KI-EINSATZ IM BANKING

FAIRNESS IST TRUMPF

Künstliche Intelligenz (KI) ist eine bedeutende Schlüsseltechnologie des digitalen Wandels in der Finanzwelt und zukünftig wichtiger Bestandteil neuer Geschäftsmodelle, Anwendungen, Prozesse und Produkte. Der Autor warnt in diesem Beitrag jedoch: Die KI ist nicht fehlerfrei, und sie sollte entsprechend überwacht werden.

Technologiegestützte Finanzdienstleistungen haben nicht nur die etablierten Anbieter (Banken, Versicherungen, Wertpapierfirmen, Pensionsfonds etc.) im Angebot, diese werden zunehmend auch von großen Nicht-Finanzunternehmen offeriert, einschließlich der sogenannten BigTechs, wie Internet-Unternehmen, Geräteherstellern, Handelsplattformen und Telekommunikationsbetreibern.

Mit der seit November 2022 frei zugänglichen Anwendung Chat-GPT von OpenAI erkennen Verbraucher und Unternehmen neue Vorteile und Anwendungsmöglichkeiten von generativer KI. Diese kann neue Inhalte in Form von Text, Bild, Video, Audio und Softwarecode generieren und ermöglicht die Kommunikation in natürlicher Sprache ohne Vorkenntnisse. Wenn zunehmend eine Software qualitativ hochwertige Inhalte in fast allen Lebensbereichen produziert, hat dies enorme Auswirkungen auf sämtliche Lebensbereiche von Verbrauchern, Unternehmen und der gesamten Gesellschaft.

KI trägt jedoch das Risiko eines systematischen Prognosefehlers (Bias) in sich. Die algorithmische Voreingenommenheit (Algorithmic Bias) in KI-Anwendungen kann den erwarteten Nutzen von KI-Anwendungen negativ beeinträchtigen und in Einzelfällen sogar das komplette System infrage stellen, z. B. bei vorliegender Diskriminierung, unlauteren Praktiken oder Verlust der Autonomie über das System.

Kreditinstitute müssen sich bei der Entwicklung von KI-Anwendungen mit dem Thema auseinandersetzen, weil die Systeme in der Regel große Mengen personenbezogener Daten analysieren, um Korrelationen zu erkennen und Zusammenhänge abzuleiten. Dabei können die Auswirkungen der Entscheidung auf den Menschen beträchtlich sein, wie beispielsweise der Zugang zu Krediten, Zinsen, Gebühren etc.

Das Problem der algorithmischen Voreingenommenheit

KI kann eine unbeabsichtigte, unerwünschte Verzerrung erzeugen und damit gegen grundlegende Rechte und/oder zu Ergebnissen und Auswirkungen führen, die als ungerecht empfunden werden. Die Ursachen sind vielfältig. So können sich die Daten nur auf eine bestimmte Gruppe oder Klasse von Objekten/ Personen beziehen, oder die Daten sind unvollständig, fehlerhaft oder gar nicht vorhanden.

Zum Beispiel stand die Suchmaschine Google in der Kritik: Sie würde bei Anzeigen für Führungspositionen weiße Männer bevorzugen im Vergleich zu afroamerikanischen Frauen. Auch Amazon musste sich mit dem Vorwurf auseinandersetzen, nachdem der in der KI-Anwendung zur Rekrutierung neuer Mitarbeitenden implementierte Rekrutierungsalgorithmus eine „geschlechtsspezifische Voreingenommenheit" gegenüber Frauen aufwies.

Im Finanzbereich besteht das Problem der algorithmischen Voreingenommenheit insbesondere im vorhandenen Datenbestand, weil dieser im Wesentlichen auf von den Kunden selbst genannten Angaben basiert. Es ist daher grundsätzlich von der Existenz eines Bias auszugehen. Die Herausforderung beim Einsatz von KI-Anwendungen besteht darin, Datenverzerrungen im Datenbestand zu erkennen und methodisch zu eliminieren.

