Lernzettel

Fairness, Gerechtigkeit und algorithmische Verzerrungen: Diagnose, Bewertung und Gegenmaßnahmen

Universität: Technische Universität Berlin
Kurs/Modul: Informatik und Gesellschaft
Erstellungsdatum: September 19, 2025

Lernzettel: Fairness, Gerechtigkeit und algorithmische Verzerrungen: Diagnose, Bewertung und Gegenmaßnahmen

(1) Grundbegriffe und Kontext. Fairness bezieht sich auf gerechte Behandlung und gleichberechtigte Berücksichtigung von Personen oder Gruppen in IT-Systemen. Gerechtigkeit umfasst normative Bewertungen darüber, wie Vorteile und Kosten verteilt werden sollten. algorithmische Verzerrungen (Bias) entstehen, wenn Modelle Entscheidungen treffen, die bestimmte Gruppen benachteiligen oder bevorzugen, etwa durch verzerrte Daten oder problematische Zielgrößen.

(2) Ursachen algorithmischer Verzerrungen.

- Datenbias: Trainingsdaten spiegeln historische Ungleichheiten oder systematische Vorurteile wider.
- Modell-bias: Modellannahmen begünstigen bestimmte Muster oder Gruppen.
- Auswertungs- und Feedback-Schleifen: Nutzung von Entscheidungen beeinflusst zukünftige Daten.
- Kontextuelle Verzerrungen: rechtliche, kulturelle oder wirtschaftliche Rahmenbedingungen verändern Wahrheiten.

(3) Diagnostische Ansätze und Messgrößen. Wichtige Konzepte, die oft zusammen verwendet werden, um Fairness zu bewerten:

- Demografische Parität (Demographic Parity): $P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1)$ für alle Gruppen (A=a), wobei $\hat{Y}$ der vorhergesagte Mutmaßungswert ist.
- Gleiche Fehlerraten (Equalized Odds): $P(\hat{Y} = 1 \mid Y = y, A = a) = P(\hat{Y} = 1 \mid Y = y)$ für alle $y \in \{0, 1\}$ und Gruppen a.
- Kalibrierung (Calibration): $P(Y = 1 \mid \hat{Y} = p, A = a)$ ist unabhängig von $A = a$ für alle Wahrscheinlichkeitswerte p.

(4) Bewertungskriterien: ethisch, rechtlich, ökonomisch.

- Ethik: Respekt vor Autonomie, Privatsphäre, Gleichheit und Würde; Minimierung von schädlichen Verzerrungen.
- Rechtliches: Datenschutz, Diskriminierungsverbote, Transparenzpflichten; Berücksichtigung von Grundrechten.
- Ökonomie: Effizienz vs. Gerechtigkeit, Akzeptanz, Reputationsrisiken, Kosten von Gegenmaßnahmen.

(5) Gegenmaßnahmen – drei Ebenen der Intervention.

- Prä-Processing (Datenebene): Datenbereinigung, Regewichtung, Entfernen sensibler Attribute unter gebotenem Kontext, Sampling-Strategien, Datenaugmentation.

- In-Processing (Modell-Ebene): Fairness-Regularisierung, constraint-basierte Optimierung, adversarielle Debiasing-Ansätze, Transparenzanforderungen in der Modellarchitektur.
- Post-Processing (Ausgabe-Ebene): Anpassung der Vorhersagen oder Wahrscheinlichkeiten, um vorgegebene Fairness-Ziele zu erfüllen (z. B. Schwellenwerte-Anpassung).

(6) Praktische Orientierung – Leitfragen für die Praxis.

- Welche Gruppen könnten durch das System benachteiligt werden und warum?
- Welche Fairness-Ziele sind legitim in diesem Kontext (Ethik, Recht, Ökonomie)?
- Wie lassen sich Datenschutz, Transparenz und Sicherheit mit Fairness-Gefährdungen in Einklang bringen?
- Welche Stakeholder sind betroffen und wie werden Interessen abgewogen?

(7) Fallbeispiel (knappe Skizze). Ein Recruiting-Algorithmus verwendet historische Einstellungsdaten. Die Analyse zeigt, dass eine Gruppe aufgrund früherer Verzerrungen weniger positive Entscheidungen erhält. Lösungsansätze: maskieren sensibler Merkmale in der Vorverarbeitung (Pre-Processing) und/oder fügen Fairnessregularisierung im Lernprozess hinzu (In-Processing); anschließendes Post-Processing, um gewünschte Parität zu erreichen. Wichtig ist, Rechts- und Ethikrahmen zu beachten und Transparenz gegenüber Bewerbenden zu wahren.

(8) Verbindung zu Kurszielen. Dieser Teil verbindet Ethik, Berufsethik, Grundrechte, Datenschutz, Sicherheit, Ökonomie, Nachhaltigkeit und wissenschaftliche Arbeitsweise mit der Analyse von soziotechnischen IT-Systemen. Die Konzepte werden auf konkrete Fallbeispiele übertragen und kritisch bewertet.

(9) Übungen (leicht- bis mittelgradig, ohne Enumerationen).

- Welche Fairness-Metrik wäre in einem Gesundheitsanwendungskontext sinnvoll, und warum?
- Welche Gegenmaßnahme würden Sie in einem Kreditvergabe-System bevorzugen: Pre-, In- oder Post-Processing? Begründen Sie Ihre Wahl.
- Diskutieren Sie potenzielle Konflikte zwischen Kalibrierung und Demografie-Parität in einem personalisierten Werbeanzeige-System.