Metadata

- Id: EU.AI4T.O1.M3.3.3t
- Title: 3.3.3 Woher kommt das Risiko?
- Type: text
- Description: Identifizieren Sie die verschiedenen Arten von Risiken
- Subject: Artificial Intelligence for and by Teachers
- Authors:
  - AI4T
- Licence: CC BY 4.0
- Date: 2022-11-15

WHERE DOES THE RISK COME FROM?

In seiner Studie über Künstliche Intelligenz[1] stellt der Europäische Parlamentarische Forschungsdienst fest: "*Es ist wichtig, darauf hinzuweisen, dass KI-Algorithmen nicht objektiv sein können, da sie - genau wie Menschen - im Laufe ihrer Ausbildung eine Art und Weise entwickeln, wie sie das, was sie zuvor gesehen haben, einordnen, und diese 'Weltsicht' nutzen, um neue Situationen, mit denen sie konfrontiert werden, zu kategorisieren.*" [deepl translation]

Schauen wir uns an, woher die Subjektivität einer KI kommt und welche Risiken damit verbunden sind.

DIE VERZERRUNG IN DATEN UND ALGORITHMEN

Wie bei jedem digitalen System stammen die in KI-basierten Plattformen verwendeten Daten aus verschiedenen Quellen und haben unterschiedliche Formate. Sie weisen verschiedene Arten von Verzerrungen auf[2]. Datenverzerrungen sind hauptsächlich statistischer Natur. Lassen Sie uns ein paar davon auflisten.

- **Stichprobenverzerrungen** sind typischerweise in den Datenwerten vorhanden. Dies ist zum Beispiel der Fall, wenn ein Einstellungsalgorithmus, der auf einer Datenbank trainiert wurde, in der Männer überrepräsentiert sind, Frauen ausschließt.
- **Stereotype Voreingenommenheit** ist eine Tendenz, in Bezug auf die soziale Gruppe zu handeln, der wir angehören. Eine Studie zeigt zum Beispiel, dass Frauen dazu neigen, auf Stellenangebote zu klicken, von denen sie glauben, dass sie als Frau leichter zu bekommen sind.
- **Omitted variable bias** (Modellierungs- oder Kodierungsverzerrung) ist eine Verzerrung aufgrund der Schwierigkeit, einen Faktor in den Daten darzustellen oder zu kodieren. Da

es beispielsweise schwierig ist, faktische Kriterien zur Messung der emotionalen Intelligenz zu finden, wird diese Dimension in den Einstellungsalgorithmen nicht berücksichtigt.

- **Die Auswahlverzerrung** ist wiederum auf die Merkmale der Stichprobe zurückzuführen, die ausgewählt wurde, um Schlussfolgerungen zu ziehen. So verwendet eine Bank beispielsweise interne Daten, um einen Kreditscore zu ermitteln, und konzentriert sich dabei auf diejenigen, die einen Kredit aufgenommen haben oder nicht, lässt aber diejenigen außer Acht, die noch nie einen Kredit aufnehmen mussten usw.

Die algorithmische Verzerrung ist hauptsächlich eine Frage der Argumentation. Solche Verzerrungen werden von KI-Ingenieuren absichtlich oder unabsichtlich eingeführt.

In der bereits erwähnten Studie des Europäischen Parlamentarischen Forschungsdienstes werden zwei konkrete Beispiele genannt: "*Betrachten Sie einen symbolischen KI-Algorithmus zur Prüfung von Bewerbungen. Er könnte die Bewerber nur auf der Grundlage ihrer Ausbildung und Erfahrung bewerten. Wenn er jedoch Faktoren wie Mutterschaftsurlaub nicht berücksichtigt oder die Ausbildung in ausländischen Einrichtungen nicht in der Weise anerkennt, wie es menschliche Auswahlausschüsse tun würden, könnte der Algorithmus Frauen und ausländische Bewerber diskriminieren.*" [deepl translation]

