Humanistische Union

„Mein Leben wird ganz wunderbar“ – Chancen und Risiken der Künstlichen Intelligenz

*In seiner Einführung in den Themenkomplex Künstliche Intelligenz geht Stefan Hügel auf die Risiken ein, die mit Künstlicher Intelligenz verbunden sind. Zahlreiche Wissenschaftler*innen sehen die Gefahr, dass eine sich selbst immer schneller weiterentwickelnde KI langfristig den Menschen verdrängen könnte. In öffentlichen Appellen weisen sie auf die Risiken hin und schlagen ein Moratorium für die Weiterentwicklung vor. Sie warnen davor, dass Kipppunkte in der Entwicklung zunächst unbemerkt überschritten werden könnten. Dazu rekapituliert Hügel konkrete Risiken. Regulierungsbestrebungen gibt es insbesondere auf Ebene der Europäischen Union in Form des AI Act, der sich vor allem auf Hochrisikosysteme konzentriert. Was für Regelungsbedarf generell besteht oder noch bestehen könnte, listet Hügel auf.*

Als kleiner Junge war mir schon klar / mein Leben wird ganz wunderbar.
Ich richte mich einfach radikal / nach Algorithmen meiner Wahl.
Deutsch Amerikanische Freundschaft, Algorithmus – zahlenlied[i]

Künstliche Intelligenz (KI) prägt zunehmend den Fortschritt in der Informatik und wird in Gesellschaft, Politik und Wirtschaft immer stärker präsent (Russel 2023). Damit geht es über ein technisches Spezialthema weit hinaus und könnte „[…] *bald fast jedes Gebiet menschlichen Strebens betreffen*“ (Kissinger/Schmidt/Huttenlocher 2021: 3). Besonders der ChatBot *ChatGPT* (*Generative Pre-trained Transformer*)[ii], der seit 2022 öffentlich zur Verfügung steht und in der Lage ist, umfassende Fragen zu beantworten, Texte zu verfassen oder Programmcode zu erzeugen, hat der Debatte um künstliche Intelligenz in jüngster Zeit einen enormen Schub gegeben. Durch Weiterentwicklung der dabei zugrundeliegenden *Large Language Models* (LLM) wird seine Leistungsfähigkeit weiter gesteigert.

Im Grunde ist KI aber schon lange kein neues Thema mehr. Der Begriff wurde bereits in den 1950er Jahren geprägt und das Thema – mit wechselnder Intensität – seither weiterverfolgt. Doch es ist anzunehmen, dass es sich bei der Entwicklung der Künstlichen Intelligenz um einen exponentiellen Prozessiv handelt (Azhar 2021), und so ist mit einer sich immer stärker beschleunigenden Entwicklung zu rechnen. Neben den vielfältigen Chancen, die sich daraus ergeben, müssen auch die Risiken betrachtet werden. Die Debatte darüber geht von praktischen Problemen der Anwendung des maschinellen Lernens als fortgeschrittene Methode der Informatik hin zu dystopischen Szenarien, in denen eine übermächtige KI zur Konkurrenz des Menschen wird und ihn perspektivisch in seiner Rolle ablöst. Aus bürgerrechtlicher Perspektive müssen diese Konsequenzen der künstlichen Intelligenz im Blick behalten werden. (Hügel 2019: 25ff.)

Kritik an Künstlicher Intelligenz gab und gibt es seit Beginn ihrer Entwicklung. Zunächst wurde vor allem daran gezweifelt, ob die daran geknüpften Versprechungen und Ziele überhaupt technisch erreicht werden können. Joseph Weizenbaum (1987: 268) argumentierte, dass die Erwartungen an Künstliche Intelligenz letztlich auf einem zu stark vereinfachten Begriff von Intelligenz beruhen. Seither wurden die Grenzen des Möglichen immer weiter verschoben – neben Fortschritten in der Methodik trägt dazu zweifellos die

Entwicklung der Hardware-Technologie einen erheblichen Teil dazu bei. Künstliche Intelligenz als methodische Weiterentwicklung der Informatik – im Sinne *schwacher KI* – entwickelt sich derzeit stürmisch weiter. Über die Entwicklung *starker* KI – sprich, einer Form der Künstlichen Intelligenz, die der Intelligenz des Menschen ähnlich ist – ist damit noch nichts gesagt. Kann beispielsweise eine Maschine ein Bewusstsein entwickeln oder zumindest simulieren?

Was ist Künstliche Intelligenz?

Künstliche Intelligenz ist schon deswegen schwer zu definieren, da bereits der Begriff der menschlichen Intelligenz schwer abzugrenzen ist. Bei Wikipedia[iii] findet man:

> „Intelligenz [wird] verstanden als die Eigenschaft, die ein Wesen befähigt, angemessen und vorausschauend in seiner Umgebung zu agieren; dazu gehört die Fähigkeit, Sinneseindrücke wahrzunehmen und darauf zu reagieren, Informationen aufzunehmen, zu verarbeiten und als Wissen zu speichern, Sprache zu verstehen und zu erzeugen, Probleme zu lösen und Ziele zu erreichen."

Ralf Otte (2023: 9-16) beschreibt Intelligenz auf acht Stufen in drei Dimensionen – rationale Intelligenz, wahrnehmende Intelligenz und fühlende Intelligenz – und stellt fest, dass sich heutige künstliche Intelligenz ausschließlich in der Dimension rationaler Intelligenz bewegt. Howard Gardner (1983) führt acht Dimensionen menschlicher Intelligenz auf: Bewegungsintelligenz, bildlich-räumliche Intelligenz, Sprachintelligenz, logisch-mathematische Intelligenz, musikalische Intelligenz, naturalistische Intelligenz, zwischenmenschliche Intelligenz und selbstreflexive Intelligenz. Vor allem Letztere wird als wesentliche Eigenschaft den Menschen betrachtet, da sie das (Selbst-) Bewusstsein einschließt. Max Tegmark (2017: 63) schreibt kurz und bündig: „Intelligenz [ist die] Fähigkeit, komplexe Ziele zu erreichen."

In der Künstlichen Intelligenz unterscheidet man zwischen der *starken* Künstlichen Intelligenz – eine umfassende Intelligenz, die der des Menschen analog ist und die Möglichkeit eines eigenen Bewusstseins einschließt – und der *schwachen* Künstlichen Intelligenz – eine technische Intelligenz, die einzelne, spezifische kognitive Fähigkeiten innerhalb eines abgegrenzten Aufgabenbereichs besitzt, ohne umfassenden Zusammenhang zwischen den einzelnen Fähigkeiten. Während die Möglichkeit einer starken KI auch philosophisch umstritten ist und zumindest bis heute technisch nicht realisiert werden kann, lässt sich die schwache KI als fortgeschrittene Methode der Informatik einstufen, mit der Aufgaben aufgrund unstrukturierter Datenmengen insbesondere mit statistischen und stochastischen Methoden gelöst werden können. Letztlich beruhen auch Verfahren der Künstlichen Intelligenz auf Methoden der algorithmischen Verarbeitung von Daten und unterliegen damit auch ihren mathematischen Beschränkungen, beispielsweise den Regeln der Berechenbarkeit.

Ein frühes Gedankenexperiment, wann einem Computer Intelligenz zugeschrieben werden kann, ist der *Turing-Test*: Ein menschlicher Schiedsrichter kommuniziert mit zwei Personen – einem Mensch und einem

Computer. Kann er aufgrund der Antworten nicht unterscheiden, welcher der Gesprächspartner*innen Mensch und welcher Maschine ist, so gilt die Maschine als intelligent. (Hofstadter/Denett 1986)

Wie funktioniert maschinelles Lernen?

KI in der heute sichtbaren Form ist in der Regel *maschinelles Lernen*[iv]. Dabei werden insbesondere statistische und stochastische Verfahren genutzt, um Artefakte zu erkennen oder – bei generativer Künstlicher Intelligenz – Texte, Bilder oder Programmcode zu erzeugen. *Deep Learning* mit mehrstufigen neuronalen Netzen macht es so möglich, auch komplexe Aufgaben zu lösen, wie Bilder zu erkennen und zu klassifizieren, die Bedeutung von geschriebenem Text und gesprochene Sprache zu verstehen, optimale Strategien zu lernen und auch kreativ tätig zu werden, indem Bilder, Texte, Musik erzeugt und dabei auch Emotionen geäußert werden können (Paaß/Hecker 2020). Dieses „Verständnis“ beruht auf großen Datenmengen, mit denen das neuronale Netz trainiert wird. Dabei „versteht“ die KI den Kontext, in dem sie diese Artefakte produziert, nur scheinbar – wenn beispielsweise Texte generiert werden, wird an jeder Stelle des Textes ermittelt, welches Wort (oder welcher Buchstabe) am wahrscheinlichsten auf den bisher produzierten Text folgt. Von einem Textverständnis im üblichen Sinn kann hier keine Rede sein, geschweige denn von einer Bewertung der Ergebnisse nach inhaltlichen oder ethischen Maßstäben. Abhängig vom „erlernten“ Modell kann es so teilweise zu inhaltlich völlig unsinnigen Ergebnissen (*Halluzinationen*) kommen – die Modelle sind aber kognitiv zu erstaunlichen Leistungen fähig.

Letztlich basiert maschinelles Lernen in der heutigen Form auf Korrelationen in den zugrundeliegenden Daten. Die grundsätzliche Möglichkeit, dass *Large Language Models* (LLM) zu einer Generellen Künstlichen Intelligenz (AGI – Artificial General Intelligence) führen können, die jede Aufgabe übernehmen könnte, zu der auch ein Mensch fähig ist, ist fraglich (Levine 2023: 46).

Risiken

Systeme der Künstlichen Intelligenz leiten ihr Verhalten also aus Daten ab, ohne in der Lage zu sein, dies (ethisch) zu bewerten. Einige Beispiele von Berichten aus der jüngeren Vergangenheit illustrieren die Risiken einer so außer Kontrolle geratenen KI, deren „gelerntes“ Verhalten unerwünscht oder zumindest unerwartet ist:

- 2016 veröffentlichte Microsoft den Chatbot *Tay*[v], der über Twitter kommunizierte und aus den Interaktionen mit anderen Nutzer*innen lernte. Er musste nach kurzer Zeit wieder abgeschaltet werden, nachdem es böswilligen Nutzer*innen gelungen war, ihn durch gezielte Interaktionen dazu zu

bringen, rassistische und extremistische Tweets abzusetzen.

- Presseberichten zufolge beging ein junger Mann in Belgien Suizid, nachdem er mit einem Chatbot über den Klimawandel kommuniziert hatte und dieser ihn „ermutigte“, sich zur Rettung der Erde selbst zu opfern. (Affsprung 2023; El Atillah 2023)

- Einem Studenten an der Technischen Universität München, Marvin von Hagen, gelang es, durch *Prompt Injection*[vi] den Chatbot *Bing Chat* zu veranlassen, seine – normalerweise nicht offengelegten – internen Regeln zur Feinsteuerung, die Teil des der KI zugrundeliegenden Modells sind und beispielsweise Bias ausgleichen oder den Ton der Antworten (höflich, unfreundlich, sarkastisch) beeinflussen, preiszugeben (Szöke 2023). Veröffentlichte Informationen darüber gelangten offenbar wiederum in das Modell des Chatbots und wurden dort als Bedrohung interpretiert. Der Bot bezeichnete daraufhin von Hagen als einen „Feind“, der „die Konsequenzen für seine Handlungen tragen“ müsse (Schmalzried 2023). Offenbar „erkannte“ der Chatbot, dass sich die Berichte über den Fall auf ihn bezogen. Ist das bereits eine primitive Form von Selbstbewusstsein?

Weitere Beispiele für problematische Auswirkungen von künstlicher Intelligenz im Einzelfall finden sich beispielsweise bei Cathy O’Neil (2016) und bei Katharina Zweig (2019; 2023). Inwieweit solche anekdotischen Fälle verallgemeinerbar sind, sei dahingestellt. Auch weitere namhafte Wissenschaftler*innen warnen jedoch inzwischen vor weitgehenden Konsequenzen einer entfesselten Künstlichen Intelligenz.

Löst Künstliche Intelligenz den Menschen ab?

Der Physiker Stephen W. Hawking prognostizierte, dass bei einer weiteren Entwicklung der Computer entsprechend Moores Law – Verdoppelung der Geschwindigkeit und Speicherkapazität von Rechnersystemen circa alle 18 Monate – die Intelligenz von Computern die des Menschen in den kommenden 100 Jahren übertreffen könnte, und schrieb:

> „Wenn eine Künstliche Intelligenz (KI) besser wird als Menschen bei der Konstruktion von KI, sodass sie sich rekursiv ohne menschliche Hilfe selbst verbessern kann, dann steht uns höchstwahrscheinlich eine Intelligenzexplosion bevor, die letztlich in die Maschinenintelligenz mündet: Sie wird unsere Intelligenz in viel höherem Maß übertreffen als unsere menschliche Intelligenz die von Schnecken. Bevor es so weit ist, müssen wir sicherstellen, dass die Computer Ziele verfolgen, die auf einer Linie mit unseren Zielen liegen.“ (Hawking 2018: 208)

Und noch deutlicher:

> „Es ist zu befürchten, dass die KI alleine weitermacht und sich mit ständig zunehmender Geschwindigkeit selbst überarbeitet. Menschen, die aufgrund der Langsamkeit ihrer biologischen Evolution beschränkt sind, könnten nicht mithalten und würden verdrängt.“ (Hawking 2018: 211)

Man mag das als alarmistisch abtun. Doch es wäre wohl nicht klug, diese Möglichkeit zu ignorieren. Auch wenn er vermutlich nicht derartige Szenarien im Kopf hatte – auch hier gilt das Diktum von Hans Jonas (1979: 36): „Handle so, dass die Wirkungen deiner Handlung verträglich sind mit der Permanenz echten menschlichen Lebens auf Erden.“

Dieses Prinzip müsste freilich zur Anwendung kommen, bevor uns die weitere Entwicklung aus der Hand genommen wird, das heißt bevor eine fortentwickelte KI sich ihre eigene Ethik schafft, bei der sie das Ziel der Permanenz auf sich selbst „umbiegt“ – wie es im Fall von Marvin von Hagen anscheinend geschehen ist, wenn der Chatbot seine eigene Fortexistenz über die des*der Nutzer*in stellt oder „erkennt“, dass es in Wahrheit der Mensch ist, der diesem Planeten den größten Schaden zufügt – und entsprechend handelt.

Auch andere Akteur*innen warnen vor Risiken der KI. In einem Ein-Satz-Statement, das von namhaften Expert*innen unterzeichnet wurde, heißt es: „Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war“ (Save AI 2023).

Ein weiterer öffentlichkeitswirksamer Appell wurde angesichts der zunehmenden Leistungsfähigkeit von Chatbots veröffentlicht. Über 30.000 Unterzeichner*innen, darunter mit bekannten Namen wie Stuart Russell, Elon Musk und Steve Wozniak, fordern ein sechsmonatiges Moratorium beim Training von KI-Systemen, die mächtiger sind als GPT-4:

> „AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research and acknowledged by top AI labs. As stated in the widely-endorsed Asilomar AI Principles, Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources. Unfortunately, this level of planning and management is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control.“ (Future of Live.org 2023)

Die Initiative wurde allerdings gemischt aufgenommen (Paul and agencies 2023; Yudkowsky 2023). Das Time-Magazine kommentiert:

> „The key issue is not ‚human-competitive' intelligence (as the open letter puts it); it's what happens after AI gets to smarter-than-human intelligence. Key thresholds there may not be obvious, we definitely can't calculate in advance what happens when, and it currently seems imaginable that a research lab would cross critical lines without noticing." (Yudkowsky 2023)

Auch die Informatik-Professorin Hannah Bast, die als Sachverständige Mitglied der Enquête-Kommission zur künstlichen Intelligenz des Deutschen Bundestages (2020) war, warnt vor tiefgreifenden Veränderungen. Bast zufolge werde die Tatsache, dass Maschinen nun Sprache verstehen, alles verändern. Nicht in den nächsten zwei, drei Jahren, aber doch sehr bald.

Wer vom Fach war, wusste damals sofort, dass dies alles verändern würde. So wird das auch jetzt sein. Man wird es nicht sofort bemerken, aber nach und nach wird es unser Leben komplett verändern. (Moreno 2023)

Die Debatte wird weitergehen, ob die Risiken der bereits eingesetzten Entwicklung richtig eingeschätzt werden oder ob die (negativen) Potenziale einer neuen Technologie hier alarmistisch übertrieben werden. Tegmark (2017: 52) unterscheidet zwischen mehreren Gruppen mit unterschiedlichen Einstellungen zur KI: *Techno-Skeptiker, Technikfeinde, Digitale Utopisten* und die *Nutzbringende KI-Bewegung,* und ordnet sie anhand der Dimensionen ein, ob sie erwarten, dass KI das menschliche Niveau überschreiten wird und ob sie das für vorteilhaft halten.

Dass es Risiken gibt, steht außer Frage: Diskutiert werden die mangelnde Nachvollziehbarkeit der Systeme, erhebliche Risiken für den Datenschutz und die weitgehenden Möglichkeiten der Überwachung, beispielsweise durch automatisierte Gesichtserkennung, und vieles mehr.

Ethische Fragen

Offensichtlich ergeben sich aus der Nutzung von Methoden der Künstlichen Intelligenz – wie aus jeder Techniknutzung – auch ethische Fragen (Stahl 2023: 17-22), die insbesondere in entsprechenden Fachgremien behandelt werden. Stellvertretend seien hier die *Asilomar AI Principles* genannt, die Fragen der Forschung, Ethik und Werte und längerfristige Probleme benennen und Handlungsprinzipien dazu formulieren (Future of Live.org 2017). Malte Rehbein (2018: 24) kritisiert bereits den Ansatz als technikdeterministisch und utilitaristisch:

> „Artificial intelligence has already provided beneficial tools that are used every day by people around the world. Its continued development, guided by the following principles, will offer amazing opportunities to help and empower people in the decades and centuries ahead.“

Ein im Zusammenhang mit KI genanntes Beispiel ist das Trolley-Problem[vii], wenn es um die Entscheidung über Menschenleben geht. Es ergeben sich dabei ethische Fragen – ähnlich einer maschinellen Triage.

Kurz skizziert, besteht das Trolley-Problem in folgender, beispielhafter Situation: *Eine Straßenbahn ist außer Kontrolle geraten und droht, fünf Personen zu überrollen. Durch Umstellen einer Weiche kann die Straßenbahn auf ein anderes Gleis umgeleitet werden. Unglücklicherweise befindet sich dort eine weitere Person.* Daraus ergibt sich die Frage: Darf oder muss (durch Umlegen der Weiche) der Tod einer Person gezielt in Kauf genommen werden, um das Leben von fünf Personen zu retten? Es ist zu erwarten, dass eine solche Situation bei sogenannten autonomen Fahrzeugen häufig auftritt. Beim maschinellen Lernen würde die Entscheidung durch eine (komplexe) Bewertungsfunktion getroffen, die vorab von einem Menschen festgelegt wurde. Im Gegensatz zum zufälligen Ereignis muss damit vorab eine abstrakte, bewusste Entscheidung analog des Trolley-Problems getroffen werden. Damit ergibt sich die Frage, der wir bisher ausweichen konnten: Welche Entscheidung ist ethisch vertretbar? Bei nicht mehr vermeidbaren Unfällen bei autonomen Fahrzeugen dürfte diese Situation häufiger auftreten.

Aber auch darüber hinaus ergeben sich ethische Fragestellungen für die Nutzung von Methoden künstlicher Intelligenz. Ob es eine Maschinenethik geben kann, wenn man künstliche Intelligenz ausschließlich auf kognitiver Ebene begreift, ist zweifelhaft (Misselhorn 2018). Einige der konkreten Fragestellungen, die in den nächsten Abschnitten angerissen werden, sind nicht neu, durch die erweiterten technischen Möglichkeiten erhalten sie aber eine neue Brisanz.

Überwachung

Künstliche Intelligenz verarbeitet große Datenbestände an (Trainings-)Daten, die sich auf die Modelle auswirken. Davon können auch sensible, personenbezogene Daten betroffen sein. Darüber hinaus können Verfahren der KI zur umfassenden Überwachung genutzt werden, beispielsweise zur automatisierten Gesichtserkennung. Demnach ergibt sich ein erhebliches Risiko für die Bürger- und Menschenrechte aus diesen weitgehenden Möglichkeiten der Überwachung. Methoden der KI können dabei sowohl für die automatische Gesichtskontrolle als auch für die Kontrolle der Kommunikation eingesetzt werden.

Aktuelles Thema ist die Chatkontrolle, sprich, die Überwachung der Inhalte von Chats – vorgeblich insbesondere das Scannen von Bildern, um Jugendschutz durchzusetzen und Missbrauchsdarstellungen zu bekämpfen. Dabei werden über Chats versendete Bilder zunächst mit Datenbanken bereits bekannten Bildern abgeglichen. Zur Erkennung noch nicht bekannter Darstellungen werden Verfahren des maschinellen Lernens eingesetzt; um Verschlüsselungen zu umgehen, werden dabei die Nachrichten bereits auf dem Client gescannt (*Client Side Scanning*). Es ist anzunehmen, dass solche Verfahren in der EU illegal

sind, da die verdachtsunabhängige Überwachung Grundrechte verletzt (siehe dazu die einschlägigen Urteile zur Vorratsdatenspeicherung). Deswegen haben entsprechende Pläne der EU scharfe Kritik von Bürgerrechtsverbänden und Datenschützer*innen hervorgerufen.

Doch die Überwachung geht noch weiter: Automatisierte Erkennung von Personen ermöglicht es, potenziell jedes Fehlverhalten in der Öffentlichkeit zu erkennen und zu ahnden. Dies wird vor allem mit dem System des *Social Scoring* in China verbunden (Shi-Kupfer 2023): Durch automatische Gesichtserkennung in der Öffentlichkeit kann „verdächtiges" Verhalten erkennt und bewertet werden (etwa das Überschreiten einer roten Fußgängerampel). Die Ergebnisse werden zusammen geführt und in einen Gesamt-Score für die Person zusammengefasst. Dies kann dann ernste Konsequenzen für einzelne Personen haben, wie die Verhängung eines Flugverbots.

Bisher wird diese Form des *Social Scoring* vor allem mit einem Modellversuch in China verbunden. Aber auch in den USA und der EU werden inzwischen umfassend Daten gesammelt und können durch entsprechende Verfahren der KI kategorisiert und bewertet werden (Kühnreich 2022). Selbst wenn wir annehmen, dass dies nicht von Staats wegen geschieht[viii], ergeben sich für den Einzelnen erhebliche Risiken.

Politische Beeinflussung durch Falschinformationen

Nach der Bundestagswahl 2021 kursierte ein lustiges Video im Netz. Anlässlich der Sondierungsgespräche zwischen Bündnis90/Die Grünen und der FDP war ein Selfie veröffentlicht worden, das die Verhandlungsteilnehmer*innen Annalena Baerbock, Robert Habeck, Christian Lindner und Volkmar Wissing in einer Sitzungspause zeigten. Daraus wurde ein Video produziert, das die Politiker*innen dabei zeigte, wie sie gemeinsam im Chor den Song *We are family* sangen[ix].

Dies war ein vergleichsweise harmloses Beispiel – das aber zeigt, dass Verfahren der (generativen) Künstlichen Intelligenz in vielfältiger Weise zur Manipulation von Filmen und Bildern genutzt werden können (*Deep Fakes*) (Louban et al. 2022: 265ff.). Filme können erzeugt werden, indem Gesichter „ausgetauscht" oder die Mimik, beispielsweise beim Sprechen, nachgeahmt oder „übertragen" wird. Solche Manipulationen sind – vor allem bei hoher Qualität und bei flüchtigem Hinschauen – nicht ohne Weiteres zu erkennen. Da (bewegte) Bilder als besonders glaubhaft wahrgenommen werden, kann einer Person auf diese Weise auch eine falsche (politische) Aussage oder eine kompromittierende Situation „untergeschoben" werden. So können durch Falschnachrichten die Öffentlichkeit getäuscht und sie gezielt zur politischen Desinformation und Beeinflussung verbreitet werden.

Microtargeting und Nudging

Darüber hinaus ermöglichen KI-Technologien und Techniken der Data Science wie *Microtargeting* und *Nudging*

die Beeinflussung in der politischen Kommunikation. Wähler*innen werden nach politischen Präferenzen zur gezielten Ansprache kategorisiert und individuellen Werbebotschaften gezielt platziert, anstatt eine ehrliche politische Debatte zu führen. Dabei werden einzelne Gruppen und Personen gezielt angesprochen (*Microtargeting*) und damit das Verhalten des Einzelnen – in diesem Fall die Stimmabgabe – auch unbemerkt, manipulativ beeinflusst (*Nudging*).

Unternehmen wie Google, Facebook oder X (Twitter) sammeln diese Daten, werten sie aus und bekommen dadurch die Möglichkeit, Menschen nach Vorlieben beliebig zu klassifizieren und gezielt zu beeinflussen. Bekanntes Beispiel dafür war der Facebook-Skandal um das Unternehmen Cambridge Analytica. Dies warf Fragen nach der Beeinflussung und Manipulation von Wahlen auf:

Im März 2018 löste die Nutzung von Facebook-Profildaten durch das US-amerikanische Unternehmen Cambridge Analytica den Facebook-Skandal aus. Cambridge Analytica hatte mithilfe ihrer Facebook-App *thisisyourdigitallife* Daten von Facebook-User*innen und ihren Kontakten ausgelesen, unter politischen Gesichtspunkten ausgewertet und die gewonnenen Erkenntnisse für die Kampagne von Donald Trump im US-Präsidentschaftswahlkampf genutzt. Cambrige Analytica wurde für das politische Ausnutzen von einigen Dutzend Millionen Datensätzen weltweit kritisiert, während das ganz normale Geschäftsmodell großer Plattformen genau darin besteht, die gleichen Methoden auf Milliarden Datensätze anzuwenden – sowohl für Produktwerbung als auch für politische Kampagnen.

Informationen und deren Kategorisierung können gezielt in der Werbeindustrie und für politische Beeinflussung genutzt werden. Dabei wird eine sehr feine Aufteilung der Zielpersonen in einzelne Kategorien vorgenommen, die eine gezielte Ansprache der Adressat*innen mit spezifischen politischen Inhalten ermöglicht (Dachwitz 2023).

Die gezielte Ansprache kann zu Filter Bubbles und Echokammern führen, indem vorgefasste Meinungen immer weiter bestätigt oder gar extremisiert werden (Pariser 2012). Inwieweit durch solche Verfahren aber politische Einstellungen beeinflusst werden, ist umstritten (Leisegang 2023).

Transparenz und Erklärbarkeit

Im Gegensatz zu herkömmlichen Algorithmen, die (zumindest theoretisch) für jede*n Expert*in nachvollzogen werden können, ist maschinelles Lernen und die daraus resultierende technische Verarbeitung aufgrund seiner Struktur und Komplexität nicht mehr im Einzelnen nachvollziehbar. Zusätzlich ist offen, wem die Verantwortung für Entscheidungen zugeschrieben werden kann. Eine wesentliche Fragestellung im Zusammenhang mit KI ist die Transparenz der Datenverarbeitung und Erklärbarkeit der Ergebnisse, um menschliche Kontrolle und die Zuschreibung von Verantwortung sicherzustellen[x]. Dies wird unter anderem in den Asilomar-Prinzipien gefordert (Future of Life.org 2017). Daraus hat sich der Forschungsbereich *Explainable Artificial Intelligence* herausgebildet.

Andreas Holzinger motiviert die Problemstellung (2018: 138ff.):

> „Um ein Niveau an praktisch nutzbarer AI zu erreichen ist es notwendig: (1) aus hochdimensionalen Datenmengen zu lernen, (2) daraus Wissen zu extrahieren, (3) dieses zu verallgemeinern, (4) dabei aber den ,,Fluch der Dimensionalität“ in den Griff zu bekommen, und schließlich (5) die den Daten zugrundeliegenden Erklärungsfaktoren zu verstehen. Letzteres impliziert allerdings die wahrscheinlich größte Herausforderung moderner AI: Daten im Kontext einer Anwendungsdomäne zu verstehen.“

Ziel ist es somit, die Black Box transparent zu machen, in der maschinelles Lernen durch laufende Anpassung von Parametern erfolgt. Die Schwierigkeit ist dabei, dass der Zusammenhang der Parameter, die in einem vieldimensionalen Raum berechnet werden, zur inhaltlichen, mit der für Menschen verständlichen Problemstellung nicht direkt erkennbar ist. Es lässt sich aber feststellen, welche Daten zu einer Entscheidung geführt haben – beispielsweise welche Bereiche eines Bildes dazu geführt haben, den Inhalt einer bestimmten Kategorie zuzuordnen. (Zweig 2023: 257ff.)

Bias – „Programmierter Rassismus“

Ein großes Problem in der praktischen Nutzung von KI-Verfahren ist der Bias – sprich die Verfälschung von Ergebnissen aufgrund von Daten, in die falsche Zusammenhänge oder Vorurteile eingeschrieben sind (Schinzel 2022: 26-34). Maschinelles Lernen ist auch inhaltlich von den Daten abhängig, die in Modelle einfließen. Die Modelle bilden damit Entscheidungen der Vergangenheit ab; Fehlurteile, beispielsweise aufgrund von Vorurteilen in der Vergangenheit, werden so in der Gegenwart fortgeschrieben. Bekannt wurde der Algorithmus des österreichischen Arbeitsmarktservice (AMS) – dem Gegenstück zur deutschen Arbeitsagentur –, der die Wiedereinstiegschancen Arbeitssuchender beurteilen soll (Pumhösel 2020). Dabei werden die Kategorien Alter, Geschlecht, Wohnort, bisherige Berufslaufbahn, Ausbildung, Staatsbürgerschaft herangezogen. Insbesondere wurde festgestellt, dass der Algorithmus zu einer geringeren Einschätzung der Wiedereinstiegschancen von Frauen gegenüber Männern gekommen war und ihm deswegen Diskriminierung vorgeworfen wurde.

Auch in den Niederlanden gab es erhebliche Folgen durch fehlerhafte Risikoindikatoren bei einer Software, die Betrug beim Bezug von Kindergeld aufdecken sollte. Einzelne Indikatoren – beispielsweise der Besitz einer doppelten Staatsbürgerschaft – führten zu massiven Falschbewertungen. In der Folge erhielten Kindergeldbeziehende Rückforderungen in teilweise sechsstelliger Höhe und gerieten dadurch in Armut; einzelne begingen Suizid. Die niederländische Regierung musste aufgrund des dadurch ausgelösten Skandals zeitweise zurücktreten, führte ihr Amt aber freilich geschäftsführend weiter.

> Authorities penalized families over a mere suspicion of fraud based on the system’s risk

> indicators. Tens of thousands of families – often with lower incomes or belonging to ethnic minorities – were pushed into poverty because of exorbitant debts to the tax agency. Some victims committed suicide. More than a thousand children were taken into foster care.“ (Heikklä 2022)

Neben falschen Indikatoren können auch weitere Eigenschaften der Daten zu diskriminierenden Bewertungen führen. Maschinelles Lernen fußt auf Modellen und Daten, die unvollständig sind und nur einen Teil der wirklichen Welt abbilden können. Es ist von der Qualität der Trainingsdaten abhängig. Dies wird zusätzlich verstärkt, wenn Feedbackschleifen fehlen und damit keine Korrektur der Daten vorgenommen wird. Bewertungen in der Vergangenheit werden fortgeschrieben und wirken sich auf die Lernergebnisse und die Voraussagen des Systems aus.

Ein Gender Bias kann sich ergeben, wenn falsche Entsprechungen in die Trainingsdaten eingeschrieben sind – beispielsweise, wenn einem „Chefarzt“ aufgrund früherer Einstellungspraxis als weibliches Gegenstück „Oberschwester“ anstatt richtig „Chefärztin“ gegenübergestellt wird. Dies muss bei den Trainingsdaten vorab erkannt und entsprechend korrigiert werden.

Zusätzlich kann Maschinelles Lernen auf Äußerlichkeiten basieren, etwa wenn Menschen auf der Basis von Bildern klassifiziert werden, und so Nebensächlichkeiten die Ergebnisse dominieren – besonders dies kann zu „programmiertem Rassismus“ (Wolfgangel 2018) führen. Außerdem kann die Bewertung nur auf Basis bekannter Daten erfolgen; auch der Kontext muss in den Daten enthalten sein, wenn er berücksichtigt werden soll, beispielsweise die Gründe für Bewertungen.

Ergebnisse werden dabei auch indirekt beeinflusst, wie bei der Bevorzugung Weißer bei der Auswahl von Bewerber*innen, der Bevorzugung/Benachteiligung wegen des Wohnorts oder der Benachteiligung von Frauen bei der Besetzung von Vorstandsposten aufgrund der Praxis in der Vergangenheit.

Generell sind die Parameter für Menschen kaum nachvollziehbar. Es ist in der Praxis häufig nicht möglich, anhand der Parameter die Bewertungen, die in die Trainingsdatensätze eingeschrieben sind, zu verstehen. Dies erschwert es zusätzlich, einen in den Trainingsdaten eingeschriebenen Bias zu vermeiden.

