Telekom CR-Wissen

Bias und Fairness in der KI

Wenn KI heutzutage in die Schlagzeilen gerät, dann nur allzu oft wegen Problemen mit Voreingenommenheit (bias) und Gerechtigkeit (fairness). Das passiert immer dann, wenn KI-Modelle systematisch bestimmte Gruppen oder Einzelpersonen benachteiligen. Die T-Labs der Deutschen Telekom AG arbeiten daher mit Magenta Austria daran, die Verzerrungen zu überprüfen, bevor ein KI-Modell zum Einsatz kommt.

Wenn die Entscheidung eines Modells von Voreingenommenheit beeinflusst ist, hat das Modell kein realistisches oder vollständiges Bild der Umgebung gelernt, in der es nach der Entwicklung eingesetzt werden soll. Wenn ein solches Modell in die Produktion geht, wird es unterschiedliche Situationen nicht gleich gut bewerten. Eine solche erlernte Verzerrung ist in der Regel darauf zurückzuführen, dass die dem Lernprozess zur Verfügung gestellte Datenbasis nicht ausreichend vollständig oder ausgewogen ist. Die Gründe für ungewollt unvollständige/unausgewogene Daten können vielfältig sein, ebenso wie die Gefahren, die damit verbunden sind: Hier gibt es die Interaction bias („Interaktionsverzerrung“), bei der zum Beispiel Menschen Interaktionsverzerrungen erzeugen, wenn sie mit KI-Systemen interagieren oder absichtlich versuchen, diese zu beeinflussen und verzerrte Ergebnisse zu erzielen. Ein Beispiel dafür ist, wenn Leute absichtlich versuchen, Chatbots schlechte Sprache beizubringen.
Die Selection bias („Auswahlverzerrung“), bei der zum Beispiel in Bewerber-Auswahlverfahren bevorzugt Männer ausgewählt werden, weil das KI Modell zuvor mit hauptsächlich männlichen Daten „angelernt“ wurde und es mit mehr weiblich konnotierten Hobbys oder Fremdsprachen im Lebenslauf nicht vernünftig umgehen konnte.
Beim implicit bias („Implizite Voreingenommenheit“) geht es um unbewußte Voreingenommenheit und das damit verbundene Risiko, dass die Technologie zum

Beispiel People of Color systematisch benachteiligt.

Intelligente Systeme machen ähnliche Fehler wie Menschen, allerdings auf automatisierte Weise und potenziell in größerem Umfang. Diese Fehler können sowohl für Einzelpersonen als auch für Unternehmen schwerwiegende Folgen haben – von Umsatzeinbußen über rechtliche Konsequenzen bis hin zu Marken- und Rufschädigung. Wenn personen-bezogene Daten verwendet werden und/oder die technologische Teilhabe verschiedener Nutzergruppen sichergestellt werden soll, muss genau darauf geachtet werden, ob Datenverzerrungen Menschen möglicherweise ausschließen oder diskriminieren können. Um gängige Verzerrungsmessungen und dazu passende Gegenmaßnahmen für Telekom-Anwendungsfälle zu bewerten, arbeiteten die T-Labs gemeinsam mit Magenta Austria an deren Churn- und Propensity-Modellen, um sicherzustellen, dass bestimmte Personengruppen nicht systematisch vernachlässigt werden.

© 2024 Deutsche Telekom AG