MARSILIUS KOLLEG

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

MENSCH UND AUTOMAT – DIE WAHRNEHMUNG UND BEWERTUNG KÜNSTLICHER INTELLIGENZ

Fellowbericht

Andreas Voss

DOI: 10.11588/fmk.2022.2.92723

MARSILIUS-KOLLEG
2021 / 2022


MARCELLUS
SKOLEBO

ANDREAS VOSS

MENSCH UND AUTOMAT

Die Wahrnehmung und Bewertung künstlicher Intelligenz

MODERNE AUTOMATEN: DER EINFLUSS KÜNSTLICHER INTELLIGENZ AUF UNSER LEBEN

Seit mindestens zwei Jahrzehnten sind Computer kaum noch aus dem Alltag der meisten Menschen wegzudenken. In den letzten Jahren hat jedoch eine neue Entwicklung zu einem sprunghaften Anstieg der Möglichkeiten der Informationstechnologie geführt: Gemeint sind die rasanten Fortschritte der sogenannten künstlichen Intelligenz (KI). Durch KI können Computer heute Aufgaben bewältigen, die noch vor wenigen Jahre als unlösbar galten. Ein prominenter Meilenstein, der diese Entwicklung symbolisiert, war der Sieg der Software AlphaGo am 9. März 2016 gegen den südkoreanischen Go-Profi Lee Sedol, der als bester Spieler der Welt galt. Wegen der hohen Komplexität des Go-Spiels und der großen Zahl der Zugmöglichkeiten versagen hier klassische Computeralgorithmen. Die Entwicklung eines Computerprogramms, das auf dem Niveau von Profispieler:innen Go spielen kann, galt deshalb lange Zeit als unrealistisch.

Ein weiteres Beispiel, welches die enorme Leistungsfähigkeit moderner KI belegt, stellt der Bereich des (teil-)autonomen Fahrens dar. Auch hier ist die Komplexität der Situation, die analysiert werden muss, und die Vielfalt der Handlungsmöglichkeiten so groß, dass erst moderne KI-Algorithmen es ermöglichen, dass sich Fahrzeuge selbstständig im Straßenverkehr bewegen. Erfolgreiche Modellversuche unterschiedlicher Firmen – auch in echten Verkehrssituationen – belegen, dass heutige Technik bereits das autonome Fahren ermöglicht. Leicht lassen sich viele andere Beispiele für den breiten Einsatz von KI finden: Sogenannte Chatbots können sinnvolle Unterhaltungen führen,[1] Empfehlungssysteme wählen Produkte aus, die

ein:e Kund:in in Zukunft benötigen könnte, vielversprechende Bewerber:innen werden von einer Software auf Basis ihrer Bewerbungsunterlagen ausgewählt, medizinische Diagnosen werden automatisch aus Röntgenbildern und anderen Daten gestellt. Diese Liste ließe sich leicht fortführen und es ist offensichtlich, dass die Zahl von KI-Anwendungen auch in Zukunft weiterhin ansteigen wird, weil die dahinterliegende Technologie immer leichter zugänglich wird.

MÖGLICHKEITEN UND GRENZEN DER KÜNSTLICHEN INTELLIGENZ

Um Möglichkeiten und Grenzen der KI besser zu verstehen, ist es wichtig, sich deren grundlegende Funktionsweise anzusehen. Die meisten modernen KI-Algorithmen basieren auf maschinellem Lernen durch sogenannte künstliche neuronale Netze. Dabei bestimmen nicht länger konkrete Programmierungen das Verhalten der KI; vielmehr lernt ein neuronales Netz in einer Trainingsphase bestimmte Entscheidungen zu optimieren. Bei einem autonomen Auto würde der Algorithmus beispielsweise lernen, so zu steuern wie Testfahrer:innen, die die KI in der Trainingsphase „beobachtet" hat. Bei einer automatischen, KI-gesteuerten Personalauswahl würde der Algorithmus lernen, die gleichen Entscheidungen zu treffen wie diejenigen, die in der Personalabteilung an der Generierung der Trainingsdaten mitarbeiten.

Aus diesem Lernprinzip werden zwei der wichtigsten Fehlerquellen von KI-basierten Entscheidungen deutlich. Erstens können Entscheidungen einer KI nur so gut sein wie ihre Trainingsdaten. Wenn Trainingsdaten systematische Fehlentscheidungen enthalten, wird dieser Fehler sich auch im künftigen Verhalten der KI wiederfinden. Diese Art von Fehlern wurde etwa bei Algorithmen der automatischen Personalselektion beobachtet. Hier wurden bestimmte Muster der Diskriminierung, die sich in unfairen Entscheidungen der Menschen in einer Personalabteilung fanden, von Algorithmen gelernt und somit fortgeführt.

Die zweite Problematik, die sich aus dem Lernprinzip der KI ergibt, betrifft den Umgang mit neuen oder ungewöhnlichen Daten. KI ist in der Regel in der Lage, gute Entscheidungen zu treffen, wenn eine zu bewertende Situation den Trainingsdaten hinreichend ähnlich ist. Bei Situationen, für die ein Algorithmus nicht trainiert wurde, kann es jedoch zu einem kompletten Versagen kommen. Dieses Prinzip kann möglicherweise Unfälle von autonomen Fahrzeugen erklären, die ungebremst auf ein deutlich sichtbares Hindernis fuhren, das für eine:n menschliche:n Fahrer:in leicht zu erkennen gewesen wäre. Die KI konnte jedoch diese Situation nicht „verstehen", weil ähnliche Situationen im Training des Algorithmus nicht enthalten waren. Man kann dieses zweite Problem auch als fehlende Extrapolationsfähigkeit eines Algorithmus bezeichnen: Die KI kann nur sehr eingeschränkt über den gelernten Bereich hinaus allgemeine Handlungs- bzw. Entscheidungsprinzipien ableiten. Diese Einschränkung der Fähigkeiten vieler Algorithmen, in neuen, unbekannten Situationen sinnvolle Entscheidungen zu treffen, stellt damit auch den Begriff der künstlichen *Intelligenz* infrage, da gerade die Fähigkeit, neue (nicht gelernte) Probleme zu lösen, im Zentrum einer Intelligenzdefinition stehen muss.

VORAUSSETZUNGEN FÜR DAS VERTRAUEN IN KÜNSTLICHE INTELLIGENZ

In meinem Marsilius-Projekt stand nun die Frage im Vordergrund, wie Menschen KI wahrnehmen und bewerten und welche Bedingungen erfüllt sein sollten, damit sie diesen modernen Automaten und ihren Entscheidungen vertrauen. Ein grundlegendes Problem stellt hier der Blackbox-Charakter von Algorithmen dar, die auf künstlichen neuronalen Netzen beruhen. Gemeint ist damit, dass der Algorithmus darauf trainiert wird, optimale Entscheidungen zu treffen, die Gründe für die Entscheidungen aber letztlich verborgen bleiben, sodass ein Verstehen der Entscheidung des

Automaten im Sinne Wilhelm Diltheys kaum möglich ist. Es bleibt also unklar, warum ein autonomes Auto eine bestimmte Geschwindigkeit wählt oder ein Computerprogramm eine:n bestimmte:n Bewerber:in für eine Stelle auswählt. Dies schränkt natürlich auch das Vertrauen in die KI bzw. in die Qualität ihrer Entscheidungen ein. Spezialist:innen für neuronale Netze wie mein Projektpartner Ullrich Koethe arbeiten daran, Entscheidungen der KI für eine menschliche Nutzerschaft erklärbar zu machen. Hier stellt sich für mich als Psychologen die Frage, ob und wie dies gelingen kann, das heißt, welche Art von automatisch generierter Erklärung tatsächlich von denen, die den Automaten nutzen sollen, verstanden und akzeptiert wird.

Neben der Erklärbarkeit von Entscheidungen spielen natürlich eine ganze Reihe weiterer Faktoren eine Rolle für das Vertrauen in einen Algorithmus. In einer aktuellen Studie ließen wir Proband:innen den Einsatz eines hypothetischen Algorithmus zur Lügendetektion im juristischen Kontext bewerten. Hier konnten neben der Transparenz des Algorithmus die Präzision (Validität) und die Fairness der Entscheidungen des Automaten als wichtige Prädiktoren seiner Akzeptanz bestätigt werden.

Wie die Arbeiten meiner Projektpartnerin Rebecca Müller belegen, waren Menschen bereits in der Antike und während des Mittelalters von realen wie fantastischen Automaten fasziniert. Mit der zunehmenden Leistung der KI werden Automaten in der Zukunft sicherlich weiterhin in vielen Bereichen des menschlichen Lebens an Bedeutung gewinnen. Die Erforschung der sich so intensivierenden Interaktion von Mensch und Automat steht aber noch ganz am Anfang.

[1] Die Google-Software LaMDA führte so überzeugende Gespräche, dass der Google-Mitarbeiter Blake Lemoine ihr im Juli 2022 ein eigenes Bewusstsein zuschrieb.