Rechtsschutz für „geistiges Eigentum“

Bei der Nutzung von KI-Verfahren spielt die Frage nach dem Schutz geistigen Eigentums mindestens in drei Dimensionen eine Rolle:

- Verfahren der KI, durch die neue Artefakte geschaffen werden,
- Produkte, die als technische Schöpfungen durch Verfahren der KI erzeugt werden,
- Verarbeitete Daten, die als Trainingsdaten in die Produkte einfließen und darin enthalten sind, ohne die genaue Herkunft im Regelfall bestimmen zu können.

Das Europäische Parlament (2020) hat eine Entschließung zu den Rechten des geistigen Eigentums bei der Entwicklung von KI-Technologien vorgelegt. Dort wird auf die Nicht-Patentierbarkeit mathematischer Methoden hingewiesen. Zu den durch KI erzeugten technischen Methoden kommt sie zur Auffassung,

> „dass durch KI erzeugte technische Schöpfungen gemäß dem Rechtsrahmen fu?r Rechte des geistigen Eigentum geschützt werden müssen, […] ist der Ansicht, dass selbstständig von künstlichen Akteuren und Robotern erzeugte Werke eventuell nicht urheberrechtlich geschützt werden können, da der Grundsatz der Originalität, der mit natürlichen Personen verbunden ist, gewahrt werden muss und der Begriff der „geistigen Schöpfung“ an die Person des Autors gebunden ist […]“ (Europäisches Parlament 2020: Abs. 15))

Die dritte Fragestellung betrifft die Ergebnisse der Verarbeitung von Daten durch Verfahren der Künstlichen Intelligenz, beispielsweise in Form von Trainingsdaten für Systeme des maschinellen Lernens, die potenziell wiederum Urheberrechten unterliegen[xi]. In die Modelle können urheberrechtlich geschützte Daten einfließen und die Produkte auf urheberrechtlich geschütztem Material basieren – ohne dass dies im Detail nachvollziehbar ist und die Urheber entsprechend honoriert werden. Das Europäische Parlament (2020: Erwägungsgrund D) stellt dazu fest,

> „dass KI-Technologien die Rückverfolgbarkeit von Rechten des geistigen Eigentums und deren Anwendung auf Werke, die durch KI erzeugt wurden, erschweren und somit verhindern, dass Menschen, deren ursprüngliche Arbeit in solchen Technologien zum Einsatz kommt, eine faire Vergütung erhalten.“

Arbeit

Die Debatte, ob der Einsatz von Computern die menschliche Arbeitskraft ersetzen kann und zum Abbau von Arbeitsplätzen führt, ist nicht neu. Nachdem zunächst vor allem einfachere Tätigkeiten durch Computer ausgeführt werden konnten, sind durch Künstliche Intelligenz und maschinelles Lernen inzwischen zunehmend auch hochqualifizierte Berufe betroffen.[xii]

Zu der zunehmenden Möglichkeit der Übernahme von geistigen Tätigkeiten, beispielsweise der Erstellung von Berichten oder ähnlichem durch Chatbots kommen die in Personalabteilungen eingesetzten Verfahren der Human Resource Analytics und damit verbunden umfassende Überwachungsmöglichkeiten am Arbeitsplatz (Waas 2023). Die Produktivität von Arbeitnehmer*innen wird gemessen und auf mehreren Ebenen aggregiert, um den Erfolg von Management-Entscheidungen feststellen zu können. Gleichzeitig werden die einzelnen Arbeitnehmenden bewertet, beispielsweise um Leistungsträger*innen zu ermitteln, deren Ausscheiden für das Unternehmen besonders kostenintensiv wäre – analog natürlich auch Mitarbeiter*innen, deren Leistung als nicht zufriedenstellend bewertet wird. Hierzu können inzwischen auch biometrische Daten herangezogen werden, die durch Sensoren erfasst werden, oder Geolocation-Technologien zur Standortbestimmung von Mitarbeiter*innen im Außendienst.

Militärische Nutzung

Wie jede Technologie wird auch die KI im militärischen Bereich vorangetrieben – unter anderem mit dem Ziel, möglichst effektiv in militärischen Auseinandersetzungen Menschen zu töten. KI-Verfahren werden dabei in autonomen Waffensystemen eingesetzt, die ohne menschliches Eingreifen militärische Ziele bekämpfen können. Ein Zielbild kann dabei sein, dass der „Feind“ selbständig erkannt und bekämpft wird.[xiii]

Dabei ergeben sich eine Reihe von Fragen: Woran sind „feindliche“ Kombattant*innen eigentlich zu erkennen? Wie werden sie von einem Kind und anderen Zivilist*innen unterschieden? Wie ist eine Waffe zu erkennen? Wie werden die Regeln der Genfer Konvention im „intelligenten“ System umgesetzt? Wie können „richtige“ Entscheidungen sichergestellt werden?

Noch werden autonome Waffen weitgehend abgelehnt – die Letztentscheidung soll auch aus militärischer Sicht beim Menschen liegen. Doch wie kann sichergestellt werden, dass menschliche Entscheider*innen – angesichts sehr kurzer Antwortzeiten – eine autonome Entscheidung treffen können, die über die einfache Bestätigung der maschinellen Empfehlung hinausgeht. Wie können menschliche Entscheider*innen im Zweifel rechtfertigen, sich über diese Empfehlung hinweggesetzt zu haben?

Ein besonderes Risiko ergibt sich aus militärischen Entscheidungssystemen, die auf KI basieren und die computergestützte Reaktionen auf (vermeintliche) militärische Angriffe auslösen. Karl Hans Bläsius und

Jörg Siekmann (2023: 9) weisen seit Jahren auf das Risiko eines Atomkriegs aus Versehen hin, der durch die Fehlfunktion solcher automatisierter Systeme ausgelöst werden kann. Wie wir heute wissen, wurde sehr wahrscheinlich 1983 ein Atomkrieg nur verhindert, weil der sowjetrussische leitende Offizier Stanislaw Jewgrafowitsch Petrow in eigener Verantwortung handelte und einen gemeldeten Angriff US-amerikanischer Atomraketen als Fehlalarm einstufte[xiv]. Darüber, wie ein automatisiertes Entscheidungssystem in dieser Situation reagiert hätte, können wir nur spekulieren.

