Diskriminierende KI?

Entstehung und Lösungsansätze von Bias in KI-Systemen

Marieke Prien, LearningCenter Hochschule Osnabrück, 15.05.2025

Inhalt

- **Definitionen *KI* und *Bias***
- **Relevanz des Themas**
- **Entstehung von Bias**
- **Lösungsansätze**
- **Fragen & Diskussion**

Was ist Künstliche Intelligenz?

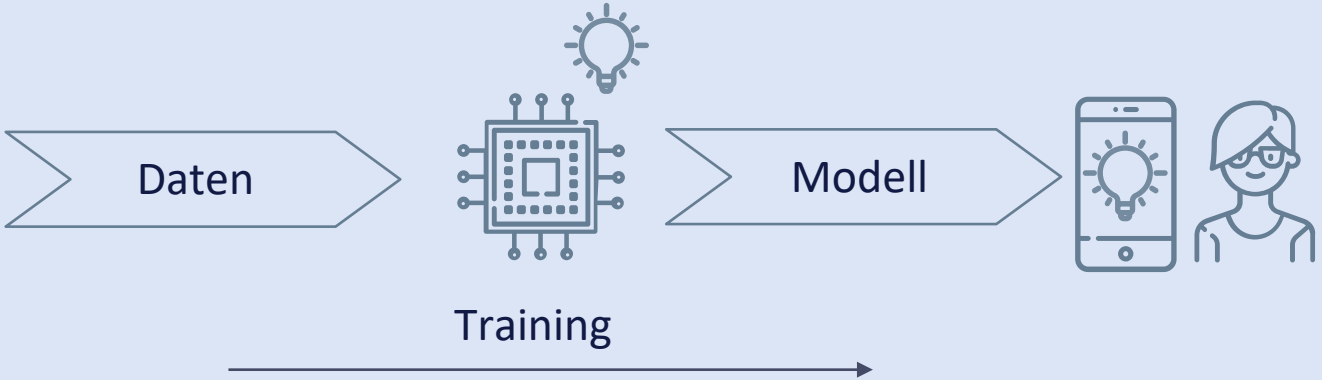
> „Künstliche Intelligenz ist die Fähigkeit einer Maschine, menschliche Fähigkeiten wie logisches Denken, Lernen, Planen und Kreativität zu imitieren.“

Europäisches Parlament, 2020

Wie arbeitet KI?

KI arbeitet häufig mit **maschinellem Lernen.**

Maschinelles Lernen ist ein **statistisches** Verfahren, bei dem **Muster** in **Trainingsdaten** erkannt werden.



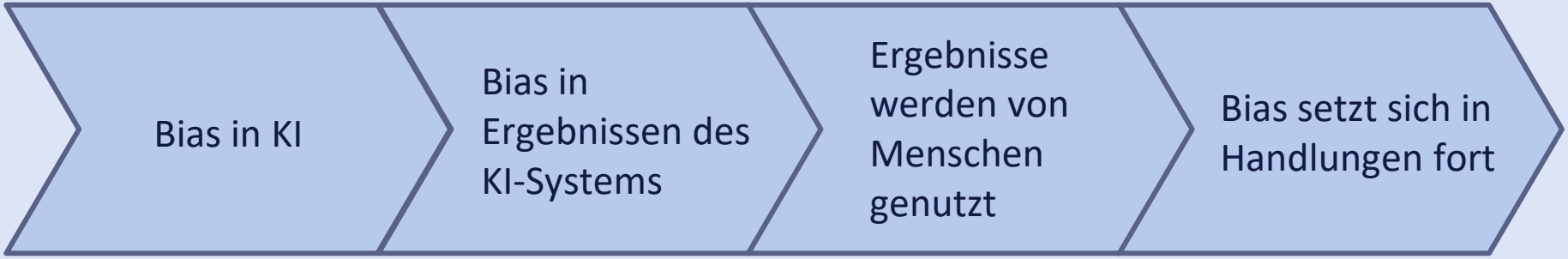
Was ist Bias?

bias (engl.): Voreingenommenheit, Vorurteil, Befangenheit,
systematischer Fehler, Verzerrung

Bias in KI sind systematische Verzerrungen, die sich im Modell zeigen.
Sie verursachen, dass die KI „voreingenommen“ ist oder „Vorurteile“ hat.

Warum reden wir über Bias?

Bias ist diskriminierend.



» Vorurteile werden reproduziert und beeinflussen Entscheidungen sowie Handlungen, wodurch Menschen ungleich behandelt werden.

Warum reden wir über Bias?

