Literature Review

Der Themenbereich, aus dem unsere Arbeit kommen soll, ist der der „Fair AI“ oder auch „AI Fairness“. Als Forschungsthema haben wir uns Bias Detection beziehungsweise Unfairness Detection in Datensätzen herausgesucht, um dieses nun allerdings weiter konkretisieren zu können, haben wir eine strukturierte Literaturanalyse durchgeführt. Verwendet haben wir dabei zwei Literaturdatenbanken, Scopus und IEEE Xplore. Die erste gestartete Suche hat 426 Ergebnisse in Scopus zurückgeliefert, da das zu viele waren haben wir unsere Search Strings nochmal angepasst. Insgesamt haben wir drei Suchen gestartet, wodurch wir kumuliert 205 Ergebnisse bekommen haben. Die Suchanfragen sehen wie folgt aus:

( TITLE-ABS-KEY ( ai AND fair AND adjust AND data ) OR TITLE-ABS-KEY ( ai AND fairness AND adjust AND data ) OR TITLE-ABS-KEY ( ai AND reduce AND bias AND adjust AND data ) OR TITLE-ABS-KEY ( ai AND ensure AND fairness AND adjust AND data ) OR TITLE-ABS-KEY ( ai AND reduce AND bias AND data ) OR TITLE-ABS-KEY ( ai AND prejudice AND adjust AND data ) OR TITLE-ABS-KEY ( ai AND prejudice AND reduce AND data ) )

( TITLE-ABS-KEY ( fairness ) AND TITLE-ABS-KEY ( ai ) AND TITLE-ABS-KEY ( data ) OR TITLE-ABS-KEY ( "ensure fairness" ) AND TITLE-ABS-KEY ( ai ) AND TITLE-ABS-KEY ( data ) OR TITLE-ABS-KEY ( ai ) AND TITLE-ABS-KEY ( fair ) AND TITLE-ABS-KEY ( data ) AND TITLE-ABS-KEY ( manipulation ) )

TITLE-ABS-KEY ( "Fair AI" )

Die erste Suchanfrage hat sich besonders auf die Begriffe „ai“, „fair“, „adjust“, „data“, „reduce“ und „bias“ konzentriert. Diese Suchanfrage hat in Scopus 92 und in IEEE Xplore 70 Ergebnisse zurückgeliefert. Für die Suche in IEEE Xplore musste die Suchanfrage der IEEE Xplore Suchsyntax entsprechend angepasst werden.

Die zweite Suchanfrage bezieht die Begriffe „ensure fairness“ und „manipulation“, anders als die erste Anfrage, mit ein, hat allerdings nur ein Ergebnis in Scopus und gar kein Ergebnis in IEEE Xplore zurückgeliefert.

Die dritte und letzte Suchanfrage beinhaltet nur die Phrase „Fair AI“ und ist damit sehr allgemein. Die Anfrage hat in Scopus 37 und in IEEE Xplore 5 Ergebnisse zurückgeliefert. Für die Suche in IEEE Xplore musste diese Suchanfrage entsprechend angepasst.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Scopus | IEEE | ∑ |
| Zurückgelieferte Ergebnisse | 94+1+37 = 130 | 70+0+5 = 75 | 207 |
| Duplikats-Eliminierung | -0 | -2 | 205 |
| Titel-Eliminierung | -115 | -66 | 24 |
| Abstract-Eliminierung | -4 | -5 | 15 |
| Verfügbarkeits-Eliminierung | -4 | -0 | 11 |
| Fulltext-Eliminierung | -3 | -1 | 7 |
| Forward & Backward Search | +1 |  | 8 |

Artikel Konzepte

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Fairness Berechnung | (Potential) Bias Detection | Mitigating Bias | Methoden für tabulare Datensätze | Anmerkungen |
| Managing Bias in AI |  | x | x |  | Sehr allgemein, geht nicht in die technische Tiefe |
| Attributing Fair Decisions with Attention Interventions |  |  | x | x | Bezieht sich weniger auf Datensätze, sondern mehr auf KI Modelle |
| Designing fair AI for managing employees in organizations | x | x | x |  | Beschreibt ein theoretisches Framework für Fair AI für Unternehmen |
| Fairness in Data Wrangling | x |  |  |  | Sehr mathematisch |
| Fairness through Awareness | x |  |  |  | Sehr mathematisch |
| Time to Assess Bias in Machine Learning Models for Credit Decisions |  | x |  |  | Beschreibt ein Verfahren für die Risikobewertung der fairen Kreditvergabe |
| Measuring and Mitigating Bias in AI-Chatbots |  | x | x |  | Entwicklung eines “Chatbot Bias Assessment Frameworks“ |
| Does Data Repair Lead to Fair Models? Curating Contextually Fair Data To Reduce Model Bias | x | x | x |  | Sehr mathematisch |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |