Künstliche Intelligenz: Wenn der Algorithmus diskriminiert

Künstliche Intelligenz wird oft als neutral und unvoreingenommen betrachtet, da sie auf Algorithmen und Daten basiert. Doch diese vermeintliche Objektivität ist trügerisch: KI-Systeme lernen aus Daten, die von Menschen geschaffen und gesammelt wurden. Wenn diese Daten bereits diskriminierende Muster enthalten, kann die KI diese Muster verstärken und sogar unbewusste Vorurteile auslösen – ernstzunehmende Diskriminierung entsteht. Wir haben mit Lorenzo Medici (Head of Development bei der a&f systems ag) im Interview über die Herausforderungen für publizierende Unternehmen gesprochen.

a&f systems: Lorenzo, kann eine Künstliche Intelligenz wirklich rassistisch sein und verschiedene Gruppen diskriminieren?
Lorenzo Medici: Gerne würde ich statt der Bezeichnung «rassistisch» den Begriff «diskriminierend» verwenden. Denn das Wesen, warum KI rassistisch sein kann, bezieht sich nicht nur auf Ethnien oder Herkunft. Die soziale Schicht, das Geschlecht etc. kann sehr schnell zu diskriminierenden KI führen. Es bleibt aber zu betonen, dass KI-Systeme nicht per se diskriminierend sind. Die Gefahr besteht darin, dass KI sich diskriminierend verhält, wenn die verwendeten Daten dies entsprechend triggern.

a&f systems: Wie ist das genau zu verstehen: «wenn die Daten dies entsprechend triggern»?
Lorenzo Medici: Um diese Frage nachvollziehbar zu erklären, muss vorab der strukturelle Aufbau eines KI-Systems kurz beleuchtet werden. Ein KI-System, auch Modell genannt, besteht immer aus einem Datenset und einem mathematischen Regelwerk, wie die Daten interpretiert werden sollen. Beide Komponenten spielen eng zusammen. Es gibt nicht ein einziges mathematisches Regelwerk, das für alle Datensets ideal ist. Je nachdem, was das KI-Modell leisten soll, wird auch das entsprechende Regelwerk gewählt. Bevor die Daten aber mit dem mathematischen Modell verschmelzen, müssen diese noch entsprechend aufbereitet werden. Diesen Schritt nennt man Merkmalskonstruktion. Das ist ein äusserst wichtiger Schritt, damit man valide Modelle erhält. In den einfachsten Fällen handelt es sich bei einem Modell um statistische Grundanalysen von Daten. Das bedeutet, dass die Daten statistisch analysiert werden, damit man über dieses spezielle Datenset gewisse statistische Erkenntnisse besitzt. Benutzt man nun dieses Modell, werden die Anfragen den Erkenntnissen gegenübergestellt. Die Antwort eines KI-Systems bezieht sich also immer auf die Erkenntnisse, die das Modell mit Hilfe der eingespeisten Daten erlangt hat.

a&f systems: Sind also die vorhandenen Daten diskriminierend?
Die Daten, die in ein System eingespeist werden, sind immer existierende Daten. Dies bedeutet, dass die Daten immer den Blick in die Vergangenheit darstellen und damit implizit auch Werte früherer Zeiten widerspiegeln. In solchen Fällen ist die Forschung in der Merkmalskonstruktion gefordert, wie diese Daten für Systeme aufbereitet werden können. Es gibt viele Varianten, warum Daten diskriminierend sein können. Richtet man den Fokus auf die Daten, erkennt man sehr schnell, dass es sich nicht immer um intervallskalierte Daten handelt, also Grösse, Länge, Gewicht etc. – Daten, die schon in einer numerischen Skala vorliegen. Sehr oft haben wir sogenannte kategoriale Daten, so wie Mann-Frau oder Haus-Bürogebäude. Diese kategorialen Daten bilden oft implizit kulturelle Werte ab. Man sieht das am Beispiel von Mann-Frau. Mittlerweile unterscheidet man mehr Geschlechter, welche vor 50 Jahren nie zur Disposition standen. Zudem unterliegen die kategorialen Daten immer einer Semantik. Die kann sich wiederum je nach Sprache strukturell ändern. Auf der anderen Seite liegen Daten auch in Text- und in Bildform vor. Damit aber ein mathematisches Modell die Daten verarbeiten kann, ist es zwingend notwendig, diese Daten so zu transformieren, dass sie in den für die Mathematik notwendigen Zahlenraum übertragen werden können. Diesen Arbeitsschritt, die Merkmalskonstruktion, muss von Datenanalysten gemacht werden. In Bezug auf Rassismus lassen sich nun verschiedene Gefahren bei der Merkmalskonstruktion erkennen:

Eine Gefahr ist der Algorithmic Bias, also die algorithmische Voreingenommenheit, die zur Verzerrung von KI-Systemen führen kann. Beispiel: Wenn das Rückfallrisiko von Straftätern in den USA bewertet wird, um z. B. Richtlinien für Haftstrafen festzulegen und das Modell dann mit historischen Daten aus Strafverfolgungsfällen trainiert wird, die bereits rassistische Verzerrungen aufweisen. Wir wissen beispielsweise, dass schwarze Menschen und Latinos in den Gefängnissen im Verhältnis zur Bevölkerungsgesamtverteilung überrepräsentiert sind. Weisse Menschen hingegen sind unterrepräsentiert. Das ist ein Fakt, jedoch werden in dieser Betrachtungsweise die sozioökonomischen Umstände, Bildungsniveau, Zugang zu Bildung etc. nicht beachtet. Das Modell, das nun aufgrund dieser Daten generiert wird, erkennt, dass das Rückfallrisiko von ethnischen Merkmalen abhängig ist und so gewissen Ethnien eine höhere Haftstrafe zuweist. Eine analoge Anwendung haben natürlich auch Versicherungen, die ihre Prämien mitunter den demografischen Daten angepasst haben. Es ist anzunehmen, dass dort möglichst umfassende Parameter in Modelle einfliessen, so dass die Staatsangehörigkeit nicht der Hauptparameter für die Prämienberechnung darstellt.

Ein anderes Beispiel ist der kulturelle Wandel. Wie vorher angesprochen, kennen wir mittlerweile eine Mehrzahl an Geschlechtern und nicht mehr nur Mann-Frau. Die Daten, mit denen man die aktuellen Modelle gebaut hat, weisen diese weiteren Geschlechter meistens nicht aus. Diese Personen werden somit durch konsequente Nicht-Beachtung ignoriert. Es wird also einige Zeit dauern, bis diese Geschlechter sich in den Daten niederschlagen. Erst dann besteht die Möglichkeit, bei einer neuen Generierung eines Modells, die weiteren Geschlechter einzubeziehen.

a&f systems: Gibt es noch andere Bias und Problematiken, die Gefahrenquellen für Diskriminierung darstellen?
Lorenzo Medici: Beispielweise der Undersampling Bias: Wenn das Datenset zu klein ist, um die Realität abzubilden. Oder der Selection Bias: Wenn die Auswahl der Daten gewisse Bereiche gezielt ausgewählt oder ausgelassen werden. Das vorher beschriebene Beispiel zeigt auf, dass wenn eine Personengruppe nicht in den Daten vorliegt, diese das Modell nicht beeinflussen können und so ausgegrenzt werden.

Es gibt zudem noch technisch begründete Hindernisse und Gefahren. Bei einer Bildanalyse sind dunkelhäutige Menschen schwerer zu erkennen. Der Grund liegt in den Analyseverfahren. Bei bestimmten Beleuchtungsumgebungen sind die Daten dunkelhäutiger Menschen weniger kontrastreich. Zudem verwenden gewisse Algorithmen die Farbintensität, um Merkmale im Gesicht zu erkennen. Bei dunkleren Menschen können diese Merkmale anders erscheinen, was die Analyse weiter erschwert. Das hat zur Folge, dass die Fehlerquelle bei der Erkennung grösser ist. Das wiederum bedeutet, dass ein System technisch eher anschlägt, da es das Gesicht nicht so genau erkennt. Die Konsequenz ist dann, dass dunkelhäutige Menschen eher als suspekt identifiziert werden als hellhäutige Menschen. Selbst wenn sich jedes Mal die Situation klärt, ist das Verhalten des Systems klar rassistisch.