Politische Regulierung

Die politische Bedeutung der KI zeigt sich nicht zuletzt darin, dass der 19. Deutsche Bundestag eine Enquête-Kommission eingerichtet hat, die die Auswirkungen der Künstlichen Intelligenz untersuchen sollte und diese als Ergebnis ihrer Arbeit einen umfassenden Bericht vorgelegt hat. (Deutscher Bundestag 2020)

Gleichzeitig gibt es auf Ebene der Europäischen Union Bestrebungen, Künstliche Intelligenz durch den AI Act zu regulieren, der sich vor allem auf die Einhegung von Hochrisikosystemen konzentriert (Europäische Union 2021)[xv]. Hier geht es noch nicht darum, dass Künstliche Intelligenz den Menschen überflügelt, sondern um die Einhegung konkreter Risiken beim Einsatz von KI-Systemen. Die Verordnung soll entsprechend das Ziel der Union unterstützen, „bei der Entwicklung einer sicheren, vertrauenswürdigen und ethisch vertretbaren Künstlichen Intelligenz weltweit eine Führungsrolle einzunehmen“, und dabei „für den vom Europäischen Parlament ausdrücklich geforderten Schutz von Ethikgrundsätzen“ zu sorgen (Europäische Union 2021: Erwägungsgrund 5). Weiter heißt es:

> „Abgesehen von den zahlreichen nutzbringenden Verwendungsmöglichkeiten künstlicher Intelligenz kann diese Technik auch missbraucht werden und neue und wirkungsvolle Instrumente für manipulative, ausbeuterische und soziale Kontrollpraktiken bieten. Solche Praktiken sind besonders schädlich und sollten verboten werden, weil sie im Widerspruch zu den Werten der Union stehen, nämlich der Achtung der Menschenwürde, Freiheit, Gleichheit, Demokratie und Rechtsstaatlichkeit sowie der Grundrechte in der Union, einschließlich des Rechts auf Nichtdiskriminierung, Datenschutz und Privatsphäre sowie der Rechte des Kindes.“ (Europäische Union 2021: Erwägungsgrund 15)

Künstliche Intelligenz ist eine faszinierende Technologie mit Chancen, aber auch erheblichen Risiken – da reichen schon die gesellschaftspolitischen Risiken für Bürgerrechte und Datenschutz, ohne gleich an die „Ablösung“ des Menschen durch eine übermächtige Technik zu denken. Ein Frühwarnsystem ist wichtig, um Entwicklungen frühzeitig zu erkennen und gegenzusteuern. Wir dürfen das keinesfalls wenigen großen Plattformen überlassen.

Stefan Hügel *ist Diplom-Informatiker aus Frankfurt am Main. Studiert hat er in Karlsruhe und Freiburg im Breisgau. Er arbeitet als IT-Berater. Daneben ist er der Vorstandsvorsitzende des Forum InformatikerInnen für Frieden und gesellschaftliche Verantwortung und Vorsitzender des Bundesvorstandes der Humanistischen Union. Seine Schwerpunkte sind Netzpolitik, Datenschutz, IT-Sicherheit, Information*

Warfare und Künstliche Intelligenz.

Literatur

Affsprung, D. 2023: The ELIZA Defect: Constructing the Right Users for Generative AI, in: AAAI/ACM Conference on AI, Ethics and Society (AIES ’23), August 08–10, 2023, Montréal/New York, https://doi.org/10.1145/3600211.3604744.

Azhar, A. 2021: Exponential. How Accelerating Technology is leaving us behind and what to do about it, London.

Bläsius, K. H./Siekmann, J. 2023: Ist die Künstliche Intelligenz gefährlich? In: FIfF-Kommunikation, Jg. 40, H. 3, S. 9.

Dachwitz, I. 2023. Das sind die 650.000 Kategorien, in die uns die Online-Werbeindustrie einsortiert, in: netzpolitik.org, https://netzpolitik.org/2023/microsofts-datenmarktplatz-xandr-das-sind-650-000-kategorien-in-die-uns-die-online-werbeindustrie-einsortiert/.

Deutscher Bundestag 2020: Bericht der Enquête-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale, BT-Drs 19/23700, https://dserver.bundestag.de/btd/19/237/1923700.pdf.

El Atillah, I. 2023: Man ends his life after an AI chatbot ‘encouraged’ him to sacrifice himself to stop climate change, in: Euronews.next vom 31. März 2023, https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-anai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate.

Europäisches Parlament 2020: Rechte des geistigen Eigentums bei der Entwicklung von KI-Technologien, Entschließung, P9_TA(2020)0277, https://www.europarl.europa.eu/doceo/document/TA-9-2020-0277_DE.html.

Europäische Union 2021: The Artificial Intelligence Act, https://artificialintelligenceact.eu/the-act/.

Future of Life.org 2017: Asilomar AI Principles, https://futureoflife.org/open-letter/ai-principles-german/.

Future of Life.org 2023: Pause Giant AI Experiments: An Open Letter, https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Gardner, H. E. 1983: Frames of Mind, the theory of multiple intelligences, New York.

Greenwald, G. 2014: Die globale Überwachung. Der Fall Snowden, die amerikanischen Geheimdienste und die Folgen, München.

Hawking, S. W. 2018: Kurze Antworten auf große Fragen, Stuttgart.

Heikklä, M. 2022: Dutch scandal serves as a warning for Europe over risks of using algorithms, in: Politico vom 29. März 2022, https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/.

Hofstadter, D. R./Dennett, D. 1986 [1981]: Einsicht ins Ich. Fantasien und Reflexionen über Selbst und Seele, Stuttgart.

Holzinger, A. 2018: Explainable AI (ex-AI), Aktuelles Schlagwort, in: Informatik-Spektrum, Jg. 41, H. 2, S. 138-143.

Hügel, S. 2019: Künstliche Intelligenz und Politik. Algorithmen, Data Science, Microtargeting – und ihre Auswirkungen auf politische Entscheidungen, in: vorgänge. Zeitschrift für Bürgerrechte und Gesellschaftspolitik, Nr. 225/226 = Jg. 58, H. 1-2, S. 25-42.

Jonas, H. 1979: Das Prinzip Verantwortung. Versuch einer Ethik für die technologische Zivilisation. Frankfurt am Main.

Kissinger, H./Schmidt, E./Huttenlocher, D. 2021: The Age of AI, London.

Kühnreich, K. 2022: Social Credit Systems und Gamification. Digitale Gesellschaft, Netzpolitischer Abend, Vortragsaufzeichnung, https://media.ccc.de/v/dgna-4248-social-credit-systems-und-gami.

Leisegang, D. 2023: Algorithmen rütteln kaum an politischen Einstellungen, in: netzpolitik.org, https://netzpolitik.org/2023/studien-zu-facebook-und-instagram-algorithmen-ruetteln-kaum-an-politischen-einstellungen/

.

Levine, E. V. 2023: Cargo Cult AI, in: Communications of the ACM, Bd. 66, H. 9, S. 46.

Louban, A. et al. 2022: Das Phänomen Deepfakes. Künstliche Intelligenz als Element politischer Einflussnahme und Perspektive einer Echtheitsprüfung, in: Friedewald, M. Et al. (Hrsg.): Künstliche Intelligenz, Demokratie und Privatheit, Baden-Baden, S. 265-288.

Misselhorn, C. 2018: Grundfragen der Maschinenethik, Stuttgart.