Super Fails | Tayminator
Wie Microsofts KI zum Nazi wurde

Insight - Amazon scraps secret AI recruiting tool that showed bias against women

How an AI grading system ignited a national controversy in the U.K.

Study finds gender and skin-type bias in commercial artificial-intelligence systems

Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.

Österreich

Jobcenter-Algorithmus landet vor Höchstgericht

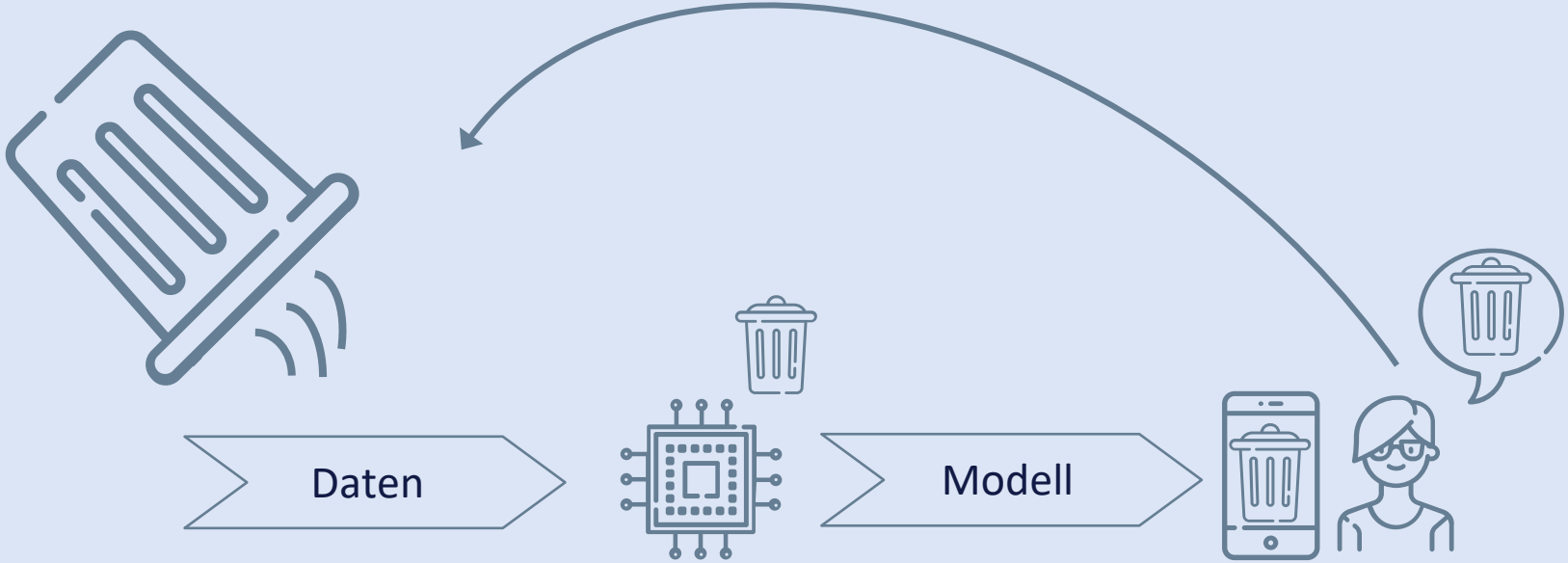
In Österreich streiten die Behörden seit Jahren über e
Arbeitssuchenden vorhersagen soll. Nun landet der Fa
Verwaltungsgericht.

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

Wie entsteht Bias?

Garbage In – Garbage Out



Wie entsteht Bias?

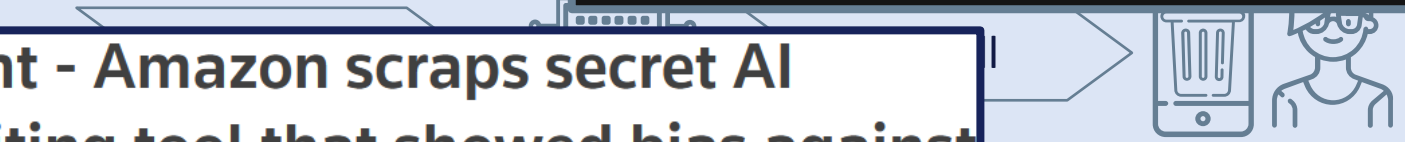
Garbage In – Garbage Out

Study finds gender and skin-type bias in commercial artificial-intelligence systems

Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.