Geschlechtsbezogener Verzerrungseffekt

Gender Bias bezeichnet das Auftreten von systematischen Fehlern aufgrund inadäquater Berücksichtigung des Aspekts Geschlecht. Folgende Fragestellungen sind dabei relevant:

1. Kann eine Gleichheit bzw. Ähnlichkeit von Frauen und Männern in bestimmten Bereichen angenommen werden, obwohl relevante Geschlechterunterschiede existieren? Der Gender Bias kann sich dadurch darstellen, dass diese Geschlechterunterschiede entweder nicht als Differenzierungsvariablen berücksichtigt werden oder nicht als mögliche Erklärungsvariablen untersucht und diskutiert werden.
2. Bestehen Unterschiede zwischen den Geschlechtern, obgleich objektiv keine bestehen, oder findet eine Überbetonung der Variable Geschlecht im Vergleich zu anderen Faktoren (z. B. Ethnie, sozioökonomischer Status) statt, die nicht gerechtfertigt ist?

Bias hinsichtlich Rasse und Alter

Bartlett et al. fanden heraus, dass schwarze und hispanische Amerikanerinnen und Amerikaner um 7,9 bzw. 3,6 Basispunkte höhere Zinsen für Konsumenten- und Immobilienkredite bezahlen.

Ein ähnliches Beispiel zeigt sich häufig bei Bilderkennungs-Software auf Basis neuronaler Netze. Bei der Rekonstruktion und Zuordnung einzelner Bildteile zu einem bestimmten Objekt oder einer Person führt die hohe Ähnlichkeit von Bildelementen – wie Hautfarbe oder Form von Nase und Augen – zu Klassifikationsproblemen und letztlich zu falschen Ergebnissen. Besonders beim Einsatz von KI bei der Personalauswahl stellt sich das Problem, um auf unbewussten Vorurteilen von Menschen basierende Diskriminierung zu überwinden. Nach Bertrand & Mullnaithan stellt das Screening von Lebensläufen bei Auswahlverfahren insbesondere eine Quelle für Diskriminierung dar.[1]

Ursachen liegen in menschlicher Voreingenommenheit

Menschliches Urteilsvermögen spiegelt sich in den Systemen wider. Menschen entwickeln Algorithmen, und somit werden die vom Entwickler bevorzugten Parameter, Konfigurationen etc. zwangsläufig im KI-Modell abgebildet und beeinflussen das Ergebnis.

Eine weitere Ursache für algorithmische Voreingenommenheit liegt im Aufbau der KI-Modelle. Es handelt sich in der Regel um Black-Box-Modelle, die als Datensenken und -quellen existieren, ohne dass es eine Erklärung dafür gibt, was darin enthalten ist. Für den Anwender bzw. die Anwenderin ist es nicht möglich, zu hinterfragen, wie die Black-Box-Modelle zu einem Ergebnis kommen. Selbst aus gleichen Input-Daten können unterschiedliche Ergebnisse resultieren.

Unzureichende Trainingsdaten begründen ebenfalls algorithmische Verzerrungen. Wenn die zum Training des Algorithmus verwendeten Daten für einige Personengruppen repräsentativer sind als für andere, ergeben sich systematisch Verzerrungen für die unterrepräsentierte Gruppe. Eine implementierte Verzerrung liegt vor, wenn ein Algorithmus sich durch einen Selbstlernprozess modifiziert.

Ein wesentlicher Aspekt betrifft die Verzerrung der Daten. Die für eine KI-Anwendung genutzten Daten stammen in der Regel aus mittels Data Mining bestehenden Datenbeständen. Fehlende, unvollständige und/oder unklare Daten können die Algorithmen bzw. die Ergebnisse verfälschen. Dies gilt auch, falls das Spektrum der ausgewählten Daten hinsichtlich der Breite bzw. Vielfalt zu eng ist.

Eine weitere Quelle für Verzerrungen, insbesondere in externen Daten, resultiert aus fehlendem Wissen hinsichtlich Qualität und Herkunft der verwendeten Daten. Datensätze können sensible Attribute enthalten, die direkt zu einer Verzerrung und Diskriminierung bei den Ergebnissen führen. Zum Beispiel reicht die Anrede einer Person aus, um ihre ethnische Herkunft zu identifizieren.