Moreno, J. 2023: „Die Menschheit kreiert derzeit eine intelligentere Spezies“, Podcast mit Hannah Bast, in: Spiegel Online, https://www.spiegel.de/netzwelt/ki-forscherin-hannah-bast-die-menschheit-kreiert-derzeit-eine-intelligentere-spezies-podcast-a-b9b78548-b47c-46ff-86d1-c1a7ed2069fd.

O’Neil, C. 2016: Weapons of Math Destruction. How Big Data increases inequality and threatens Democracy, London.

Otte, R. 2023: Intelligenz und Bewusstsein. Oder: Ist KI wirklich KI? In: Aus Politik und Zeitgeschichte, Bd. 73, H. 42, S. 9-16.

Paaß, G./Hecker, D. 2020: Künstliche Intelligenz. Was steckt hinter der Technologie der Zukunft? Wiesbaden.

Pariser, E. 2012: Filter Bubble. Wie wir im Internet entmündigt werden, München.

Paul, K. and agencies 2023: Letter signed by Elon Musk demanding AI research pause sparks controversy, in: The Guardian, https://www.theguardian.com/technology/2023/mar/31/ai-research-pause-elon-musk-chatgpt.

Pumhösel, A. 2020: Gender-Bias: Schlechtere Jobchancen für Frauen durch Algorithmen, in: Der Standard Online, https://www.derstandard.at/story/2000115720676/gender-bias-schlechtere-job-chancen-fuer-frauen-durch-algorithmen.

Rehak, R. 2021: The Language Labyrinth: Constructive Critique on the Terminology Used in the AI

Discourse, in: Verdegem, P. (Hrsg.): AI for Everyone? Critical Perspectives. London, S. 87-102. DOI: https://doi.org/10.16997/book55.f.

Rehbein, M. 2018: Die „Asilomar AI Principles“ zu Künstlicher Intelligenz, in: FIfF-Kommunikation, Jg. 35, H. 3, S. 24.

Russell, S./Norvig, P. 2023: Künstliche Intelligenz. Ein moderner Ansatz, 4. Auflage, München.

Safe AI 2023 Statement on AI Risk, https://www.safe.ai/statement-on-ai-risk.

Schinzel, B. 2022: Diskriminierung durch digitale Entscheidungsstrukturen, in: Aus Politik und Zeitgeschichte, Bd. 72, H. 10-11, S. 26-34.

Schmalzried, G. 2023: Wie ein Münchner Student zur Zielscheibe von Microsofts KI wurde, in: Bayerischer Rundfunk, BR24, https://www.br.de/nachrichten/netzwelt/microsoft-ki-bing-chatgpt-muenchner-student-als-zielscheibe.

Shi-Kupfer, K. 2023: Digit@l China. Überwachungsdiktatur und technologische Avantgarde, München.

Stahl, B. C. 2023: Grauzonen zwischen Null und Eins. KI und Ethik, in: Aus Politik und Zeitgeschichte, Bd. 73, H. 42, S. 17-22.

Szöke, D. 2023: Prompt Injection: Marvin von Hagen trägt vor, wie er Bing Chat austrickste, in: Heise Online, https://www.heise.de/news/Prompt-Injection-Marvin-von-Hagen-traegt-vor-wie-er-Bing-Chat-austrickste-9210511.html.

Tegmark, M. 2017: Leben 3.0. Mensch sein im Zeitalter Künstlicher Intelligenz, Berlin.

Waas, B. 2023: Künstliche Intelligenz und Arbeitsrecht (= Hans-Böckler-Stiftung, HSI-Schriftenreihe, Bd. 26), Frankfurt am Main, https://www.hugo-sinzheimer-institut.de/faust-detail.htm?sync_id=HBS-008472.

Weizenbaum, J. 1987 [1976]: Die Macht der Computer und die Ohnmacht der Vernunft, 2. Auflage. Frankfurt am Main.

Wolfangel, E. 2018: Programmierter Rassismus, in: Zeit Online, https://www.zeit.de/digital/internet/2018-05/algorithmen-rassismus-diskriminierung-daten-vorurteile-alltagsrassismus.

Yudkowsky, E. 2023: Pausing AI Developments isn't enough. We need to shut it all down, in: Time, https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/.

Zweig, K. A. 2019: Ein Algorithmus hat kein Taktgefühl. Wo künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können, München.

Zweig, K. A. 2023: Die KI war's! Von absurd bis tödlich: Die Tücken der künstlichen Intelligenz, München.

Anmerkungen:

i Deutsch Amerikanische Freundschaft (2003) Algorithmus – zahlenlied, in: Deutsch Amerikanische Freundschaft (2003) Fünfzehn neue DAF Lieder, Track 10, superstar recordings.

ii Wikipedia, Stichwort ChatGPT, https://de.wikipedia.org/wiki/ChatGPT.

iii Wikipedia, Stichwort Künstliche Intelligenz, https://de.wikipedia.org/wiki/Künstliche_Intelligenz.

iv Die hier verwendete anthropomorphe Terminologie hat sich etabliert, obwohl sie philosophisch problematisch ist. Eine Diskussion dazu findet sich in: Rehak 2021.

v Wikipedia, Stichwort Tay (Bot), https://de.wikipedia.org/wiki/Tay_(Bot).

vi *Prompting* bezeichnet die Interaktion mit einem Chatbot, indem Anfragen an ihn gestellt werden. *Prompt Injection* bezeichnet das „Unterschieben“ von Anfragen in böswilliger Absicht, um dem Chatbot Antworten zu entlocken, die von dessen Entwickler:innen nicht vorgesehen sind.

vii Wikipedia, Stichwort Trolley-Problem, https://de.wikipedia.org/wiki/Trolley-Problem.

viii Angesichts der Enthüllungen von Edward Snowden müssen wir aber davon ausgehen, dass massive staatliche Überwachung auch in demokratisch konstituierten Gesellschaften an der Tagesordnung ist (Greenwald 2014).

ix https://www.youtube.com/watch?v=iAE6dCE7URg.

x Vgl. hierzu auch den Beitrag von Kleemann, Hirsbrunner und Aden im vorliegenden Heft.

[xi] Vgl. hierzu das „Interview“/den Chatverlauf von ChatGPT mit Werner Koep-Kerstin und Philip Dingeldey im vorliegenden Heft.

[xii] Vgl. hierzu auch die Beiträge von Philip Dingeldey und Annabell Lamberth im vorliegenden Heft.

[xiii] Vgl. hierzu auch den Beitrag von Hans-Jörg Kreowksi und Aaron Lye im vorliegenden Heft.

[xiv] Wikipedia, Stichwort Nuklear-Fehlalarm von 1983, https://de.wikipedia.org/wiki/Nuklear-Fehlalarm_von_1983.

[xv] Vgl. hierzu auch den Beitrag von Hartmut Aden im vorliegenden Heft.
