Spickzettel Biases in LLM

Unter dem von Kahnemann und Tversky (1974) geprägten Begriff *Bias* versteht man systematisch auftretende Wahrnehmungsverzerrungen oder Abkürzungen im Denken (*Heuristiken*). Herm, Janiesch und Fuchs (2022) konnten in einer strukturierten Literaturanalyse die unten aufgeführten Biases innerhalb von KI-Ausgaben identifizieren.

Label Definition Bias – Problematische oder unscharfe Definitionen von Klassen/Labels führen zu verzerrten Ergebnissen.
Optimization Bias – Die Optimierungsfunktion bevorzugt bestimmte Ergebnisse (z. B. Genauigkeit statt Fairness).
Specification Bias – Modell oder Fragestellung ist falsch oder unvollständig formuliert.
Feature Selection Bias – Wichtige Merkmale fehlen oder irrelevante Features werden ausgewählt.
Aggregation Bias – Daten unterschiedlicher Gruppen werden zu stark verallgemeinert.
Classifier Bias – Der gewählte Algorithmus bringt systematische Verzerrungen mit sich.
(Hyper-)Parameter Bias – Modellparameter begünstigen bestimmte Ergebnisse.
Dataset Bias – Das gesamte Dataset ist nicht repräsentativ.
Sample Bias – Stichprobe bildet die Realität nicht korrekt ab.
Annotation Bias – Subjektivität oder Fehler bei der Datenbeschriftung.
Measurement Bias – Messmethoden liefern systematisch verzerrte Werte.
Temporal Bias – Daten sind zeitlich veraltet oder nicht übertragbar.
Historical Bias – Alte gesellschaftliche Diskriminierungen spiegeln sich in den Daten.
Distribution Bias – Ungleichgewicht in der Häufigkeit bestimmter Gruppen oder Merkmale.
Demographic Bias – Bestimmte Gruppen sind über-/unterrepräsentiert.
Social Influence Bias – Verhalten wird durch Trends, Meinungen oder Gruppendruck verzerrt.
Fairness Bias – Modellentscheidungen sind unfair gegenüber bestimmten Gruppen.
User Bias – Nutzer bringen ihre eigenen Vorurteile ein.
User Interaction Bias – Verzerrungen entstehen durch die Art, wie Nutzer mit dem System interagieren.
Personal Bias – Subjektive Meinungen oder Perspektiven einzelner Personen prägen das Ergebnis.
Evaluation Bias – Die Bewertungsmethode bevorzugt bestimmte Ergebnisse.
Cause-Effect Bias – Korrelation wird fälschlich als Kausalität interpretiert.
Co-Occurrence Bias – Zufällige Häufungen werden als Muster fehlinterpretiert.
Masking Bias – Wichtige Effekte werden durch Aggregation oder Modellierung verdeckt.
Uncertainty Bias – Unsicherheit wird nicht korrekt berücksichtigt.
Inherited Bias – Verzerrungen werden aus Vorgängermodellen übernommen.
Funding Bias – Finanzierungsquelle beeinflusst Fragestellung, Datenauswahl oder Ergebnisse.
Team Bias – Zusammensetzung des Teams prägt Entscheidungen und blinde Flecken.

Herm, L.-V., Janiesch, C., & Fuchs, P. (2022). *Der Einfluss von menschlichen Denkmustern auf künstliche Intelligenz – Eine strukturierte Untersuchung von kognitiven Verzerrungen. HMD Praxis der Wirtschaftsinformatik, 59*(4), 556–571. https://doi.org/10.1365/s40702-022-00844-1 .
Tversky, A., & Kahneman, D. (1974). *Judgment under uncertainty: Heuristics and biases. Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124