Ein zusätzlicher Gesichtspunkt betrifft die Verzerrung durch die Aufteilung der Quelldaten in Trainings- und Testdaten. Eine Verzerrung kann entstehen, wenn das KI-Modell auf die Testdaten oder den Test trainiert wird. Verzerrungen können durch die Wiederverwendung von Code bei der Implementation entstehen. Es besteht die Gefahr, dass durch die einfache Übertragung von Code sich unbeabsichtigt Verzerrungen in anderen Systemen darstellen. Der übertragende Code ist möglicherweise unvollständig oder passt nicht in den eingefügten Kontext.

Zudem können technische Rahmenbedingungen wie Speicher- und Rechenkapazitäten technisch begründete Verzerrungen verursachen. So ist bspw. die Anzahl der auszuführenden Iterationen technisch limitiert, mit der Konsequenz, dass das KI-Modell nur eine bestimmte Anzahl von Objekten vollständig identifizieren kann.

Lösungsansätze gegen Verzerrungen in KI-Projekten

Verzerrungen in KI-Systemen betreffen den gesamten Entwicklungsprozess, ausgehend von der Selektion und Aufbereitung des Datenbestands über die Entwicklung, Implementierung und Anwendung. Die zuvor genannten Problemfelder erfordern mehr als nur technische Lösungen. Wichtiger sind Methoden und Prozesse zum Verstehen und Messen von „Neutralität".

Wie lässt sich feststellen, wann ein System fair genug ist, um freigegeben zu werden, und in welchen Situationen ist eine vollautomatische Entscheidungsfindung überhaupt zulässig? Allerdings stellt sich auch hier die grundsätzliche Frage, was neutral ist; denn jeder Mensch hat eine andere Wahrnehmung.

Was können Sparkassen und Banken tun, um Verzerrungen in KI-Projekten zu berücksichtigen? Zielführend ist die Entwicklung eines praktischen Leitfadens für KI-Projektteams, um mögliche Verzerrungsprobleme in KI-Systemen zu erkennen und zu bewerten.

Grundlegende Empfehlungen zur Erhöhung der Neutralität in KI-Lösungen sind:

- Alle Mitarbeitenden in KI-Projektteams kennen das Problem der Verzerrung und haben methodische Kenntnisse zur Identifikation entsprechender Auffälligkeiten.
- Nutzung qualitativ hochwertiger und vielfältiger Daten für Training und Test.
- Strikte Trennung von Training und Test. Dies umfasst auch die involvierten Personengruppen bzw. Endbenutzer und -benutzerinnen.

- Etablierung von Prozessen für Benchmarking-Verfahren, zur Auswahl und Validierung von Daten, zur Qualitätskontrolle sowie spezifische Anleitungen, was zu tun ist, wenn Zweifel an der Interpretation der Ergebnisse des Algorithmus bestehen.
- Durchführung von Querprüfungen zur Erkennung auffälliger Muster und Verzerrungen in Algorithmen.
- Einsatz von KI-Tools, mit denen sich Algorithmen und Modelle bewerten und auf mögliche Unfairness überprüfen lassen.

FAZIT

KI-Anwendungen müssen für alle Beteiligten transparent sein und die möglichen unbeabsichtigten Folgen von KI und ihrer Nutzung hinsichtlich Fairness, Verantwortlichkeit, Transparenz und Erklärbarkeit berücksichtigen: Unterscheidet das Modell zum Beispiel zwischen bestimmten sozialen Gruppen oder Klassen? Erwarten wir ähnliche Ergebnisse für verschiedene Untergruppen? Verstehen wir, wie das Modell funktioniert? Kann das Modell die erzeugten Ergebnisse erklären? Nur wenn die Ergebnisse nachvollziehbar sind, sind diese vertrauenswürdig. Unternehmen und KI-Entwickelnde benötigen ein tiefes Verständnis über algorithmische Verzerrungen und die damit verbundenen Risiken und Lösungen.

Autor

Prof. Dr. Dirk Neuhaus, MBA, ist Professor für Informationssysteme in Finanzdienstleistungsunternehmen an der Hochschule für Finanzwirtschaft & Management (HFM) in Bonn. Er forscht auf dem Gebiet der Künstlichen Intelligenz.

1 Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. American Economic Review, 94(4), 991-1013. https://www.doi. org/10.1257/0002828042002561.