Information »KI und Bias«

Bias in der KI bezieht sich auf systematische Fehler in den Vorhersagen oder Entscheidungen, die von KI-Modellen aufgrund von Vorurteilen in den Trainingsdaten gemacht werden.

Daten-Bias: Wenn die Daten, die zum Trainieren eines KI-Modells verwendet werden, nicht repräsentativ für die Realität sind und das Modell verzerrte Vorhersagen macht.

Algorithmischer Bias: Wenn beispielsweise Algorithmen, die zur Vorhersage von Kriminalität entwickelt wurden, aufgrund von historischen Polizeidaten, die rassistische Vorurteile widerspiegeln, bestimmte ethnische Gruppen überproportional ins Visier nehmen.

Bestätigungs-Bias: Wenn KI-Modelle dazu neigen, Vorhersagen zu machen, die ihre ursprünglichen Trainingsdaten bestätigen, auch wenn neue Daten etwas anderes nahelegen.