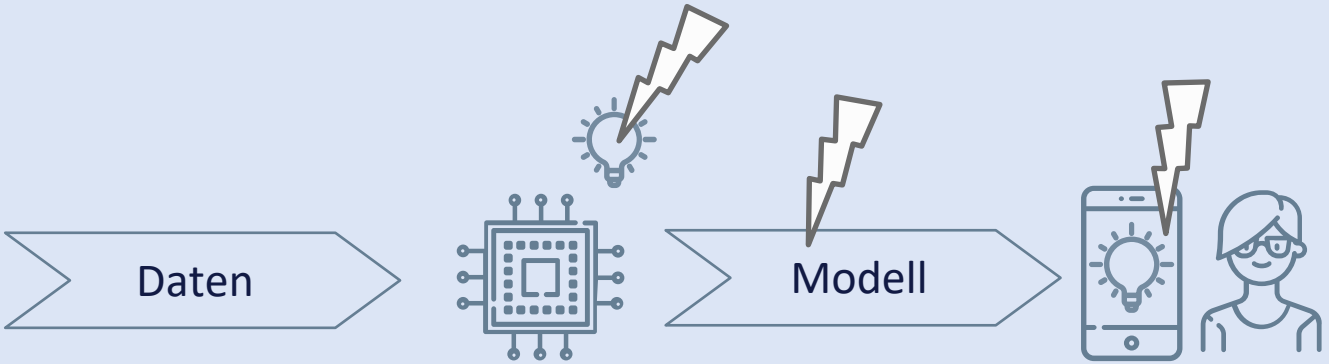
Super Fails | Tayminator
Wie Microsofts KI zum Nazi wurde

Insight - Amazon scraps secret AI recruiting tool that showed bias against women



Wie entsteht Bias?

Algorithmen und Modelle



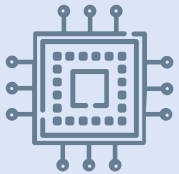
Wie entsteht Bias?

Algorithmen und Modelle

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

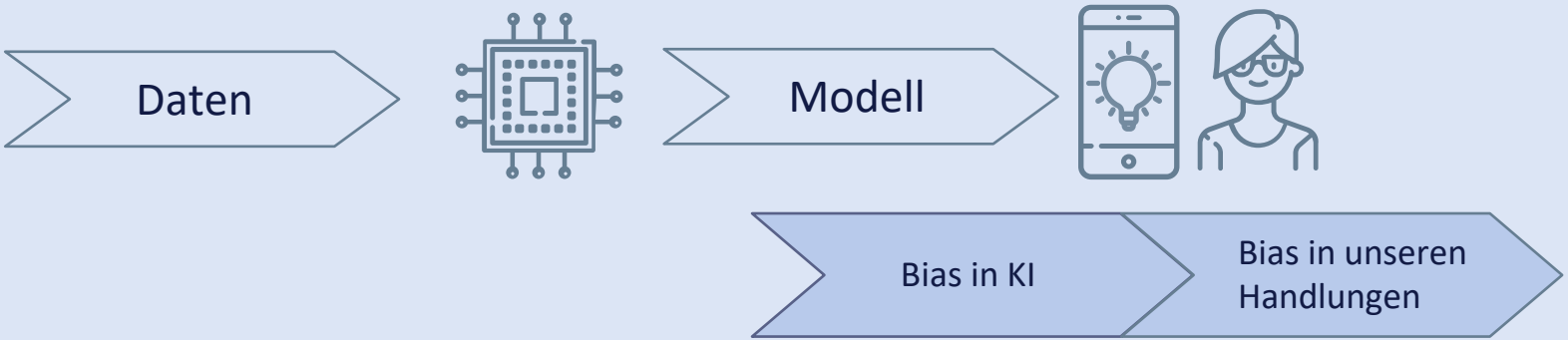
Daten



How an AI grading system ignited a national controversy in the U.K.

