Versuchsplan

# Beschreibung des Artefakts

Das Thema unserer Arbeit soll aus dem Themengebiet der Fair AI kommen, wir haben uns dabei für die Ermittlung von Verzerrungen (Bias) und Quellen potenzieller Diskriminierung in tabellarischen Datensätzen entschieden. Dieses Thema ist eher technischer Natur, daher haben wir uns für einen design-orientierten Forschungsansatz für den Verlauf dieser Arbeit entschieden. Geplant ist, ein Artefakt zu entwickeln, das tabellarische Daten und je nach Ausprägung des Artefakts einen Kontext des Datensatzes als Input nimmt und als Output gefundene Probleme im Datensatz im Bezug auf Fairness zurückgibt. Sollten keine Probleme gefunden worden sein, wird dementsprechend eine Meldung zurückgegeben, dass es in diesem Datensatz kein Problem mit Diskriminierung oder anderen Arten von Unfairness zu geben scheint. Der als Input übergebene Kontext wird durch das Artefakt begrenzt werden, also wird es nur eine gewisse Auswahl von Datensatzarten geben, die von dem Artefakt akzeptiert werden, für alle anderen Datensätze, die als Input übergeben werden, die keiner der vom Artefakt vorgegebenen Kategorie angehören, kann kein zuverlässiges Resultat garantiert werden. Die folgende Abbildung fasst die beschriebene Funktion des Artefaktes visuell nochmal zusammen.

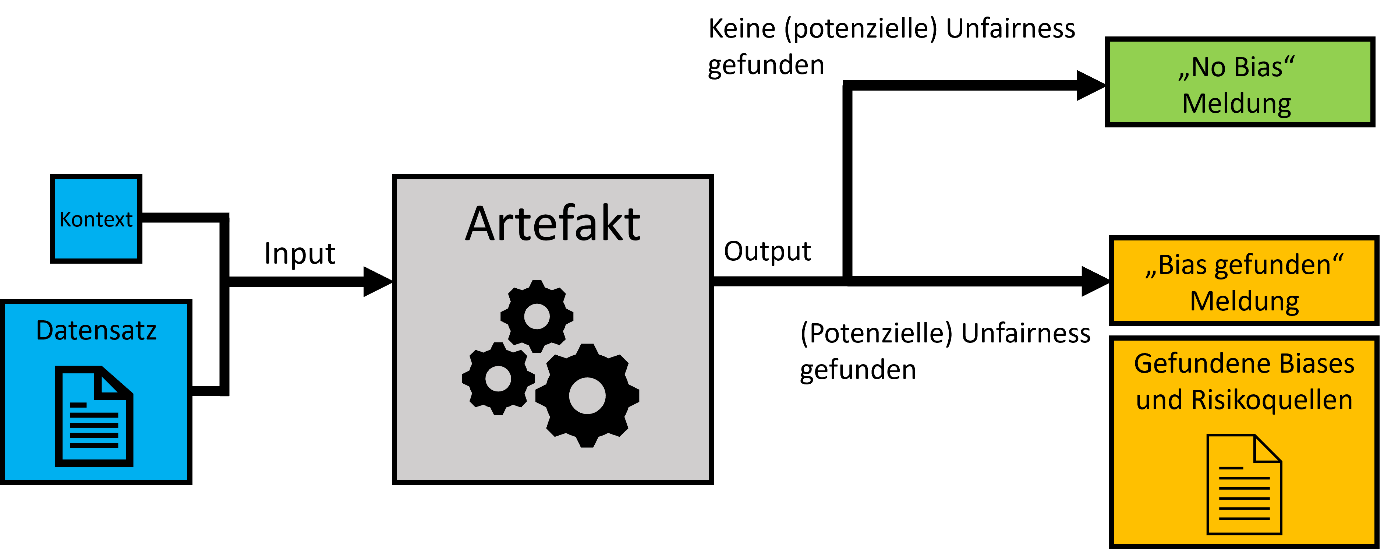


Abbildung Artefaktsfunktionsweise (abstrahiert)

# Geplante Experimente

Die Funktionsweise und den Nutzen des beschriebenen Artefakts gilt es natürlich zu ermitteln und zu testen, dafür haben wir uns einige Experimente erdacht. Wir unterscheiden nach zwei Schemata für die Experimente.

Beim Ersten übergeben wir dem Artefakt einen Datensatz, von dem wir wissen, dass in ihm mindestens eine Art von Verzerrung oder Unfairness enthalten ist, zusammen mit seinem Kontext. Das Ergebnis des Outputs wird darauffolgend mit dem Wissen über die Verzerrungen und Unfairness des Datensatzes verglichen. Sollten wir beispielsweise einen Datensatz zur Kreditvergabe, in dem Frauen und dunkelhäutige Personen diskriminiert werden, indem sie keine bis kaum Kredite bekommen, nehmen und dem Artefakt mit dem Kontext „Kreditvergabe“ übergeben, so wird darauffolgend das Ergebnis des Artefakts mit der bekannten Ungerechtigkeit des Datensatzes verglichen. Wenn das Ergebnis dem bekannten Wissen entspricht, spricht das für die Funktionsweise und den Nutzen des Artefakts.

Das zweite Schema benötigt für eine Ausführung des Experiments zwei Datensätze, einen, der wie beim ersten Schema mit bekannten Verzerrungen und Unfairness versehen ist und eine Form des Datensatzes, der von seinen Verzerrungen und seiner Unfairness bereinigt wurde. Der erste Teil des Experiments folgt dem Schema, das zuerst beschrieben wurde, im zweiten Teil wiederholen wir den ersten Teil, nur verwenden wir nun den bereinigten Datensatz. Sollten in beiden Fällen die Ergebnisse gleich sein, der bereinigte Datensatz aber garantiert bereinigt sein, so würde das gegen die Funktionsweise und den Nutzen des Artefakts sprechen und Optimierungsbedarf aufzeigen, im Idealfall jedoch würde das Artefakt beim bereinigten Datensatz die Meldung ausgeben, dass keine Verzerrungen gefunden wurden. Sollten wir nicht in der Lage sein, solche Datensatzpaare zu finden, so würden wir das Experiment abwandeln, indem wir nur faire Datensätze nehmen und das Ergebnis mit der passenden Meldung vergleichen.

Beide Schemata erfordern Datensätze, über die wir bestimmtes Metawissen benötigen. Sollte es nicht möglich sein, solche Datensätze zu finden, so würden wir selbst Datensätze erzeugen, wobei das die Aussagekraft des Experimentes stark reduzieren würde. Nach Datensätzen werden wir auf Kaggle.com suchen.

Der nächste Schritt der Versuchsreihe ist es, abhängig davon, wie durchführbar die davor gehenden Experimente waren, dem Artefakt Datensätze zu geben, über die keine Metainformationen über Ungerechtigkeiten, Verzerrungen und Unfairness vorliegen.