die Datenschutz-Grundverordnung (DSGVO) schreibt vor, dass die Datenverarbeitung transparent sein muss (Erwägungsgründe 39, 58, 78 DSGVO und Artikel 5 Absatz 1 Buchstabe a und Artikel 12 DSGVO), gewährt das Recht auf Auskunft und das Recht, über das Vorliegen einer automatisierten Entscheidungsfindung informiert zu werden (Artikel 13-15) sowie das Recht, nicht ausschließlich einer automatisierten Entscheidungsfindung unterworfen zu werden (Artikel 22), und führte die Datenschutz-Folgenabschätzung ein, um die Risiken für die Rechte und Freiheiten der betroffenen Personen und die Risiken der Verarbeitungstätigkeit zu bewerten (Artikel 35). Dieses Recht gilt jedoch nicht, wenn die betroffene Person ausdrücklich in die automatisierte Entscheidungsfindung eingewilligt hat (Artikel 22 Absatz 2 Buchstabe b) DSGVO).



Wichtiger als die Regulierung der Algorithmen ist in diesem Stadium die Regulierung der Daten, die den Algorithmen zugeführt werden.

Ziel ist es, "*faire* Daten" zu sammeln, die ein Gegengewicht zu gesellschaftlichen Vorurteilen bilden können.

Anstatt die Daten zu blockieren, schlagen einige Wissenschaftler vor, sie *zu verändern*. Die Vorverarbeitung der Daten kann durch eine allgemeine Anpassung der Eingabedaten oder durch eine Anpassung der Zielvariablen (z. B. Geschlecht) erfolgen.



4. Nachträgliche Strategien zur Beseitigung von Diskriminierung: individuelle Rechte oder Gruppenrechte?

- Gerichte (durch gerichtliche Überprüfung)
- Datenschutzbehörden (z. B. durch Datenschutz-Folgenabschätzungen)
- "FDA für Algorithmen"?
- "ein individuelles Recht, KI-Entscheidungen anzufechten"?



Die blinden Flecken der Algorithmen und die Bedeutung der Zentrierung auf den Menschen

- Gibt es ein "Recht auf eine menschliche Entscheidung"?
- KI kann relevante Teile des menschlichen Intellekts nachahmen und sogar übertreffen. Dennoch sind menschliches logisches Denken, menschliche Wahrnehmung und die Anwesenheit des Menschen bei der Ausübung von *Ermessen* (was etwas anderes ist als Willkür) oder Urteilsvermögen im Einzelfall weiterhin erforderlich.



Warum?

- Der Grund dafür ist, dass nur der Mensch die Realität vollständig begreifen kann. Die KI führt logische Operationen aus, die abstrakt von der Realität (der Außenwelt oder der realen Welt) sind. Auf diese Weise erschafft die KI ihre Realität innerhalb der "Realität", auch wenn es sich dabei immer um eine begrenzte Realität handelt, die ontologisch nicht ihr *Wesen* erreichen kann. Der Mensch hingegen *erlebt und erfasst die Realität*.



- Es handelt sich nicht um einen sinnlosen Wettbewerb zwischen Mensch und Maschine. Wir brauchen kein menschliches Eingreifen in die KI, weil "wir, die Menschen" besser sind.
- Wir brauchen menschliche Führung, weil wir Menschlichkeit, Ethik und Moral nicht aus öffentlichen und privaten Entscheidungsprozessen ausklammern können, die unser Leben tiefgreifend beeinflussen können.