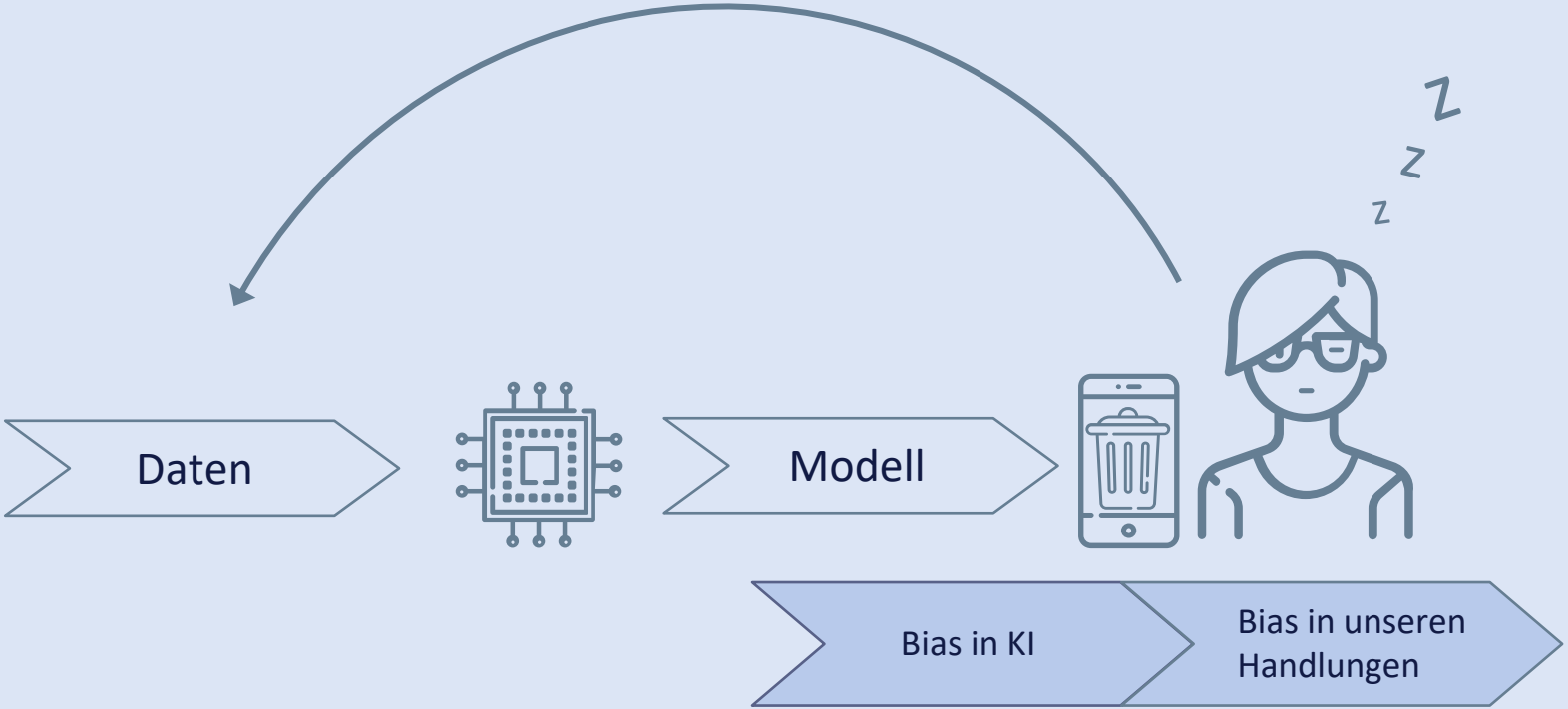
Wie entsteht Bias?

Anwendungsfehler



Wie entsteht Bias?

Anwendungsfehler



Wie entsteht Bias?

Anwendungsfehler

Super Fails | Tayminator
Wie Microsofts KI zum Nazi wurde

Daten

Österreich

Jobcenter-Algorithmus landet vor Höchstgericht

In Österreich streiten die Behörden seit Jahren über ein System, das Jobchancen von Arbeitssuchenden vorhersagen soll. Nun landet der Fall vor dem obersten Verwaltungsgericht.

Handlungen

Rechtliche Schritte gegen Bias

AI Act

- Risikomanagementsystem — Artikel 9
- Repräsentative, hochqualitative Daten — Artikel 10
  - Nutzung personenbezogener Daten, um Bias zu verhindern — Artikel 10, Abs. 5
- Transparenz & Bereitstellung von Informationen über KI-System — Artikel 13
- Menschliche Aufsicht im gesamten Prozess — Artikel 14
- Überwachung & Meldung von Vorkommnissen — Artikel 72 & 73
- Recht auf Erläuterung für Betroffene — Artikel 86

Technische Lösungsansätze

Ad Hoc: Interpretierbarkeit/White Box-Modelle

Arbeitsweise der Modelle ist im Voraus bekannt.

- Basieren auf logischen Regeln
- Entscheidungsfindung ist deutlich

IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offenses	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest.	

Rudin, 2019

Technische Lösungsansätze

Ad Hoc: Interpretierbarkeit/White Box-Modelle

Vorteile

- Transparenz
- Nachvollziehbare Ergebnisse
- Fehlersuche und –behebung ist einfacher
- Reduzierung von Risiken
- Reduzierung von Kosten

Nachteile

- Vereinfachungen
- Häufig geringere Genauigkeit
- Braucht mehr Datenaufbereitung
- Begrenzte Flexibilität bei neuen Aspekten
- Kommerzielle Interessen

Technische Lösungsansätze

Post Hoc: Erklärbarkeit/XAI

Arbeitsweise der Black Box-Modelle wird im Nachhinein erklärt.

- **LIME** (Local Interpretable Model-Agnostic): Verändert den Input und beobachtet den Output für einzelne Instanzen; erstellt aus den Erkenntnissen ein interpretierbares Modell.
- **SHAP** (SHapley Additive exPlanations): Berechnet, welchen Beitrag zur Vorhersage einzelne Merkmale liefern.

Technische Lösungsansätze

Post Hoc: Erklärbarkeit/XAI

Vorteile

- Effiziente Black Box-Modelle können weiter genutzt werden

Nachteile

- Unklar, ob Erklärung komplett richtig ist

Weitere Lösungsansätze

Überprüfungen, Tests und Metriken

- Fairness
- Robustheit

Vielfältige, repräsentative Datensätze

- Erweiterung der Trainingsdaten
- Erkennen und Entfernen verzerrender Korrelationen

Weitere Lösungsansätze

Diverse Entwicklungsteams

- Teammitglieder aus verschiedenen Gruppen
- Teammitglieder mit verschiedenen Expertisen

Bewusstseinsbildungen

- Schulungen
- Eigenverantwortung bei der Entwicklung und Nutzung von KI

Algorithms are vulnerable to inherit the flaws of the system they are designed to fix.

Hecht, 2020

If you don’t confront the social issues involved, no amount of technology is going to improve a situation. We can’t solve social problems with engineering solutions.

Tse, Esposito, Goh, 2019

Diskussion

- Wird in der Hochschulforschung darauf geachtet, dass erhobene Daten divers sind? Wo liegen Schwierigkeiten, wo gibt es Potenzial für Verbesserungen?
- Ist es ethisch in Ordnung, wenn wir absichtlich Bias zeigen oder generieren zu Lernzwecken?
- Was sind weitere Beispiele von Bias-reproduzierenden Inhalten um uns herum?
- Wie gehen wir um, bzw. sollten wir umgehen, mit historischen Forschungsdaten die Verzerrungen enthalten?
- Wir alle haben Vorurteile in uns - können wir Bias überhaupt verhindern? Wir könnten wir das erreichen?
- Gehen die Vorgaben des AI Acts weit genug? Oder zu weit?
- Für wen muss eine Erklärung verständlich sein? Nur für Expert*innen oder für alle Nutzer*innen einer KI?
- Was brauchen Sie für vertrauenswürdige KI – wann vertrauen Sie KI?

Quellen

in chronologischer Reihenfolge

- Definition KI: https://www.europarl.europa.eu/topics/de/article/20200827STO85804/was-ist-kunstliche-intelligenz-und-wie-wird-sie-genutzt abgerufen 23.04.2025
- Microsoft Chatbot Tay: https://www.arte.tv/de/videos/105664-001-A/super-fails-tayminator/ abgerufen 24.04.2025
- Amazon Bewerber*innen Algorithmus: https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/ abgerufen 23.04.2025
- AI grading system in UK: https://towardsai.net/p/artificial-intelligence/uks-failed-attempt-to-grade-students-by-an-algorithm abgerufen 25.04.2025
- Gender and Skin-type Bias: https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 abgerufen 23.04.2025
- Jobcenter Österreich: https://netzpolitik.org/2021/oesterreich-jobcenter-algorithmus-landet-vor-hoechstgericht/ abgerufen 23.04.2025
- Machine Bias: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing abgerufen 25.04.2025
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206-215.
- AI Act Explorer: https://artificialintelligenceact.eu/ai-act-explorer/ abgerufen am 07.05.2025
- LIME & SHAP: https://www.datacamp.com/de/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models abgerufen 28.04.2025
- Hecht, Yannique (2020). UK's Failed Attempt to Grade Students by an Algorithm. *Towards AI:* https://towardsai.net/p/artificial-intelligence/uks-failed-attempt-to-grade-students-by-an-algorithm%20 abgerufen am 23.04.2025
- Tse, T. C., Esposito, M., & Goh, D. (2019). The AI Republic: Building the nexus between humans and intelligent automation. S.l.: Lioncrest Publishing.
- Folienvorlage von Slidesgo; Icons von Flaticon

Marieke Prien

m.prien@hs-osnabrueck.de

Diskriminierende KI? Entstehung und Lösungsansätze von Bias in KI-Systemen