"*Betrachten wir nun ein ähnliches KI-Tool im Rahmen des ML-Paradigmas (maschinelles Lernen). Solche Algorithmen finden ihre eigenen Wege, um zu erkennen, welche Art von Kandidaten in ihren Trainingsdaten ausgewählt wurden. Wenn es in der Vergangenheit strukturelle Verzerrungen bei der Auswahl gab - zum Beispiel Rassendiskriminierung - kann der Algorithmus diese lernen. Selbst wenn Daten über Nationalität oder ethnische Zugehörigkeit aus den Daten entfernt werden, ist ML in der Lage, Stellvertreter für zugrunde liegende Muster in anderen Daten wie Sprachen, Postleitzahlen oder Schulen zu finden, die gute Prädiktoren für die ethnische Zugehörigkeit sein können.*" [deepl translation]

DIE DREI FACETTEN DES ALGORITHMISCHEN RISIKOS

Das algorithmische Risiko kann auf drei Arten charakterisiert werden[3].

- Erstens gibt es die **algorithmische Enge**, die sich auch auf Meinungen, kulturelles Wissen oder sogar kommerzielle Praktiken beziehen kann. In der Tat konfrontieren die Algorithmen den Internetnutzer je nach seinem Profil und den integrierten Parametern mit denselben Inhalten, obwohl der Grundsatz der Fairness gewahrt bleibt. Dies ist der Fall bei Nachrichtenempfehlungsseiten wie Facebook oder Produktempfehlungsseiten wie Amazon.
- Die zweite Facette des algorithmischen Risikos hängt mit der **Kontrolle aller Aspekte des Lebens eines Individuums** zusammen, von der Regulierung der Informationen für Investoren bis hin zu seinen oder ihren Essgewohnheiten, Hobbys oder sogar seinem Gesundheitszustand. Diese Verfolgung des Individuums suggeriert eine Form der Überwachung, die dem Wesen der individuellen Freiheit zuwiderläuft.

- Der dritte Punkt betrifft die **mögliche Verletzung der Grundrechte**. Insbesondere die algorithmische Diskriminierung, die als ungünstige oder ungleiche Behandlung im Vergleich zu anderen Personen oder anderen gleichen oder ähnlichen Situationen definiert wird, die auf einem gesetzlich ausdrücklich verbotenen Grund beruht. Dies umfasst die Untersuchung der Fairness (*Fairness*) von Ranking- (Sortierung von Personen, die online nach einem Job suchen), Empfehlungs- und Prognoselernalgorithmen. Das Problem der diskriminierenden Voreingenommenheit durch Algorithmen betrifft verschiedene Bereiche wie Online-Einstellungen, Gerichtsentscheidungen, Entscheidungen von Polizeistreifen oder Schulzulassungen.

WIE GEHT MAN MIT DATEN UND ALGORITHMISCHEN RISIKEN UM?

Für R. Schwartz & al.[4] ist "*Voreingenommenheit weder neu noch einzigartig für KI, und es ist nicht möglich, das Risiko der Voreingenommenheit in einem KI-System auf Null zu bringen*". In der Zwischenzeit ist die Erkenntnis, dass KI-Agenten von Natur aus subjektiv sind, eine entscheidende Voraussetzung dafür, dass sie nur für Aufgaben eingesetzt werden, für die sie gut gerüstet sind.

Die EPRS-Studie schließt mit mehreren Empfehlungen für den Einsatz von KI-basierten Anwendungen:

- Verstehen Sie Voreingenommenheit und Subjektivität
- Vermeiden Sie Anwendungen, die über die Fähigkeiten der KI hinausgehen
- Vermeidung von Anwendungen mit unerwünschten Auswirkungen
- Aufrechterhaltung der menschlichen Autonomie
- Suche nach Lösungen für Probleme, nicht nach Problemen für Lösungen
- Überlegen, was wir wirklich von KI wollen

---

1. Artificial intelligence: How does it work, why does it matter, and what can we do about it ? - Philip Boucher, Scientific Foresight Unit (STOA) - ISBN: 978-92-846-6770-3 - Union Européenne, 2020 ↩
2. Algorithms, Data and Bias: Public Policy Needed, Anne Bouverot, Thierry Delaporte, 2019 ↩
3. Article in French: D'où vient le risque ? Des données et des algorithmes - Serge Abiteboul, Thierry Viéville, 2020 ↩
4. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence - Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, NIST Special Publication 1270 , 2022 ↩