a&f systems: Zahlreiche Unternehmen nutzen bereits Künstliche Intelligenz. Wie könnte Diskriminierung bei der Verwendung verhindert werden?
Lorenzo Medici: Wenn wir von Diskriminierung sprechen, müssten wir im gleichen Atemzug auch über Bevorteilung sprechen. Das eine existiert nicht ohne das andere. Dass wir in einer Welt leben werden, in der niemand bevorteilt wird, ist eher unwahrscheinlich. Klar, von dieser Idealvorstellung darf nicht abgesehen werden und so sind die Datenanalysten sehr gefragt. Man muss sich bewusst sein, dass die Datenanalysten ebenfalls mit ihrem eigenen kulturell geprägten Wertesystem die Daten betrachten. Zudem können auch Firmen, welche diese KI-Systeme erstellen, Kriterien/Parameter setzen. Was gleich die Fragen aufwirft, ob Werte, welche sich durch die moderne Kommunikationstechnologie sehr schnell und stark verbreiten, hauptsächlich Firmeninteressen der Ersteller verfolgen? Die Politik hat die Notwendigkeit erkannt und ist diesbezüglich auch schon aktiv geworden. Welche Werte von wem vertreten werden, wird sicher sehr spannend bleiben, da KI-Anbieter auch von USA, Europa und Asien kommen. Die Unterschiede dieser Kulturen und ihren Werthaltungen sind offensichtlich. Es bestehen Anstrengungen, im ethischen Bereich den kleinsten gemeinsamen Nenner auf internationaler Ebene zu finden Höchstwahrscheinlich werden makroökonomische Überlegungen die Treiber sein, damit dieser gemeinsame Nenner gefunden wird.

a&f systems: Welche potentiellen Gefahren birgt der «KI-Rassismus» für Medienschaffende und Redaktionen?
Lorenzo Medici: Nun das Grundproblem liegt darin, dass falsche Informationen publiziert werden. Informationen, die aus einem KI-System kommen, sollten ausserhalb der Künstlichen Intelligenz verifiziert werden. Man muss sich bewusst sein, dass ein KI-System mit dem Parameter «Temperatur» die Wahrscheinlichkeitsverteilung der Generierung von Texten beeinflusst. Je kleiner dieser Wert ist, desto stringenter verhält sich das System und wird daher immer die gleichen Antworten liefern.

Je näher gegen 1, je vielfältiger werden die Antworten. Es wird also kreativ, was bei faktenbasierten Artikeln nicht unbedingt gewünscht ist; insbesondere, wenn Begriffe ausgetauscht werden, die zwar eine mehr oder weniger kongruente Bedeutung haben, aber ungenau sind. Das kann zu falschen Aussagen führen.

a&f systems: Wie kann diesen Gefahren vorgebeugt werden?
Lorenzo Medici: Durch Recherche und Verifikation, welche durch die Redakteure gemacht werden muss.

a&f systems: Wir haben nun über Gefahren und Risiken von KI gesprochen. Gibt es denn auch Positives, das den Medienschaffenden dient?
Lorenzo Medici: Ja sicher. Zum einen sind KI-Recherchen viel breiter. Die Redaktion kann mittels KI-Suchen in Themengebieten starten, die Resultate liefern, welche bei konventionellem Suchen nicht geliefert würden, da die Modelle auch Konnotationen machen können. Das erleichtert es den Medienschaffenden, an Hintergrundinformationen zu gelangen. Die aber sicher immer verifiziert werden müssen.

Auf der anderen Seite ist die NLG (Natural Language Generation) mittlerweile auf einem sehr hohen Niveau. Das erleichtert und beschleunigt das Schreiben eminent. In der Bildbearbeitung gibt es schon unzählige Anwendungen, die helfen, Bilder zu bearbeiten oder nach Bildern zu suchen. Es ist unbestritten, dass diese Technologie unsere Arbeit nachhaltig verändern wird.

In diesem Zusammenhang ist ja auch die a&f systems aktiv. Wir binden KI-Systeme in unsere Systeme ein, so zum Beispiel im Publishing Circle. Das ist auch unbedingt notwendig, um den steigenden Anforderungen des Marktes gerecht zu bleiben.

a&f systems: Herzlichen Dank für das aufschlussreiche Interview und den spannenden Einblick, Lorenzo!

Lorenzo Medici, lic. phil. I (Soziologie/Mathematik)

Lorenzo Medici ist seit 2018 als Head of Development bei der a&f systems ag tätig. Sein Hauptfokus liegt auf dem Management von kundenspezifischen Entwicklungen und dem Erstellen von Softwarearchitekturen. Zu seinen Fachgebieten gehören Softwareentwicklung, Projektmanagement sowie auch die Wechselwirkung zwischen technologischen und gesellschaftlichen Entwicklungen. Zudem ist er auch Geschäftsführer und Inhaber der Projektmanagement Medici AG und ist an der Hochschule Luzern Gastdozent und Prüfungsexperte.

Ihr Ansprechpartner:
Stefan Schärer
Head of Sales & Marketing,
Co-Owner, Member of the Executive Board
sschaerer@a-f.ch

a&f systems ag
Grenzstrasse 3b
6214 Schenkon
+41 41 925 71 11
info@a-f.ch
www.a-f.ch

a&f systems gmbh
Eleonorenstraße 20
D-30449 Hannover
+49 511 89 880 494
info@a-f.de
www.a-f.de

20230712.D.LAY